

Proceedings of a Workshop on Statistics on Networks

Scott T. Weidman, Editor, Committee on Applied and Theoretical Studies, National Research Council

ISBN: 0-309-65703-2, 470 pages, CD-ROM, (2007)

This free PDF was downloaded from:

<http://www.nap.edu/catalog/12083.html>



Visit the [National Academies Press](#) online, the authoritative source for all books from the [National Academy of Sciences](#), the [National Academy of Engineering](#), the [Institute of Medicine](#), and the [National Research Council](#):

- Download hundreds of free books in PDF
- Read thousands of books online, free
- Sign up to be notified when new books are published
- Purchase printed books
- Purchase PDFs
- Explore with our innovative research tools

Thank you for downloading this free PDF. If you have comments, questions or just want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, [visit us online](#), or send an email to comments@nap.edu.

This free book plus thousands more books are available at <http://www.nap.edu>.

Copyright © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

Proceedings of a Workshop on Statistics on Networks

Scott T. Weidman, Editor

Committee on Applied and Theoretical Statistics

Board on Mathematical Sciences and Their Applications

Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

This study was supported by Grant #H-98230-05-1-0019 between the National Academy of Sciences and the National Security Agency. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

Copies of this report on CD-ROM are available from the Board on Mathematical Sciences and Their Applications, 500 Fifth Street, N.W., Room 960, Washington, D.C. 20001.

Additional copies of this CD-ROM are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, <http://www.nap.edu>.

Copyright 2007 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON APPLIED AND THEORETICAL STATISTICS

EDWARD J. WEGMAN, *Chair*, George Mason University
DAVID L. BANKS, Duke University
AMY BRAVERMAN, Jet Propulsion Laboratory
EMERY N. BROWN, Harvard Medical School
ALICIA CARRIQUIRY, Iowa State University
THOMAS COVER, Stanford University
KAREN KAFADAR, University of Colorado at Denver
KATHRYN B. LASKEY, George Mason University
MICHAEL LESK, Rutgers University
THOMAS LOUIS, Johns Hopkins University
DOUGLAS NYCHKA, National Center for Atmospheric Research
LELAND WILKINSON, SPSS, Inc.

Staff

Board on Mathematical Sciences and Their Applications (BMSA) Workshop Organizers:

SCOTT WEIDMAN, BMSA Director
BARBARA WRIGHT, Administrative Assistant

PROGRAM COMMITTEE

DAVID L. BANKS, *Chair*, Duke University
EMERY N. BROWN, Massachusetts General Hospital
KATHLEEN CARLEY, Carnegie Mellon University
MARK HANDCOCK, University of Washington
RAVI IYENGAR, Mount Sinai School of Medicine
ALAN F. KARR, National Institute of Statistical Sciences
ROBERT D. NOWAK, University of Wisconsin-Madison
WALTER WILLINGER, AT&T-Research

NOTE: Funding for this workshop and its proceedings was generously provided by the National Security Agency.

BOARD ON MATHEMATICAL SCIENCES AND THEIR APPLICATIONS

C. DAVID LEVERMORE, *Chair*, University of Maryland
MASSOUD AMIN, University of Minnesota
MARSHA J. BERGER, New York University
PHILIP A. BERNSTEIN, Microsoft Corporation
PATRICIA F. BRENNAN, University of Wisconsin-Madison
PATRICK L. BROCKETT, University of Texas at Austin
DEBRA ELKINS, General Motors Corporation
LAWRENCE CRAIG EVANS, University of California at Berkeley
JOHN F. GEWEKE, University of Iowa
DARRYLL HENDRICKS, UBS AG
JOHN E. HOPCROFT, Cornell University
CHARLES M. LUCAS, AIG (retired)
CHARLES F. MANSKI, Northwestern University
JOYCE R. McLAUGHLIN, Rensselaer Polytechnic Institute
JILL PORTER MESIROV, Broad Institute
ANDREW M. ODLYZKO, University of Minnesota
JOHN RICE, University of California at Berkeley
STEPHEN M. ROBINSON, University of Wisconsin-Madison
GEORGE SUGIHARA, University of California at San Diego
EDWARD J. WEGMAN, George Mason University
LAI-SANG YOUNG, New York University

Staff

SCOTT WEIDMAN, Director
NEAL GLASSMAN, Senior Staff Officer
BARBARA WRIGHT, Administrative Assistant

Preface and Workshop Rationale

On September 26 and 27, 2005, the Committee on Applied and Theoretical Statistics of the National Research Council conducted a 2-day workshop that explored statistical inference on network data so as to stimulate further progress in this field. To encourage cross-fertilization of ideas, the workshop brought together a wide range of researchers who are dealing with network data in different contexts. The presentations focused on five major areas of research: network models, dynamic networks, data and measurement on networks, robustness and fragility of networks, and visualization and scalability of networks.

Disciplines such as biology, social sciences, and telecommunications have created different kinds of statistical theory for inference on network data. The workshop was organized to draw together experts from the various domains and to facilitate the sharing of their statistical, mathematical, and computational toolkits. The ubiquity of networks and network data created a challenging environment for the discovery of common problems and techniques.

The overall goals of this report, which is produced only on a CD and not in printed form, are to improve communication among various communities working on problems associated with network data and to increase relevant activity within the statistical sciences community. Included in this report are the full and unedited text of the 18 workshop presentations, the agenda of the workshop and a list of attendees (Appendix A) and biographical sketches of the speakers (Appendix B). The presentations represent independent research efforts on the part of academia, the private sector, federally funded laboratories, and government agencies, and as such they provide a sampling rather than a comprehensive examination of the range of research and research challenges posed by massive data streams.

This proceedings represents the viewpoints of its authors only and should not be taken as a consensus report of the Board on Mathematical Sciences and Their Applications or the National Research Council.

Contents

Keynote Address, Day 1

- Network Complexity and Robustness 2
John Doyle

Network Models

- Neurons, Networks, and Noise: An Introduction 62
Nancy Kopell
- Mixing Patterns and Community Structure in Networks 74
Mark Newman
- Dimension Selection for Latent Space Models of Social Networks 97
Peter Hoff

Dynamic Networks

- Embedded Networked Sensing (Redux?) 121
Deborah Estrin
- The Functional Organization of Mammalian Cells 146
Ravi Iyengar
- Dynamic Network Analysis in Counterterrorism Research 169
Kathleen M. Carley

Data and Measurement

- Current Developments in a Cortically Controlled Brain-Machine Interface 189
Nicho Hatsopoulos
- Some Implications of Path-Based Sampling on the Internet 207
Eric D. Kolaczyk
- Network Data and Models 226
Martina Morris

The State of the Art in Social Network Analysis

- The State of the Art in Social Network Analysis 255
Stephen P. Borgatti

Keynote Address, Day 2

Variability, Homeostasis per Contents and Compensation in Rhythmic Motor Networks 271
Eve Marder

Dynamics and Resilience of Blood Flow in Cortical Microvessels 292
David Kleinfeld

Robustness and Fragility

Robustness and Fragility 318
Jean M. Carlson

Stability and Degeneracy of Network Models 343
Mark S. Handcock

Visualization and Scalability

Characterizing Brain Networks with Granger Causality 376
Mingzhou Ding

Visualization and Variation: Tracking Complex Networks Across Time and Space 396
Jon Kleinberg

Dependency Networks for Relational Data 425
David Jensen

Appendixes

A Workshop Agenda and List of Attendees 450

B Biographical Sketches of Workshop Speakers 455

Keynote Address, Day 1

Network Complexity and Robustness

John Doyle, California Institute of Technology

DR. DOYLE: I am going to try to set the stage for this meeting. The work that I am going to talk about is the result of a lot of collaborations. I certainly won't give justice to those people here. All the good ideas and all the good work that I am going to talk about is done with them, so when I say "we" I mean "they." A lot of them are here today; two of them have posters and Jean Carlson is going to talk later. Walter Willinger is here. I am going to talk a lot about work with them.

There are buzzwords related to network complexity—network-centric, systems biology, pervasive, embedded. What I am interested in is the core theory challenges that underlie those common themes. I am going to be a little narrow in my focus in the sense that I am going to concentrate on biological networks and technological networks: I want to stick with stuff where we mostly know how the parts work, so that means not much social networks. There is a remarkably common core of theoretical challenges, so from the math-stat side I think there are some really common themes here. There has been recent dramatic progress in laying the foundation, yet there has also been amazingly, at the same time, a striking increase in what I would call unnecessary confusion. I will talk a little bit about both of these.

One of the common themes I am going to talk about is the fact that we see power laws all over. I think a lot of people at this workshop are going to be talking about that, because it is a common trait across all advanced technology and biology. Another common theme is that many of these systems in biology and advanced technology are robust yet fragile. What I mean by that is they work really well most of the time but that they fail occasionally, and when they do fail it can be catastrophically. We will hear more about that from other speakers and poster presenters. What I am going to do today, though, is talk about motivation for new theory and also education. We need to educate each other about even some fairly elementary aspects of statistics and mathematics. To get this started on the broadest level possible, I am going to stick with stuff that only assumes the background provided by a Caltech undergraduate education.

Let's start with some data. The figures below show the 20th century's hundred largest disasters worldwide. What I have plotted here are three kinds of disasters: technological disasters in tens of billions of dollars, natural disasters in hundreds of billions of dollars (and natural disasters can be even bigger), and power outages over some period of time and tens of millions of customers. What you see on a log-log chart are roughly straight lines with slopes of minus one.

Figure 2 shows the raw data. I have just taken the event sizes and plotted them; the worst events are along the bottom, while the more common events are higher up and to the left. The significance of this is that the worst event is orders of magnitude worse than the median event, which we are learning much more about this last year. It also means that demonstrations of these behaviors (events of the type represented in the upper left corner of Figures 1 and 2) and reality (which includes the events toward the lower right) can be very different, both in scale and in character, with the more common behaviors being robust and the rare behavior displaying the great fragility of these systems. When we take and build our new network-centric embedded-everywhere systems that some of us in this room are going to provide to all of us, they will degrade, and the reality may be much, much worse. This is also an indication of robust yet fragile. The typical behavior is much smaller than the worst case, and the worst case is very, very bad.

20th Century's 100 largest disasters worldwide

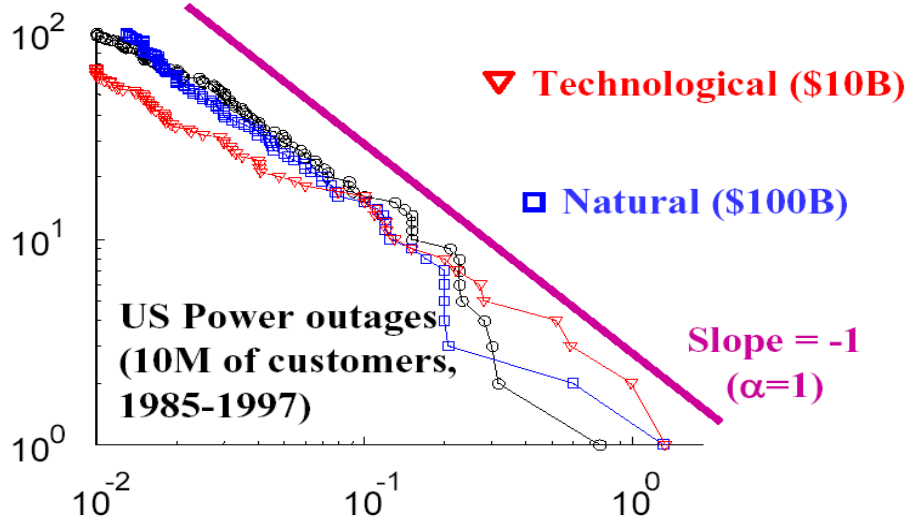


FIGURE 1

Raw data, *not* statistical

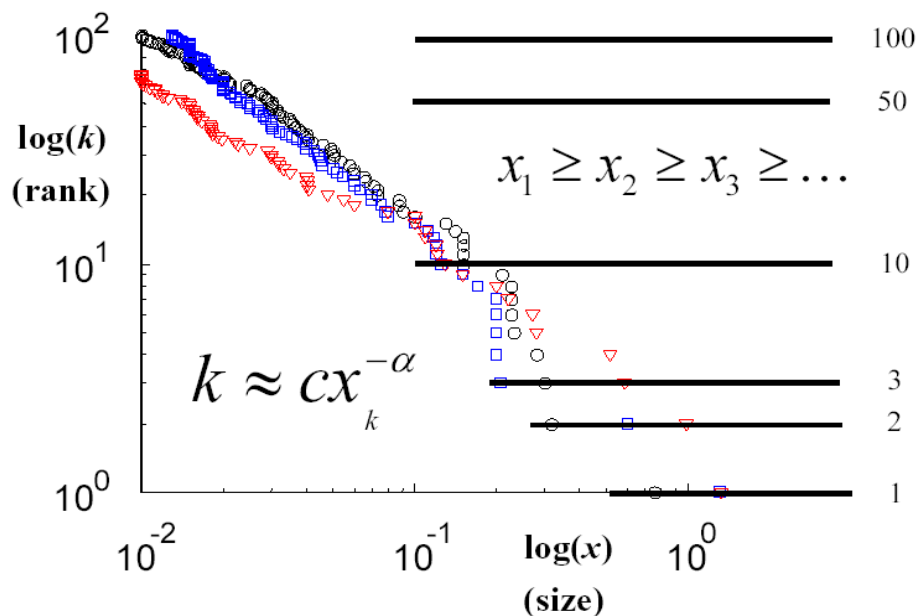


FIGURE 2

I am not going to focus so much on disasters. Jean Carlson is going to talk about that tomorrow, but I do want to get some of the ideas behind these kinds of statistics. They are either all you study or you won't study it at all, and there is not much conversation back and forth between those groups of people. The reason you get straight lines is that logarithms of a power law give you straight lines, and that is the slope alpha. That just describes the data, and I don't assume any kind of model. If I said I am going to think of this as samples from a stochastic process, that is a model. What I have written down is just data and I have done no statistics on it. The fact that I have drawn a line there is just to help your eye visualize things. One important thing to do is to look at data. One of the things that isn't often done in papers is people don't show you their data in a way that you can make judgments about the statistics they have made. We will see that there are lots of errors that arise because of this.

Power laws are really ubiquitous. Why is that and what do probability theory and statistics tell us about power laws? Well, if I were to ask you why are Gaussians so ubiquitous you would say it is central limit theorem. If I asked you why the exponential distribution is so popular or so common you would probably talk about the Markov properties or the marginalization properties, which are essentially, mathematically the same thing. They both occur in situations of low variability, so expect to see Gaussians and exponentials all over the

place.

When it comes to systems with high variability, power laws actually have even more strong statistical properties than Gaussians or exponentials. In fact, they have all that Gaussians and exponentials have, and even more. They are in some sense—I think Walter Willinger coined this phrase—more normal than normal. We should not call Gaussians normal; we should call power laws normal. They arise for no reason other than the way you measure data, and the way you look at the world will make them appear all over the place, provided there is high variability.

Much of science goes out of its way to get rid of high variability. We study in our labs low-variability phenomena, but high variability is everywhere. With the former, you get convergent moments, while the latter corresponds to divergent moments. Typical notions of moments and mean and variance don't mean anything. We talk about the coefficient of variation in a Gaussian. That is a well-defined concept in an exponential; it is actually 1. In power laws, of course, it is divergent. The issue is that power laws are not exceptional, but the really important issue is low versus high variability. What about mechanisms? We have large disasters because we have uncertain environments, we put assets at risk, and we only devote a certain amount of resources to ameliorating those risks, and thus sometimes those resources get overwhelmed. So, large events are not really surprising. If we were just to plot the large power outage of August 14, 2003, it is completely consistent with this data. The ramifications of the attacks on September 11, 2001, as a technological disaster, are not off the map. Hurricane Katrina is not off the map. These are all more or less consistent with the largest event. Like I said, I am not going to discuss disasters very much. High variability is much more fundamental and is very ubiquitous, particularly in highly engineered or evolved systems. Power laws are more normal. They shouldn't be thought of as signatures of any specific mechanisms any more than Gaussians are. It also means that their statistical properties lead people to find them where they aren't, through statistical errors. They are very abundant in networks, and I will come back to try to talk about in biology, in particular, why they are everywhere in biology.

Because of all these properties, there are major errors that are not just isolated instances, but typical in high-impact journals in science, presumably because mathematicians and statisticians play little or no role in writing, reviewing, or on the editorial boards of these journals. One of these errors is variability in power laws. They are badly misestimating the slope or basically, even more profoundly, misunderstanding where high variability comes from in the first place. A typical mistake that is made, taking data that are actually exponential and plotting them in such a way that suggests a power law, is exemplified in the following series of figures. I numerically generated random data with the little Matlab code shown, using Matlab's random

number generator (actually a pseudo-random number generator). I generated exponentially distributed data, and on Figure 3's semi-log plot you see it is nearly a straight line. I checked that I had written my Matlab program right, and that the Matlab random number generator works. It is linear on a semi-log plot because, obviously, you take logs of an exponential and you get linear with slope $-a$.

Finding power laws in low variability data

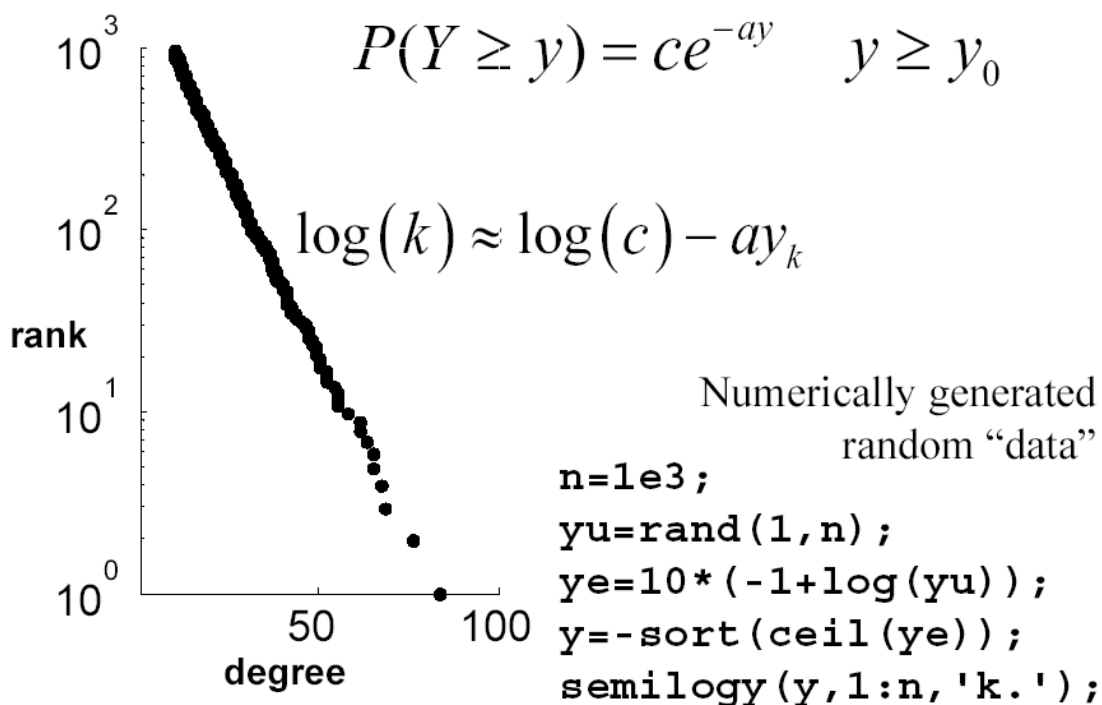


FIGURE 3

Instead of these rank plots you might think that we could just look at frequencies. Frequencies are obviously related to the rank as shown in Figure 4, even in this case where they are integers. You could write a little program to calculate the frequency plot and do a frequency plot and, lo and behold, a power law as shown in Figure 5, slope 1.5. All of a sudden, I think I have a power law here. In this case I know I don't, but in some sense I have made a classic error. The idea is that I just plotted in a silly way—again, you are differentiating data—it is noisy and you are going to get problems. I don't want to belabor this point too much because it is well understood why this happens, there is no mystery about it; it is just a matter of proper education. The statistics community has to explain this to people.

$$P(X \geq x) = ce^{-ax} \quad \text{rank}$$

$$p(X = x) = \tilde{c}e^{-ax} \quad \text{frequency}$$

$$= ce^{-ax} - ce^{-a(x+1)}$$

$$= c(1 - e^{-a})e^{-ax}$$

```

u=unique(y);
nu=length(u);g=0*u;
for k=1:nu
    g(k)=sum(y==u(k));
end
figure(2);
loglog(u,g,'bo');
    
```

FIGURE 4

$$P(X \geq x) = ce^{-ax}$$

$$p(X = x) = \tilde{c}e^{-ax}$$

```

u=unique(y);
nu=length(u);g=0*u;
for k=1:nu
    g(k)=sum(y==u(k));
end
figure(2);
loglog(u,g,'bo');
    
```

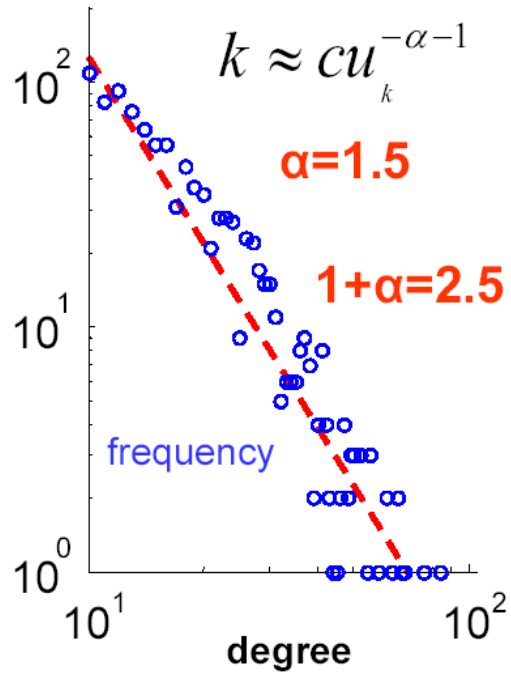


FIGURE 5

$$P(X \geq x) = cx^{-\alpha}$$

$$p(X = x) \approx \alpha x^{-\alpha-1}$$

$$P(X \geq x) = x^{-\alpha} \quad x \geq 1$$

$$p(X = x) = x^{-\alpha} - (x+1)^{-\alpha} \quad x \geq 1$$

$$= x^{-\alpha} - x^{-\alpha} \left(1 + \frac{1}{x}\right)^{-\alpha} \quad x \geq 1$$

$$= x^{-\alpha} \left(1 - \left(1 + \frac{1}{x}\right)^{-\alpha}\right) \quad x \geq 1$$

$$= x^{-\alpha} \left(\frac{\alpha}{x} + \dots\right) \quad x \geq 1$$

$$= \alpha x^{-\alpha-1} \quad x \gg 1$$

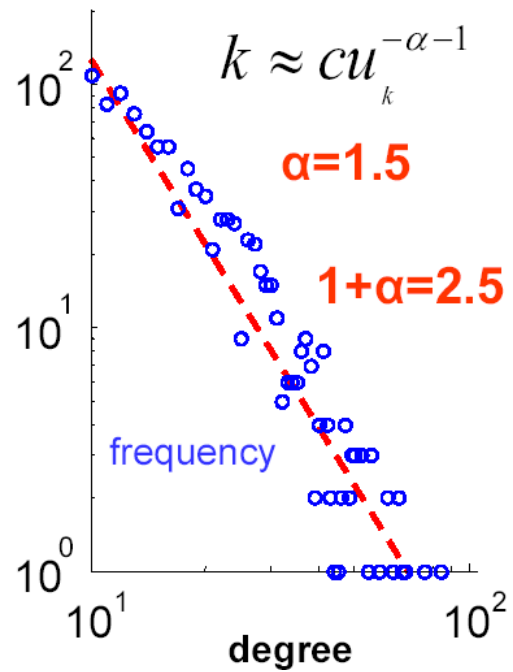


FIGURE 6

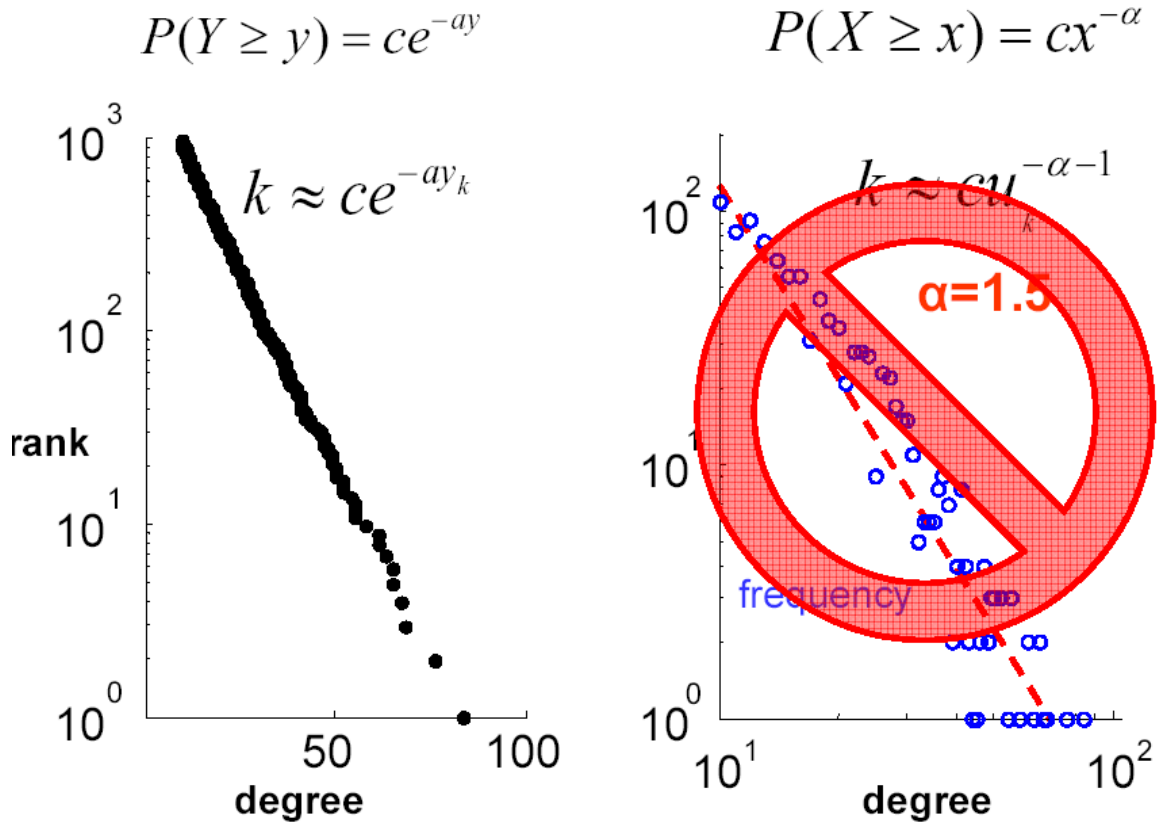


FIGURE 7

We really have got to not do that. You would think that it was an isolated instance but, in fact, it is almost always the case that you will see the plots on the right above are systematic errors. There are a bunch of networks that have been claimed to have power laws in them, and the data that was presented actually didn't.

Figure 8 comes from a paper on protein interaction power law, an article from *Nature*. I got the data and re-plotted it, but this is basically a plot that appears in a supplement. The paper concluded that it is a power law with a slope 1. It doesn't even have finite mean or variance, if you think of it as a probability model. What does it really have? Roughly it has exponential. Again, just to comment, this is an undergraduate sort of error. If a Caltech undergraduate did this they wouldn't last long. Of course, this is in *Nature* and in some sense—I have reviewed a few papers where I have said, no, you have got to plot it—they say, no, if we don't plot it the way we do on the right, then the editors won't accept them. I said we have got to change that.

Protein-Protein Interaction (PPI) networks

Node degree distribution of all interactions in 'filtered yeast interactome'

Han, J.-D et al (2004).
Evidence for dynamically organized modularity in the yeast protein-protein interaction network.
Nature, 430, 88-93.

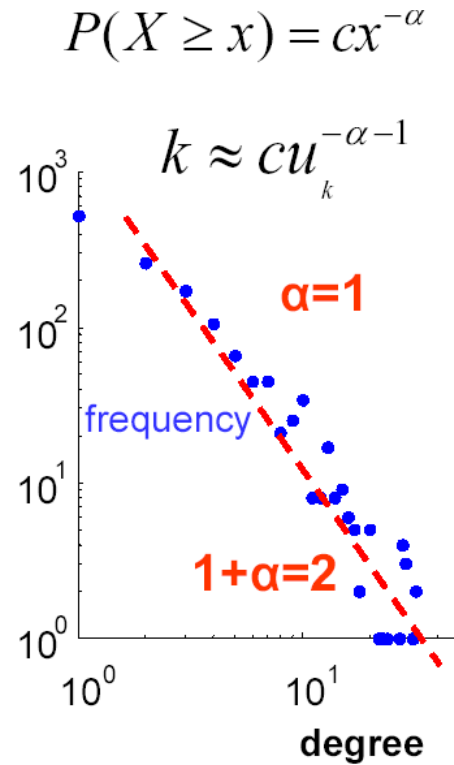


FIGURE 8

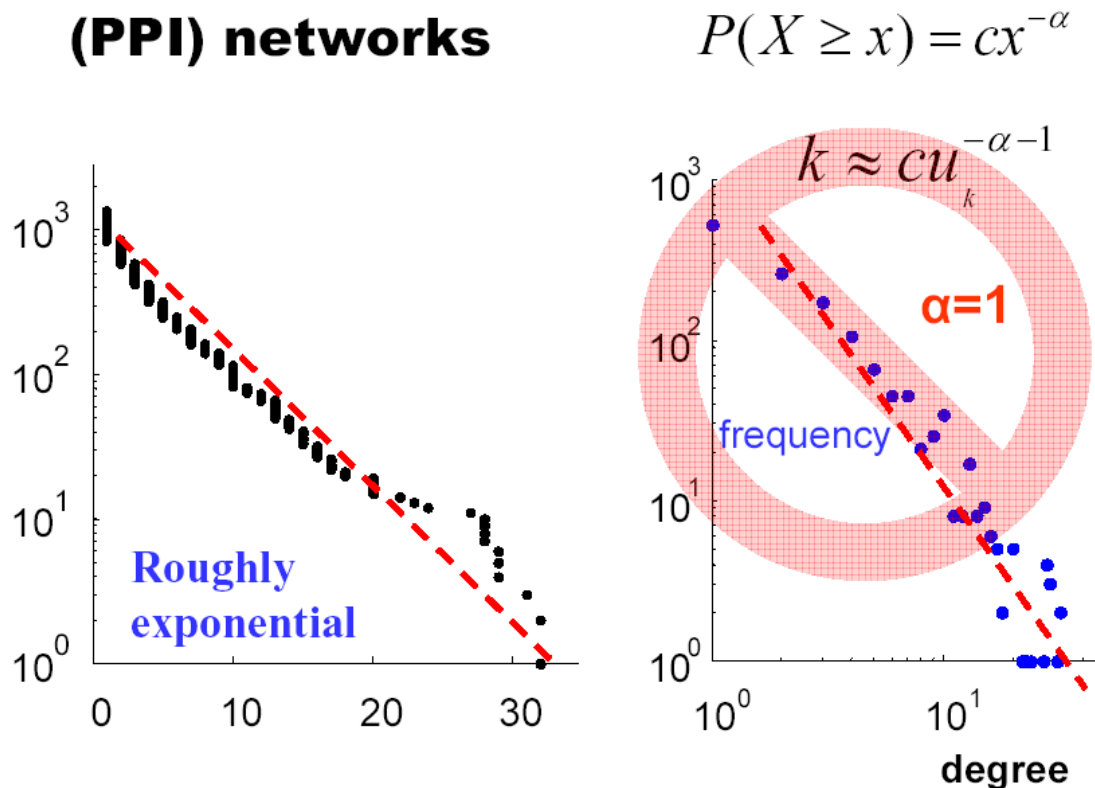


FIGURE 9

Figures 9-13 deal with an analogous error, from *Science*, does the power grid have a power law? No, it's almost a perfect exponential. You can't get a better exponential than that and yet it was called a power law. You can ask, what about the World Wide Web? Let's re-plot to make sure of what they were doing. They did logarithmic binning. You might think that helps. In this case, it really might be a power law, but you get the wrong slope. For a power law, slopes of 1.1 versus 1.7 are really different. You are off by orders of magnitude. So, once you get into power laws, the slopes are really important—again the point is, it doesn't exactly look like a power law, the real data. That is not the big deal. The big deal is it does have high variability. The Web has high variability. In fact, when you look at the Internet, practically everything has high variability. It is almost impossible to find something that doesn't have high variability, although Internet routers provide an exception. If you look at the router size frequency plot, you could come up with a power law whereas, in fact, it is clearly exponential in the tail because there is an excess of small-degree routers. That is probably because, either at the edge of the network there are a lot of degree-one routers or, in fact, it is probably an artifact of the way the data was taken, and you have to check this, but it is certainly not a power law. These errors are typical, but the big one, I think, is badly misunderstanding the origins of high variability, because high

variability is everywhere.

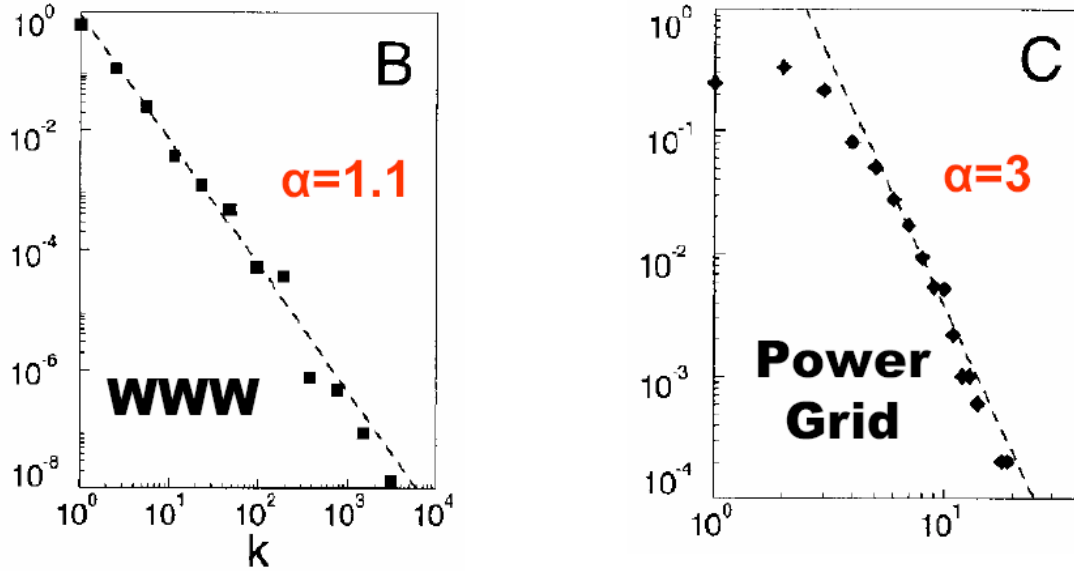


Fig. 1. The distribution function of connectivities for various large networks. (A) Actor collaboration graph with $N = 212,250$ vertices and average connectivity $\langle k \rangle = 28.78$. (B) WWW, $N = 325,729$, $\langle k \rangle = 5.46$ (6). (C) Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{\text{actor}} = 2.3$, (B) $\gamma_{\text{www}} = 2.1$ and (C) $\gamma_{\text{power}} = 4$.

FIGURE 10

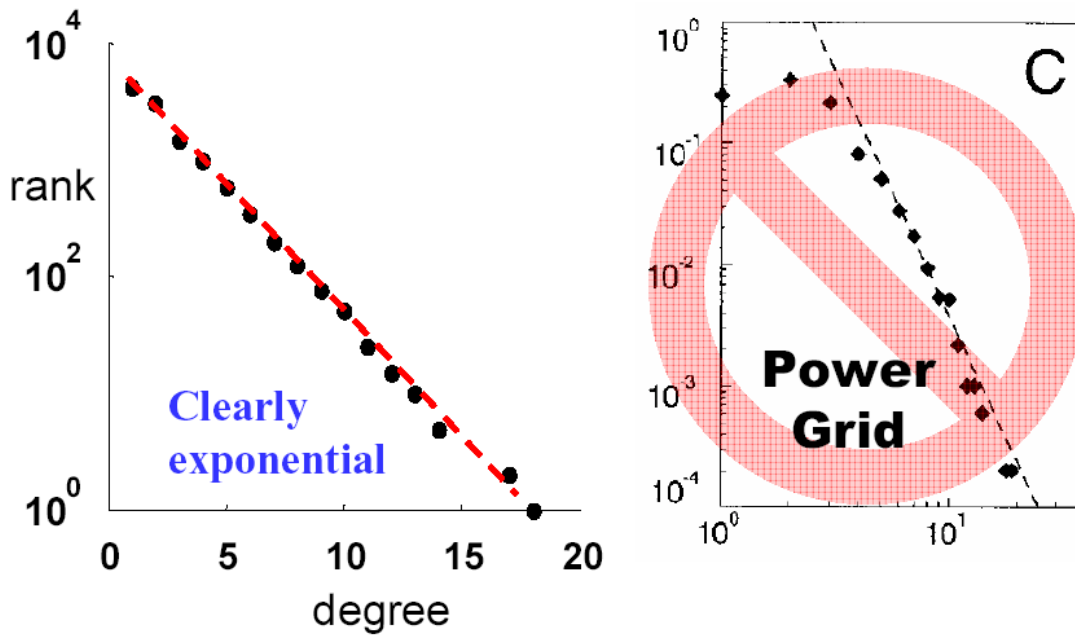


FIGURE 11

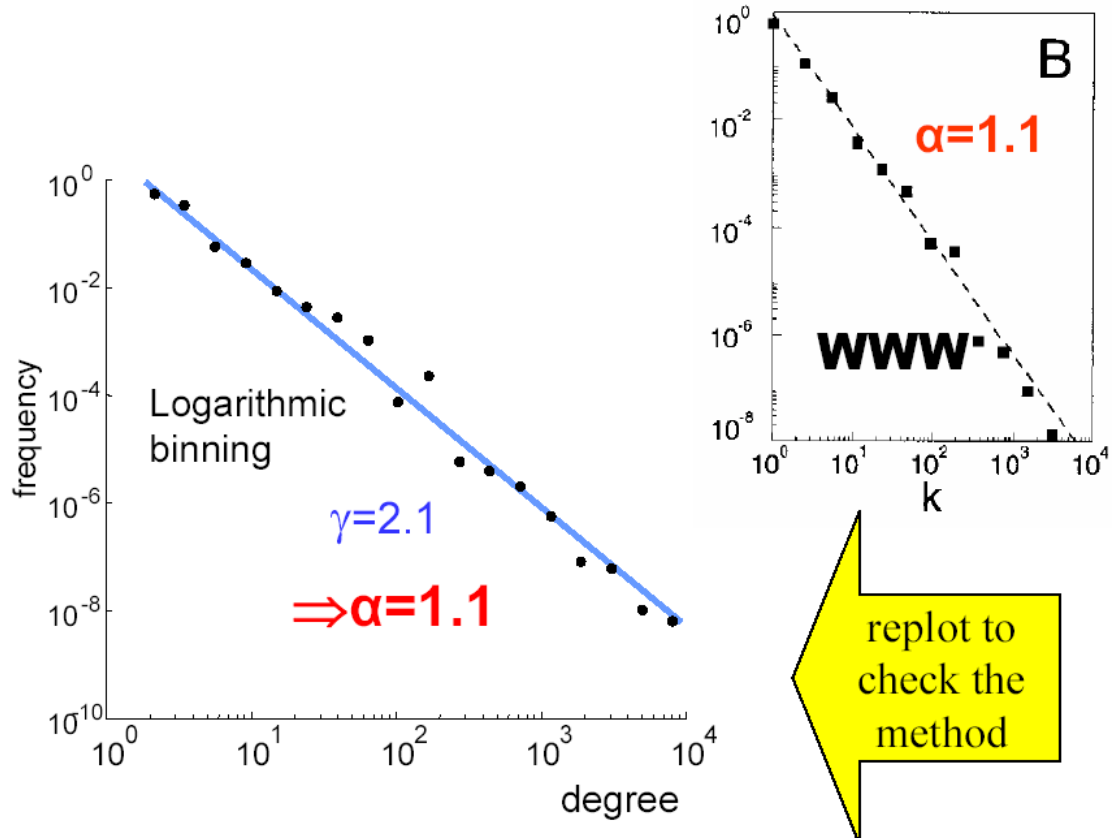


FIGURE 12

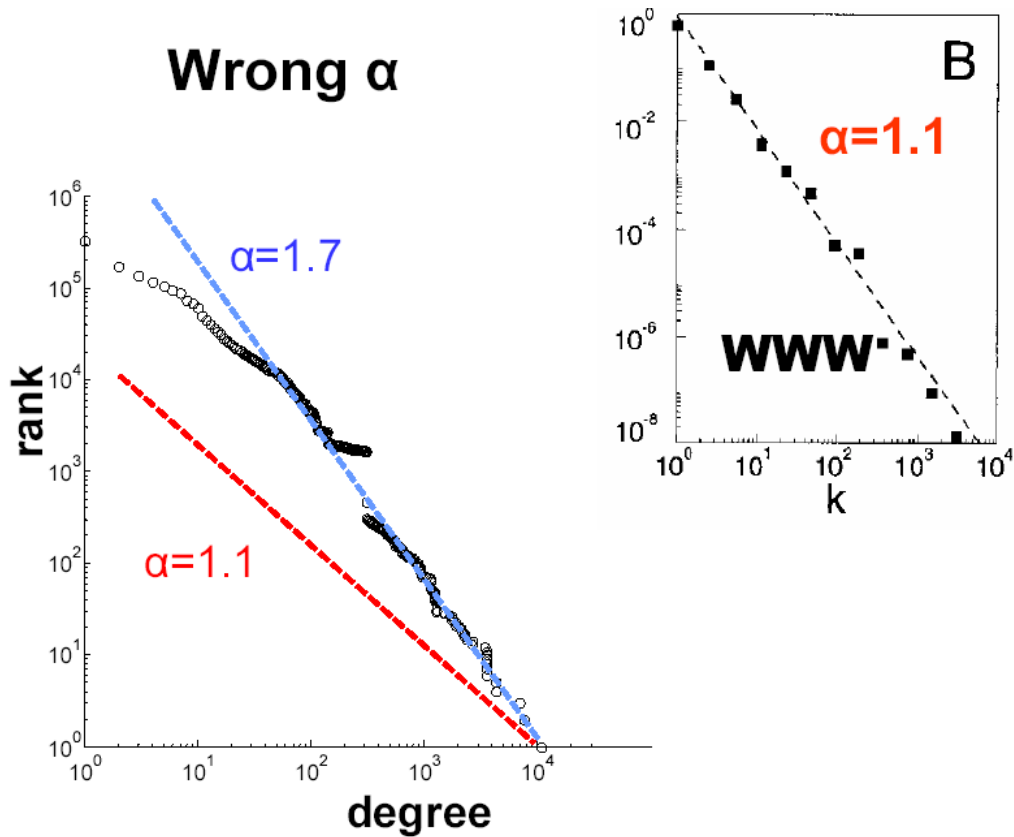


FIGURE 13

Is it even possible for a router topology to have high variability and no degree? You could do it. Figure 14 is a chart of Abilene, which is the Internet 2, and Figure 15 is actually a weather map taken on November 8. There are a lot of Internet experts here; if you look at the Abilene weather map, you might find this picture quite peculiar. If you ask why, it is because it is running at about 70 to 80 to 90 percent, in some places, saturation. You never see that. Why is that? This is the day we broke the land speed record, and we were using this part of the network, and nobody even noticed.

Internet2 Abilene Weather Map

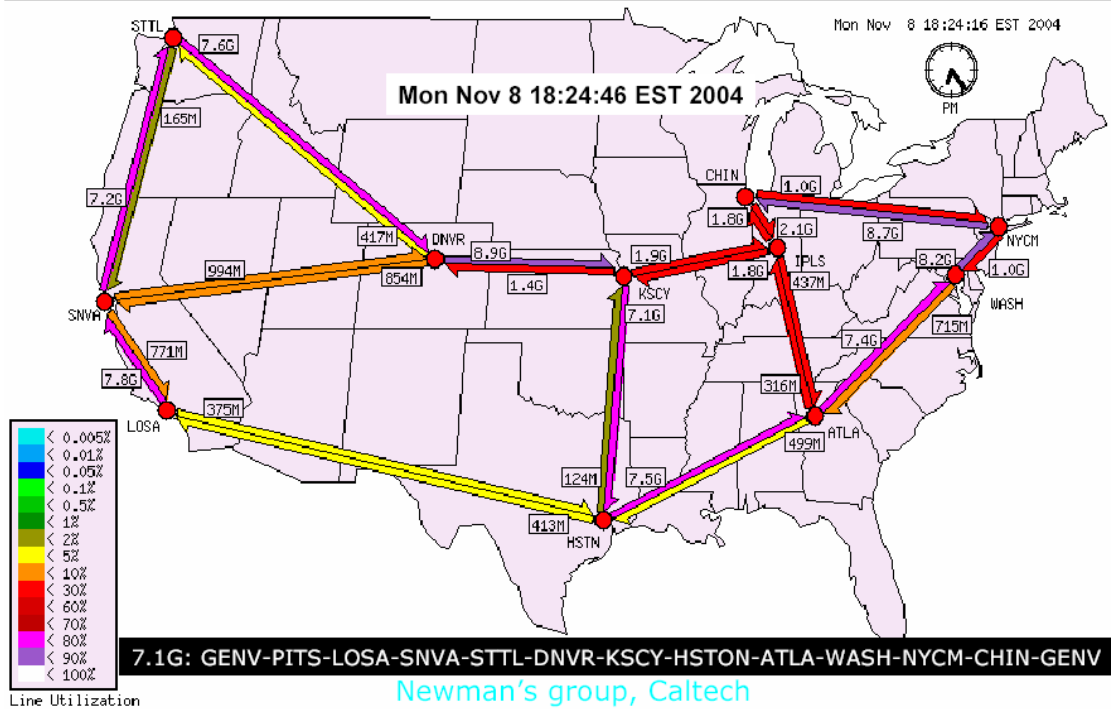


FIGURE 14

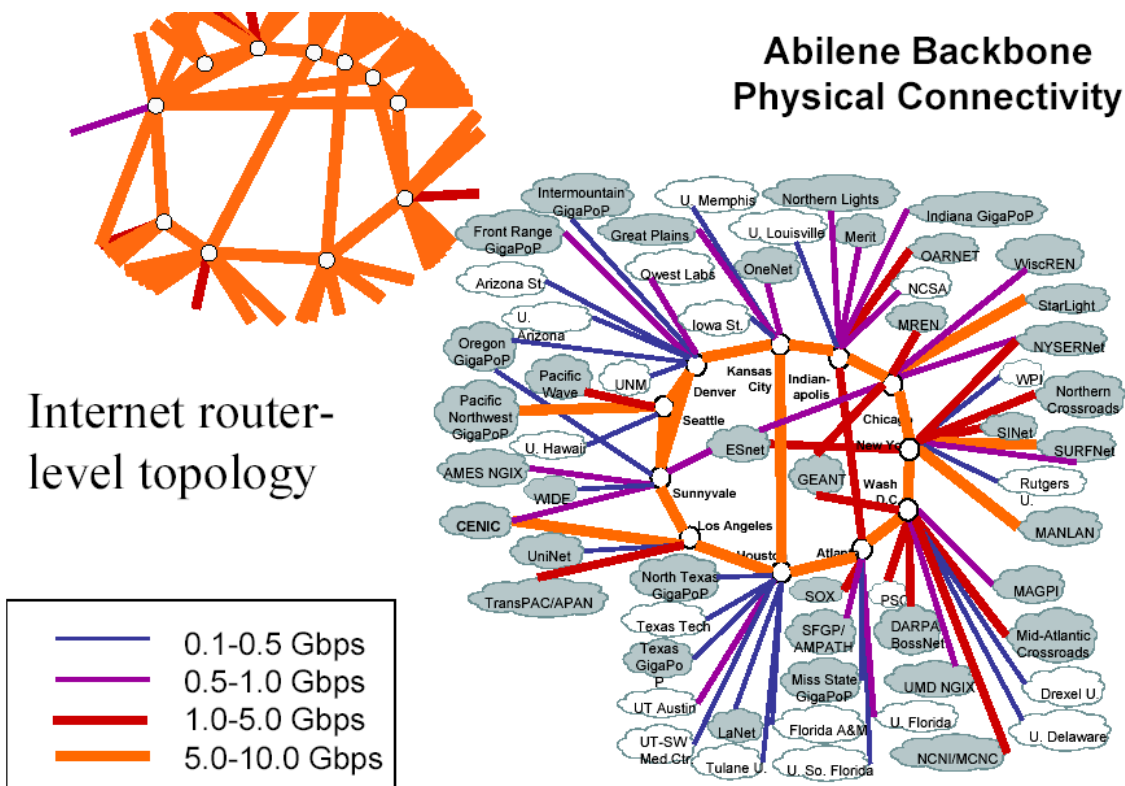
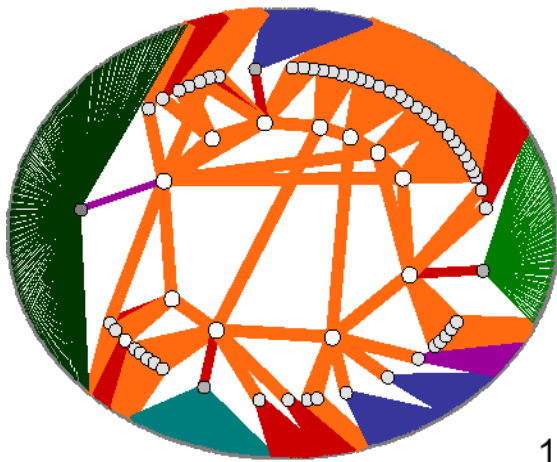


FIGURE 15

Here is a little advertisement for the FAST protocol. Figure 16 shows the topology. Everybody in the U.S. academic system uses this network, and it is typical of an ISP backbone. There are lots of these. We are going to pull out that backbone and make a greatly compressed network out of it by just adding edge systems. Yes, you can create a network that has a power law degree distribution, and the next slide shows such a power law. But following that, Figure 16 is another graph with exactly the same degree distribution. These couldn't be more different, so the degree distribution of a network tells you almost nothing about it, particularly if it is high variability. If it is high variability it gives you more possibility to do more and more bizarre things, and you can quantify the extent to which, in the space of graphs, these are at opposite extremes. Again, they couldn't be more different. On the left we have a low-degree core and high-performance robustness, which is typical of the Internet. On the right-hand side you have got these high-degree hubs, horrible performance and robustness. If the Internet looked anything like the right, it wouldn't work at all. Of course it would fix our problem with spam, because you wouldn't get e-mail; you wouldn't get anything. So, the Internet looks nothing like that, and yet, apparently, a lot of scientists think it does. Certainly no one in the Internet community does. In

fact, there is high variability everywhere.



Low degree
mesh-like core

High variability
edge systems

Completely different
networks can have the
same node degrees.

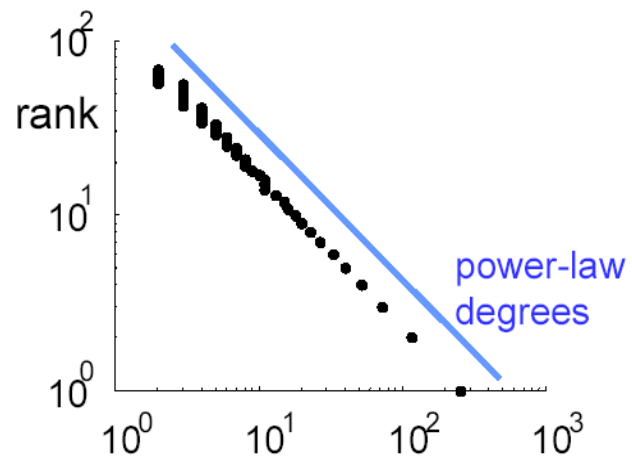


FIGURE 16

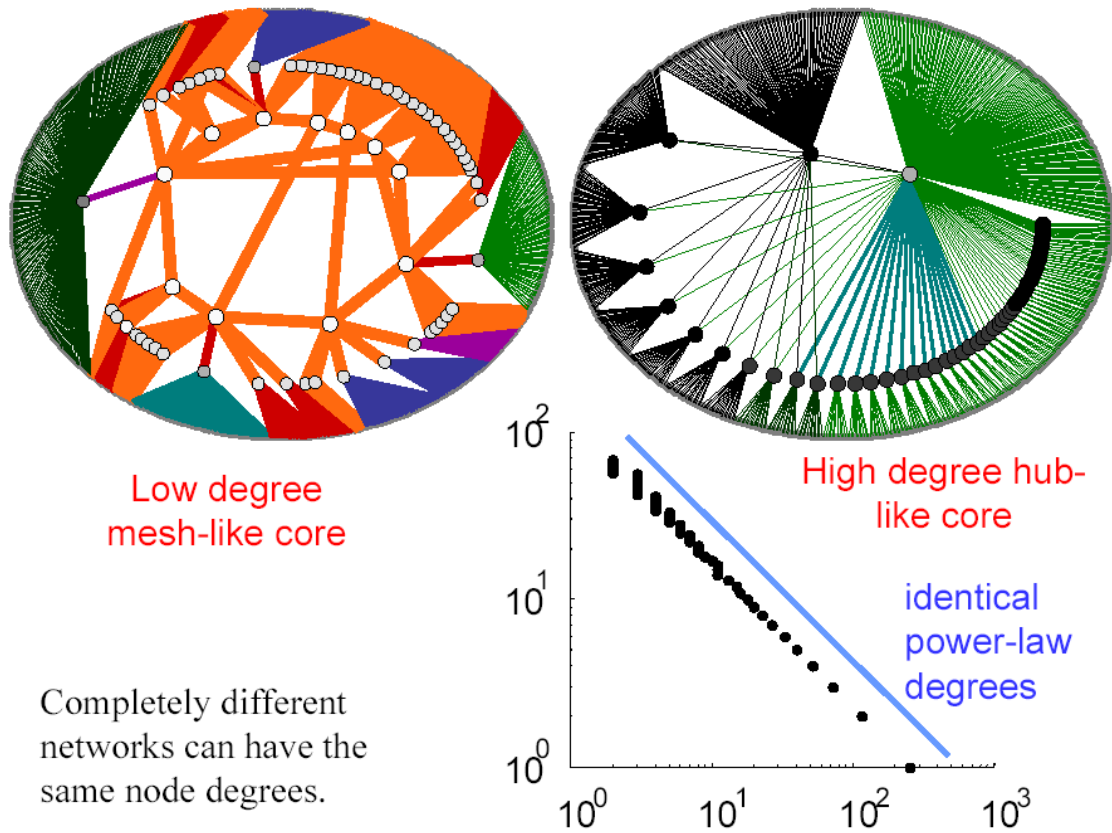


FIGURE 17

I have shown you a bunch of cases. You have to work pretty hard to find an example on the Internet or an example in biology that doesn't have high variability. Ironically, the weird thing is, it is like those with low variability have been found and then presented as having power laws—very strange. Anyway, there are enormous sources of high variability, and it is connected with this issue of robust yet fragile, and I want to talk about that.

Everything I am going to say should appear obvious and trivial provided you remember what you studied as an undergrad either in biochemistry, molecular biology, or engineering textbooks. I don't know if everybody has had this undergraduate level, but I am going to assume only undergraduate level. You have to continue to ignore what you read in high-impact journals on all these topics. We have already seen that for power laws, but you have to temporarily ignore most of these things. Maybe you can go back and think about it later.

Imagine that your undergraduate biochemistry and molecular biology textbook was just okay, and maybe like *Internet for Dummies* or whatever you read about the Internet. Maybe other people will talk about it. Figure 18 is a typical graph we get of a metabolism, where we have

enzymes and metabolites. Again, I am going to hope that you know a little bit about this, because I will talk about it at a very simple level.

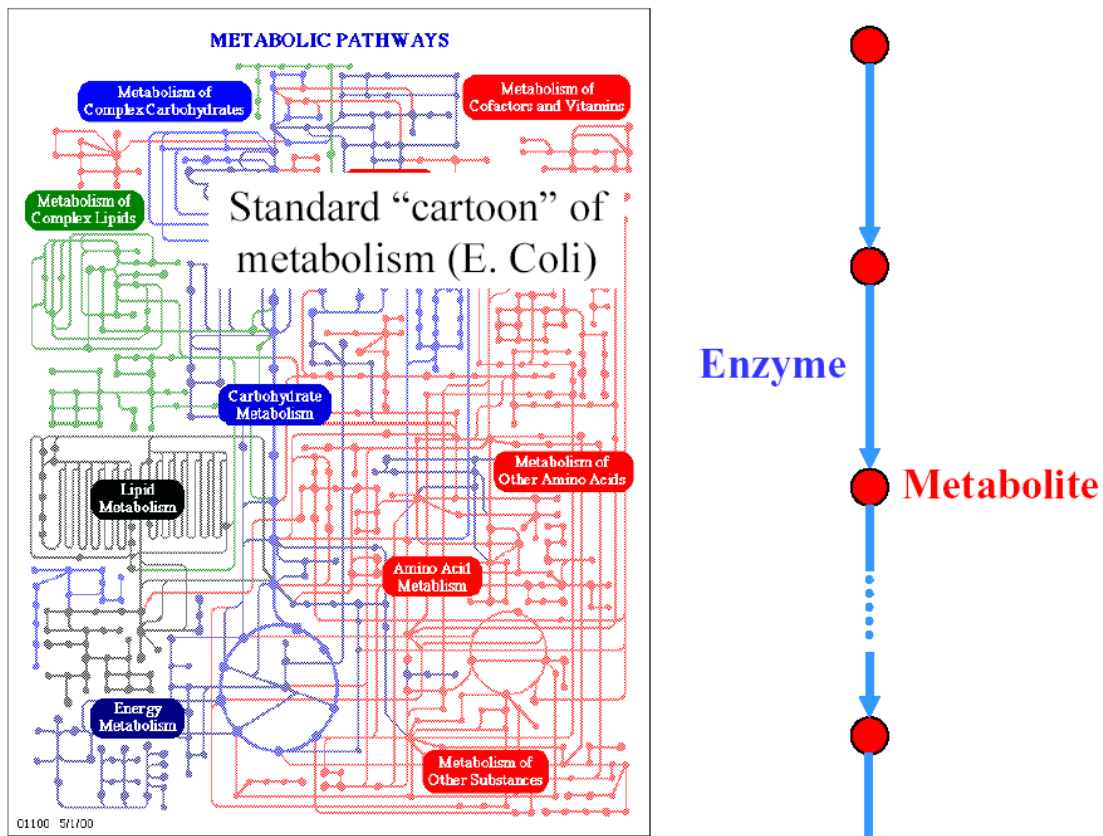


FIGURE 18

In Figure 19, I zoom in on the center part of this, which is glycolysis. Experts in the room will notice that I am doing a little weird thing like an abstract version of what bacteria do, and I am going to stick to bacteria since we know the most about them.

Precursors

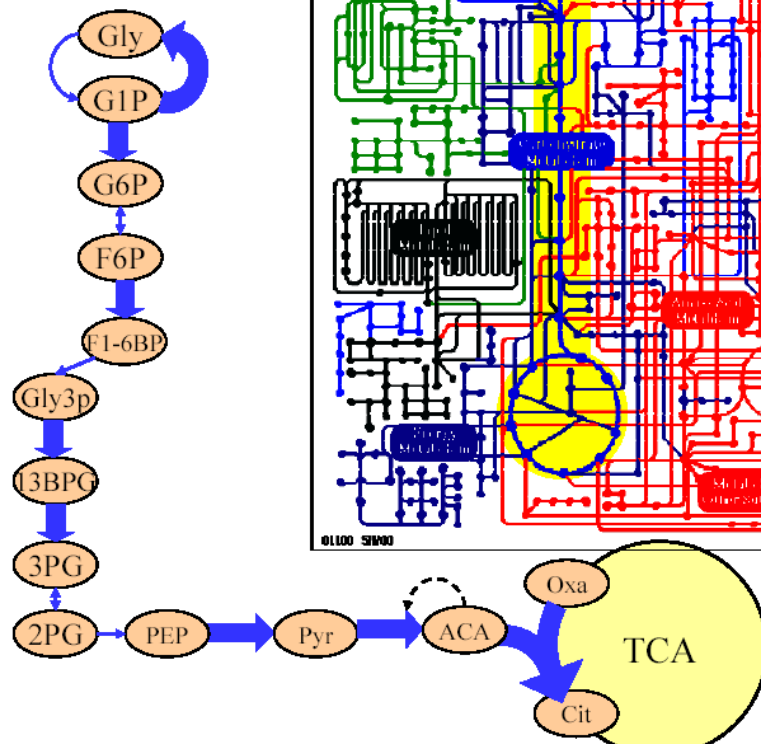
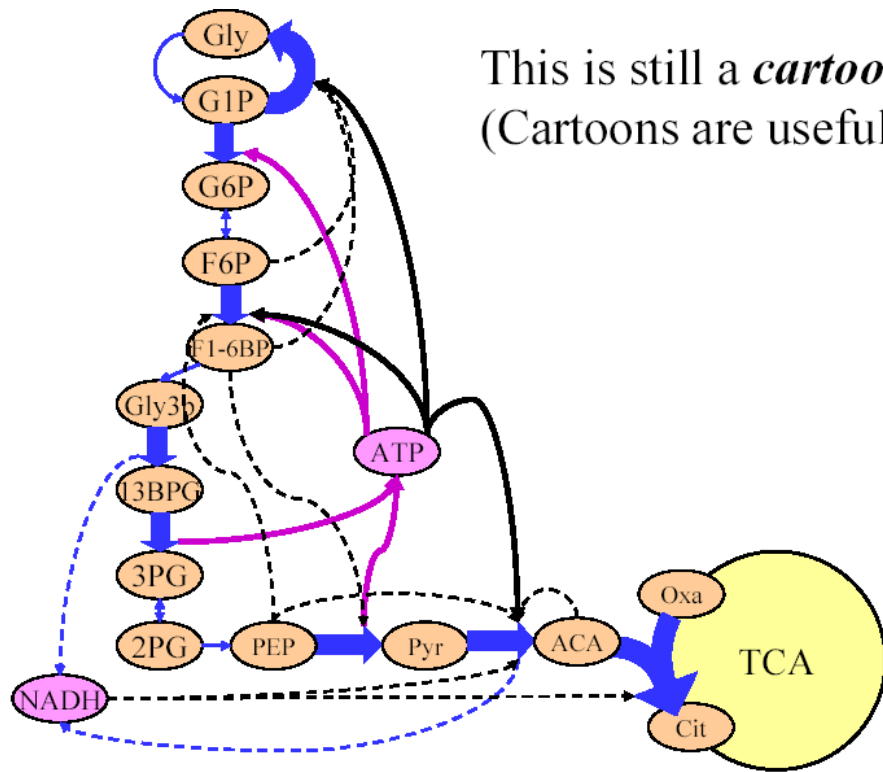


FIGURE 19

What is left out on a cartoon like this? First of all, it is autocatalytic loops, which are fueling loops. In this case you make ATP at the end, but you have to use it up at the beginning, and you will see that isn't quite right, but it gets you the spirit. There are also regulatory loops. Inserting them, we get a more complicated cartoon, shown in Figure 20. Cartoons are useful, but you have to understand what this means. Those lines mean very different things. They certainly wouldn't be comparable, but the problem is if you put all the lines in this thing it would just be black with lines as in Figure 21. People don't draw it that way simply because if you put all the action in you wouldn't see anything but black lines everywhere, but you would know what is missing.



This is still a *cartoon*.
(Cartoons are useful.)

FIGURE 20

If we drew the feedback loops the diagram would be unreadable.

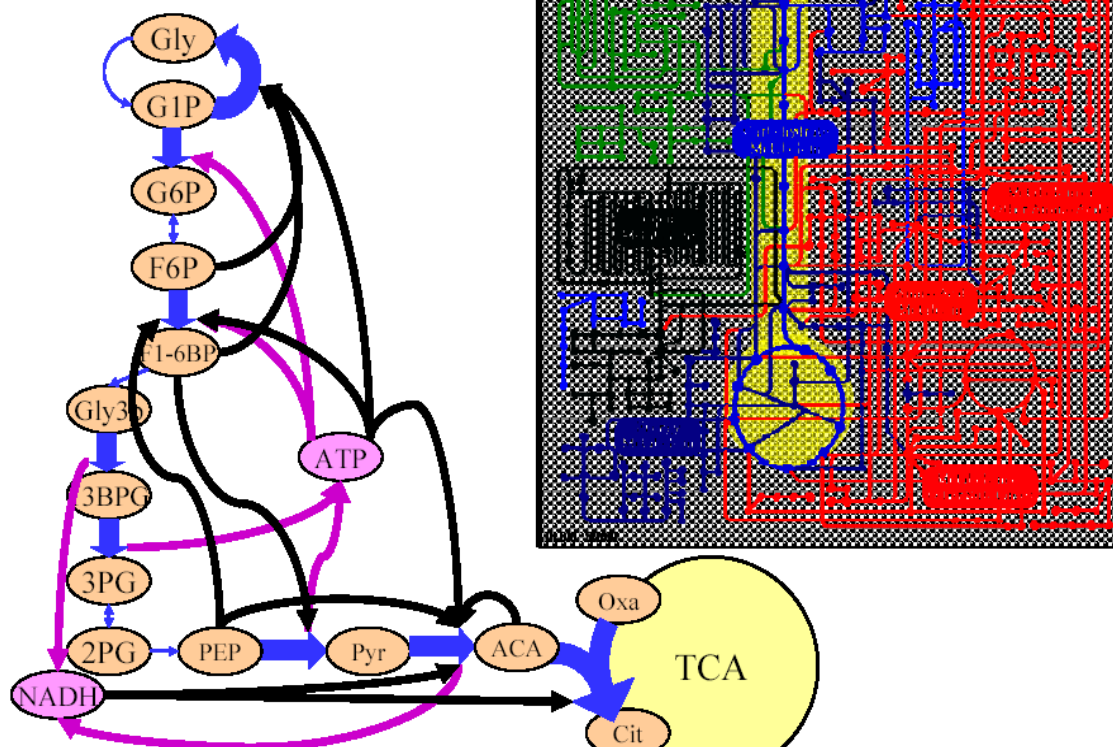


FIGURE 21

Now look at Figure 22: The thing on the left can be modeled to some extent as a graph, but the item on the right cannot; there isn't any sense in which this is a graph. It is not even a bipartite graph. You can't think of it that way; it makes no sense. An enormous amount of confusion results from trying to make this a graph; it is not a graph. At the minimum, you could think about stoichiometry plus regulation.

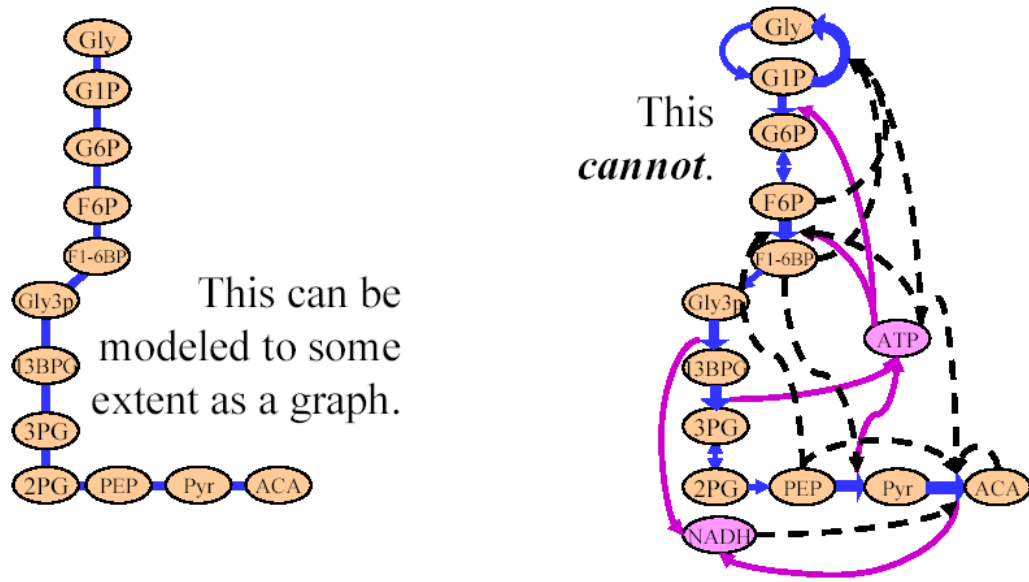


FIGURE 22

Here is what I am talking about: the change in concentrations of mass and energy in the cell is a product of a mass and energy balance matrix, which we usually know exactly, and a reaction flux component which we don't know: It depends on the concentrations.

$$\frac{d}{dt}(\text{Mass \& Energy}) = \begin{bmatrix} \text{Mass \&} \\ \text{Energy} \\ \text{Balance} \end{bmatrix} \begin{bmatrix} \text{Reaction} \\ \text{flux} \end{bmatrix}$$

FIGURE 23

Figures 24-26 show elements need to be represented by these two terms. The stoichiometry matrix could be represented as a bipartite graph or a matrix of integers. Again, this is standard stuff. The reaction flux has two components, and that is one of the things I want to talk about. First of all, you have regulation of enzymes. The enzyme levels themselves, which are the enzymatic reactions, are controlled by transcription, translation, and degradation. On top

of that they had another layer of control, which is allosteric regulation of the enzymes themselves. Enzyme levels are controlled, and the enzymes themselves are controlled. The rates of the enzymes themselves are controlled also. There are two layers of control, and you might talk about how those two layers of control relate to the two layers of control on the Internet, which is TCP and IP, why they are separated, and what they do.

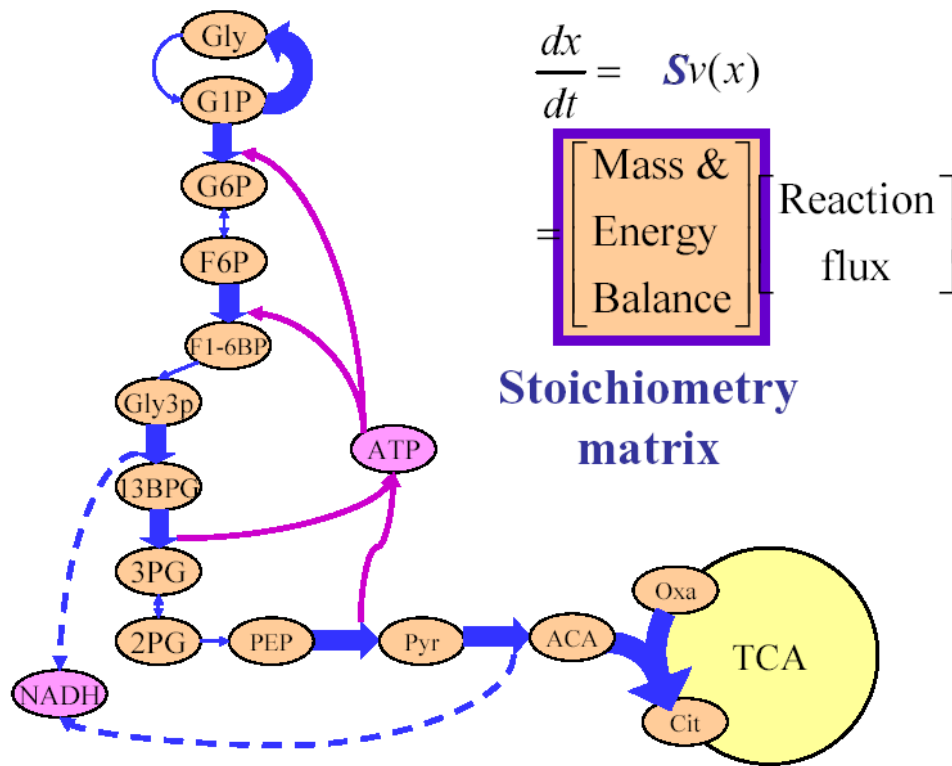


FIGURE 24

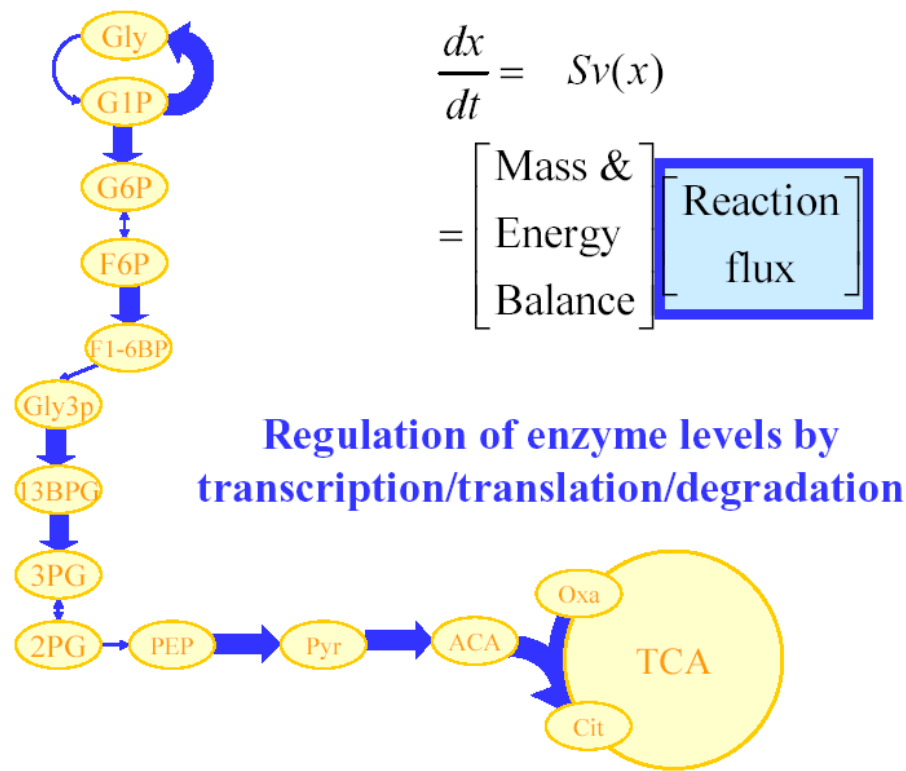


FIGURE 25

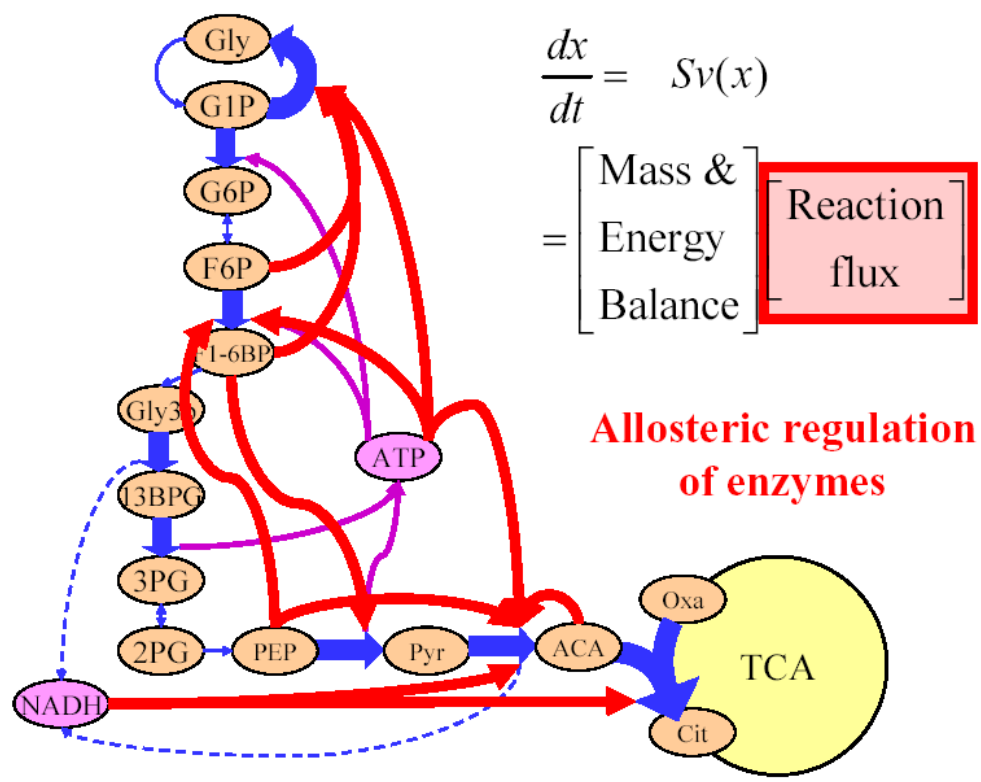


FIGURE 26

What I want to talk about is that these systems have universal organizational structures and architecture. As shown on Figure 27, one of these we will call the bow tie, which is the flow of energy and materials, and another is the hour glass, which is the organization and control. These both coexist and we will see how they fit together. I want to look at Figure 26 and talk about precursors and carriers, again, the standard terminology. Precursors are those things connected by the blue lines, the basic building blocks for the cell, and the carriers are things like ATP, NADH, carriers of energy, redox and small moieties, again, standard terminology.

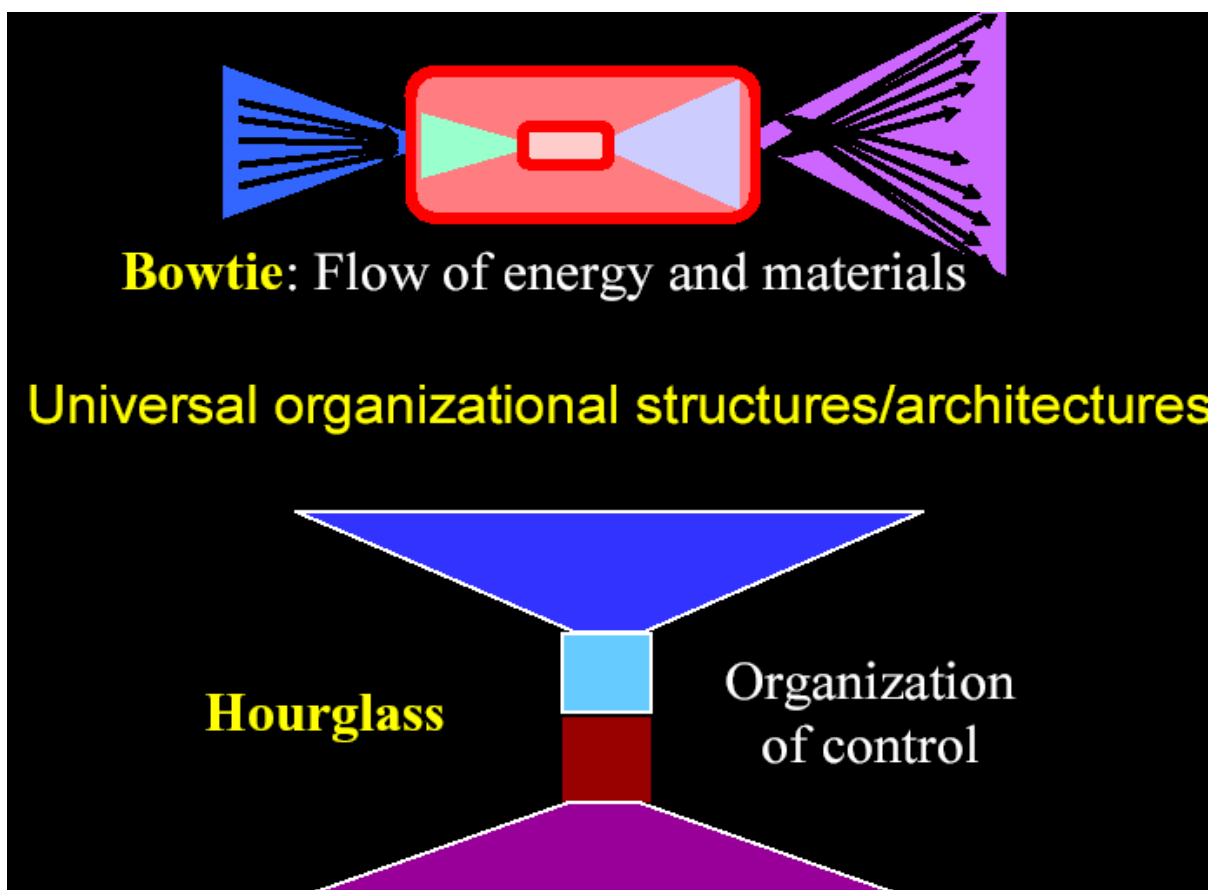


FIGURE 27

Figures 28-32 illustrate how to map catabolism into this framework. What catabolism does is take whatever nutrients are available and break them down to make precursors and carriers. Those are then fed into biosynthesis pathways to create the larger building blocks of the cell. This is core metabolism.

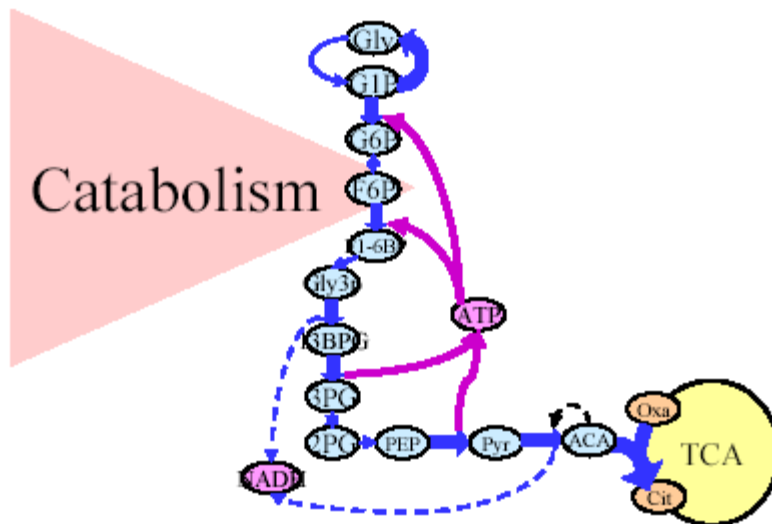


FIGURE 28

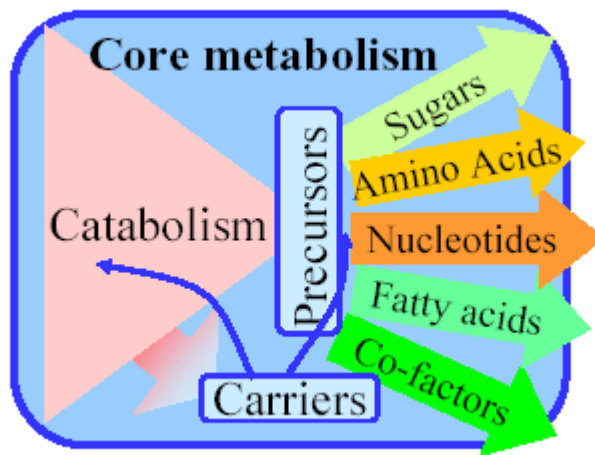


FIGURE 29

That core metabolism then feeds into the polymerization and complex assembly process by which all of these things are made into the parts of the cell, as shown in Figure 30 below.

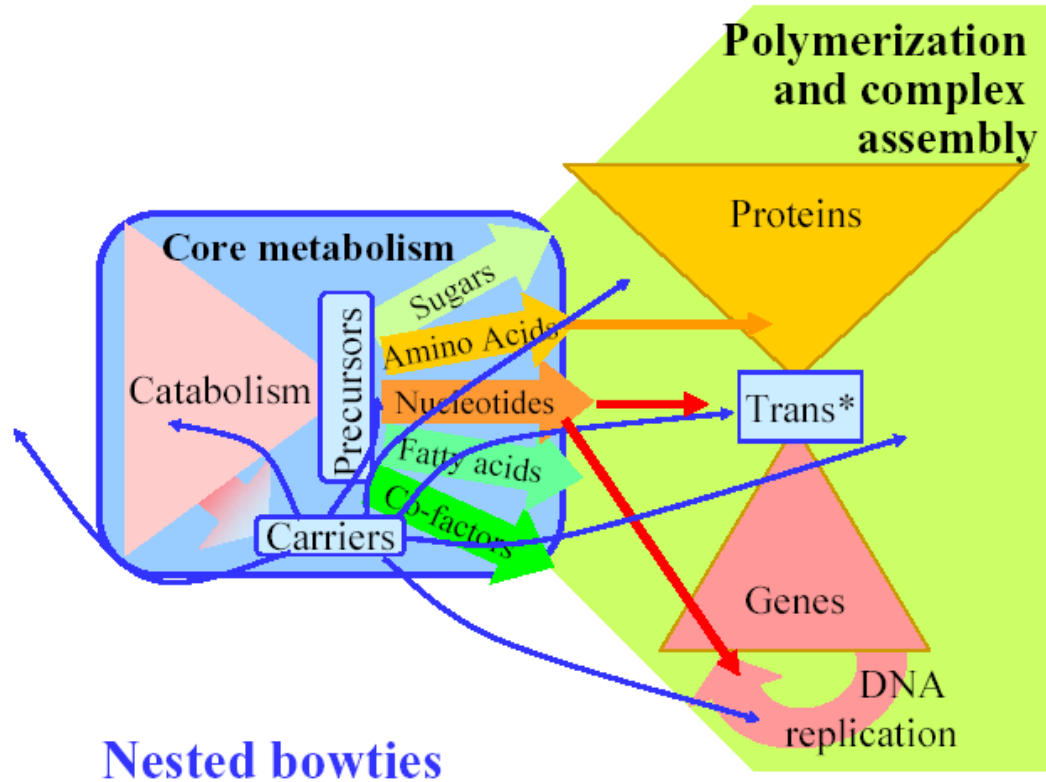


FIGURE 30

Of course, if you are a bacterium you have to do taxis and transport to get those nutrients into catabolism, and then there are autocatalytic loops, because not only are those little autocatalytic loops inside for the carriers but, of course, you must make all the proteins that are the enzymes to drive metabolism, and then there are lots of regulations and controls on top of that. So, schematically, we get the nested bow ties in Figure 31.

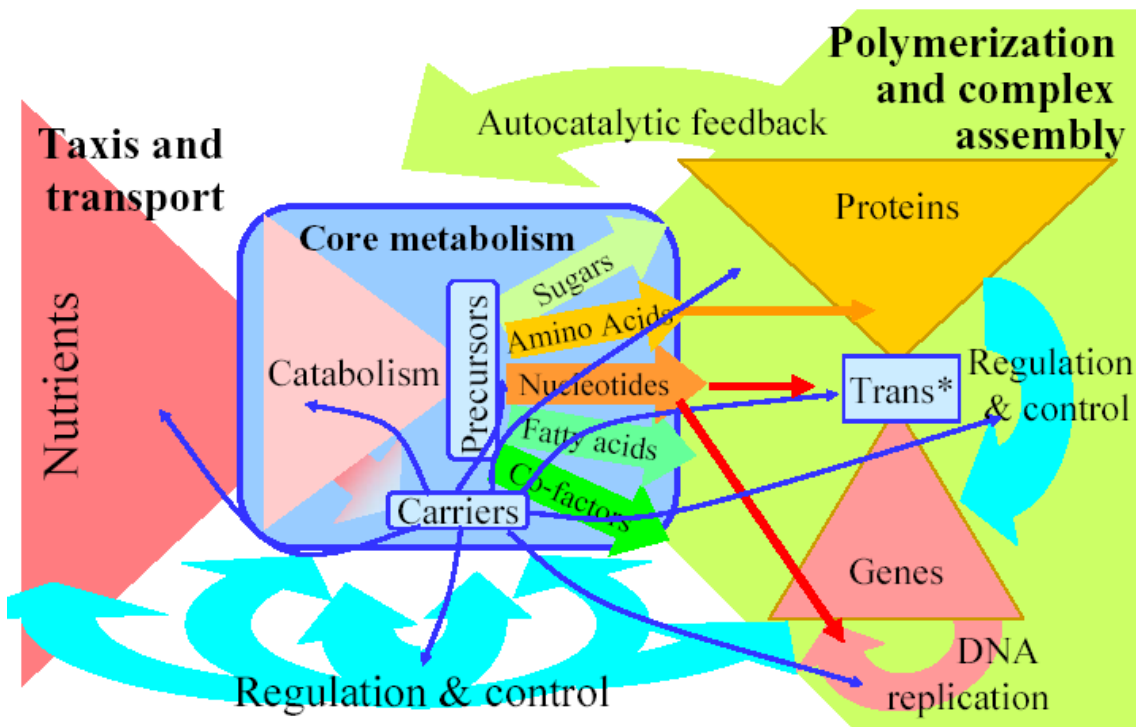


FIGURE 31

What is important about this is that you have huge variety at the edges and basically no variety in the middle. What do I mean by that? If you look at all the things that bacteria encounter in the environment, all the chemicals, all the ions and all the possibilities, you couldn't list them all; it is basically infinite. What do they eat? Well, they eat nearly anything. I mean anything that has a reasonable redox status, some bug will eat it. Since we have cultured only 1 percent of all the bacteria in the world, we have no idea. They probably eat just about anything. They eat uranium, which is a cool thing to do, too. There is a huge variety in what they will eat.

What are these final building blocks? There are about 100, and they are the same in every organism on the planet, every cell on the planet, exactly the same, but there are only about 20 in the middle, again, the same in every cell on the planet, and those are the precursors and the carriers, about 12. Again, it depends on how you count. I am counting AMP, ADP and ATP all as the same. It is like a little charged up battery, so that is how I am counting. If you want to count every one of them, you will get 30 or 40, but it is a small number. When you go on to this side again you get an almost unlimited number of macromolecules. I am grossly underestimating, but in a given cell you will have on the order of a million distinct macromolecules that are produced by the polymerization and complex assembly. What you have is this huge variety. If you go and look at almost any kind of data from these systems—time constants rates, molecular

constants, fluxes, variable, everything you look at—you'll find a huge variability, variability over orders of magnitude. If you take this data in a particular way you are likely to get power laws, again, simply for the same reason that, if there was low variability, you would likely find Gaussians.

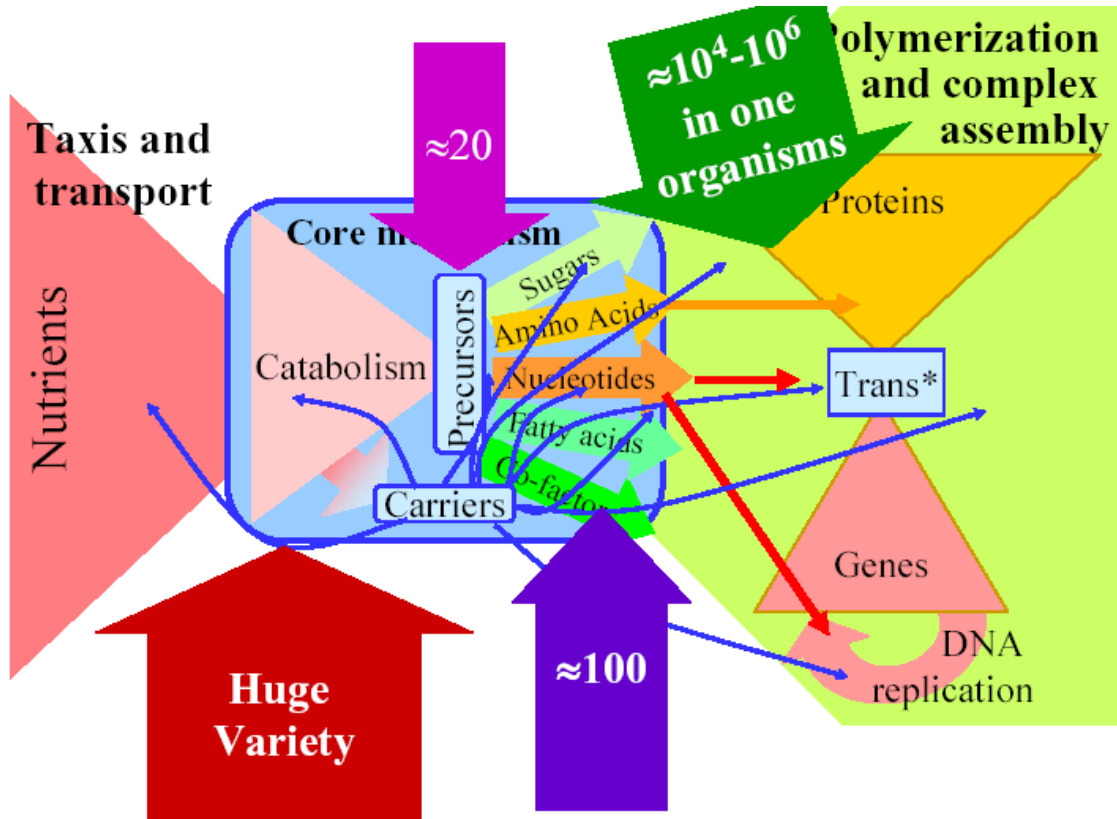


FIGURE 32

There is a very limited critical point of homogeneity. It is homogeneity within the cell, and within every cell on the planet, so this critical point of homogeneity is essential. What you get is this nest of bow ties. We have this big bow tie and inside it you have little bow ties, and you have also regulation and control on top of it. The question is why? Why this structure? What is it about this? What does it confer? Since every cell on the planet has the same architecture, you could say it is just an evolution accident or it is an organizational principle. Which is it? What do you compare it with? Hopefully, we will find something on Mars that will have different handedness, and then we know that it didn't have common descent and then we can actually compare them. But in the meantime, we don't have much to compare. What to compare with and how to study it? I think what you compare it with is technology. How do you

study it? You use math, which is why we are here, I assume. It turns out we build everything this way, so if we look at manufacturing, this is how we build everything. We collect and import raw materials, you make common currencies and building blocks, and you undergo complex assembly.

As an example of an engineered system, let's look at power production, as shown in Figure 33 and Figure 34, the bowtie figure abstracted from it. There is a huge number of ways we make power and lots of ways we consume power, at least in the United States. When I come to the East Coast I don't have to worry about whether my laptop will work. It works because you have this common carrier providing power. If you go to Europe, power is carried in a little different way. Gasoline is a common carrier, ATP, glucose, proton motive force. This is the way we build buildings. There is a huge variety of things you can make stuff from, a huge variety of ways you can do it, and in the middle there is more or less a standard set of tools.

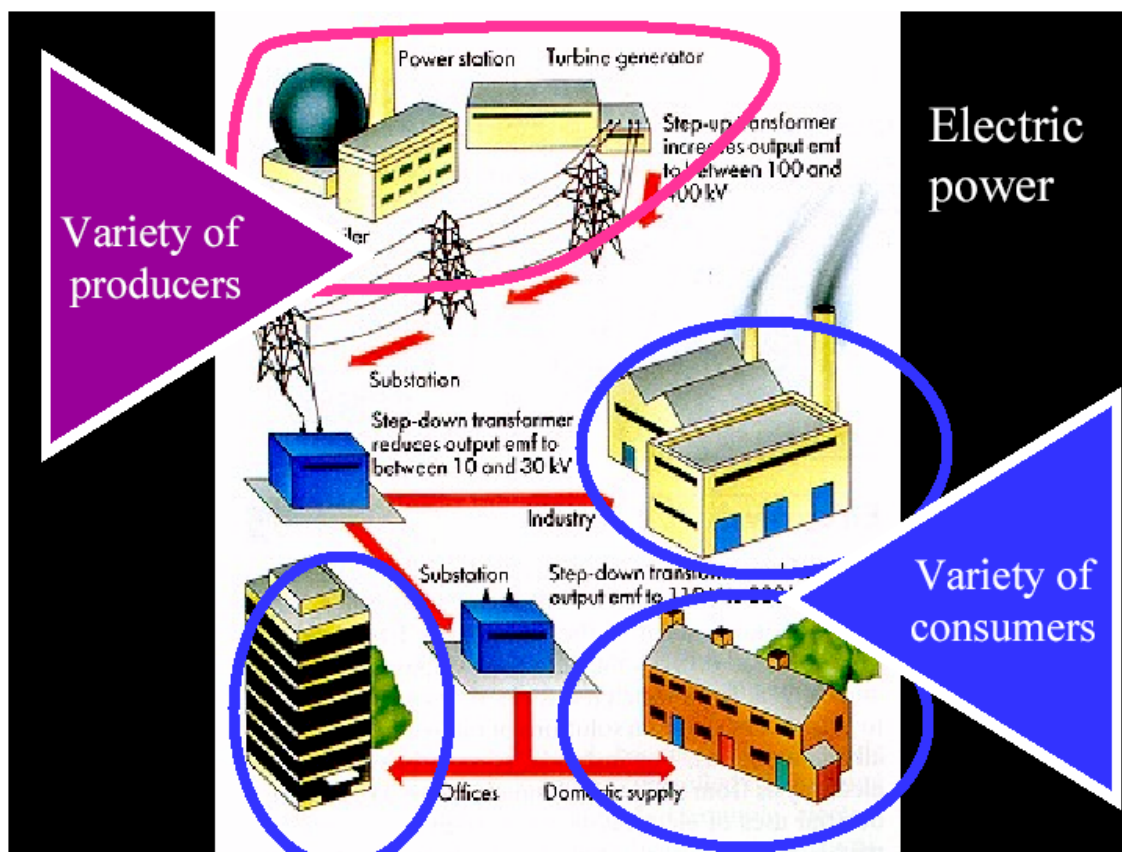


FIGURE 33

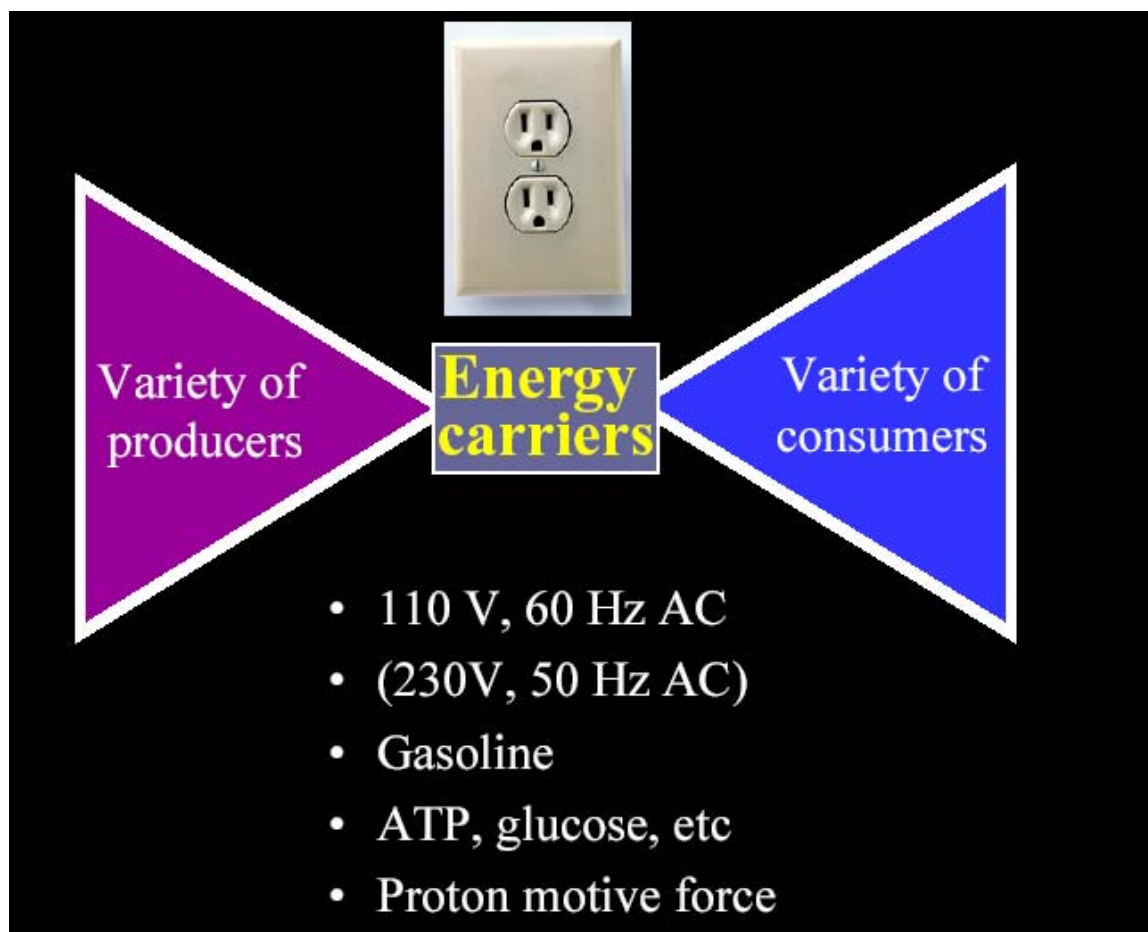


FIGURE 34

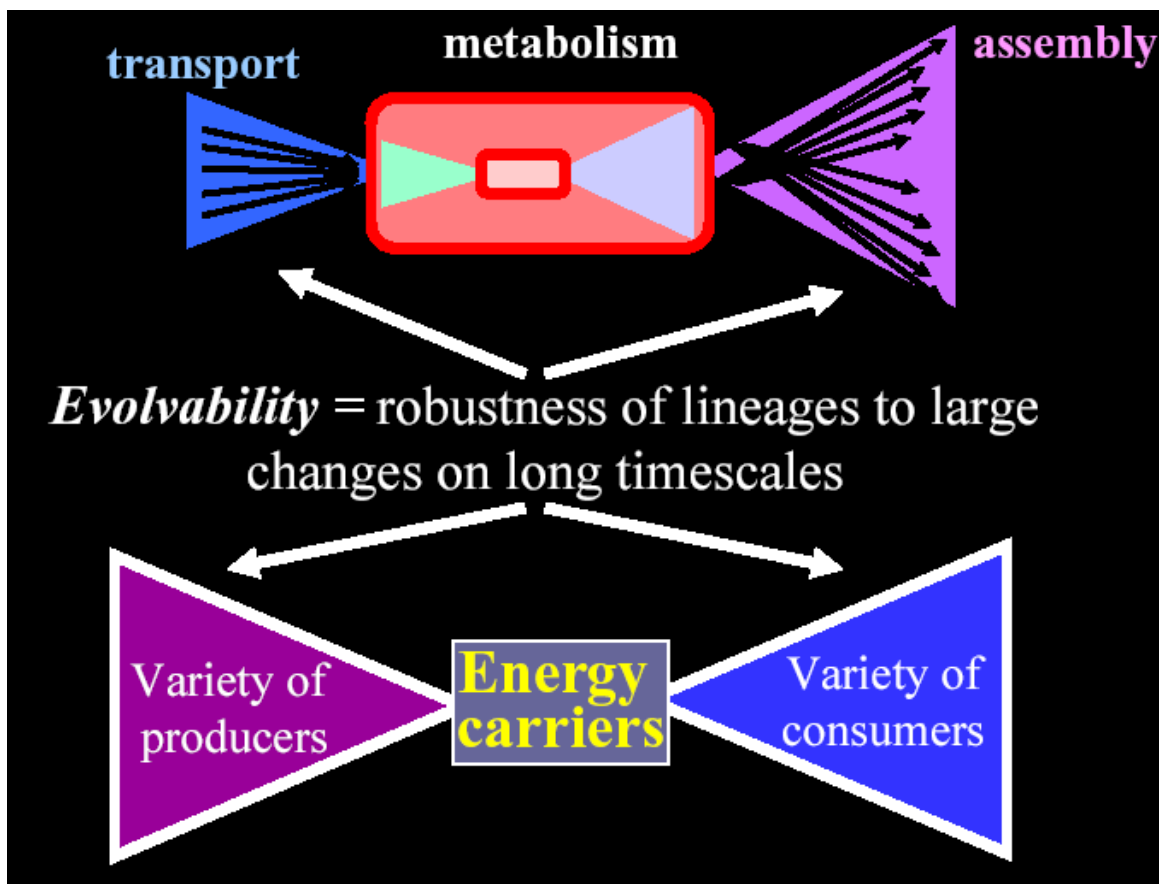


FIGURE 35

With the engineered system, we can go back and see how it used to be. What we look back and see is the evolution of protocols, getting a glimpse of evolving evolvability. Evolvability is the robustness of lineages to large changes in long time scales. I want to think of robustness and evolvability as just a continuum. We will see there is no conflict between them. The idea that a system being very robust makes it not evolvable is an error. The problem is that all this stuff in the middle is usually hidden from the user. The user sees the hardware and they see the applications, but they don't see the protocols in between. It is easy to make up all sorts of nonsense about how it works. This is a universal architecture only in advanced technologies. It has only been since the industrial revolution that we did things this way.

In biology you get extreme heterogeneity; it is self dissimilar; it looks different at every scale. If you zoom in on the middle, the core is highly efficient, but the edges is where you get robustness and flexibility because that is where the uncertainty is. This is shown schematically in Figure 36. If you look at the core you get highly efficient, very special purpose enzymes that are controlled by competitive inhibition and allostery, and small metabolites. It is like this machine is sitting there and the metabolites are running through it—a big machine, little metabolites. At

the edges it is exactly the opposite: you get robustness and flexibility, but you have general purpose polymerases, and they are controlled entirely differently, usually by regulated recruitment. So, you have an entirely different control mechanism, very different styles, and again, this is all standard undergraduate biochemistry, but it is an important observation. If you think about the distinction between this and riveting, you have a machine that sits on the floor and makes rivets. It just sits there and spits out rivets; it doesn't move and you control it by a knob on the back. Once you take the rivets and you want to rivet an airplane, you now take the rivet gun and move around on the airplane. You take the rivet to the system. It is very much similar to core metabolism versus complex molecular assembly. Again, there are these universal arcs.

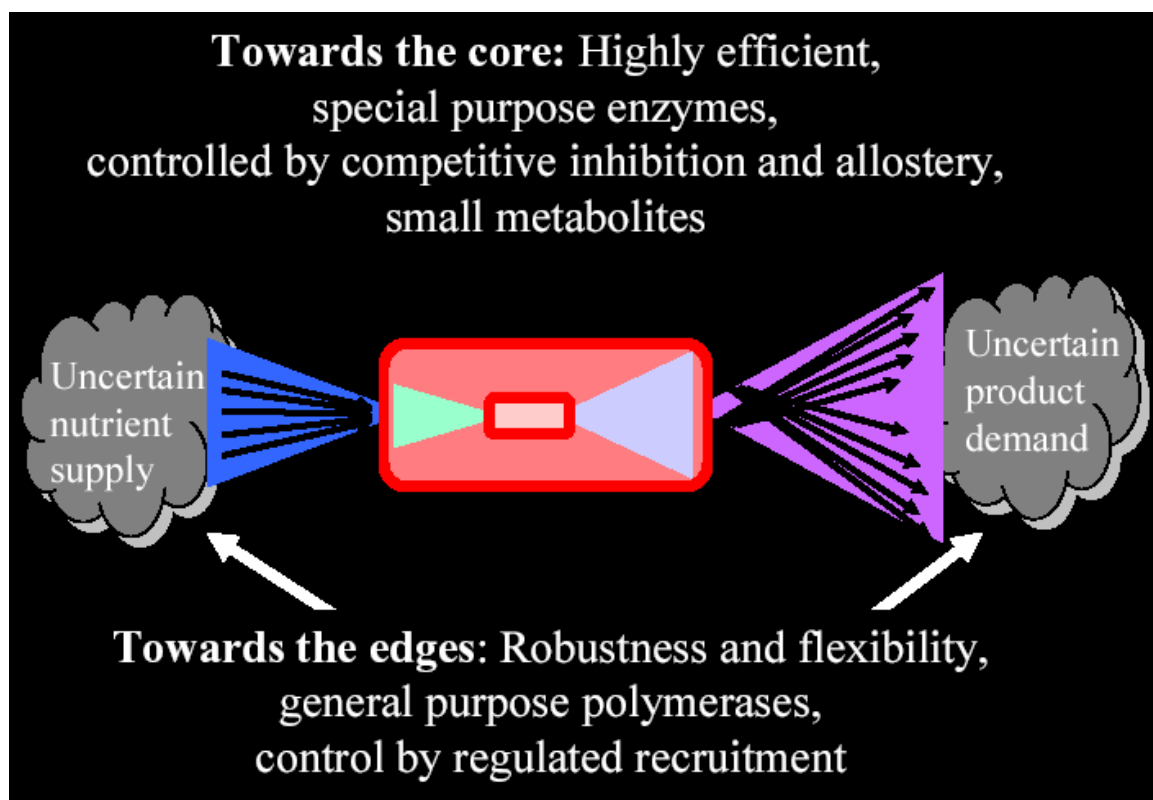


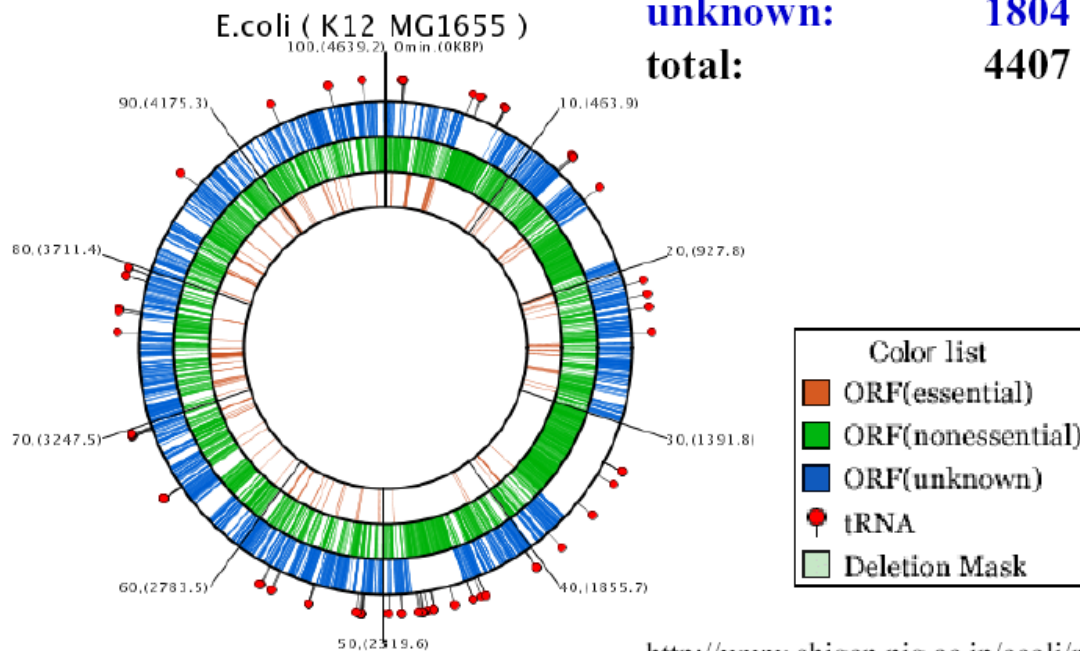
FIGURE 36

Figure 32 isn't to scale. To show its high variability you may ask how would I scale it? We could scale it by flows. We could scale it by whatever. Suppose we scale it by the genome. We can think of this as a gene network, and it holds for *e. coli* with its 4,400 genes (see Figure 37, which is not exactly up to date). About half of those *e. coli* genes have known function. Of the ones with known function, about 10 percent are called essential. Again, I am using the biological notion of essential, which just means if you take out that gene, you can't keep it alive

on life support no matter what. If you scale up that way it turns out the whole genome, to a first approximation, is regulation and control. This is a radically different scaling than I started out with. All the complexity is down here at the bottom. That is not unusual.

Gene networks?

essential: 230
nonessential: 2373
unknown: 1804
total: 4407



<http://www.shigen.nig.ac.jp/ecoli/pec>

FIGURE 37

In technology it is the same way. If you take a modern car, a modern high end Lexus, Mercedes or so on, essentially all the complexity is in these elaborate control systems. If you knock one of those out what you lose is robustness, not minimal functionality. For example, brakes are not essential. If you remove the brakes from a car the car will still move, particularly in a laboratory environment. What will happen if you take it out on the road? You are more likely to crash. There is a standard joke about the biochemist and the geneticist and their brakes. This is not because brakes are redundant; brakes and seat belts are not redundant. In some sense they are, but no engineer would think of them as redundant. They are layered control systems and that is the same story with cells. These non-essential knockouts aren't due to redundancy. That is a fiction. If you knock out this stuff you lose robustness not minimal functionality, and then there are a few things in here that, if you knock out, they are often lethal. If redundancy was

a strategy that bacteria adopted for robustness where would they put it? In the core, but they don't, for the most part, so that is not the explanation. The explanation is straightforward, straight out of standard undergraduate biochemistry. This supplies materials and energy; this supplies robustness. Robustness contributes to complexity much more than the basic functionality so we have to understand this bottom part. We can't just look at the stoichiometry if we want to understand this.

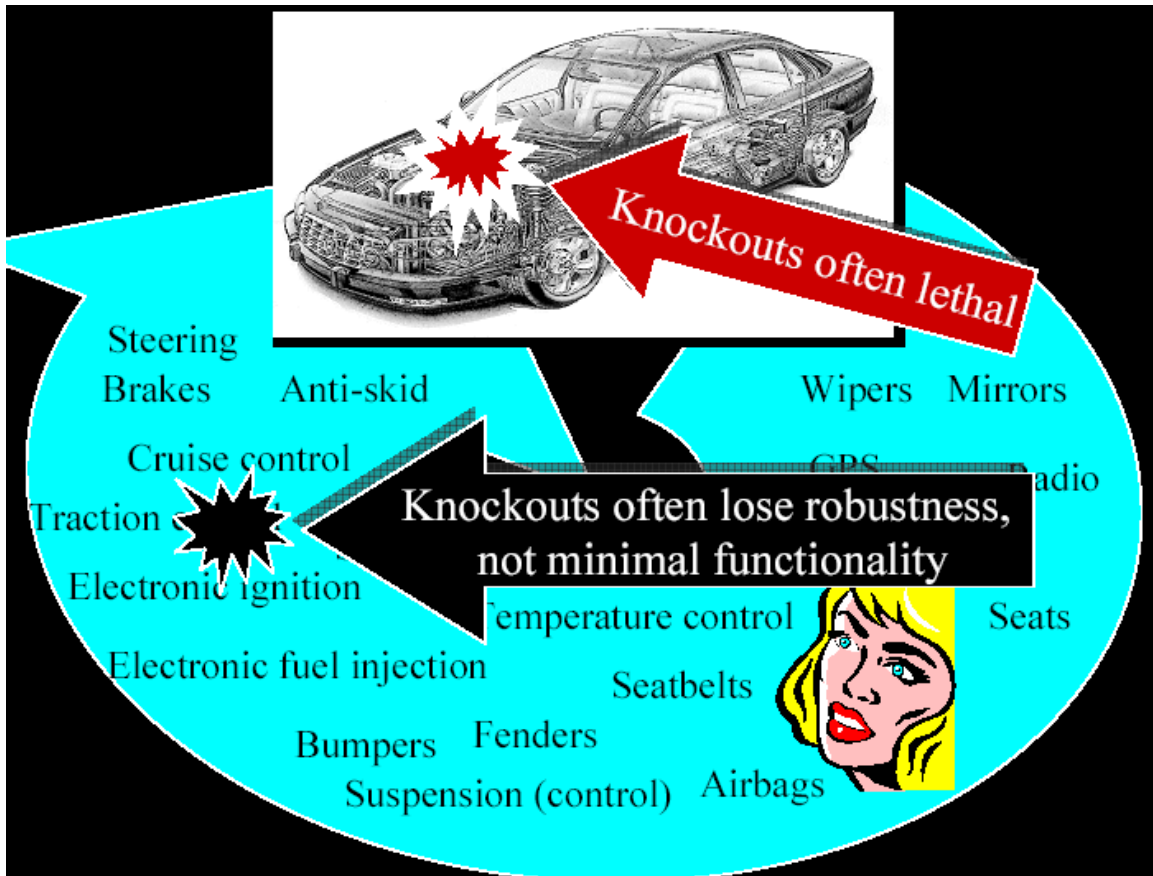


FIGURE 38

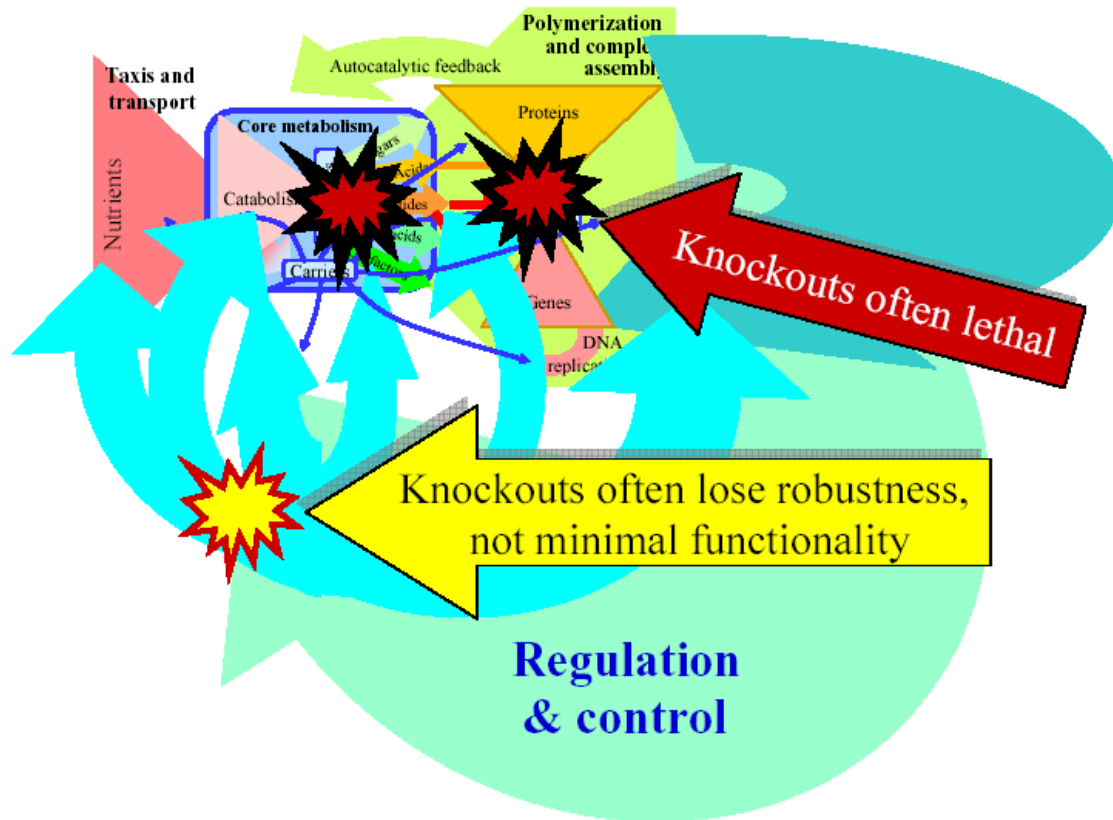


FIGURE 39

We are familiar with the idea that we have this exploding internal complexity, even though it is supposed to appear very simple. Our cars appear simpler. It used to be if you wanted to be a driver you had to be a mechanic. Then you could be a mechanic as an option. Now you can't be a mechanic. You can't fix your own car. In fact, the mechanic can't fix your car without a bank of computers. So, these systems have become much more robust. Your modern high-end luxury vehicle that has all of these control systems will almost never have a mechanical failure. How does it fail? The answer is software. Mercedes Benz automobiles now have terrible maintenance records, but they have eliminated mechanical failures. They have replaced them with software errors, and now they have worse performance than ever. Imagine that some day all our cars on the freeway turn left. That is why we call it robust yet fragile. They are extremely robust most of the time but when they fail they fail big time. What is the nightmare? The possibility is that we will never sort this out and biology might just accumulate the parts list and never really understand how it works. In technology we might build these increasingly complex systems and it will have increasingly arcane failures. Apparently, nothing in the sort of orthodox view of complexity says this won't happen. I believe it won't happen if we do the math right, but we

need new math to do this right.

What are the greatest fragilities of this architecture? It is hijacking, because the presence of common parts makes it easier to develop the ability to hijack. If you have a common user interface anybody can come in and hijack it. I can come in, plug in, and steal your energy. Viruses come and plug right in, and they don't have to worry about which *e. coli* they are in. They are all the same with respect to this; viruses only have to worry about the immune systems. So, what you see is that these things have very characteristic fragilities.

Let me briefly talk about eukaryotes. In fact, let's talk about us for a minute. I don't know much about them but since we are staying at the undergraduate level I am probably okay. There are plenty of people in the room to fix all the things I say wrong. What is the fragility of us as a whole? It is obesity and diabetes. At a physiological level we have a bow tie architecture with glucose and oxygen sitting in the middle, as shown in Figure 40 below.

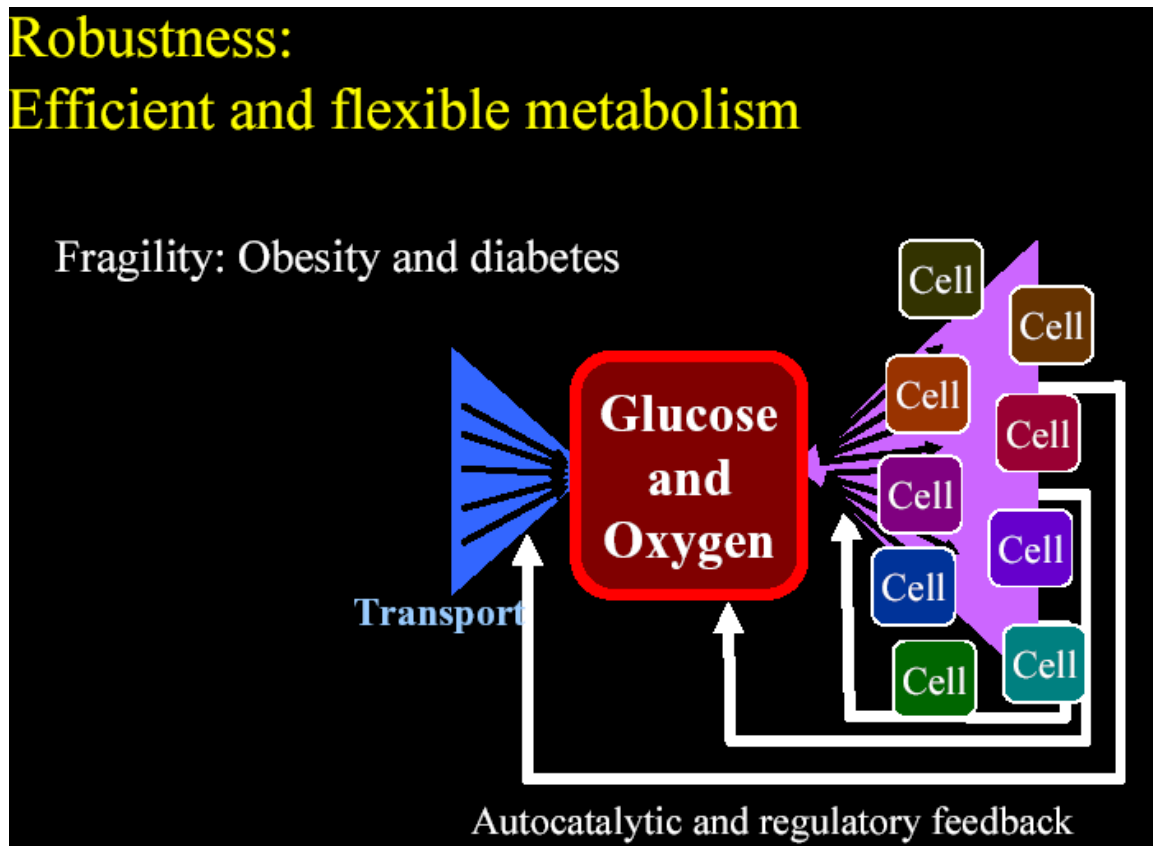


FIGURE 40

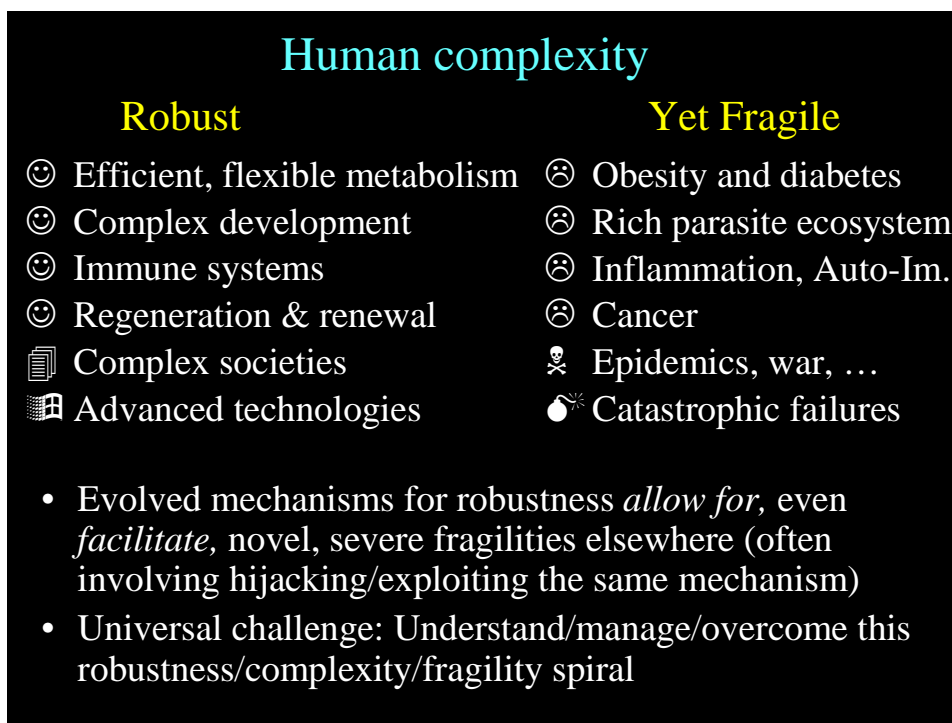


FIGURE 41

What happens is if you look at all the problems on the right (Figure 41), they are the direct consequences of mechanisms on the left. So, we have efficient flexible metabolism, great 500,000 years ago when you are starving between trips to hunting or whatever. You put that in a modern environment and you are not running around any more and you get obesity and diabetes. You need complex development in order to create us. That makes a rich parasitic ecosystem. You then have an immune system, so now you have horrible autoimmune diseases, some of the worst diseases. One form of diabetes is caused by autoimmune disease. You need regeneration and renewal in the adult and that, of course, can get hijacked. Hijacking is the greatest fragility. You get cancer; you get complex societies and we have epidemics. Perhaps others will talk about that. It looks like robustness and fragility are a conserved quantity in the sense that when you try to make something more robust, you make something else more fragile, and that is actually true. If I have time I may say something about it. Again, there is an undergraduate-level story that you can talk about, that there is a conservation law for robustness and fragility.

Let's look a little bit at transcripts for regulation for just a second. Some of my favorite work is this work on network motifs by Uri Alon and his co-workers. What is a motif? Figure 42 shows a transcriptional regulatory motif for heat shock. What you have is the gene which codes for the alternative signal factor that then regulates a bunch of operons. I am going to focus on an abstract version of an operon. What does this do? Typically, your proteins are supposed to be

folded, and this is a reason why you have fevers, because your body is trying to unfold the proteins in the bacteria that are attacking you. That is probably not all there is to it but, as I said, there are people here who can fix the things I say wrong. What is happening is you have chaperons that both fold nascent proteins and refold unfolded proteins, and you also have proteases that degrade particular aggregates and then they can be reused. This is a system and if this is all you had, you could survive heat shock provided you made enough of these things.

Network motifs in the transcriptional regulation network of *Escherichia coli*

Shai S. Shen-Orr¹, Ron Milo², Shmoolik Mangan¹ & Uri Alon^{1,2}

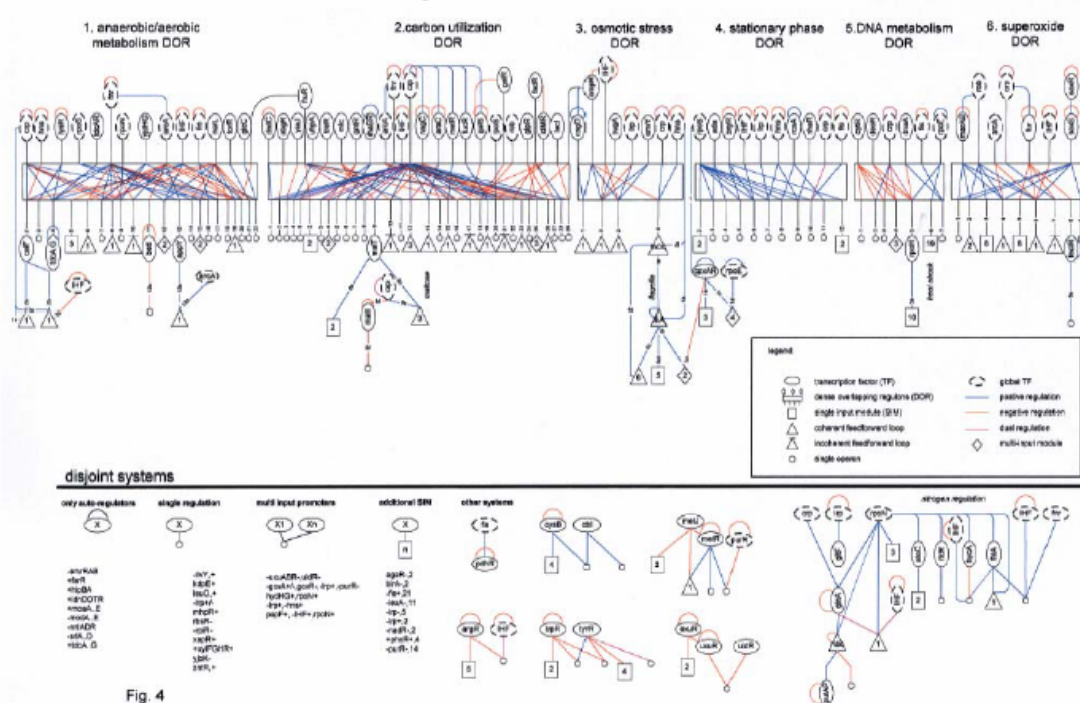


Fig. 4

FIGURE 42

That is not how *e. coli* does it; what *e. coli* does is builds elaborate control systems, as shown next. First of all, the RNA itself is sensitive to heat; it unfolds and is transcribed more rapidly under the presence of heat. That makes more sigma 32 that makes more chaperons. In addition, the chaperons sequester the sigma 32. It has hydrophobic residues so it looks like an unfolded protein. If the chaperons suddenly get busy having to fold unfolded proteins, the sigma 32 is free to go make more, a negative feedback loop, a very cleverly designed one, and then there is an additional feedback loop associated with degradation. So, there is an elaborate control system that is involved in protein protein interactions. What you see here are two layers of

control. There is the transcription regulation on the left, and there is this protein-protein interaction, which allosteric regulation of enzymes is a similar thing on the right. You have got these reaction fluxes, regulation enzymes, and then you have got this two level control, allosteric regulation of enzymes.

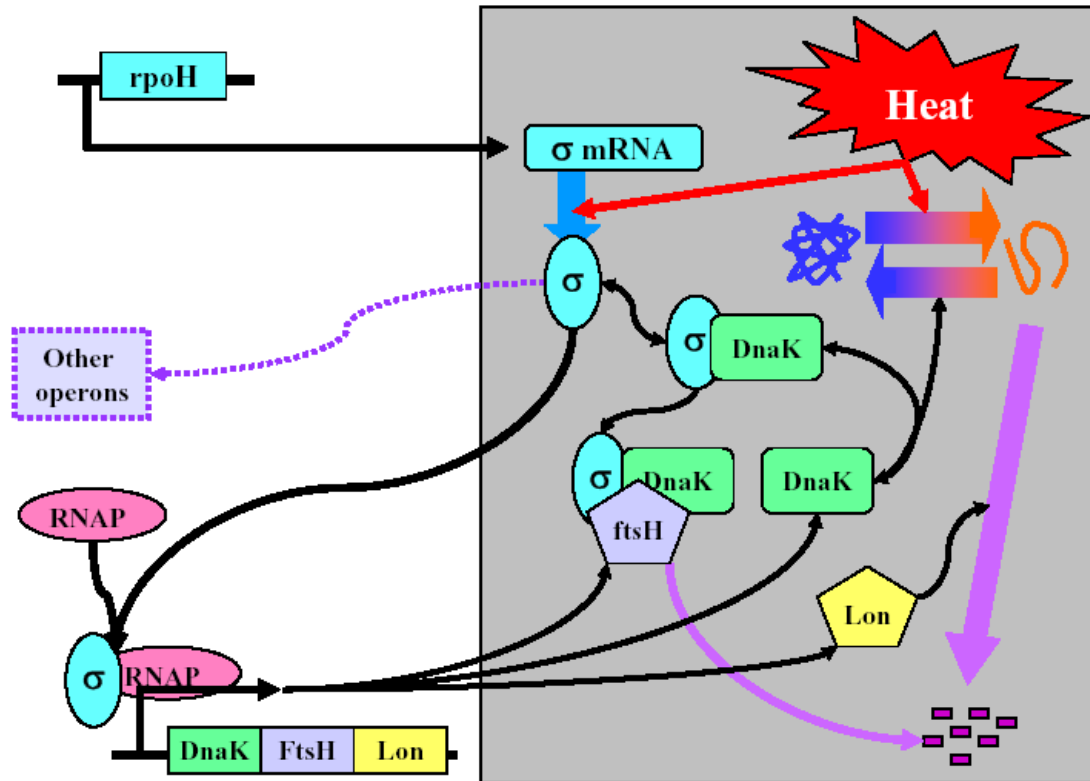


FIGURE 43

What I want to do is compare that with the Internet. Hopefully, you are familiar with the Internet hourglass shown below. You have a huge variety of applications and a huge variety of linked technologies, but they all run IP under everything, IP on everything.

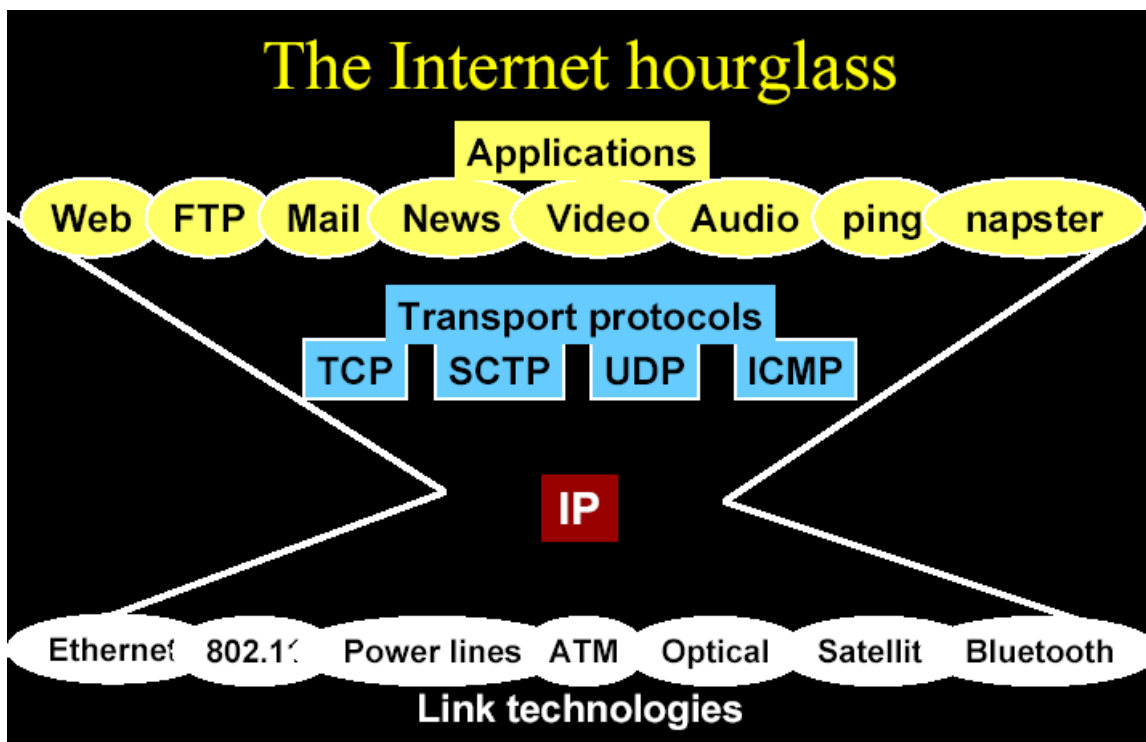


FIGURE 44

You have this common core that is universal, every computer on the planet, more or less. What is universal about this? How does it compare with biology? Well, you have a huge variety of components. Think of all the different things you could build a bacteria from. It is an almost unlimited list. Then there is a huge variety of environments in which bacteria can live. The same thing is true for the Internet, a huge variety of applications at the top and a huge variety of different physical components at the bottom. If you are going to build architecture you have got to deal with enormous uncertainty at both edges. Again, it's similar to the bow tie, but now we are talking about it a little bit differently: it is the control system and not the flow of material and energy.

If you are bacteria you pick a genome, or you go to the Cisco Systems catalog and you make your network. You now have a physical network, but it doesn't work because it doesn't do anything yet. What does it need? It needs this feedback control system or it doesn't work. Then, it is interesting.

Hourglass architectures

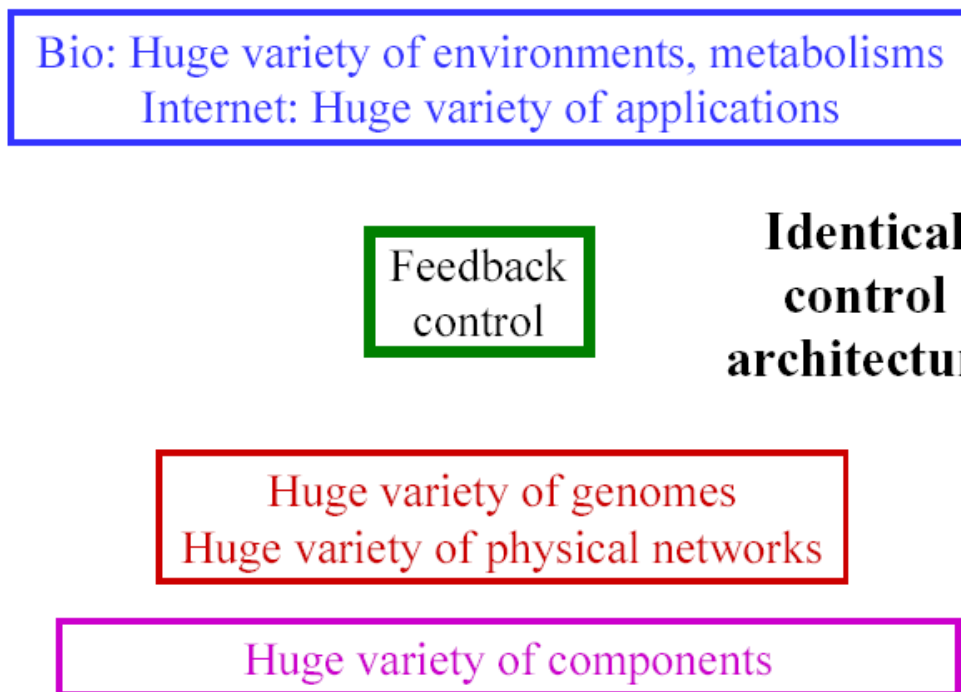


FIGURE 45

There is huge variety everywhere except in the feedback control. This architecture fits every cell on the planet, every computer on the planet. You have got the hardware at the bottom: routers, links, servers, or DNA, genes, enzymes. You pick a raw physical network or you pick a raw genome. Then to turn it into a network you need to have routes on it so you have an IP route rerouting. It has two things: it has a feedback control system that deals with uncertainty below it. You put it together, you run IP on it, and it gives you routes. It now makes you a very robust network to whatever is going on at the lower levels. If you had a very benign environment that is all you would need. You simply run your applications right on top of IP; you don't need anything else. On the biology side you need transcriptional regulation to create the enzyme levels that are necessary to run the cell. In a benign environment this is all you would need.

In fact, though, you are usually in an uncertain environment with complicated applications coming and going, supply and demand of nutrients and products and, as a consequence, on the Internet you need TCP, which gives you robustness of supply and demand and packet losses, you do congestion control and it activates re-transmission of lost packets. The resulting comparison is shown on Figure 46. On the right-hand side you have allosteric regulation, all that post-translational modification, all the different things that you do in protein-

protein interactions and regulation. This all happens in the cytoplasm. It makes you robust to changes in supply and demand in a much shorter time scale. What you see is a layer of control that gives you primarily robustness to what is below a layer of control that gives you robustness to what is above, and you stick those two together, and you get a very robust plug-and-play modularity.

TCP/IP	Metabolism/biochem
5:Apps: Supply and demand of packet flux	5:Apps:supply and demand of nutrients and products
4:TCP: robustness to changing supply/demand, and packet losses	4:Allosteric regulation of enzymes: robust to changing supply/demand
3:IP: routes on physical network, rerouting for router losses	3:Transcriptional regulation of enzyme levels
2:Physical: Raw physical network	2:Genome: Raw potential stoichiometry network
1:Hardware catalog: routers, links, servers, hosts,...	1:Hardware catalog: DNA, genes, enzymes, carriers,...

FIGURE 46

It isn't an accident that TCP and IP are separate and do different things, and it is not an accident that we have—I have used allosteric controls as a stand in for all the different ways that you can do modifications of proteins, again, hidden from the user. We don't know all that is going on. When you use your computer, you don't see this going on. Again, it is easy to make up fantastic stories about these. What is important about this is you have a vertical decomposition of this protocol stack. Each layer independently follows the rules, and if everybody else does good enough it works. The Internet does it this way and so does biology. Also, what you have is a horizontal decomposition. Each level of decentralized and asynchronous, so there is no centralized control. It all runs based on these protocols, so you get this bow tie picture.

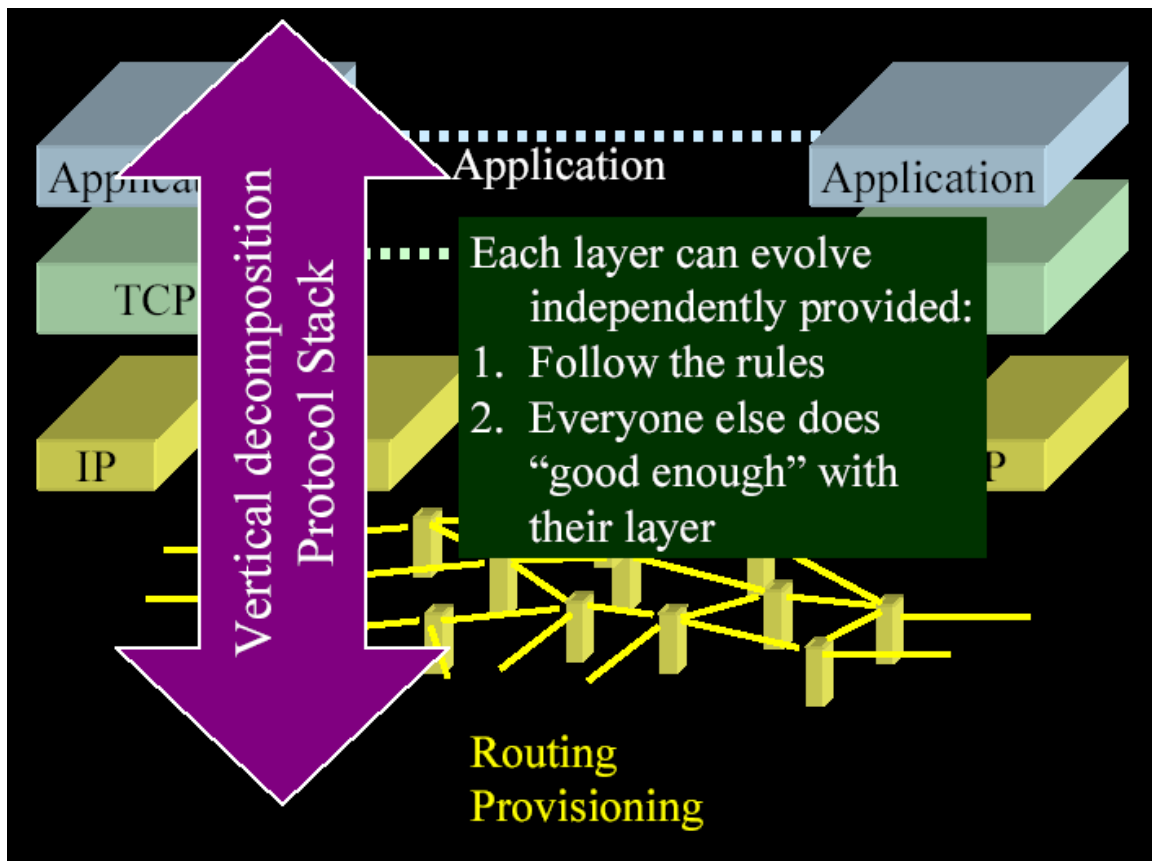


FIGURE 47

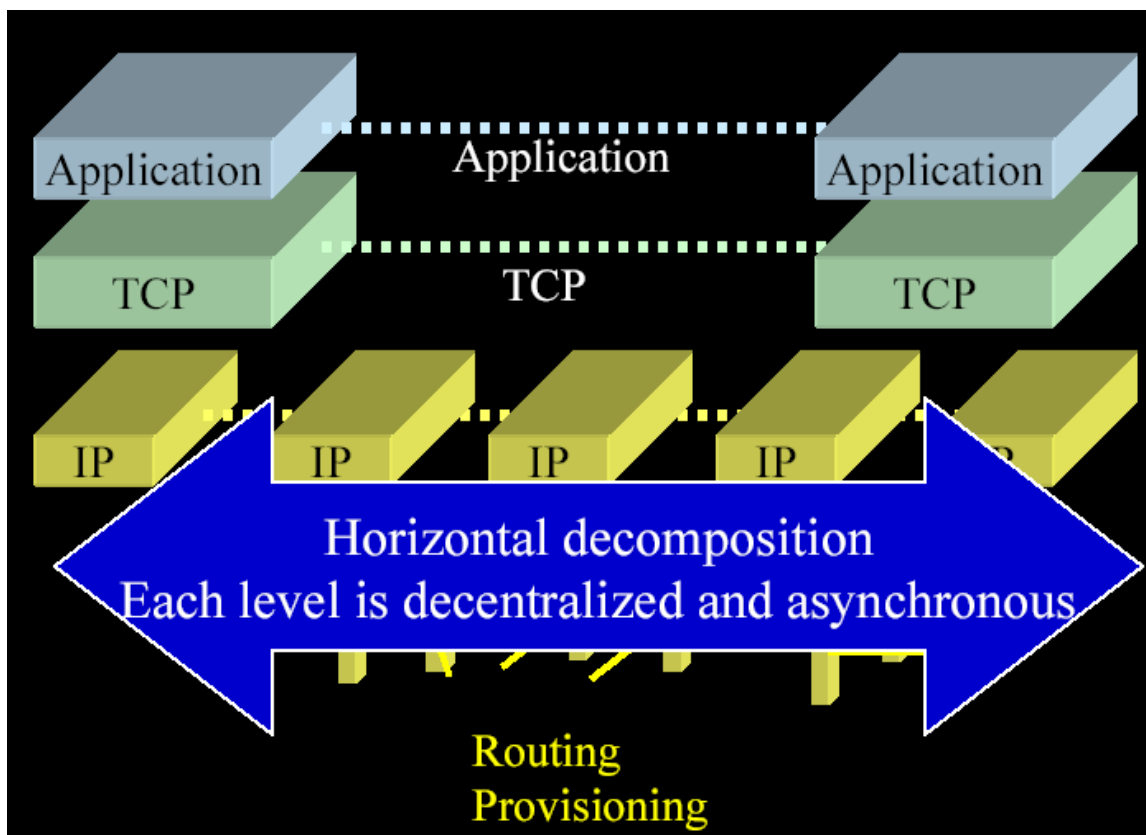


FIGURE 48

Where does the protein and hourglass fit together? It is the regulation and control that are organized as an hourglass. If you are thinking about this, how does this relate to the Internet? The cell metabolism is an application that runs on this bow tie architecture—it is the bow tie architecture that runs on this hourglass. They sit together in the same cell. Right now we don't do a lot of this in technology. We keep these things fairly separate. We are starting to build control systems that run on top of IP networks. Something I thought I would never work on is a disaster, but it is going to happen whether I like it or not. We are going to have to make it work. What this does is gives you robustness on every time scale. Again, there is no trade off in these architectures between robustness in the short run and robustness in the long run.

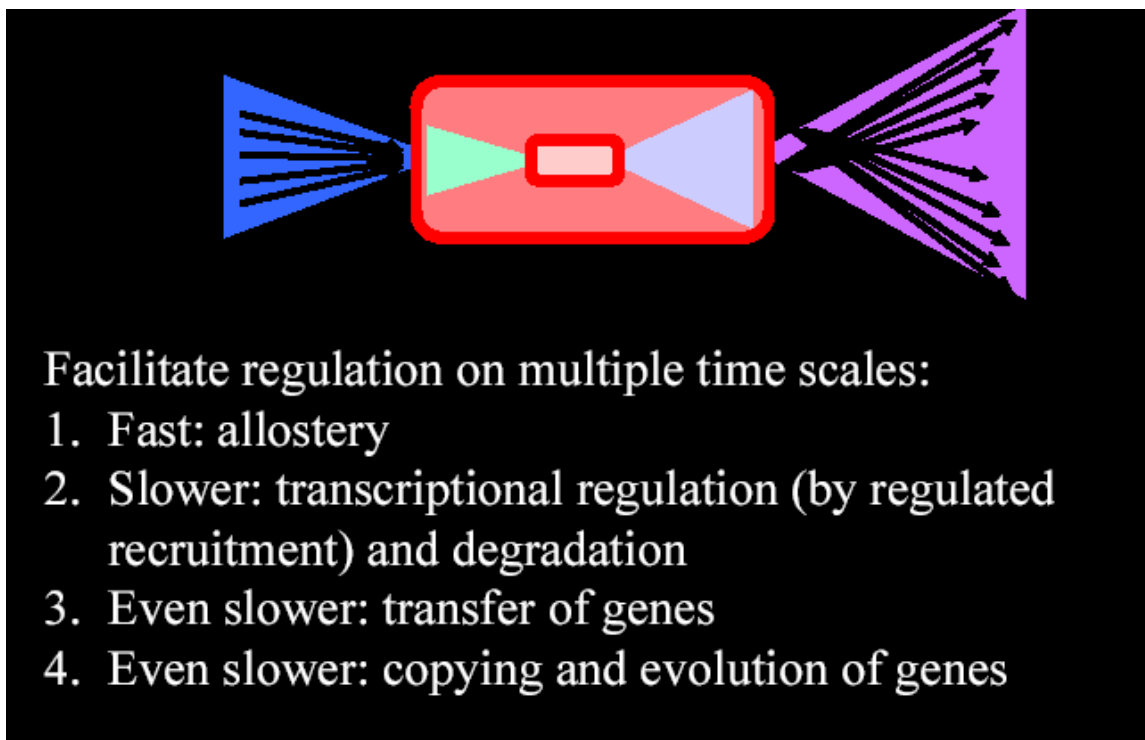


FIGURE 49

Robustness/evolvability on many timescales

- **Shortest: File transfers from different users**
 - Application-specific navigation (application)
 - Congestion control to share the net (TCP)
 - Ack-based retransmission (TCP)
- **Short: Fail-off of routers and servers**
 - Rerouting around failures (IP)
 - Hot-swaps and rebooting of hardware
- **Long: Rearrangements of existing components**
 - Applications/clients/servers can come and go
 - Hardware of all types can come and go
- **Longer: New applications and hardware**
 - Plug and play if they obey the protocols

FIGURE 50

Viruses and worms: What is the worst form of spam? It is the spam that looks like real e-mail. So for the worst attack on a complex system, how does a virus get into a cell or into our network? It has to look okay and it gets through. Those are the attacks that we need to worry about most and on the Internet file transfers you have got navigation, you have got congestion control, and you have ack-based re-transmission. On a little broader time scale, if routers fail you can reroute around failures, you can do hop swaps and rebooting of hardware. On longer time scales you can rearrange the components. On the longest time scale you can bring in entirely new applications and hardware, and it plugs right in. So, on every time scale you get enormous robustness created by this architecture. You also get fragility on every time scale. If the end systems don't run the congestion control there is nothing to prevent the whole network from collapsing. If everybody turns off TCP tomorrow we are dead. The worst thing is fail-on of components, not things failing off. If you take a sledgehammer to the Internet nothing happens; it is not a big deal. If you fail-on, or hijack, that is the worst thing that can happen. Distributed denial of service is an example, black holing, and other fail-on cascades. So things failing on are the worst things. In fact, the worst thing is to nearly obey the protocols and deviate subtly. What does that do? Suppose a nutrient fluctuates? All you have to do is control the pathway that feeds

the core and everything else works. You don't have to do control except locally. That is why these centralized control schemes work. A new nutrient comes in, the whole regulatory apparatus comes in, it comes over here, kicks in transcription, translation, makes a new protein, all of a sudden, that comes back, that is all regulated and control, you have got a solid catalytic feedback, you make a new pathway and you eat that nutrient. Nothing else had to change. If you lose a router it is no big deal. You have all these regulatory mechanisms to protect you with that. You are going to incorporate new components.

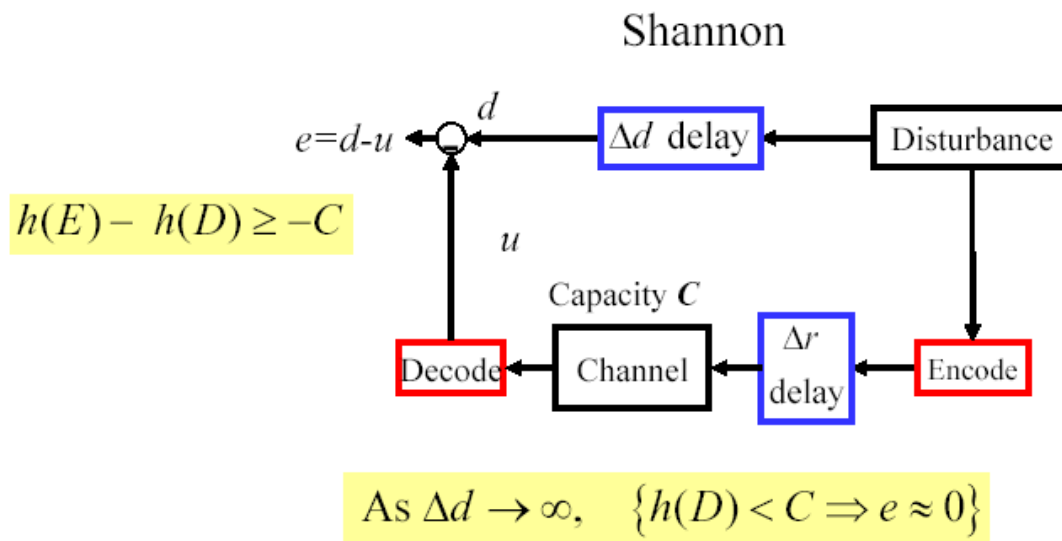
Suppose you need a pathway but it isn't there. You are a bug and somebody is dosing you with antibiotics, and you want to consume it and you can't; this is no big deal. If somebody else has it, you can grab it and stick it in so bacteria can do this. This is why it can acquire antibiotic resistance on such a fast time scale. We can't do this; we have this lineage problem. If we could do this, I know I would look like some combination of Michael Jordan and Lance Armstrong, not like this. We can't; we are stuck with the genes our parents give us and we can't go grab more, although I guess you guys are working on it. On longer time scales the idea is that you can create copies and then evolve them, and they work the whole time. They work the whole time because they run the same protocols. It makes it easy to evolve. Also, they are fragile on time scales. The fluctuations of demand and supply can actually exceed regulatory capabilities so you have heard of glycolytic oscillations. You can actually kill bacteria by simply fluctuating the amount of glucose in its environment, and it is killed because its control system reacts improperly and, again, in well known ways.

On the short term we have failed on components. You get indiscriminate kinases, loss of suppression, and loss of tumor suppression. Again, the worst thing is fail-on. The worst thing for kinase is not to stop phosphorylating its target, but to phosphorylate targets that it is not supposed to. That is a major cause of cancer. Failing on is a much bigger problem in these systems. In fact, this hijacking that we talked about, that is diabetes and obesity. It turns out the same mechanisms responsible for robustness and most perturbation allows possible extreme fragility failures. So, what we are seeing is extreme variabilities in all the measurable quantities, but also this extreme variability and robustness and fragility, and it is created by these architectures. High variability everywhere and, of course, once there is high variability, you can find power laws if you are looking. These are universal organization structures and architectures.

I have a whole story on signal transduction. Time is getting late, and I knew that was going to be the case so I put it at the end. What I talked about was the theoretical foundations a little bit. We want to get into some discussion, but let me just do a little bit of math staying at the undergraduate level. What are the theoretical foundations for studying these kinds of systems?

They are the most rigorous theories we have around; the most sophisticated and applicable theories come out of computation, computer science, control theory, and information theory. Those subjects have become fragmented and isolated. The weird thing was when I started working on the Internet I realized here was the biggest communication network ever built, and it doesn't use anything from Shannon, more or less, at the core, the TCP/IP. And that the people who work on TCP/IP and the people who work on forward error correction never talk to each other. It makes some sense to me, but we have got to get that all back together. There have been attempts to create new sciences, and they have pretty much been failures. The problem is we don't need new sciences; we need new mathematics; we need integrated mathematics.

Recent progress has really been spectacular. I want to give you an undergraduate example of the kind of progress, and I will try to do this very quickly and, again, it will assume that you have a background on the topic I am talking about. The standard Shannon story—actually, this is the way Shannon first presented it, which is not the way it is usually presented in textbooks, if you go back and look at his original papers—supposes you have a disturbance that is going to create some error, and you have a process of encoding and delay through a channel with capacity C , then decode. Then you are going to try to construct a signal to cancel that disturbance. (see Figure 51). That is treating communications like a control problem. This is actually how Shannon first thought about it. He said the entropy reduction is limited by the channel capacity, standard story, and the other big result is that, if the delay of a disturbance is long enough—if the delay goes to infinity—and the entropy of the disturbance is less than the channel capacity, then you can make the error zero. These are the big results in Shannon.



1. Hard bounds
2. Achievable (\Leftarrow assumptions)
3. Solution decomposable (\Leftarrow assumptions)

FIGURE 51

Hard bounds: The bounds are achievable under certain assumptions, and there is a decomposition theorem that is well known in coding theory about how to achieve that. In control theory there is the corresponding thing which is a little bit less known, and I am beginning to realize the reason it is less well known is because people like me teach this stuff really badly. So, everybody takes their undergraduate course in controls. How many people took an undergraduate course in controls? It is always the course you hated most. I don't understand this, and it is a pathology and a virus that I have. We have got to get it out of some of this stuff.

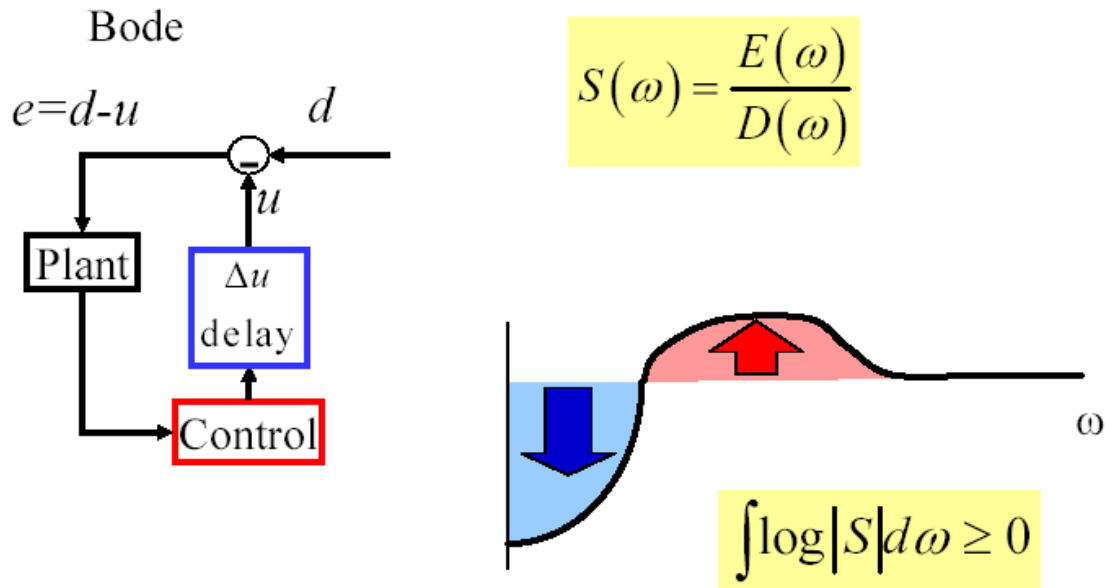


FIGURE 52

Anyway, control theory has a similar thing. You take D and E , you take the transforms and you can define this thing called a sensitivity function. If you think of the other thing as a conservation law based on channel capacity, you can't exceed channel capacity here. You have another conservation law that says that the total sensitivity is conserved. So, I have this demo, shown in the following illustration. It is harder to move the tip of the pointer around if it is up than if it is down. If I want to go around here, I can move it around real easy—same parts but way harder. This theorem tells us why. It turns out that there is an instability it adds to the problem, and it turns out that the instability gets worse as it gets shorter, and eventually I can't do it at all, and you can do the experiment yourself at home. You can check out this theorem at home—same thing, hard bounds, achievable solution, decomposable.

Control demo

$$\int_0^\pi \log |S(\omega)| d\omega \propto \log |1/L|$$

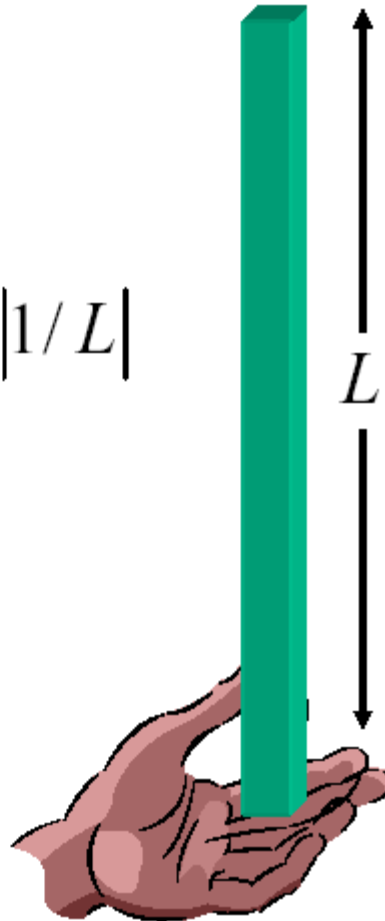
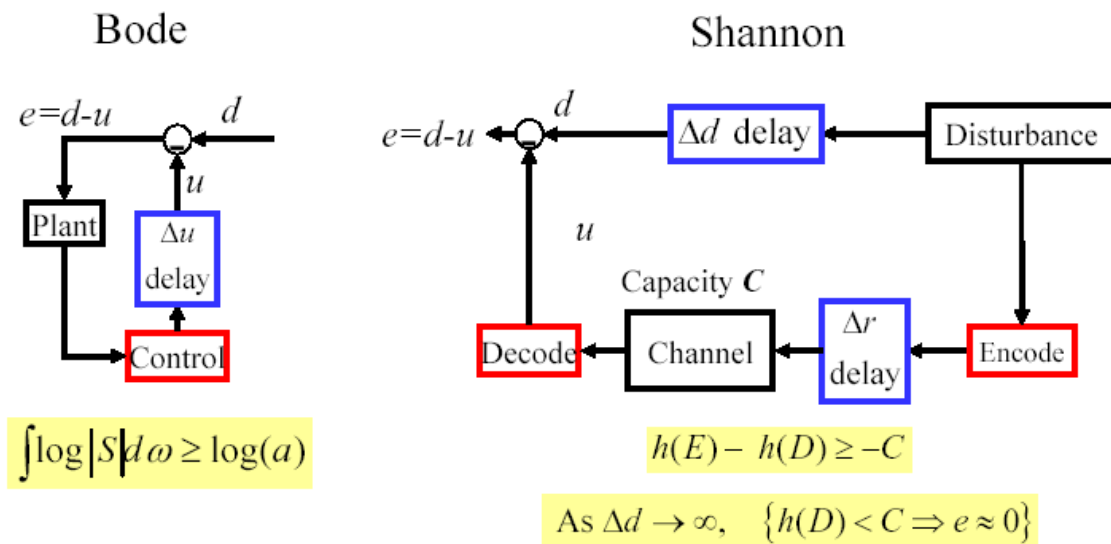


FIGURE 53

These two things have been sitting staring at us for 60 years, which is a total embarrassment. What happens if you stick them together? What would be the best thing that could possibly be true? You just stick them together? There is the cost of stabilization, there are the benefits of remote sensing and there is a need for low latency. That can't possibly be right, because it is so obvious you would have thought some undergraduate would have done it 50 years ago. The other weird thing is Shannon worked for Bode at Bell Labs. They actually sat and talked about this stuff together. Not only that but it is a hard bound; it is achievable under certain assumptions, with the usual, you first prove it for the added Gaussian case, and then the solution is decomposable under certain assumptions. It is possible to unify these things, and there are actually some undergraduate results that come out right away. This hasn't been published yet, but it is going to appear at the next conference on decision and control; it has been submitted to IEEE Transactions on Control, and so on.



1. Hard bounds
2. Achievable (\Leftarrow assumptions)
3. Solution decomposable (\Leftarrow assumptions)

FIGURE 54

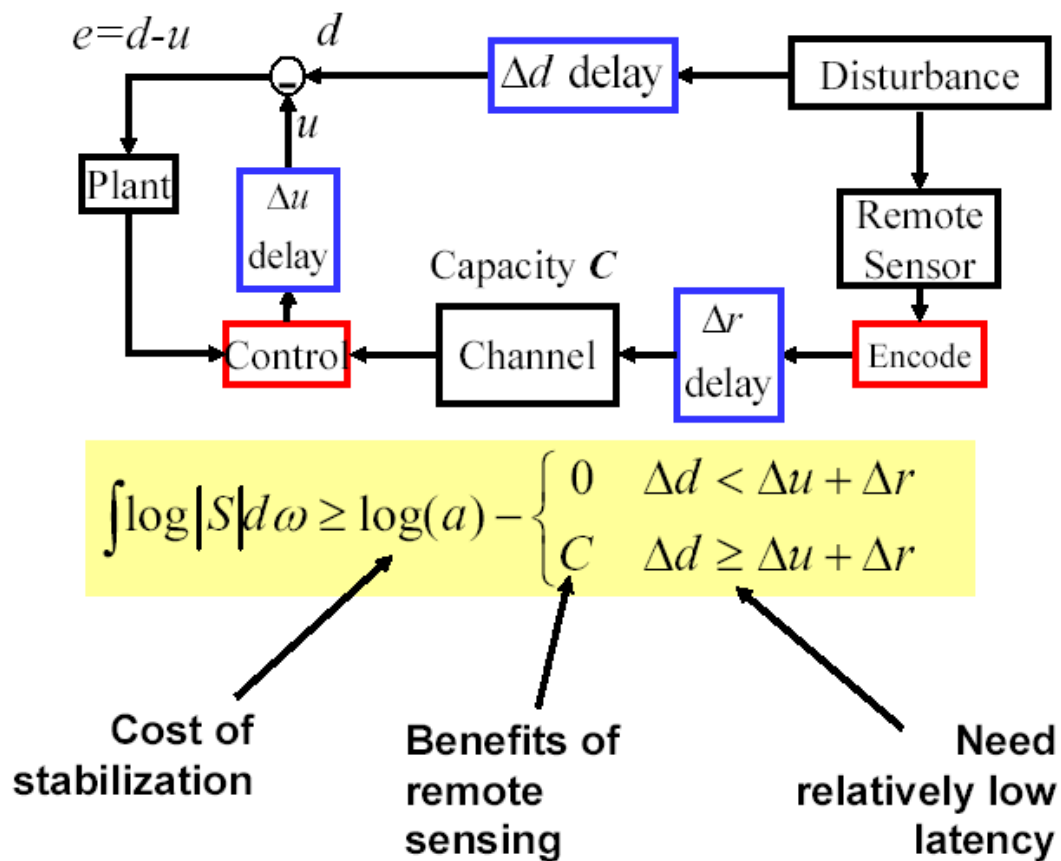
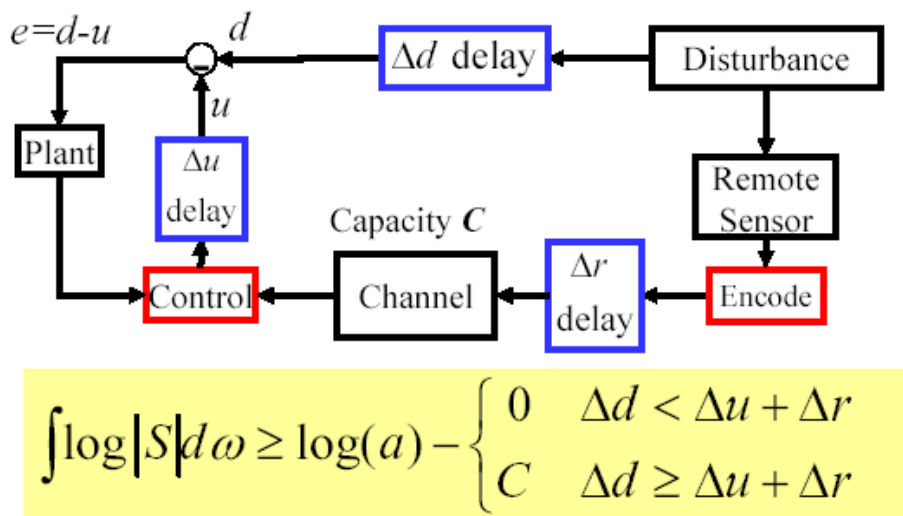


FIGURE 55



Claim (or irresponsible speculation?):

1. Biological complexity is dominated by the tradeoffs which are captured (simplistically) in this theorem.
2. Ditto for techno-networks.

FIGURE 56

Here is a claim or probably, more properly, irresponsible speculation. A lot of the complexity we see in biology is dominated by dealing with this trade off, and the deal for techno-networks. We are starting to use communication networks to do control. Biology already does that. Biology is latency driven everywhere. What does the Internet manage? Really latency. We need to have a latency theory of communications. Go to the FAST web site and read about this. This is pretty cool. It is a case where a very interesting new theory was done, global stability with arbitrary numbers, arbitrary areas and delays so you get very great robustness, global stability of a nonlinear system in the presence of time delays. I never thought I would see this. If I had to do it, I wouldn't. I have good students who did this. One of the co-authors of the paper is here, Lun Li, and you can talk to her about it. She, I think, actually understands it. Anyway, it has been very practical and it has led to the theory of new protocols that, again, here this is macho stuff; this is breaking the world record.

What is the theory? The way you connect these things is it is really a constrained optimization problem. It starts out as constrained optimization, but then you really have to redo optimization theory to get it to work in this layer decomposition way. There is actually a coherent theory for the first time of these layered, decentralized, asynchronous protocols. It is a

nascent theory. That little theorem I showed you before is an example of the kind of unified theory that is now coming along, but it is very much in its beginnings because we had these stovepipe areas chugging along for 50 years with the smartest people I know cranking out theorems in an isolated way. We have to go back 50 years and start putting them back together again. It looks doable and as I tried to show you, we are going to have to think about high variability all over the place. The high-variability statistics is going to be a dominant issue, and it is not necessarily a bad thing. In our disaster world it is a bad thing, but a lot of times this high variability can be exploited. TCP exploits at a huge amount. My cell phone story, if that wasn't there, TCP wouldn't work. So, it is a situation where we have got to be able to learn to exploit that more systematically. I think you are probably going to hear about bits and pieces of this from others throughout the next couple of days. I will stop and see if you have any quick questions.

QUESTIONS AND ANSWERS

QUESTION: A very quick question. In your earlier plot you showed for the power law, you said that apparently it is not power law, but more like exponential. It looks like it is a better fit with the power law than exponential. How do you explain that?

DR. DOYLE: It is a statistical error. So, it is a case where you simply are making a mistake. Again, I didn't really explain the details. This is a standard thing. The idea is that the thing on the right was a differentiated version of the thing on the left. The simple way of thinking about it is, you differentiate it and you create all this noise. There are two advantages to the picture on the left, the rank plot. First of all, I show you the data in its raw form, and you can say whether or not you think it is a power law. There are no statistics done. I show you the raw data. Then you can do a little bit of statistics by just sort of eyeballing the straight line. The idea is that you shouldn't think of a straight line it has to fit. The idea is that we use mean and variance to describe all sorts of low-variability phenomena. We know it isn't really Gaussian and we don't check the other moments. We need similarly to find better ways, in a broader sense, to describe high-variability data. All of a sudden you get means of 1 and variances of 100—it is not a useful statistic, it doesn't convert. You take the data over again. Means and variances don't mean anything in this high-variability world. What you need to do robust statistics is a solvable power law, but you would want to use that even when it wasn't a power law, just like you use mean and variance, so there are robust ways to do this.

This is all decades old statistics, and we don't teach it very well. That is one of the

challenges that we need to do a better job, just of teaching these things. The problem is science has so focused on getting rid of high variability. High variability was thought to be a bad thing, like you are doing bad experiments. As Jean will point out, high variability exists all over in nature. In the laboratory, it only exists in very exotic circumstances, so it is associated with exotica.

DR. HANDCOCK: I would like to pick up on that question a little bit, about the identification and characterization of power laws. I think one issue that needs to be discussed, or at least I would like to hear some discussion more of, is the case where statisticians in particular have looked at questions for a long period of time, but they haven't really been used, that knowledge has not been used in other areas of science.

DR. DOYLE: That is absolutely right, it has not been used.

DR. HANDCOCK: In this area, I think it is particularly important. Clearly, statisticians have had a lot to say about curve fitting and other issues like that, but what I am most reminded of is the recurrence of these debates. The classic example to my mind is Morris Kendall's 1961 address to the Royal Statistical Society, where he essentially lambastes previous work done in a half a century before that time on essentially very similar questions here.

DR. DOYLE: This is a very old story, a very old story.

DR. HANDCOCK: I would like to hear people say more, how can the work of statisticians be more recognized and just routinely used by other sciences, to avoid occurrences of this kind?

DR. DOYLE: I will tell you that my experience has been to give them good software tools do it right and make it easy for them to do it right. If you go to Matlab and get the stat tool box, it's all low variability. So, my experience has certainly been, if we want to get good, new robust control theory into the hands of people, you make software.

If you want to get biologists to be able to share models, you have to have a systems biology mark-up language, you have to make software. So, you have to turn your theories into software, and it has got to be usable, it has got to be user friendly. So, we use Matlab and call the SVD, and how many of us know actually how the SVD works, sort of?

Well, you don't need to. The point is, if you do it right, you need to know what it does, but you don't need to know exactly how it does it. We need to do the same thing for high-variability statistics. It is one thing to lambaste everybody about it. It is another thing to try to create the tools, and show people, teach people. This is the thing we have got to do, we have got to teach people. There are people out there who are discovering this high-variability data everywhere. We are going to hear a lot about it in the next two days, and it is everywhere. The

irony is, because of its strong statistical properties, you end up finding it where it isn't, and the stats community can make a lot of contribution. The problem is that I am not sure anybody studies this any more. It is this old, old topic, old classical stability laws; it was hot in the 1920s, it was hot in the 1950s, and maybe it will come back. I think it should come back.

DR. KLEINFELD: Implicit in your talk was this notion that in engineering systems things are components, and also implicit in your talk and sort of in the world is that, the way biology seems to design things, they are very much in modules. Is there some sort of theorem that says that one should really build things in modules, like some complexity diverges or . . .

DR. DOYLE: Yes and no. I would say we have nascent, little toy theorems that suggest that these architectures—here the point is, you don't have modules unless you have protocols. If you look around, you discover the modules first, you see them, you see the parts, but they are meaningless unless there are protocols. None of this stuff would work or hook together unless they had protocols. What we can prove now is that some of these protocols are optimal in the sense that they are as good as they can be.

For example, TCP properly run achieves a global utility sharing, that is fair sharing among all the users that use it. What that says is that you can get the same performance as if you had a centralized non-modular solution, but you can get it with a protocol that is both robust and evolvable. Now, is there any other way to do it? We don't know. You prove that this is optimal and that this is robust. So, there is a lot more work to be done. I mean, it is only the last few years that we have even had a coherent theory of how TCP/IP works. So, this is quite a new area. I think that is an important question. How much further can we go in proving properties? What we need to design now, as engineers, we need to design protocols. We need to design protocols that are going to run our world, and we need to do them and make them robust and verifiable, and scalable. Now we don't do that very well.

In biology, what we are doing is reverse engineering the protocols that evolution has come up with. Unfortunately, the good news is, it looks like it uses more or less the same protocols. So, it is not going to be incomprehensible. It could just be this infinite parts list. If we just made a parts list of everything on the Internet, it would be completely bewildering, we would never make any sense of it. Because we know the architecture, and because we know the protocols, we know what everything is doing. We have got to do the same thing with biology.

REFERENCE

Han, J.D., et al. 2004. "Evidence for Dynamically Organized Modularity in the Yeast Protein-Protein Interaction Network." *Nature* 430.

Network Models

Neurons, Networks, and Noise: An Introduction

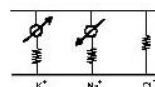
Nancy Kopell, Boston University

DR. KOPELL: As Emery said, my background is dynamical systems in neuroscience, but in a sense I am here to lobby for more work at the interface between sophisticated dynamical systems and sophisticated statistics because I think that is a major underdeveloped area. I have more questions than I have answers for that.

General Mathematical Framework:

Hodgkin-Huxley Equations

$$c \frac{dv}{dt} = -\sum I_{ion} + D \nabla^2 v - \sum I_{synapse}$$



$$I_{ion} = g m^j h (v - V_R)$$

Conductance \times Electromotive force

m and h satisfy

$$\frac{dx}{dt} = (x_{\infty}(v) - x) / \tau_x$$

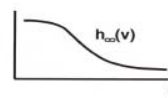
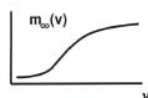


FIGURE 1

My assignment from Emery was to talk about what people study when they study networks of neurons. Generally, as shown in Figure 1, they start with the Hodgkin-Huxley equations or variations of them, which describe the electrical activity of a neuron or even a tiny piece of a neuron. One can even think of a neuron spread out in space, itself; a single neuron as a network. These equations come from electrical circuit theory, and the major equation there is conservation of current, so it is talking about ionic currents that are going across the cell membrane and through little channels.

Each ionic current is given by Ohm's law, $E = IR$. That is nice and undergraduate. The ionic current I_{ion} shown in Figure 1 is the electromotor force divided by the resistance but, in the neural world, they are very optimistic: they don't speak of resistance, they speak of conductance, which is 1 over

resistance, and that is how we define I_{ion} in Figure 1. What makes this a little bit more complicated than standard undergraduate electrical circuit theory is that this conductance is not a constant, but rather it is a product of so-called gating variables, which talk about how these pores in the membrane open and close, and these gating variables themselves are dynamic variables that depend on the voltage. The voltage depends on the gates and the gates depend on the voltage. It is all highly nonlinear. There are many characteristic time scales in there. Even when you are talking about a bit of a single neuron, you are talking potentially about a very large number of dimensions. Now, we hook cells up via a large number of synapses, which I will tell you about in a minute, and one can have an extremely large network to be dealing with. On top of all this is something that I am not going to describe but that I am hoping that Eve Marder will. That has to do with the layers of control that were talked about so wonderfully in the first lecture. That is, on top of all of this, there are neuromodulators, which you can think of as changing on a longer time scale all of the parameters that are in here, and which are themselves changed by the activity.

Synaptic Currents

Cells communicate via synapses:

Produce cross-membrane currents after spikes

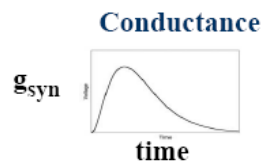
Synaptic current: $I_{syn} = g_{syn} E = \text{conductance} \times \text{driving force}$

Chemical synapse: Driving force $= v_{post} - V_{syn}$

Reversal potential V_{syn} depends on the kind of synapse:

Excitatory: $v_{post} - V_{syn} < 0$

Inhibitory: $v_{post} - V_{syn} > 0$.



Electrical Synapses: $I_{syn} = g_{el} (v_{post} - v_{pre})$

Full equation: $d v_{post} / dt = \dots - I_{syn}$

FIGURE 2

How do you hook up two cells? How do you hook up many cells? There are many ways that neurons talk to one another. The major way is via so-called synapses. Here I am talking about two kinds

of synapses, chemical and electrical. To begin with, just think about a pair of cells. When the presynaptic cell spikes it unleashes a set of events that leads to another current in the post-synaptic cell. Once again, this other current is conductance times the driving force. For chemical synapses, this driving force is the difference between the voltage of the post-synaptic cell and something that has to do with the nature of the synapse, as shown in Figure 2. Very crudely, one can think of the synapses as being either excitatory or inhibitory, and that will determine the sign of what this driving force is. For an excitatory synapse the current will drive the cell to a higher voltage, which will make it more likely for the cell to be able to produce an action potential and, for an inhibitory one, it does the opposite. There are also electrical synapses which depend on the difference in the voltages between the post-synaptic and pre-synaptic cells.

Dynamics of Single Cell

Caricature: (conservation of current)

Integrate and fire: $dv/dt = I - v$

Voltage v tends to I .

If $I >$ “threshold”, action potential (spike) occurs when v reaches “threshold”.

v is then reset to value < 0 .

Quadratic I&F: $dv/dt = I + a(v-v_0)(v-v_1)$



(Hodgkin-Huxley equations describe details of spike)

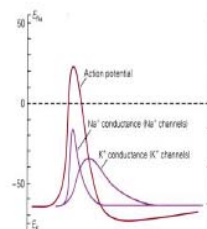


FIGURE 3

Figure 2 gives just a sense of how things are hooked up in these networks. The way to think of it is that all of these currents get added to the equation for conservation of current, so it is electrical circuits with bells and whistles. This can get extremely complicated. There is a fair amount of work in the literature in which people say they can't handle it and go to something a little bit simpler. The standard simpler thing that people work with is the so-called integrate-and-fire neuron in which most of what I said is ignored. You still have the conservation of current; dv/dt is something that looks linear but it isn't. The reason that it isn't linear is that there is a threshold in there. It is a voltage built up to some predetermined

spot that the person working with this decides. One pretends that there is an action potential which happens, a spike, and then the cell gets reset to another voltage value. The graph in Figure 3 shows how the voltage builds up and then decays in an action potential. This is the simplest kind of caricature with which people work. There are other kinds of caricatures in which the voltage isn't just put in by hand, but it is in the equations and a trajectory decays back to rest if it doesn't reach this threshold and, if it does reach threshold it goes on and produces this action potential.

Once you have a network like this, there are general questions that everybody in the field wants to know. What are the dynamics of these networks? They can be local, in which cells are connected to all or a lot of the cells in the whole network. They can be spatially extended, as they would be in a neural tissue. They can be deterministic, or they can have noise associated with them, and they can have inputs that can be with or without spatiotemporal structure. In the study of these networks without inputs, what one thinks about basically is the background activity of a particular part of the nervous system. When we think about putting inputs in, we are thinking about, given the dynamics of the network and the absence of inputs, given that background activity, how it is going to process things that actually do have structure?

Where does statistics come into all of this? Well, data are noisy. One builds these nice nonlinear deterministic models which are extremely complicated. How do you begin to compare what you see in the dynamical systems with what you can get out? In general people mostly make qualitative comparisons. What one would like is to be able to compare on a much deeper level what is going on. That is why I say I am here to lobby for a combination of dynamical systems and statistics. I want to give you a couple of examples of this, and it comes from the things I am most interested in at the moment, which have to do with rhythms in the nervous system that are found in electrical activity of the brain and can be seen in EEGs and MEGs. What I am interested in is how those rhythms are used in cognition. That is much too large a question to be able to talk about here, but what I plan to do is give you some examples that I think will illustrate some of the statistical issues that come up.

I have two examples. The first one is a real baby example; it is two cells. To think about how you can get rhythms in a complicated situation, you have to think about coherence and what creates coherence. The simplest situation about coherence has to do with two cells, and that is what I am starting with. There are various mathematical ways of thinking about coherence of cells; I am using one in Figure 4 that I think is especially well adapted to statistics. Assume you have a neuron, and its parameters are such that it would like to keep spiking periodically if nothing else happens. Now, it gets some input in the form of another spike, and that input does something. One of the things that it does is to change the timing of the cell that receives that input. One can describe what these inputs are doing in terms of the so-called spike-time response curve. What the schematic in Figure 4 is telling is that if the input comes in, say, 60 milliseconds after the receiving cell has spiked, there will be a negative advance or delay. If the

input comes much later in the cycle, the receiving cell will spike considerably sooner than it would have.

If you have such a spike-time response curve, with a little bit of algebra—this is even high school—you can figure out, if you have two cells that are put together, that are talking to one another, what will those two-cell networks do? The mathematics that describe what it will do is called a spike-time difference map, which maps the timing difference between the two cells at a given cycle onto the timing difference between them at the next cycle, so that you can see if they will synchronize. It is simple algebra that takes you from one to another to get that map. Once you have such a map it turns out to be standard 1-dimensional map theory to tell us whether the cells will synchronize or not, and starting from what initial conditions. It is the zeroes of this that are telling you the phase lag at which you will have synchronization, and the slopes are telling you something about whether or not that steady state is going to be stable or not. It turns out for the example in Figure 4 that there is stable synchrony and stable anti-phase.

Can Understand Synchronization Via Timing Maps (Acker, White, NK)

Spike time response due to excitation: P is time advance of next spike due to input (from other cell)

$$t_1^* = t_1 + T - P(t_2 - t_1)$$

$$t_2^* = t_2 + T - P(t_1^* - t_2)$$

Spike-time difference map F :



$$\begin{aligned} \Delta^* &= \Delta + P(\Delta) - P(T - \Delta - P(\Delta)) \\ &= \Delta + F(\Delta) \end{aligned}$$

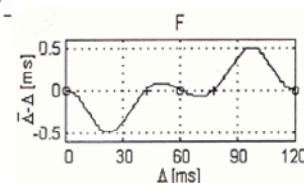
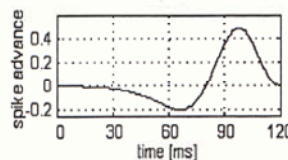


FIGURE 4 Acker, C.D., Kopell, N., and White, J.A. 2003. Synchronization of strongly coupled excitatory neurons: Relating network behavior to biophysics, *Journal of Computational Neuroscience* 15:71-90.

Let's add some statistics to it. The really nice thing about spike-time response curves is that you can actually measure them in the lab using something called a dynamic clamp, which is a hybrid network between real biological cells and in silico. You can take a real biological cell—Eve Marder is an expert and one of the people who first developed this technique—and you can inject into it a given signal at a given time that you want, and you can see what that cell does. You can measure it and, not surprisingly, you get something or other, but it is awfully noisy and, depending on the signal that you have, you will get something else.

Stochastic Spike-Time Response Methods

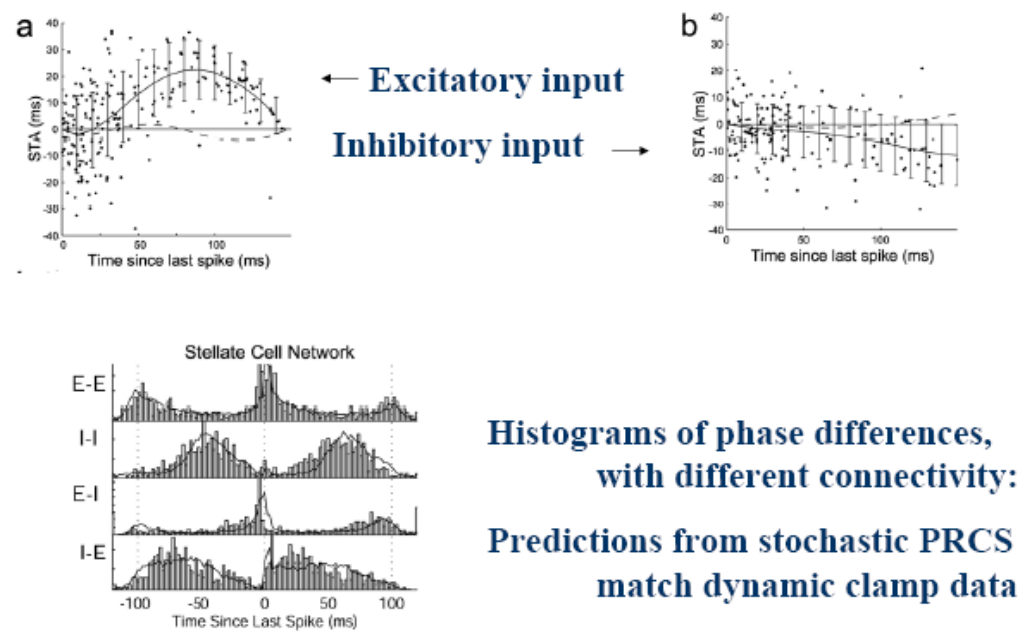


FIGURE 5 Netoff, T.I., Banks, M.I., Dorval, A.D., Acker, C.D., Haas, J.S., Kopell, N., and White, J.A. 2005. Synchronization in hybrid neuronal networks of the hippocampal formation, *Journal of Neurophysiology* 93:1197-1208.

Here is the question: What can you read from that stochastic spike-time response curve shown in Figure 5? When you have a spike-time response curve, before you create the spike-time difference map, can you read off whether it will synchronize or go into anti-phase? By synchrony I mean zero phase. For the particular deterministic spike-time response curve in Figure 5 it turns out that, if you hook up two identical cells with that, they will synchronize, not go into anti-phase. Is there any more information that

is in the scatter? The answer is yes; namely, if instead of modeling this as something that is fit, you make it random with a Gaussian distribution whose variance is the measured variance here. You can predict what will happen statistically to the network when you actually hook up these two cells to real cells with the dynamic clamp. When cell fires you look up the spike-time response curve and realize it should advance this much with a certain random amount, and you make the other cell fire at that, and you keep doing that. That gives you the outlines of what is here. You can then take the two cells and, with the dynamic clamp, inject (whenever one cell fires) a certain current into the other cell corresponding to the synapse, and it will do something with that, according to its spike-time response curve. You see that the histogram of phase differences, with different ways of connecting those same two cells, turns out to match very well the predictions from the stochastic phase response curves, or spike-time response curves. The punch line of this example is that when one takes into account the statistical structure that is in the data here, you actually can make much better predictions about what the network will be doing than if you simply take the fit of it and you do the mathematics with that. The deterministic fit predicts just that things will synchronize, which will be a spike over here and nothing else, whereas using the extra structure here is giving you the whole histogram of all of the phase differences.

That is a simple example of putting together dynamical systems and statistics. Now I am getting to one where the questions get a lot more complex, and it is a lot less under control, and it is where I think there is room for a huge amount of work from other people. There are many questions you can ask about these very large networks, but one question I am very interested in concerns the formation of so-called neural ensembles. I am definitely not using the word cell assembly, which many of you may be more familiar with. Cell assembly, in the literature, means a set of cells that are wired up together for whatever reason, and tend to fire together. By neural ensemble, I mean a set of cells, whether or not they happen to be wired up together, that are firing, at least temporarily, almost synchronously with one another. So, it is a more general notion than cell assembly. It is widely believed, although not universal, that this kind of neural ensemble is extremely important for the functioning of the brain and cognition. Among other reasons for the importance of synchrony is the idea that this synchrony temporarily tags the cells as working on the same issue, that they are related now for this input, for this particular cognitive task, as noted in Figure 6.

Cell assemblies tend to change rapidly in time. So, cells would be firing together for perhaps a fraction of a second, and then on to another cell assembly. Think of it like a dance. These kinds of subsets are also important as a substrate for plasticity because, when cells do fire together there is a chance for them to really wire together and become what is known as a cell assembly. Lots of people believe that this kind of synchronous activity is extremely important. There is a smaller group, which includes me, that believes the so-called gamma rhythms in the brain are important for the creation of

those neural ensembles. The gamma rhythm is the part of the EEG spectrum that is roughly 30 to 90 hertz, depending on who you speak to and the phases of the moon. This is all very contentious. There is a huge amount to be said about gamma rhythms and where they come from and why people think it is important to early sensory processing, to motor control, and to general cognitive processing.

Large Networks and Neural Ensembles

Hypotheses: Neural ensembles are important for

- Tagging temporarily synchronous subsets of cells as “related”.
- “Synchronous” subsets are substrate for plasticity.
- The gamma rhythm (30-90 Hz) is important for the creation of neural ensembles.
 - Gamma rhythms are potentiated by attention.

PING (Whittington et al.,
J. Physiol. 1997)

**Pyramidal-interneuron
network gamma, with
heterogeneity**

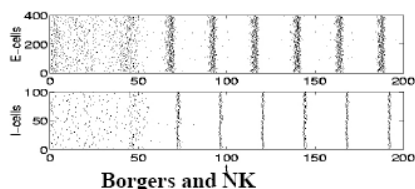


FIGURE 6 Whittington et al. 1997. Pyramidal-interneuron network gamma, with Heterogeneity, *Journal Physiology*.

Since I have less than four minutes left I am going to cut to the chase and give you a sense of where the statistical issues come from. To give you background on what we are putting the noise on, I am going to start with, as I did in the other example, something that has no noise whatsoever in it. This is one way of creating a gamma rhythm known as pyramidal interneuron network gamma shown in Figure 6. Pyramidal cells are excitatory; these interneurons are inhibitory; I am making a network out of the kinds of things I showed you right at the beginning. The simplest way to do this is with one excitatory cell and one inhibitory cell. The excitatory cell fires and makes the inhibitory cell fire. The inhibitory cell fires and inhibits the excitatory cell, which just stays silent until the inhibition wears off.

Persistent (Vigilance) Gamma Rhythm

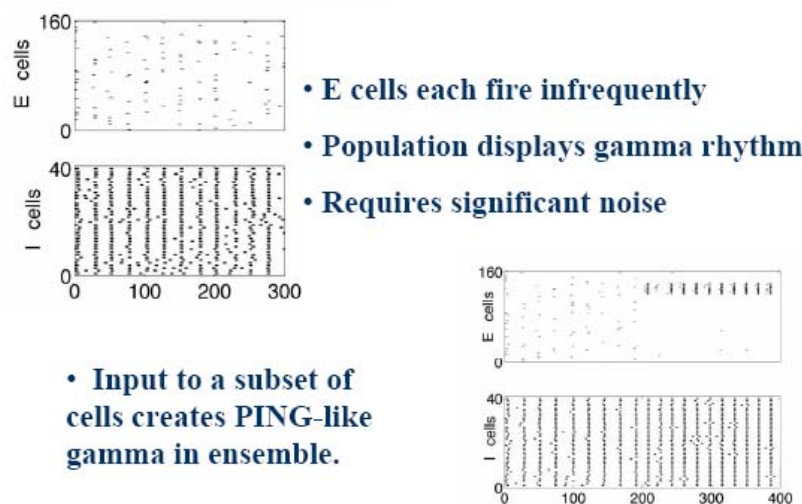


FIGURE 7

If you have a large network of this—and the example that leads to Figure 7 is not that large, it is 160 pyramidal cells, 40 inhibitory cells—it self-organizes to create this gamma rhythm. Exactly how that happens involves mathematics and probability theory when there is heterogeneity like this. That mathematics has been worked out by Christian Borgers and me. It is not the essence of what I want to say so I am going to go on from there. Roughly, the most important part is that the common inhibition leads to this self organization. Now, we add in some noise and I am going to put that noise on top of a slightly different parameter regime in which the E cells don't fire very much and they really need this drive in order to fire. You can then get a kind of gamma rhythm in which excitatory cells are firing very sparsely. These graphs in Figure 7 are raster plots. You can see that any given E cell is firing very sparsely.

You still have a population rhythm, and in our hands at least, it requires a lot of noise to make it happen. There seems to be a lot of noise in the physiological situations that correspond to this. One can produce this in slice, and many of us including this gang believe it is this kind of activity that is associated with a kind of background attention, a kind of vigilance that enables you to take inputs that are at low level and be able to perceive them much better than if this background activity were completely

asynchronous. It is related to the kind of activity that I was talking about before. Now if there is input into a subset of the excitatory cells here, it will create a cell assembly in which the cells that are involved are producing a kind of PING rhythm and everything else here is more or less suppressed.

In the larger story that I am not telling, this corresponds to inputs to a specific set of cells that are doing neural computations and that are responding in the presence of this kind of background activity. What are some of the issues here? I think there are a huge number, but I will just mention three.

First there is the question of whether there really are cell assemblies of this sort, and gamma rhythms associated with them. In a recent conversation that I had with Uri Eden, who works with Emery Brown, I found he is coming up with a way of being able to decide that question. You look at an animal that is behaving, say, running down a track, and you are looking at a place cell in the hippocampus that is supposed to be firing when a cell is in a particular spot on a maze. The question is, if you look at the relationship between where the cell is, can you make an association with some covariate, which in this case would be where the cell is, and when the cell is producing these gamma rhythms, or is having its firing related to a potential field potential that is producing local gamma rhythms? You can potentially see with statistics whether or not this gamma rhythm is actually encoding when the cells are doing what they are supposed to be doing. You would expect to see more gamma rhythms when the cells are actually firing in their correct positions, and Eden believes he has a new way of thinking about that. This has to do with looking at data and also with looking at the models themselves. The persistent vigilance models are generally large networks, their outputs are very complicated, and it is very hard to be able to change a lot of parameters and say what the network is actually doing.

A second major issue for which one needs statistics is to find the appropriate statistics to be able to actually describe what the network is doing. For instance, how does the firing of the cells depend on parameters of the system and this extra layer of control that I mentioned in neuromodulation? What happens when you add inputs to all of this? How do statistics of firing and network connectivity affect the response to inputs? At the moment, we don't have any statistics like this other than the standard off-the-shelf, and that doesn't do it.

The final major question is one that confuses me the most, and that has to do with plasticity and the so-called statistics of experience, in which one needs to use statistics to begin with, even before you phrase what the question is. The point is that for real animals living in the real world there are natural inputs, both auditory and visual, which make some cell ensembles more likely. It is not just the kind of thing that an experimenter would give to an animal. This makes some cell assemblies much more likely to occur than others, and one of the things we would like to know is what does that do to the creation of cell assemblies, what does that do to network connectivity, and what would that do to the response to inputs. In a sense the input to this theory is the statistics of experience, and the output of the theory is statistical,

but it is all through dynamical systems models that would tell you how one converts to the other.

QUESTIONS AND ANSWERS

QUESTION: One question is, usually at a single unit firing level, the rhythm is not very obvious.

DR. KOPELL: One should be looking at the correlation between the single unit and the local field potential.

QUESTION: Like in a network model, how to go about from the single unit that is the basic building block, to a local field potential kind of measurement? Can that be done within the same mathematical network?

DR. KOPELL: Yes, but once again, that is contentious. The physics of what the local field potential is actually measuring is still not wholly agreed upon by everyone. So, people do this, but it is not clear how correct it is.

DR. KLEINFELD: Nancy, where do you think all this noise is from?

DR. KOPELL: Some of it is noise that is generated in the channels.

DR. KLEINFELD: Shouldn't that fall sort of like \sqrt{N} ?

DR. KOPELL: Again, there is the issue of high variability versus low variability, and there is the question of how many inputs are averaged before one gets the output. When you are thinking in terms of the square root of N , you are sort of assuming all of this noise will cancel out. I think it doesn't, and there seem to be major noise generating networks, at least within axons, of some pyramidal cells, that will actively work to produce this noise. It is presumably doing something really important here. I don't think the noise is something that should be treated as something that will be averaged out.

DR. KLEINFELD: There is this textbook notion of 10^4 inputs per cell, and then these tremendously strong sort of depression schedules for synapses, and then there are results from people like Alex Thompson that suggest that a couple of synapses have large PSDs, like 10 millivolts instead of hundred of microvolts. Just to put closure on what you are saying, is that what you are thinking, the very large number of inputs to itself is really a few preferred inputs that are dominating the drive to a neuron, and it is because of these small numbers that you get the high variability?

DR. KOPELL: That is certainly consistent with this, yes.

REFERENCES

FIGURE 4. Acker, C.D., Kopell, N., and White, J.A. 2003. Synchronization of strongly coupled excitatory neurons: Relating network behavior to biophysics, *Journal of Computational Neuroscience* 15:71-90.

FIGURE 5. Netoff, T.I., Banks, M.I., Dorval, A.D., Acker, C.D., Haas, J.S., Kopell, N., and White, J.A. 2005. Synchronization in hybrid neuronal networks of the hippocampal formation, *Journal of Neurophysiology* 93:1197-1208.

Mixing Patterns and Community Structure in Networks

Mark Newman, University of Michigan and Santa Fe Institute

DR. NEWMAN: I am going to be talking mostly about social networks because that is mostly what I know about. I hope that some of the ideas I talk about will be applicable to some of the other kinds of networks that some of the others of you work on.

First, I want to mention some of my collaborators, one of whom, Aaron Clauset, is in the audience. The others are Michelle Girvan, Cris Moore, and Juyong Park.

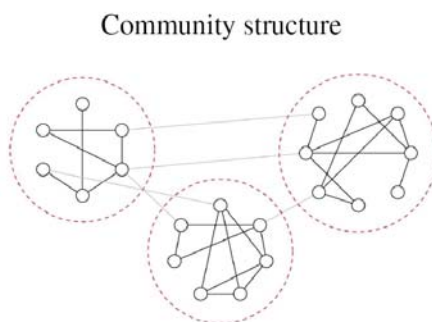


FIGURE 1

The story I am going to tell you starts with an old idea from network analysis. We believe that many of the networks we study divide up naturally into groups.

Suppose I am talking about a social network. I will assume that you know what a social network is. It is a network where the nodes are, for example, people and they are connected together by some social interaction such as they're friends, or they have business acquaintanceships, or they have been in close physical proximity recently. Choose the definition that is relevant to your particular problem. So, you have a network of vertices and edges as shown in Figure 1, and we are going to assume that it divides up naturally into communities, groups, clusters, or whatever you would like to call them. First of all, one thing that we might be interested in doing is finding those groups, if they exist in the first place, and if they do exist, determine what kind of groups they are. For example you could be interested in a social network and be looking at groups of people in a network. You could be interested in the World Wide Web, and be looking for groups of related Web pages. You could be looking at a metabolic

network and be looking for functional clusters of nodes. There are a variety of situations in which you might be interested in this.

Finding groups in networks is an old problem. It is something that computer scientists, in particular, have looked at for many decades. But there is a difference between what we want to do with social networks, for example, and what computer scientists have traditionally done. First of all, in the computer science problems, traditionally when you are looking for the groups in networks, there are some additional constraints that we don't have, such as you know how many groups you want beforehand. A standard computer science problem might be that you have a bunch of tasks and you want to divide them up over many processors on a computer, and you know how many processors you have. You know how many groups you want to divide them into. Very often you might want the groups to be of roughly equal sizes. That is a typical constraint in the kinds of algorithms that computer scientists look at because, for example, you want to load balance between many processors in a computer. In the problems we are looking at that is not often true.

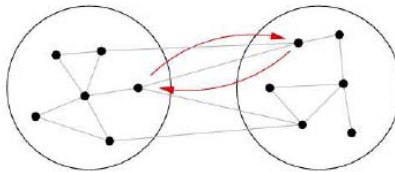
What is more, there is a fundamental philosophical difference compared to the traditional approach of partitioning a computer network, in that we are assuming here that the social networks we are looking at naturally divide into these groups somehow. In the kinds of problems you want to solve in computer science, often you are just given the network and you want to find whatever the best division of it is into groups, such that you have a bunch of these tightly knit groups, and there are lots of connections within the groups, and only a few connections between the groups, and you do the best you can with whatever network you are given. We are taking a slightly different point of view here, and assuming from the outset that there is some good division for some good sociological reason, for example, that your social network naturally divides up into these groups, and we would like to find the natural divisions of the network. And those natural divisions might involve dividing it into some small groups and some large groups, and you don't know how many groups there will be beforehand, and so forth.

There are old algorithms for doing this from computer science, like the one shown in Figure 2, which you might be familiar with, the so-called Kernighan-Lin algorithm from the 1970s. In this algorithm, if you want to divide a network into two parts, you can start by dividing it any way you like, essentially at random, and then you repeatedly swap pairs of vertices between the two parts in an effort to reduce the number of edges that run between the left-hand group and the right-hand group in Figure 2. You continue doing that until you can't improve things any more, and then you stop. This is typical of the kinds of things that traditional algorithms do. It works very well but it has some constraints. First of all, you need to know the sizes of the groups

you want beforehand, because when you swap two vertices, you don't change the size of the groups. Each group gets one vertex and loses one vertex.

Kernighan-Lin algorithm

- Greedy optimization algorithm
- Maximizes number of within-community edges by swapping vertex pairs



Pros	Cons
fast	bisects only fixed size communities

Furthermore, the Kernighan-Lin algorithm only bisects. It starts off with two groups, and it ends up with two groups. If you want more than two groups, then you would have to do this repeatedly, repeatedly bisect, divide into two, and then divide the two into two, and so forth. However, there is no guarantee that the best division into three groups is given by finding first the best division into two and then dividing one or the other of those two groups.

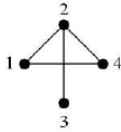
Something that I like better is shown in Figure 3, spectral partitioning, a rather elegant method that is based on matrix-algebra-type ideas. This is also something that has been worked on since the 1970s. In this approach, we take the so-called graph Laplacian, a matrix representing the network in which the degrees of the vertices are down the diagonal. The degree of a vertex is how many connections it has. There is a -1 off the diagonal in each position corresponding to an edge in the network. Minus 1 here means that there is an edge between vertex 2 and vertex 1. This is a symmetric matrix in an undirected network, and it has an eigenvector of all 1s with an eigenvalue of 0.

Spectral partitioning

Graph Laplacian is the matrix L where

$$L_{ij} = \begin{cases} \text{degree of } i & \text{for } i = j, \\ -1 & \text{for } i \neq j \text{ and } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Lowest eigenvector:



$$= \begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ -1 & -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 0.$$

FIGURE 3

If you have a block diagonal matrix, like the one shown in Figure 4, one that divides completely into two separate communities with no connections between them, then it has two eigenvectors with eigenvalue 0, one with all the 1s corresponding to the first block and one with all the 1s corresponding to the second block. And, of course, any linear combination of those two is also an eigenvector with eigenvalue 0.

Block-diagonal Laplacian:

$$\begin{pmatrix} \square & 0 \\ 0 & \square \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = 0,$$

and $(0, 0, \dots, 1, 1)$, and any linear combination:

$$a(1, 1, \dots, 0, 0) + b(0, 0, \dots, 1, 1).$$

FIGURE 4

If you have a network that is approximately block diagonal, that is precisely the sort of

thing we are talking about—in other words, it has two communities, two separate groups, but they are not quite separate, there are a few connections between them, so we have a few off diagonals here and here. Then, as shown in Figure 5, you still have one eigenvector with eigenvalue 0, and the second one is slightly perturbed and has a slightly positive eigenvalue. You can prove that it is always positive. By looking at the properties of that second eigenvector you can deduce what those two blocks must have been. It is actually very simple. You just look at the signs of the elements of those eigenvectors and assign vertices to the two groups according to their sign. This gives a very nice and principled way of bisecting a network into two groups. In this case, you don't have to specify what the size of the two groups is. You just find—if your matrix really is approximately block diagonal with two blocks—what those blocks are. But this algorithm only bisects, and it is a rather slow algorithm, since it involves finding the leading eigenvectors of the matrix. Typically for large networks you do this by the Lanczos method because it is a sparse matrix, but it is still a pretty slow thing to do.

- For a *nearly* block-diagonal Laplacian, this will be an approximate eigenvector, and must be *orthogonal* to $(1, 1, 1, \dots)$
- Hence $a > 0$ and $b < 0$, or *vice versa*.

Second eigenvector has form $(+, +, \dots -, -)$. So we can find blocks by looking at signs.

Pros	Cons
any size communities	bisects only slow

FIGURE 5

Looking particularly at social networks, but potentially for other networks, we have been thinking about other methods of solving this problem for the kinds of networks we are looking at. Many people have worked on this in the last decade or so. I am going to tell you about some stuff that we have done, just because I was involved in it and I like it, but of course it is certainly not the only thing that has been done.

Figure 6 shows one method we looked at; this is work that I did with Michelle Girvan, who is at the Santa Fe Institute. Suppose you have a network like the one in Figure 1, and it divides into groups where there are many edges within the groups and only a few edges between the groups. The idea is going to be, suppose I live within one of the dotted circles and I want to

drive to a node within one of the others. I would then go along one of the edges within my circle to get to the edge that connects the groups, which is like a bridge. In fact, if I lived anywhere within my dotted circle and I wanted to drive anywhere within another circle, I am going to end up going over that bridge. If you live in San Francisco and you want to get to the East Bay, you would go over the Bay Bridge, and that is the way you have to go. That means there is a lot of traffic on the Bay Bridge because everybody goes that way. These bridges between communities are things that are going to have a lot of traffic if we consider, in some simulation kind of a way, traffic moving around from every part of the network to every other part of the network. A simple concept that captures this idea of betweenness was first formulated by Lin Freeman, who is also here somewhere.

So, the idea here is that we find the shortest path between every vertex in the network and every other vertex in the network, and then we count how many of those shortest paths go along each edge in the network. This is like a primitive measure of how much traffic will be flowing along each edge of the network. Then, because some of these edges are the few edges that lie between the communities, those ones are going to get a lot of traffic. We identify the edges between the communities as those with the highest value of this betweenness measure, and then we remove them from the network. Once you remove them and those edges there are gone, you are left with the three clumps, and these are your communities.

Betweenness algorithm

- Calculate number of shortest (geodesic) paths between every vertex pair that go along each edge
- Find the edge with the largest score
- Remove it
- Recalculate the betweenness scores and repeat

FIGURE 6

It is a little bit more complicated than that in practice. The principal problem is if you have two edges between two communities, like the two at the top in Figure 1, then there is no guarantee that they will both get a high value of betweenness. It is certainly the case that the sum of the number of paths that go along one and the number of paths that go along the other should be high but, under certain circumstances, it is entirely possible that one of them could be high and

one of them could be low. If you remove the edges with the high betweenness, you wouldn't remove all the ones that connect groups, only some of them. In order to get around that, you have to find the edge with the highest betweenness, remove it, and recalculate this betweenness measure—this traffic flow measure. If I remove one edge and recalculate, any traffic that previously had to go along that edge will now go along the other bridge instead. So, even if it didn't have a high count before, it does have a high count now. It is crucial that you do this recalculation step at each step of the algorithm, and that slows things down a bit. That is the primary disadvantage of this method. In fact it takes a moderately long time to do the calculation.

We did come up with a nice way of speeding up the calculation. Naively you would expect this to be an $O(n^4)$ calculation. But some of the calculations can be reused, and you can get it down to an $O(n^3)$ calculation, but that is still prohibitively slow for very large networks. You couldn't do this on the World Wide Web graph or something like that, but it could typically be done on the sizes of social networks we are looking at. It works up to tens of thousands of vertices.

This works quite well and I will give you some examples. Some of these are whimsical examples, but there is a reason for that. Figure 7 shows a famous network from the social science literature. At the top is a network from a study done in the 1970s of a karate club at a U.S. university. The vertices represent students in this club who were practicing their karate, and connections between them represent friendships as determined by direct observation of the members of the club by a sociologist who was working with them. This is an interesting example. The sociologist, Wayne Zachary, studied this club for two years and fortuitously it turned out during the course of these two years that a dispute arose between two factions within the club, and eventually the club split in two and one half went off and formed their own club.

The theory is that you should be able to spot that coming split in this network, even though this network here was constructed before the split occurred. In fact, you can do that. The white circles and the gray squares represent the factions as they were after the split. All we did was to take this network and feed it through our algorithm. The output of the algorithm is a so-called dendrogram, which is a tree as shown in (b) in Figure 7. It doesn't matter too much what it indicates. The main thing to notice is that it splits the network into two groups, and the coding of the circles and the squares is the same as in the network in (a). You can see that it has very neatly found the group of the squares and the group of circles. There is only one error over there, vertex number three, which is the one right in the middle there, on the boundary between the two factions. Maybe it is understandable why it got that one wrong.

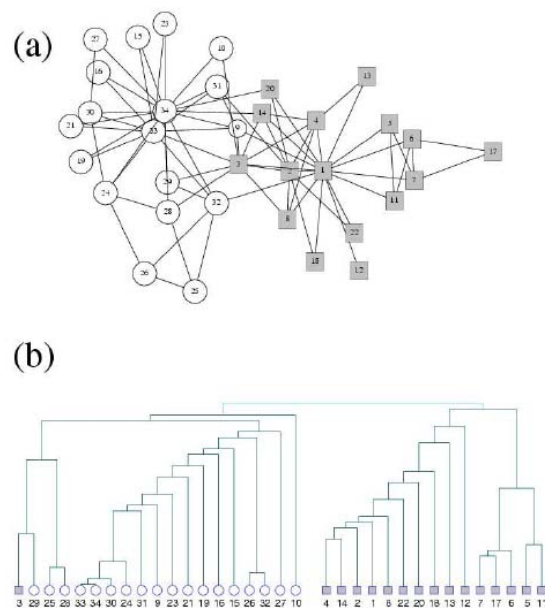


FIGURE 7

This is a network that was derived from observations made before the split, and yet clearly the network has in it information about what is going to happen in the future when the thing does split up. In a way, it is predicting what is going to happen, although it is predicting after we knew what the answer already was. Obviously, we wouldn't have put this example up here if it didn't work.

Figure 8 gives another example. I am at the University of Michigan, and so I need to have an example about football. This is a picture—one of these dendrograms, one of these trees—that shows what happens when you feed the schedule of games for division 1-A college football into our algorithm. As you might know, college football is played in conferences, which are groups of schools that play together on a regular basis. If you feed the schedule in, it picks out the conferences very nicely. They are color coded by the different shapes and colors in this figure. The groups that the algorithm finds, as you can see, pretty much correspond to the conferences. In a way, this is a silly example but we look at examples like this for a good reason, which is they are cases where we think we know what the answers should be. You want to have applications to real world networks to show that this really is working, but you also want, when you are developing algorithms, to apply it to something where you can tell, after you did it, whether it is actually giving a sensible answer. This algorithm is always going to give some

answer. If you plug in some network and it spits out an answer and you can't tell whether it is sensible, then it didn't really get you anywhere. Many of the examples we look at when we are developing the algorithm are ones where we think we know what it ought to be doing before we start.

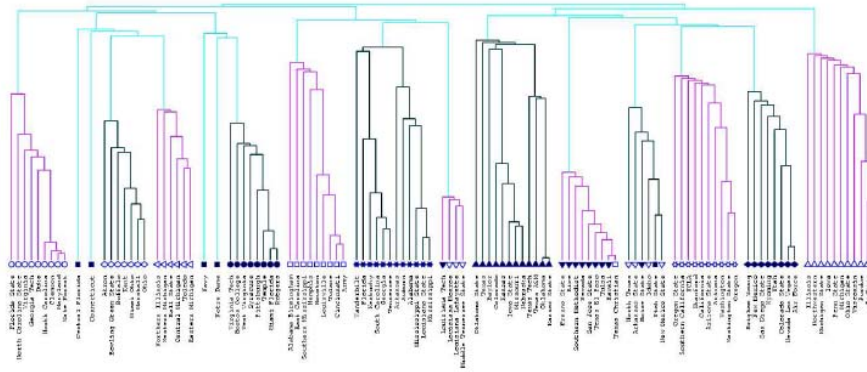


FIGURE 8

Figure 9 gives one more example. In this case this is a food web, a non-social network example, a food web of marine organisms in the Chesapeake Bay. The interesting thing about what happens when we feed this one into the algorithm is that it appears to split the food web into the pelagic organisms, the ones that live in the upper layers of the water, and the benthic ones, the ones that live in the mud down at the bottom. It appears to be indicating that this ecosystem is divided into two weakly interacting systems with not very much going on between them, so it certainly does have applications other than the social networks.

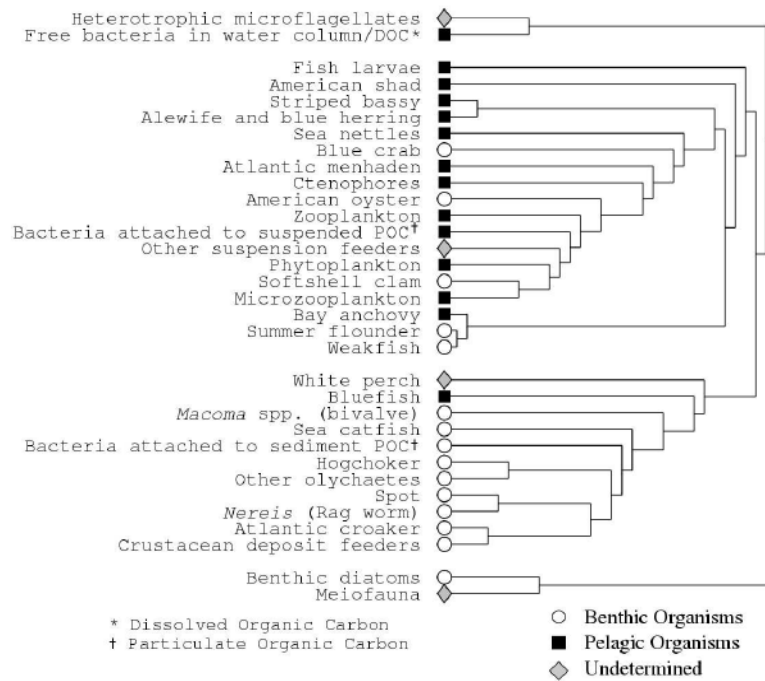


FIGURE 9

One of the differences between what we do and what computer scientists do is to know whether the solution that we got to a problem is actually a good one. We don't want to just get the best solution, whatever it is, we also want to know if the solution we got is no good; in other words, if there wasn't really any community structure there in the network in the first place. To do that, we use a measure that we call modularity. The idea with modularity is basically to measure what fraction of the edges in our network fall within the groups that we find. What we would like is that most of the edges are within groups, and only a few of them between groups. However, if you only use that, it is not a good measure of what we want, because then you could put all the vertices in a single group together, and all the edges would be within that group, and you would get this 100 percent modularity, but that is obviously a trivial result, it is not a good division of the network. What we use instead is the fraction of edges that fall within the groups minus the expected fraction of edges that would fall within those groups.

I can write down exactly how we calculate that if you like, but the basic idea you can see by examining that trivial partition. If I put all the vertices in one group together, then all the edges would be within that one group, but it is also expected that all the edges will be in that one group. Thus, you get 100 percent minus 100 percent, so the measure defined in the previous

paragraph is 0 rather than a high score. So we think of modularity as something which is high if the number of edges within groups is much greater than we would expect on the basis of chance. What we can do is take one of these dendrograms, which represents the order in which the network gets split into smaller and smaller pieces as we remove the edges, and plot this modularity against it, as shown in Figure 10. What we typically find is that the modularity has a peak somewhere and then it falls off, and this peak indicates where the best split would be in this modularity sense, the best split in terms of getting most of the edges within groups. That is shown by the red line in Figure 10. The value there at the peak indicates how good a split it is. This is a measure that goes strictly from 0 to 1. It is about 50 percent modularity in Figure 10, which is pretty good. We usually find that there is something significant going on if it is anywhere above about 30 percent, and the highest values we have seen in real networks are about 80 percent.

Modularity

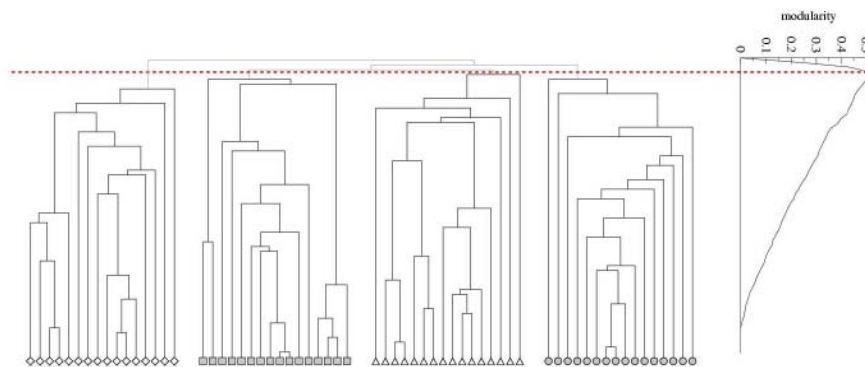


FIGURE 10

Figure 11 shows what happens if you apply this method to the karate club network. It actually finds two different things. It has a peak for the split into two groups that I talked about before, the one we know about. It has another peak for a split into five groups, and it looks like that second peak is actually a little bit higher. Maybe there is some other smaller-scale structure going on there in the networks that we didn't know about before.

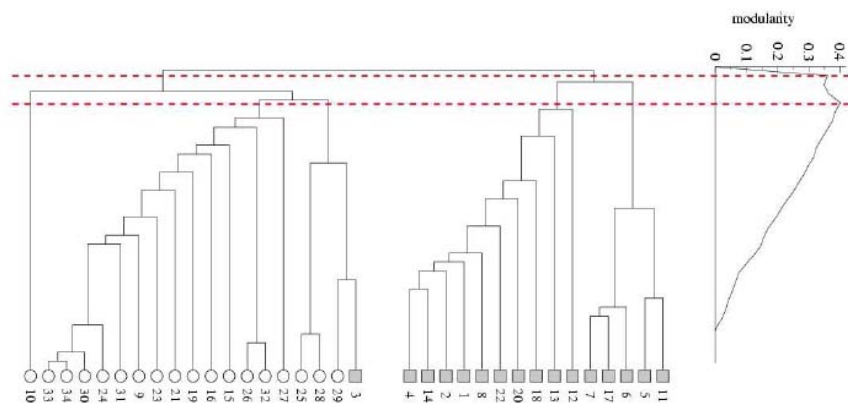


FIGURE 11

Once you have this modularity, it allows you to say what the division of a network into groups is. Now I can draw a picture like Figure 12, where I have taken a network and have colored the groups by whatever corresponds to this maximum modularity. This particular one is a network of collaborations, meaning co-authorships, between scientists at the Santa Fe Institute. The colors here correspond rather nicely to subject divisions within the institute, so it looks as though it is giving something sensible.

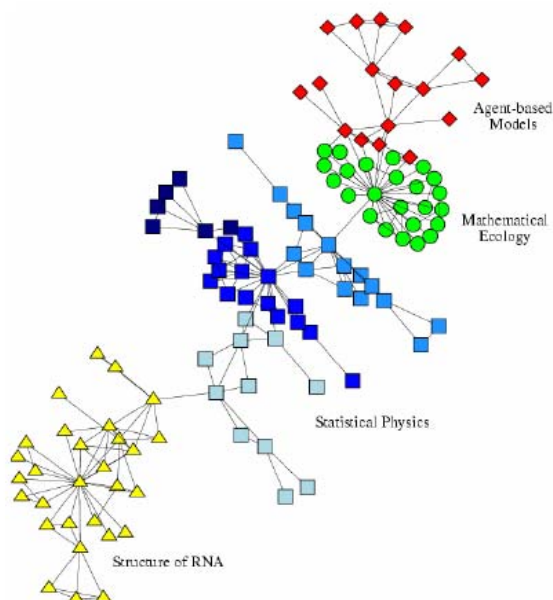


FIGURE 12

Figure 13 is a fun network. This one is characters from Victor Hugo's lengthy and rather boring novel, *Les Miserables*. It is a very big, thick book and has a lot of characters, and what I have done is joined up any two characters that were in the same scene at any point during the novel. This is another one of those examples where you think you know what should be happening. We know the plot of the novel, and so you can look at this and see if it is giving sensible results. There are well-known groups of characters that are getting lumped together here.

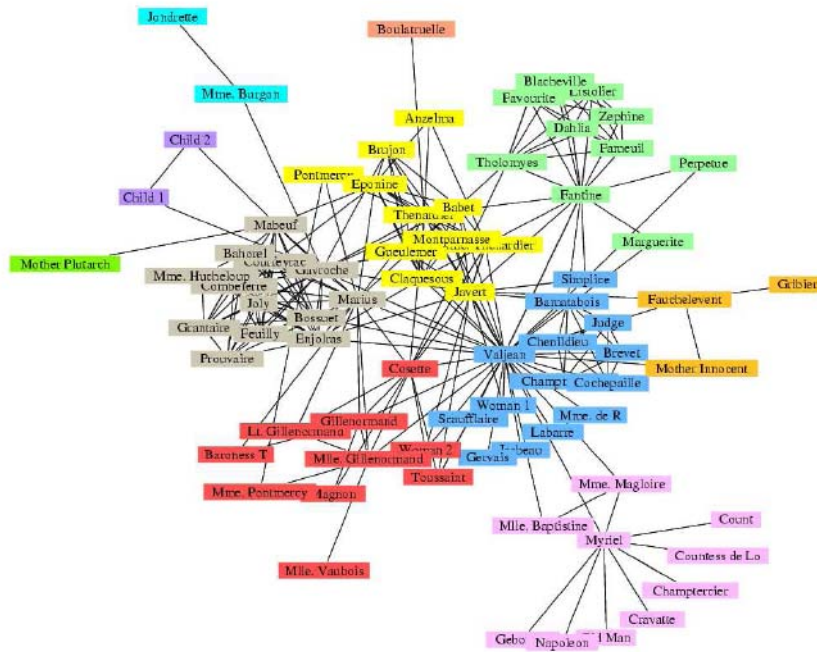


FIGURE 13

Figure 14 is another example, taking it one step further. Suppose I am given a network like the one shown in (a). Again, this is a collaboration network of who has written papers with whom. In this case, it is people in the physics community who work on networks, so it is a network of network people. If you were just given that, you can't pick out that much structure. If you feed it into this algorithm that finds the groups in the network, then you begin to see some structure. It turns out that the groups it finds, indicated by the colors in (b), pretty much correspond to known groups of collaborators at known institutions. Taking it one step further, you can take this and make this meta network shown in (c), where there is a circle corresponding to each of the groups in (b), and the edges between them represent which groups have collaborated with which other groups. I have even made the sizes of the circles and the sizes of the edges

represent the number of people and the number of collaborations. I think that this is the sort of thing that could be very useful for analyzing these very large network data sets that we are getting now.

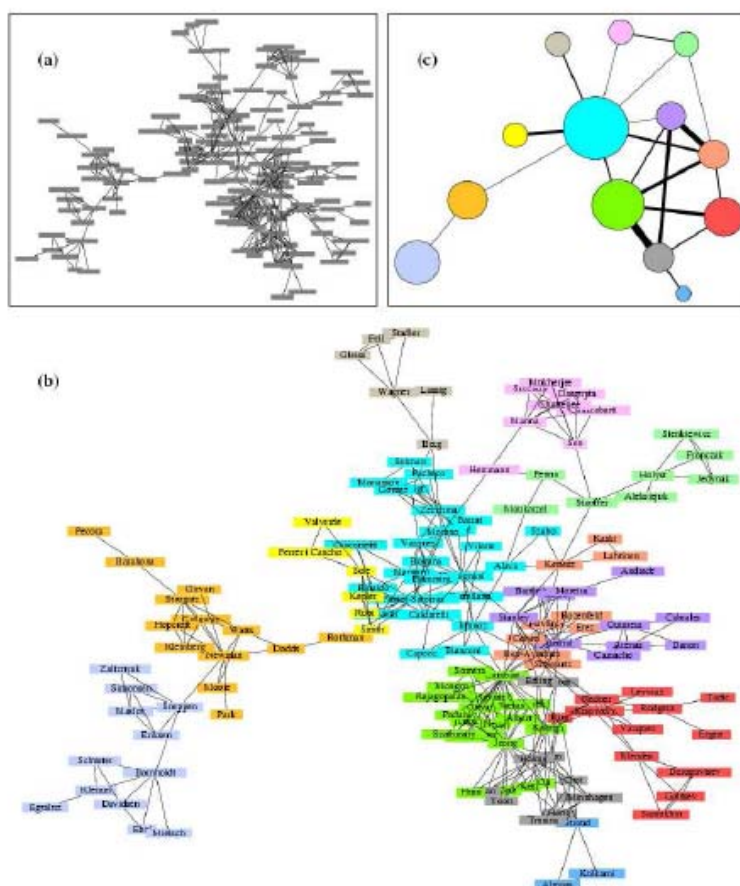


FIGURE 14

We are getting these data sets that are so big that there is no way you can draw a picture on the screen. In this case you can, but we are getting ones with millions of points now, and it would be very nice to have an automated way to coarse grain it and get the big picture that you see in panel (c) of Figure 14. Then, if you want to know the details, you can zoom in on one of those circles and see what the smaller-scale structure is. In theory, you can do this repeatedly and have several layers of this so that you can visualize things more readily, seeing the biggest structure, and the middle-scale structure, and the small-scale structure all separately when you couldn't plot the whole network on the page because it would all just come out looking like a mess.

In the last few minutes I want to tie in this idea of group structure with another topic, the network of friendships between school children. Our data are taken from something called the U.S. National Longitudinal Study of Adolescent Health, the AdHealth study. Jim Moody, who is here, is heavily involved in this. When the data for friendships in a particular high school are plotted, it shows clearly two groups of kids in the school, the white kids and the black kids, and the network of friendships is quite heavily divided along race lines. This is not a surprise to sociologists; this is something that people have known about for decades. What is happening is very clear when you draw these network pictures. This is the phenomenon that they call homophily or assortative mixing, that people tend to associate with other people that they perceive as being similar to themselves in some way. It could be race, it could be educational level, income, nationality, what language you speak, what age you are, almost anything. Practically, you name it, people associate according to that. You can also have disassortative mixing, which is when people associate with people that are different from themselves in some way.

Here are some examples. Figure 15 shows a matrix of couples that I took from a paper by Martina Morris, who I also saw here. This is from a study that was done in San Francisco, and they interviewed a lot of couples and tabulated what ethnic groups they came from. You can see that the fraction down the diagonal represents people who are in a couple with somebody from the same ethnic group as them. It is much larger than the fractions off the diagonal.

		women				a_i
		black	hispanic	white	other	
men	black	0.258	0.016	0.035	0.013	0.323
	hispanic	0.012	0.157	0.058	0.019	0.247
	white	0.013	0.023	0.306	0.035	0.377
	other	0.005	0.007	0.024	0.016	0.053
b_i		0.289	0.204	0.423	0.084	

FIGURE 15

Figure 16 is a picture of the ages of husbands and wives from the National Survey of

Family Growth, showing their age at the time they got married. What you should pick out from this is that most of the dots fall along the diagonal, indicating that people tend to marry people about the same age as them. Of course, marriage is not really a network. I assume people are only married to one person at a time. I am using it here as a proxy for friendship. People become friends before they get married.

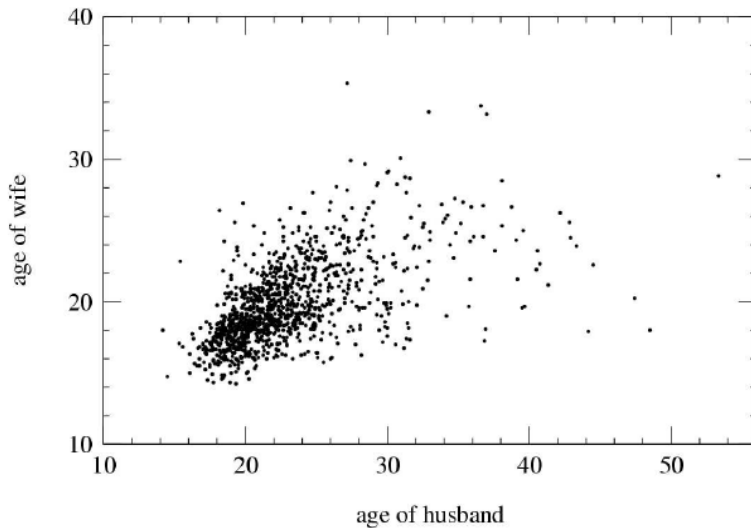


FIGURE 16

I am going to look at one type of assortative mixing. People tend to associate with other people like themselves. One way in which they can do that is if people who have many connections in the network tend to be connected to other people with many connections: gregarious people hanging out with other gregarious people and the hermits with other hermits. That would be assortative by degree. Remember, degree is how many connections you have.

In social networks, the gregarious people are hanging out with the gregarious people, and for all the other kinds of networks, at least the ones I have looked at, you have disassortative mixing, where the high-degree nodes are connected to the low-degree nodes. This could be a bad thing, for example, if you were concerned about the spread of disease. Certainly it is a bad thing if a person with many connections gets a disease, because they could pass it on to many other people. It is even worse if the people that they are connected to are, themselves, people with many connections, because then it spreads very easily amongst the people with many connections and, hence, could be much worse than if those people had fewer connections.

It is very easy to spot the difference between these two types of networks. Networks that

are assortative by degree tend to get all the people with large numbers of connections stuck together, and they form a big clump in the center of the network with a sort of peripheral ring of lower-connected people around them. This is shown on the left in Figure 17. In the disassortative case, you tend to get star-like structures, where you have a person with many connections connected to a bunch of people with only a few connections. These things are very distinctive. You start to be able to spot them; you just draw a picture of the network on the page and see that it is a disassortative network: it has these star-like structures in it.

This has practical repercussions. I just mentioned the one about disease. Figure 18 shows an analytic calculation that represents the same kind of thing. What it is representing basically is that disease spreads more easily if the network is assortative than if it is disassortative, and sadly, we find that most social networks are assortative. So, that could be a bad thing.

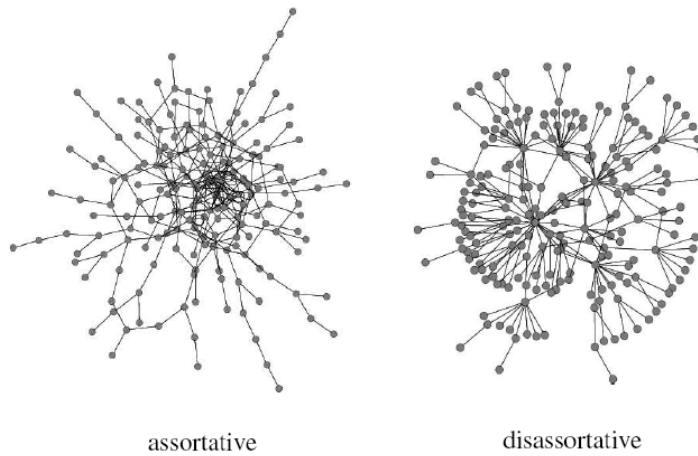


FIGURE 17

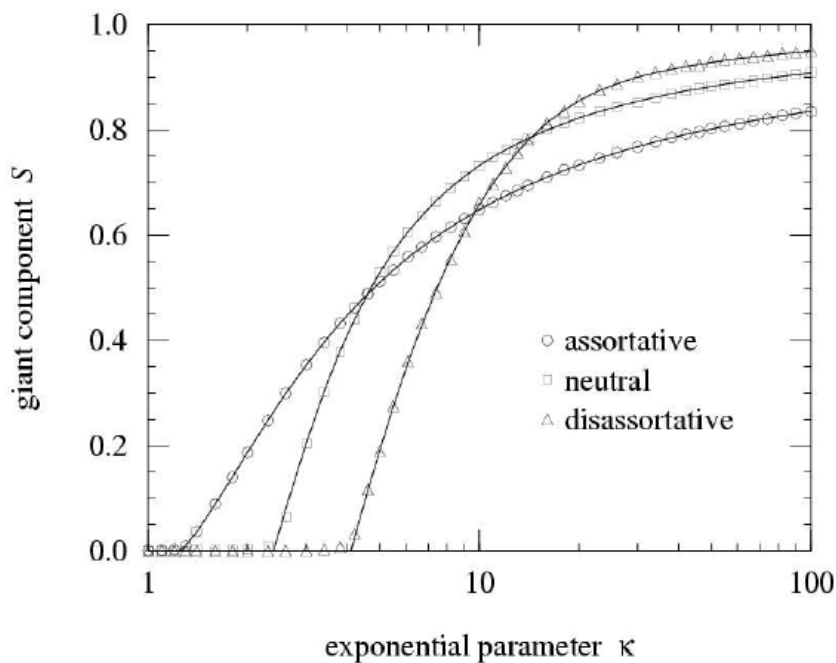


FIGURE 18

Here is another example, dealing with the robustness of networks. One thing people have been interested in the context of the spread of disease is how robust the network is. If I want to break a network up so that it is not connected any more, how should I do that? Which nodes in the network should I look to first? For example, you might imagine that you had a vaccination campaign, and you wanted to vaccinate certain people against a disease. Who should you vaccinate against the flu? Usually they vaccinate old people but, in a way, that is crazy because in many cases old people hardly talk to anybody, so they are unlikely to pass the disease on. You vaccinate them to prevent them from getting the disease, but you don't have the knock-on effect of preventing the people that they would have infected from getting the disease because there were no such people. Rather, you should be targeting school children or people who have many contacts. Not only can you protect them, but they are likely to spread it to other people and you can protect the people they would spread it on to.

So, the question is who should be targeted? Figure 19 shows on the vertical axis the size of the largest component. Roughly speaking, that is the maximum number of people that can catch this disease in the worst possible case. For the horizontal axis I am removing vertices, and removing precisely the ones that have the most connections, so I am trying to get the ones that

will pass the disease on to the most people. In the assortative cases, we have to remove a much larger fraction of the vertices—preventing the spreading by our vaccination campaign—before the size of the epidemic goes to zero, while well-targeted vaccination can have more impact in the disassortative case. Again, unfortunately, the assortative case, which is common in real social networks, is much more robust to the removal of these vertices. The hand-waving explanation for that is they are all clumped together in this big clump in the middle, and you are removing a lot of people from the same spot, all in the middle of the network, and you are leaving the rest of the network untouched. Because they all clump together, you are attacking one local bit of social space and not all the rest of it.

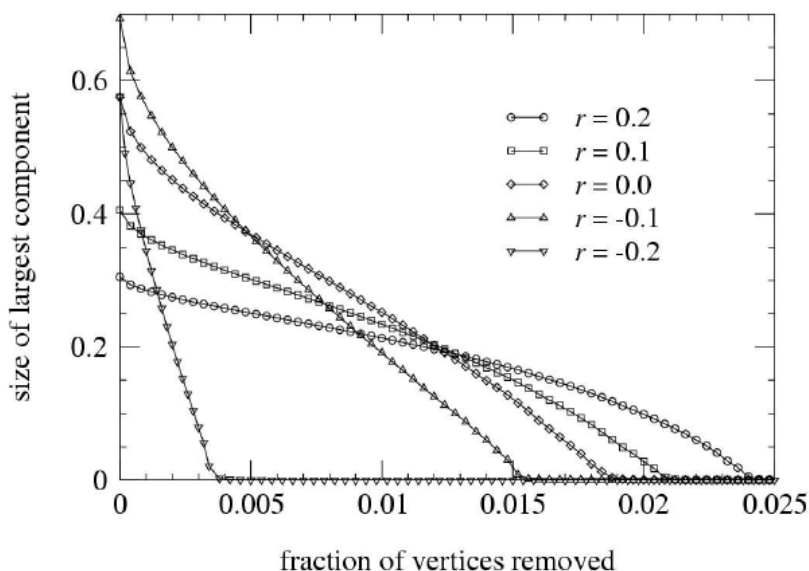


FIGURE 19

To tie this in with what I started to talk about at the beginning, I believe these two things could be connected. It is an interesting pattern that you have with disassortative non-social networks and assortative social networks. How could that be? There are two sides to the answer to that question, and I will give you one of them. How could you get assortativity in social networks? I believe that it comes from this idea of community structure. This is my explanation. If you have people in a bunch of different groups, as in Figure 20—here the groups are not connected together at all, but I will change that in a moment—then someone in a small group can only have a few friends because they are in the small group. What is more, the people they are connected to can only have a few friends because they are in the same small group. If you are in a large group, then you can have lots of friends, and your friends can have lots of friends. Just by

virtue of the fact of being divided into groups of varying sizes, you can get this assortative mixing by degree, the people in those small groups with other people in small groups, and the same for large groups.

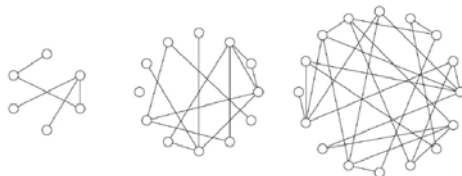


FIGURE 20

You can actually make a slightly more sophisticated model of this as follows: Figure 21 shows a bunch of people, A through K, and groups 1 through 4 that they belong to. This is the so-called bipartite graph, which I am going to project down onto just the people. In panel (b) you see just the people, and an edge between two people if they share a group in (a). Then I am going to put in just some fraction of those edges, because I am assuming that not everybody knows everybody else in the same group. Just because you are in a group together doesn't mean you know everybody in that group, so only some fraction of the edges are present here. This makes a simple model of what I showed on Figure 20, and this is actually something that we know how to do mathematically—take a bipartite graph, do a one-mode projection onto just one half of the bipartite graph, and then take out some fraction of the edges. In physics that is what we would call bond percolation. We know how to do each of those steps analytically, and we can solve this model and prove that it does always give you positive correlation between the degrees.

In theory that could explain how you get this positive correlation in social networks. I don't have time to talk about the negative half. I would be happy to answer any questions, but I have run out of time and I will stop talking now. Thank you very much.

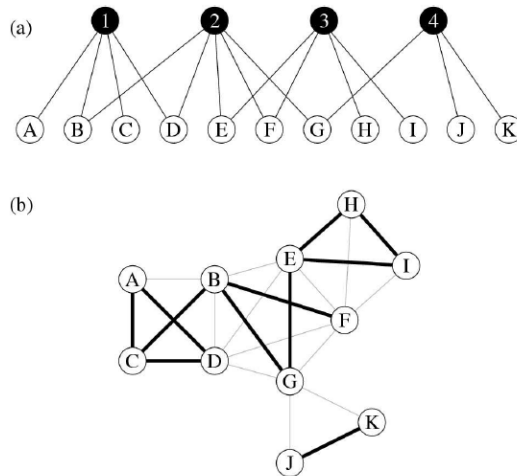


FIGURE 21

QUESTIONS AND ANSWERS

DR. BLEI: I have a quick question. Your modularity measure, where did that expectation come from? I must have missed it.

DR. NEWMAN: The expected number of connections? Basically what we do is, calculate the expected number conditional on the degree sequence of the network and nothing else. You make a model in which you say, I know the assignment of the vertices to the groups, and I know the degree of each vertex, and then, based only on that, it is actually relatively straightforward to just calculate what the expected number of edges would be in them.

One could certainly use a more sophisticated model if you knew other information about the network, such as correlations between the degrees of vertices, but in that particular measure that we looked at, we did not do that. So, it was only conditional on group membership and the degrees.

DR. HANDCOCK: Obviously, looking at clustering as a canonical problem in social networks, my question is, how does your method compare to other algorithmic methods such as block modeling, generalized block modeling, and also model-based methods such as Nowicki and Snijders's method.

DR. NEWMAN: That is a good question. So, there are a lot of methods that have been used in sociology before, some of which worked quite well. For me, the advantage of our method over some of these algorithmic methods is not necessarily how well they work but sort of the transparency of the rationale. I feel like there is sort of a principle for why this thing should work. So, it is clear, for example, if this might not be the appropriate method for your problem, or why it might not work in a particular case. I think that it is important to understand those kinds of things when you are doing statistics on networks, because you don't want to be sort of plugging the things into a block box. You want a nice simple method so you can see whether you think it will work. However, there are, of course, other very simple methods that have been proposed by sociologists that work well, like standard hierarchical clustering method, you have to propose some measure of similarity between vertices before you can do the hierarchical clustering of them, and of course, other very simple methods that have been proposed by sociologists that work well, like standard hierarchical clustering methods. I think those have many of the same advantages as ours does, but they work in very different ways. For example, in your typical additive hierarchical clustering method, you have to propose some measure of similarity between vertices before you can do the hierarchical clustering of them, and of course the results that you are going to get out depend on the particular similarity measure that you propose and then, again, the good think about that is that it allows you to see whether that would be appropriate for your problem or not. The bad thing is that it might not be, because you have to make some choice there about the things you are doing.

I would say that our method is doing many of the same things that these other methods are doing, and those other methods might be more appropriate, indeed, for some problems, but we also think our method is more appropriate for some other ones. As I mentioned many people have worked on this and there are many other interesting methods, some of which I really like, that have been proposed recently, which I didn't have time to talk about.

DR. BROWN: Let's take one final question.

DR. FEINBERG: The stories you told for the examples, which is the rationale in your answer to Mark, don't quite get at a dimension that was part of the story. That is, are these models really generating the graph or are they simply observational descriptive models for the graph the way it appears? When you told the story about how you might create the graph and build bridges, that has a dynamic generative characteristic, but that is not what you did.

DR. NEWMAN: No, it is not. One could imagine doing that kind of thing. In some sense, this last model that I mentioned here is doing that kind of thing. No, we didn't do any of those sorts of things. I think that there is definitely room for doing these sorts of things here.

One could, in theory, generate a network like this and then ask, how well does the network generated fit the real world data? That would be the forward problem, and we are not doing that in most of the cases we were talking about. You are absolutely right; we are really doing the backward problem.

REFERENCES

Morris, M. 1995. "Data Driven Network Models for the Spread of Infections Disease." *Epidemic Models: Their Structure and Relation to Data*. Cambridge, U.K.: Cambridge University Press.

U.S. Department of Health and Human Services, National Center for Health Statistics. 1995. *National Survey of Family Growth, Cycle V*. Hyattsville, MD.

Dimension Selection for Latent Space Models of Social Networks

Peter Hoff, University of Washington

DR. HOFF: I am going to talk about estimating parameters in generative models of social networks and statistical models for relational data. What I mean by relational data is data where you have data in an array, a 2-dimensional array, and you have row objects and column objects. The row objects could be a variety of things. They could be tumor tissue samples, and the columns could be genes or proteins that are being expressed. The rows could be countries, and the columns could also be countries and you are measuring interactions between countries, and I am going to have a data analysis example of that.

Inferential goals in the regression framework

Let $y_{i,j}$ be the network measurement on (i, j) , and $x_{i,j}$ be a vector of explanatory variables. If $y_{i,j}$ is binary, then maybe we have

$$Y = \begin{pmatrix} 1 & 0 & \text{NA} & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & \text{NA} & 1 & \text{NA} & \dots \\ 0 & 1 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad X = \begin{pmatrix} x_{1,2} & x_{1,3} & x_{1,4} & x_{1,5} & \dots \\ x_{2,1} & & x_{2,3} & x_{2,4} & x_{2,5} & \dots \\ x_{3,1} & x_{3,2} & & x_{3,4} & x_{3,5} & \dots \\ x_{4,1} & x_{4,2} & x_{4,3} & & x_{4,5} & \dots \\ x_{5,1} & x_{5,2} & x_{5,3} & x_{5,4} & & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Consider a basic generalized linear model

$$\log \text{odds}(y_{i,j} = 1) = \beta^T x_{i,j}$$

A model can provide

- a measure of the association between X and Y: $\hat{\beta}$, $\text{se}(\hat{\beta})$
- predictions of missing or future observations: $\Pr(y_{1,4} = 1 | Y, X)$

FIGURE 1

In social network analysis the row labels are the same as the column labels, where we are measuring relationships between these nodes. I am going to use the old workhorse of multivariate analysis, the singular value decomposition, to analyze these types of data. Just a notation here: suppose we have measurements as shown in Figure 1, where y_{ij} is the measurement in the i -th row and the j -th column; we could think of it as the measurement of the relationship

between row object i and column object j . In a social network data set these are typically 1 and 0 indicating the presence of a tie or not. Often when we are analyzing data sets what do we want to do? We want to look at patterns in the data set, perhaps relate the network data to some explanatory variables, these x 's on the right in Figure 1, and perhaps make predictions and so on.

Let's start off with a very naive strawman model. Consider what you might first do if you saw these data. You might say, binary data: I am going to fit a logistic regression relating these explanatory variables to my network observations. You model the log odds of there being a tie between two nodes as $\beta^T x_{i,j}$. So, β is measuring the relationship between these variables. A model like this can provide a measure of the association between X and Y , via regression coefficients and standard errors. You can also make predictions of missing or future observations. A lot of times you might have missing data in there and you want to predict what those data are from the observed data. Often what you want to do after this is to know what sort of patterns there might be after controlling for these explanatory variables. The very naive logistic regression model says you have all these data, and I am going to model these data as being independent, and the log odds of each tie is given by this $\beta^T x_{i,j}$.

This type of model does not recognize some very obvious things that we know might be present from multivariate analysis. It doesn't recognize that $y_{i,j}$ and $y_{i,k}$ have something in common. That is, they have the row object in common. They have the sender of the tie in common. This model ignores potential within-node homogeneity, so its node i or row i has some characteristic about it. That characteristic might play a role in all of its relationships in that whole row. Similarly, you might have homogeneity within a column. The flip side of homogeneity within a node is across-node heterogeneity. Those are two sides of the same coin. This is summarized in Figure 2.

The effects of homogeneity

Logistic regression model:

$$\Pr(Y|X, \beta) = \prod_{i \neq j} \theta_{i,j}^{y_{i,j}} (1 - \theta_{i,j})^{1 - y_{i,j}}$$
$$\log \frac{\theta_{i,j}}{1 - \theta_{i,j}} = \beta' x_{i,j}$$

This model does not recognize that $y_{i,j}$ and $y_{i,k}$ have something in common, i.e. the sender i . The model ignores potential within-node homogeneity / across-node heterogeneity. Ignoring these can interfere with our inferential goals:

- $\{y_{i,j}\}$ are not statistically independent: parameter estimates and confidence intervals for β are suspect.
- The model will have sub-par predictive performance: $y_{i,j}$ may give more information about $y_{i,k}$ than does $y_{i,m}$.

FIGURE 2

Models that ignore this type of homogeneity can wreak havoc with our inferential goals. If we have this type of heterogeneity, the $y_{i,j}$'s are not statistically independent, so parameter estimates and confidence intervals for these coefficients are suspect. More importantly, if we are trying to make predictions, a type of model like this will have very bad predictive performance. Because we know that potentially $y_{i,j}$ may give more information about, say, $y_{i,k}$ than does some observation from something in a different row and a different column. We would like to have models which have this sort of possibility for within-node or within-row or within-column homophily, as was talked about in the previous talk. I guess Mark lifted an overhead from Martina's talk; I am going to lift an overhead from Mark Handcock's paper. Figure 3 shows a network and how within-node homogeneity or across-node heterogeneity might be manifested in this network. One of the first things we might observe is that some of these nodes have just one ingoing or one outgoing tie, whereas some other nodes have many, many, many ties.

How is within-node homogeneity manifested in network structure?

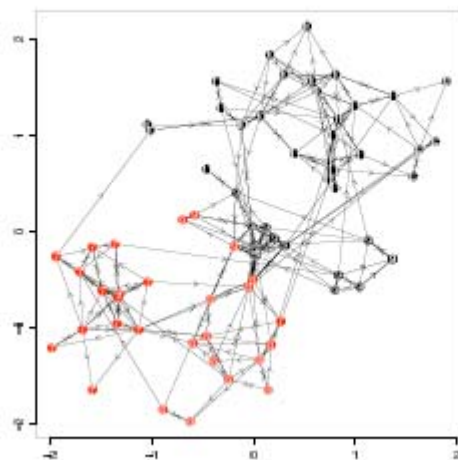


FIGURE 3

The first thing we might want to do to model this type of heterogeneity is to add something to our logistic regression model, and this is perhaps the first sort of random effect or latent variable model that is used for these types of data. We want to add some parameter to model, say, the out-degree of node i and the in-degree node j , or the expansiveness of object i and the sort of popularity of node j . If you include these in your model, you can fit these parameters, and almost every data set that I have ever played around with, when you do this, it dramatically improves the fit of the model, both within sample and out of sample. It dramatically improves how well you are fitting the data, but also if you try to do a measure about a sample prediction using cross validation, you see that adding these parameters tends to dramatically improve the fit. That's the obvious thing that a statistician would do as a first pass, but in network data there is just an additive structure on this log odd scale. There is probably a lot more going on, and we know this from studying the sociology of these social networks.

Some of the other types of things we see that could be caused by within node homogeneity, as was discussed in the previous talk, you can imagine that there are a bunch of nodes, and they all have similar characteristics, which you don't observe.

What you do observe in the network is that maybe all these nodes are connecting to one another. You might imagine that you have these clusters of nodes, and perhaps that was caused by the fact that a bunch of nodes have some sort of common characteristics that they are clustering upon. We see these things called transitivity. Transitivity is the concept that, if one node ties to

another node, if node i ties to node j , and node j ties to node k , then more often than not there is a tie between i and k .

Those are the sorts of possibilities we want to allow for in our models of network data or relational data. How might we incorporate this in a statistical model? Well, these things can't be represented by row and column effects. So, maybe we need something more complicated than that. You might say, let's just fit an interaction term. That is what a statistician might think, looking at these data. You have row effects, column effects, and any deviations from that would be fit with an interaction. You don't have enough data to fit a single separate parameter for every cell in the array. What people have done, or at least these few people here, is try to come up with reduced-rank models for these interactions, trying to simplify this interaction.

Latent variable models

Perhaps the first "latent variable" model was similar to a p_1 -type model

$$\text{logodds}(y_{i,j} = 1) = \beta'x_{i,j} + a_i + b_j$$

a_i and b_j induce within-node homogeneity that is additive on the log-odds scale.

Inclusion of these effects in the model can dramatically improve

- within-sample model fit (measured by R^2 , likelihood ratio, BIC, etc.);
- out-of-sample predictive performance (measured by cross-validation).

But this model only captures heterogeneity of outdegree/indegree, and can't represent more complicated structure, such as clustering, transitivity, etc.

With $n \times (n - 1)$ observations, maybe we can fit a richer model than the $(2 \times n)$ -parameter row/column effects model:

$$\text{logodds}(y_{i,j} = 1) = \beta'x_{i,j} + a_i + b_j + \gamma_{i,j}$$

- $\gamma_{i,j} = \alpha \times 1(\text{class } i = \text{class } j)$ (Nowicki and Snijders 2001)
- $\gamma_{i,j} = -|z_i - z_j|$ (Hoff, Raftery and Handcock 2002)
- $\gamma_{i,j} = z_i'z_j$ (Hoff 2005)

FIGURE 4

One such model is the sort of stochastic block structure or latent class model by Nowicki and Snijders, and this was the idea that was mentioned before, that maybe nodes belong to some class, and nodes form ties with people in the same class or category preferentially. This is shown in Figure 4. The important thing to realize in these types of models is that this class variable is not observed. You see these data, and then you try to use the concept of a latent class variable to model the structure that you see. If you see a network and there are lots of clusters, you can represent that cluster structure with the latent class model. These models can describe the

patterns that you see in the network. You don't see the class a priori.

Another model is to imagine that these nodes are lying in some unobserved latent social space, and that ties are formed preferentially between nodes that are close by one another. This type of structure can actually represent statistically the sort of transitivity we saw earlier and the clustering we saw earlier. More recently I have grown quite fond of a different way to represent a similar idea, which is through latent interproduct models. The idea here is that every row has a set of characteristics as a row, every column has a set of characteristics as a column, and that the relationship between a row and a column are strong, if the row and column latent characteristics match or line up. This is that type of model. This is a model that I used as a symmetric model, and it is useful for relationships that are kind of positive in a symmetric way.

Latent SVD model

The $\gamma_{i,j}$'s form an $n \times n$ matrix $\{\gamma_{i,j}\}$. Recall that every $n \times n$ matrix has a singular value decomposition:

$$\begin{aligned} \{\gamma_{i,j}\} &= \mathbf{U}_{n \times K} \mathbf{D}_{K \times K} \mathbf{V}'_{K \times n} \\ \gamma_{i,j} &= \mathbf{u}_i \mathbf{D} \mathbf{v}'_j \end{aligned}$$

where

- K is the rank of $\{\gamma_{i,j}\}$;
- $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_{K \times K}$, $\mathbf{D} = \text{diag} \{d_1, \dots, d_K\}$;
- $\mathbf{u}_i = \mathbf{U}_{[i, \cdot]}$, $\mathbf{v}_j = \mathbf{V}_{[\cdot, j]}$.

There is a large literature on the SVD for relational data... **Interpretation:**

- $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ are latent row characteristics ,
- $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ are latent column characteristics.
- if latent characteristics vectors are similar, then $\gamma_{i,j}$ is large:

$$\gamma_{i,j} = \sum_{k=1}^K d_k u_{i,k} v_{j,k}.$$

FIGURE 5

This is all leading to what I am going to talk about now, the singular value decomposition, coming up with a model-based approach for the singular value decomposition, and incorporating it into network models. As sketched in Figure 5, the idea here is that we are going to try to represent this structure in a way that there is some sort of network interpretation to the model—so that we can actually estimate the parameters in that way. Many of you are familiar with the singular value decomposition. It basically is just that any matrix has a decomposition into a set of what are called, say, left singular vectors, which I am denoting \mathbf{U} , right singular

vectors, which I am denoting \mathbf{V} , and they have weights. So, this matrix \mathbf{D} is a diagonal matrix of positive elements. You can write the i,j -th element of this matrix as $U_i^T \mathbf{V}_j$ with some weights, given by \mathbf{D} . The idea is this type of model, this sort of decomposition has been used for relational data for many decades. There is a large literature based on it. The basic idea is that this $\gamma_{i,j}$ is some measurement of the relationship between i and j , and we are decomposing that relationship into saying there are some row effects for unit i , some column effects for unit j , and the relationship is strong if those two vectors are in the same direction, if they are similar in that way.

The nuts and bolts here, K is the rank of this matrix of gammas. The \mathbf{U} s and the \mathbf{V} s are typically constrained to be orthogonal matrices and orthonormal matrices.

The interpretation here, again, is that for each of the rows we have latent row characteristics, for each of the columns we have latent column characteristics.

The model says, let's try to represent the relationship between these nodes using these latent characteristics. I want to point out that this sort of representation can represent the types of structures that we see in social networks. It can represent heterogeneity of out-degrees and heterogeneity of in-degrees. It can represent transitivity and it can represent clustering, and it is a well known mathematical beast, so it is nice in that way.

I want to point out this earlier model we talked about here in Figure 6, this having row and column effects. This is just a very simple special case of a rank-2 matrix. The singular value decomposition is more general than that.

Row-column effects are a special case

$\gamma_{i,j} = a_i + b_j$ can be written in this form as a rank-2 matrix:

$$\{\gamma_{i,j}\} = \begin{pmatrix} a_1/||a|| & 1/||\mathbf{1}|| \\ a_2/||a|| & 1/||\mathbf{1}|| \\ \vdots & \vdots \\ a_n/||a|| & 1/||\mathbf{1}|| \end{pmatrix} \begin{pmatrix} ||a|| ||\mathbf{1}|| & 0 \\ 0 & ||b|| ||\mathbf{1}|| \end{pmatrix} \begin{pmatrix} 1/||\mathbf{1}|| & 1/||\mathbf{1}|| & \dots & 1/||\mathbf{1}|| \\ b_1/||b|| & b_2/||b|| & \dots & b_n/||b|| \end{pmatrix}$$

FIGURE 6

How I spent my summer vacation

Let Y be an $m \times n$ data matrix (suppose $m \geq n$). Consider a model of the form

$$Y = UDV' + E$$

where

- $K \in \{0, \dots, n\}$;
- $U \in \mathcal{V}_{K,m}, V \in \mathcal{V}_{K,n}$;
- $D = \text{diag}\{d_1, \dots, d_K\}, d_k > 0$;
- E is a matrix of independent normal noise.

I have recently developed a method to calculate

- $p(\{u_1, \dots, u_K\}, \{v_1, \dots, v_K\}, \{d_1, \dots, d_K\} | Y, K)$
- $p(K | Y)$

FIGURE 7

Figure 7 shows how I spent my summer vacation. If you have a data matrix Y , it is well known how to decompose this matrix Y into a set of left singular values, right singular values, and so on. I didn't want to just do that decomposition. I need to incorporate the singular value decomposition in a statistical model, and I also want to figure out what the dimension is and what an appropriate dimension is. Can we represent things with a rank 2 matrix, do we need rank 3 matrix and so on?

Here is a set up. We observed there is some data matrix Y and it is equal to some rank X mean matrix. Remember, any rank X matrix has a decomposition of this form, plus Gaussian noise. The idea is that we get to observe this Y . We know that Y is equal to some mean matrix plus noise, and we would like to make inference on what the mean matrix is. In other words, we want to make inference on what the U s are, the V s are, and the singular value D s are. That is actually not too hard of a problem. The really hard problem was figuring out the dimensions. If any of you have worked with trying to make inference on the dimension of a model you know that is kind of hard.

Prior distribution for \mathbf{U}

1. $\mathbf{z}_1 \sim \text{uniform}[m\text{-sphere}]$, set $\mathbf{U}_{[,1]} = \mathbf{z}_1$;
2. $\mathbf{z}_2 \sim \text{uniform}[(m - 1)\text{-sphere}]$, set $\mathbf{U}_{[,2]} = \text{Null}(\mathbf{U}_{[,1]})\mathbf{z}_2$;
- ⋮
- K . $\mathbf{z}_K \sim \text{uniform}[(m - K + 1)\text{-sphere}]$, $\mathbf{U}_{[,K]} = \text{Null}(\mathbf{U}_{[,1]}, \dots, \mathbf{U}_{[,K-1]})\mathbf{z}_K$.

The resulting distribution for \mathbf{U} is the uniform (invariant) distribution on the Steifel manifold, and is exchangeable under row and column permutations.

Prior conditional distributions:

$$\begin{aligned} (\mathbf{U}_{[,j]} | \mathbf{U}_{[,1-j]}) &\stackrel{d}{=} \text{Null}(\mathbf{U}_{[,1]}, \dots, \mathbf{U}_{[,j-1]}, \mathbf{U}_{[,j+1]}, \dots, \mathbf{U}_{[,K]})\mathbf{z}_j, \\ \mathbf{z}_j &\sim \text{uniform}[(m - K + 1)\text{-sphere}] \end{aligned}$$

FIGURE 8

I know that there are statisticians in the audience, so very quickly I am going to do a little statistics. We are going to use the tools of Bayesian inference to estimate these parameters. Figure 8 says just a little bit about the prior distribution for \mathbf{U} , and it is going to be the same as the prior distribution for \mathbf{V} . Remember, \mathbf{U} is this vector of latent characteristics for the rows. We are constraining in our model the \mathbf{U} and \mathbf{V} to be orthonormal matrixes. How do we come up with a prior distribution for an orthonormal matrix? Here is one way to do it. In many cases the first thing you think of turns out to be the thing you want. Remember, we are saying that \mathbf{U} is an orthonormal matrix, and that means that the columns of \mathbf{U} are all unit vectors and are all orthogonal to each other. How do we generate such a thing? What we can do first is sample—each column is a point on the sphere; it is a point on the N -dimensional sphere. We sample the first one, \mathbf{z}_1 , uniformly from the N -dimensional sphere, and we set the first column of \mathbf{U} equal to that vector. \mathbf{U}_2 has to be orthogonal to \mathbf{U}_1 . What we can do is sample \mathbf{z}_2 uniformly from the $(N-1)$ -dimensional sphere, and then use the null space of \mathbf{U}_1 , multiply that by \mathbf{z}_2 , and you get something that is a point on the sphere, and it is orthogonal to \mathbf{z}_1 , and so on. You keep doing that and finally you have created your orthonormal matrix. The result is that you can find out the characteristic function of this thing, and it turns out to be the uniform distribution on the set of orthonormal matrices, which is called the Steifel manifold.

More importantly, this probability distribution is exchangeable, which basically means that the distribution of column j , given the other columns, is the same as the distribution of \mathbf{U}_k , given the first $k-1$ vectors. This is going to be a tool which allows us to do parameter estimation.

There is a lot of really interesting multivariate analysis which goes on here under the hood, which I am going to gloss over. Basically, the conditional distribution of the j^{th} column, given the others has to be orthogonal to all the others, but it is uniform to the space that is orthogonal to the others. So, you get that U_j , given the others, is equal to a random point on the $(m - K + 1)$ -dimensional sphere, times the null space of the others.

Figure 9 shows more of the math. I tried to make it more transparent by using colors. How the inference usually works is that we have the data, we have Y . We have the prior distributions for all the parameters of interest, and then we combine those to get the posterior distribution of the parameters we want. We want to know what the probability distribution is for the unknown parameters, the \mathbf{U} , the \mathbf{V} , and \mathbf{D} , given the data Y .

Actually, as long as we can find the conditional distribution of each object, given all the other objects, we can make parameter estimation. The result is that everything works out very nicely and you get the posterior distribution the j^{th} column. Suppose you are interested in the j^{th} column of the latent row characteristics. The posterior distribution is proportional to the prior distribution times the probability of the data given the j^{th} column, as shown in Figure 9. This is just through the Gaussian model for Y . Remember the prior says that the j^{th} column has to be orthogonal to the other columns. It is orthogonal to all the other columns, but then it has some non-uniform distribution about where that lies. I would be happy to go over this in more detail later.

Posterior distribution for U

Let $\mathbf{E}_{-j} = \mathbf{Y} - \mathbf{U}_{[-j]} \mathbf{D}_{[-j,-j]} \mathbf{V}_{[-j]}$. Then

$$\begin{aligned} \|\mathbf{Y} - \mathbf{UDV}'\|^2 &= \|\mathbf{E}_{-j} - d_j \mathbf{U}_{[j]} \mathbf{V}_{[j]}\|^2 \\ &= \|\mathbf{E}_{-j}\|^2 - 2d_j \mathbf{U}_{[j]}' \mathbf{E}_{-j} \mathbf{V}_{[j]} + \|d_j \mathbf{U}_{[j]} \mathbf{V}_{[j]}\|^2 \\ &= \|\mathbf{E}_{-j}\|^2 - 2d_j \mathbf{U}_{[j]}' \mathbf{E}_{-j} \mathbf{V}_{[j]} + d_j^2. \end{aligned}$$

so

$$\begin{aligned} p(\mathbf{Y}|\mathbf{U}, \mathbf{V}, \mathbf{D}, \phi) &= (2\pi\phi)^{-mn/2} \exp\left\{-\frac{1}{2}\phi\|\mathbf{E}_{-j}\|^2 + \phi d_j \mathbf{U}_{[j]}' \mathbf{E}_{-j} \mathbf{V}_{[j]} - \frac{1}{2}\phi d_j^2\right\} \\ p(\mathbf{U}_{[j]}|\mathbf{Y}, \mathbf{U}_{[-j]}, \mathbf{V}, \mathbf{D}) &\propto p(\mathbf{U}_{[j]}|\mathbf{U}_{[-j]}) \exp\{\phi d_j \mathbf{U}_{[j]}' \mathbf{E}_{-j} \mathbf{V}_{[j]}\} \end{aligned}$$

$$\begin{aligned} (\mathbf{U}_{[j]}|\mathbf{U}_{[-j]}, \mathbf{Y}, \mathbf{V}, \mathbf{D}) &\stackrel{d}{=} \text{Null}(\mathbf{U}_{[1]}, \dots, \mathbf{U}_{[j-1]}, \mathbf{U}_{[j+1]}, \dots, \mathbf{U}_{[K]}) \mathbf{z}_j, \\ p(\mathbf{z}_j) &\propto \exp\{\mathbf{z}_j' \boldsymbol{\mu}\}, \quad \boldsymbol{\mu} = \phi d_j \mathbf{N}'_{-j} \mathbf{E}_{-j} \mathbf{V}_{[j]} \end{aligned}$$

FIGURE 9

The hardest part of all this was trying to evaluate the dimension. As I said before, we were saying with this model that basically each row has some latent characteristics, each column has some latent characteristic, and that is going to be our model for how these things interact. We know that if we have an $N \times N$ matrix, we don't want to say that each person has a vector of N characteristics, and that model is saturated. There is going to be too much noise and variability there. Also, we might think that we need more than one or two dimensions, so we would like to evaluate the dimension of the latent characteristics and select a model in a principled way.

Basically, the problem reduces down to this problem shown in Figure 10. This problem is asking, is \mathbf{Y} equal to a rank 1 matrix plus noise, or is \mathbf{Y} equal to, say, a rank 0 matrix plus noise? Trust me, if you can answer this problem you can address the problem of what the dimension is. If you know how to go from zero to one you know how to basically go from k to $k+1$. You need to evaluate the relative probabilities of model 1 to model 0, and that essentially boils down to the hard part of evaluating the probability of the data under model one relative to model zero. Basically, you have to do some really complicated integration to do this, but it is fun for some people.

Evaluating the dimension

It can all be done by comparing

$$M_1 : Y = d\mathbf{u}\mathbf{v}' + E$$

$$M_0 : Y = E$$

To decide whether or not to add a dimension, we need to calculate

$$\frac{p(M_1|Y)}{p(M_0|Y)} = \frac{p(M_1) p(Y|M_1)}{p(M_0) p(Y|M_0)}$$

The numerator of the Bayes factor is the hard part. We need to integrate the following w.r.t. the prior distribution:

$$\begin{aligned} p(Y|M_1, \mathbf{u}, \mathbf{v}, d) &= (2\pi)^{-nm/2} \exp\left\{-\frac{1}{2}\phi\|Y\|^2/2 + \phi d\mathbf{u}'\mathbf{Y}\mathbf{v} - \phi d^2/2\right\} \\ &= p(Y|M_0) \times \exp\{-\phi d^2/2\} \times \exp\{\mathbf{u}'[d\phi\mathbf{Y}]\mathbf{v}\} \end{aligned}$$

FIGURE 10

I had knee surgery this summer, and I think this problem wouldn't have gotten done if I hadn't had the surgery. After the surgery I had to spend every day lying in my backyard with a pad of paper. I couldn't go out and do fun things; I had to sit there and do integrals. The hard part is you have to integrate the expression at the bottom of Figure 10. Basically we are saying the probability of the data under model 1, you can actually write it as the probability of the data under model 0, and then a part that sees whether or not adding these vectors will contribute to explaining the variation in Y . You have to integrate these over the uniform distribution on the sphere for \mathbf{U} and \mathbf{V} , and that is difficult, involves some Bessel functions and stuff like that.

I am now going to show how this might actually work in practice. For those of you who have taken some multivariate data and said, ah, I have genes on this thing on the rows and I have tumors on the columns, and I want to see the effects going on here, maybe you have looked at the singular value decomposition of your data to try to sort out what is going on in the data. If you have ever looked at plots like Figure 11, basically we decomposed the matrix and looked at the singular values. The singular values are saying how much variation there is in the matrix in each dimension. If you had a data matrix Y and it was just equal to noise, and if you have played around with these things, you know that it is going to look like a decaying graph.

A toy example

$$Y_{40 \times 30} = UDV' + E = \sum d_k U_{[:,k]} V'_{[k,:]} + E$$

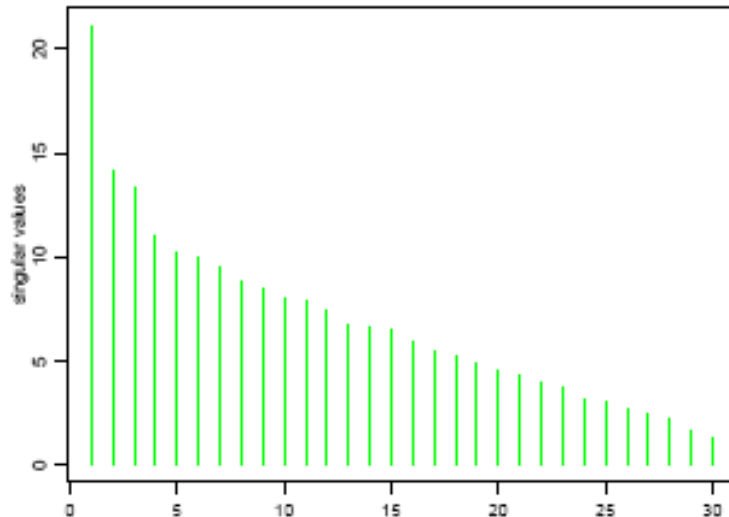


FIGURE 11

I simulated some data using some rank k . Does anybody want to hazard a guess as to what the rank of my mean matrix was? It probably wasn't zero. I have a vote for four. That seems to be popular. Why might it be four? If you have played around with these things you know that you are looking for gaps in the magnitude from one singular value to the next. You know it is definitely more than zero. In Figure 11 you can see a big gap between 1 and 2, so it is at least 1, but then there is this big gap between 3 and 4, and then there is another little gap between 4 and 5, and then it seems to be monotonically going down. So 4 is a really good guess; the actual answer is 5, as shown in Figure 12. This goes to show you how well our eye is at doing these things.

A toy example

$$Y_{40 \times 30} = UDV' + E = \sum d_k U_{[:,k]} V'_{[k,:]} + E$$

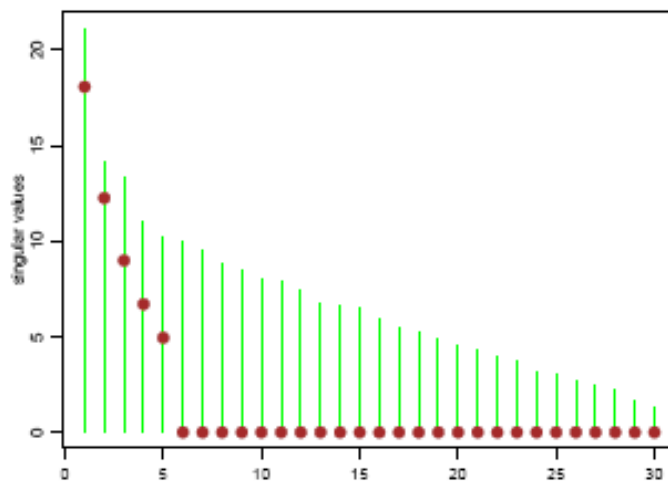


FIGURE 12

Figure 13 shows the actual singular values that I used to generate the data matrix. Four is a fair guess because the fifth non-zero singular value is smaller than the others. I used the Bayesian machinery. You can get a posterior distribution for the mean matrix, and then you can look at the posterior mean of all the singular values. If you add up all the blue, it is measuring the total amount of variation that the model picked up in the data. Here is a posterior distribution on the rank of a matrix. This should say that Bayesian methods are good, because you had your intuition about what the rank was, and the Bayesian estimate matched up to that to a great degree. The posterior distribution says there is a very good chance that there are four or more dimensions to this mean matrix. The two highest probability ranks are four and five, as you might expect.

A toy example

$$Y_{40 \times 50} = UDV' + E = \sum d_k U_{[:,k]} V'_{[:,k]} + E$$

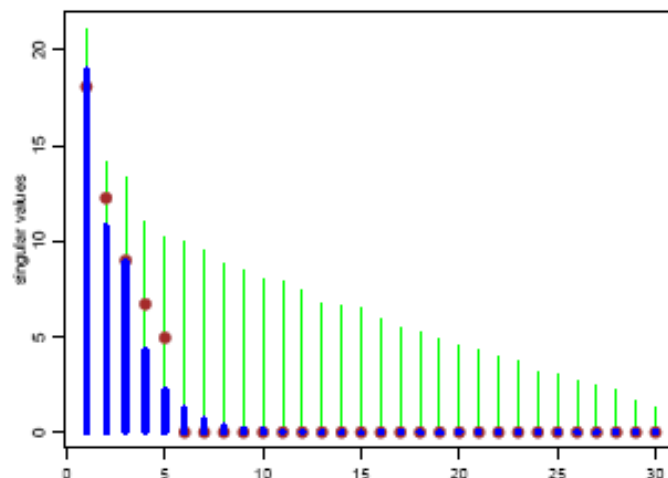


FIGURE 13

One other thing about doing Bayesian inference, if you do the singular value decomposition—here is my little plug for Bayesian analysis—suppose I told you the rank was 5, what is your least-squares estimate of the mean matrix, given that you know the rank is 5? Well, the least-squares estimate is actually the first five singular values and singular vectors, so you would just be taking the top of the green bars for the first singular values there. If you take the true matrix that was used to generate the data, tab all those entries, and plot them against the least-squares estimate, you get the green pattern shown on the right side of Figure 14.

A toy example

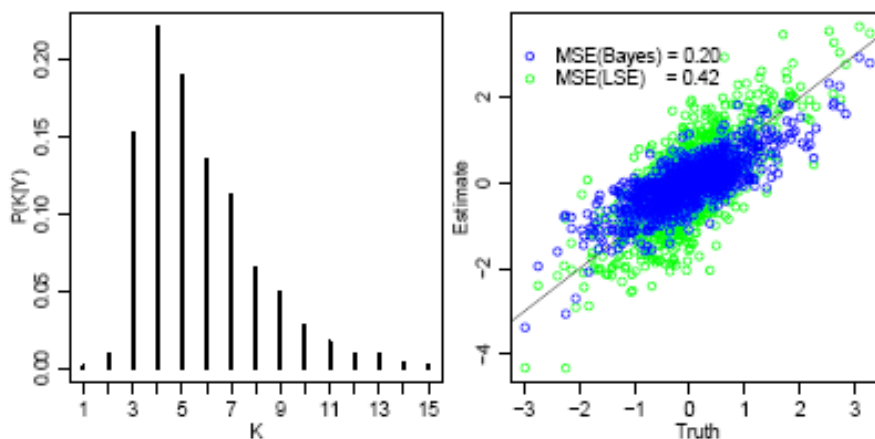


FIGURE 14

If you take the Bayesian estimate you get something with a much lower mean square. What is happening here is the singular value decomposition is taking our data matrix Y and it is looking at the projections onto these subspaces. Data matrix Y is equal to the mean matrix plus error so the singular value decomposition is projecting all the error as well onto the subspace. It is probably best to point out that looking at the range of the truth is -3 to 3 , and the range of the singular value decomposition is essentially -4 to 4 . That is not atypical of these types of problems that sort of a least-squares estimate overestimates the variability in the matrix.

The reason why I went through all that work was because I wanted to embed the concept of the singular value decomposition model into more complicated models for social networks. I wanted to incorporate this into very general models for all sorts of relational data, not just Gaussian data, not just binary data, but data of all sorts. The way I do that is with a generalized linear model. Figure 15 shows the general model for relational data. We have a sociomatrix Y . All of this generalizes to the case where Y is an $n \times m$ matrix; it doesn't have to be a square matrix.

Returning to networks

General model for relational data : Let

- Y be an $n \times n$ sociomatrix of possibly valued relations;
- X be an $n \times n \times p$ array of explanatory variables.

Our latent variable model relating X to Y is

$$\begin{aligned} \Theta &= X\beta + Z \\ Z &= UDV + E \\ g(E[y_{i,j}]) &= \theta_{i,j} \quad \text{i.e.} \\ g(E[y_{i,j}]) &= \beta'x_{i,j} + u_i'Dv_j + e_{i,j} \end{aligned}$$

For example, some potential models are

- If $y_{i,j}$ is binary, $\log \text{ odds } (y_{i,j} = 1) = \beta'x_{i,j} + u_i'Dv_j + e_{i,j}$
- If $y_{i,j}$ is count data, $\log E[y_{i,j}] = \beta'x_{i,j} + u_i'Dv_j + e_{i,j}$
- If $y_{i,j}$ is continuous, $E[y_{i,j}] = \beta'x_{i,j} + u_i'Dv_j + e_{i,j}$

FIGURE 15

We have a set of explanatory variables X , and our model, relating X to Y , is as follows: we have what we call a predictor, θ , and it is equal to a regression part plus a latent structure part. This latent structure part is just this model that we talked about, the singular value decomposition model. The generalized linear model is not that our observed data is equal to $X\beta + Z$, but some function of the mean is equal to $X\beta + Z$. Some function of the mean of $y_{i,j}$ is equal to the regression part, a latent structure part, and lack of fit. In this context, you can fit binary data where you model the log odds of the observations as being equal to the structure. If you have count data you might model things on the log scale and, if it is continuous, you might model it on the additive scale, using the identity link.

I should point out that a lot of data that people have been talking about is skewed data. I wouldn't call it high-variability data, but I would call it skewed data. You have not just the first and second moments there, you have the third and fourth moments, and those are really the things that can actually represent that skewed structure. That can also be incorporated in here, although with more difficulty. You have to decide on what scale you are making the measurements. Actually, you could do continuous data and use a log link to model skewed data.

An example of what I am going to show in the next two slides is some international conflict data. The model for how this really is generated is debatable, but it is a very nice data set because we can see the structure. Mark pointed this out in the last talk. It is good to work on

problems where you can tell if the answer makes sense or not. I have some colleagues in political science who are interested in modeling conflicts between nations. The data is from Mike Ward and Xun Cao, and what they measure is the number of militarized disputes initiated by i with target j over a 10-year period. This involves actual militarized disputes and conflicts as well as threats and a bunch of other things that they have recorded. They want to relate these to a bunch of covariates, so they have a bunch of theories about what causes nations to have conflicts with others and what things might ameliorate conflicts. There is this concept of democratic peace that democratic nations don't have conflicts with each other. There is this idea that trade inhibits conflicts and so on. They like to evaluate these things and look at the structure that is left unexplained by these covariates. Figure 16 provides an overview.

International Conflict Network, 1990-2000

11 years of data on international relations and national characteristics
(thanks to Mike Ward and Xun Cao)

- $y_{i,j}$ = number of militarized disputes initiated by i with target j ;
- $x_{i,j}$ a 10-dimensional covariate vector containing
 1. log distance
 2. log imports
 3. number of shared intergovernmental organizations
 4. log population of initiator
 5. log population of target
 6. log gdp of initiator
 7. log gdp of target
 8. polity score of initiator
 9. polity score of target
 10. polity score of initiator \times polity score of target

Model: $y_{i,j}$ are independent Poisson random variables with log-mean

$$\log E[y_{i,j}] = \beta' x_{i,j} + u_i' D v_j + \varepsilon_{i,j}$$

FIGURE 16

We fit this into our model. We say that we are going to model the mean of each observation on the log scale with this regression part plus this latent structure. I was very happy to see the posterior distribution on the number of dimensions of the latent structure as shown in Figure 17. Why was I happy? It's because it is easy to plot 2-dimensional latent structures. You can plot 3-dimensional latent structure, but this k could have been anywhere between 0 to perhaps 130. If the dimension of the latent structure was 73, I wouldn't be able to give you a very good

picture at the end of what is going on, or you could look at the first few dimensions of the latent structure. I should also point out that if you play with these data enough, you know that there is no way that the mean structure has dimension less than two. There are lots of countries that are very active in conflict, and a lot of countries that are attacked a lot; you know that you are going to have row and column effects, and that is a 2-dimensional structure. Therefore, you know the dimension has got to be at least two.

International Conflict Network, 1990-2000: Parameter Estimates

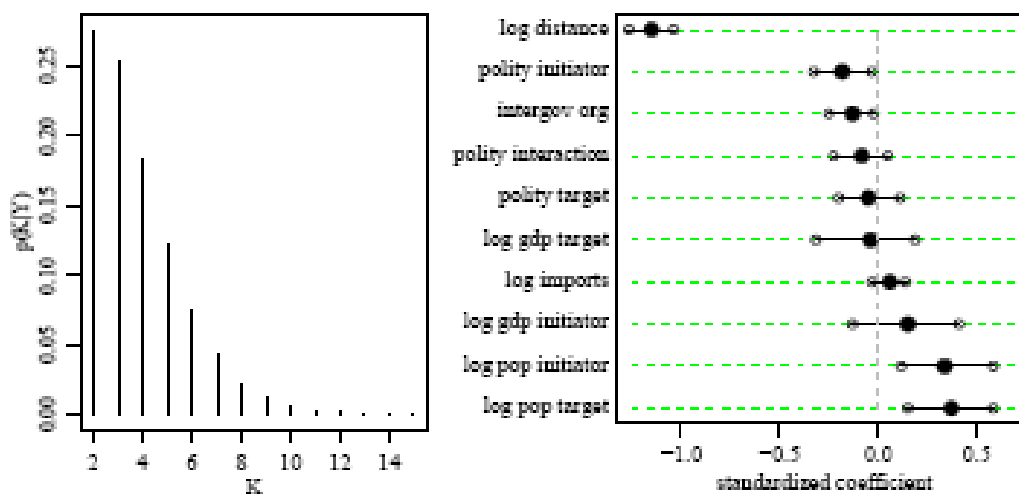


FIGURE 17

The right half of Figure 17 shows some confidence intervals on the regression parameters. There are some obvious things to point out. As distance increases, the amount of conflict decreases. People have conflicts with countries that are near them, and population plays a big role. Populations of countries that are in conflicts tend to be higher than average.

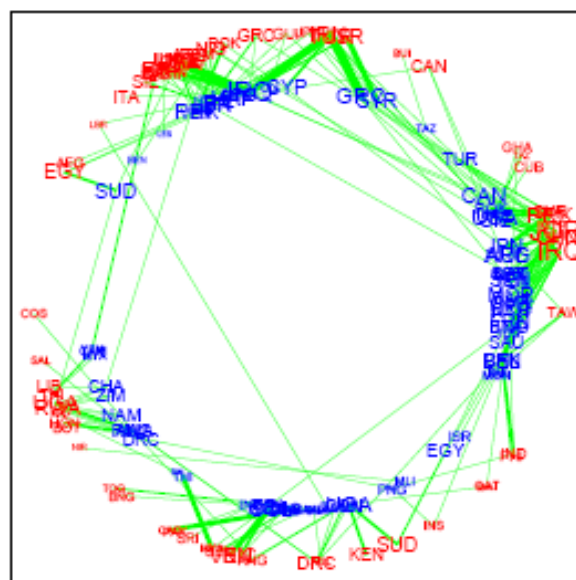


FIGURE 18

Figure 18 shows the latent structure. I will finish by explaining this plot. Recall the model. The model for each country has a 2-dimensional vector of characteristics as an attacker and a 2-dimensional vector of characteristics as a target, so that each one of these countries has a vector as an aggressor. What I have done here is plot the direction of each country's vector in red and that is this outer circle here. The magnitude of their plotting character is the magnitude of their vector in that direction. The inner circle is the direction of each country's 2-dimensional latent vector as a target. Then I have drawn these green lines indicating the strength of the number of conflicts between the countries. So, you see somewhat obvious things because we know the names of the rows and the names of the columns. Those are the main structures there.

Here is where I see if I made myself clear. If there wasn't any latent structure, can anybody guess what this plot would look like, or what would be different about this plot? Where would the green line be? If there were no two dimensional latent structure, there would be no relationship between the actual links, or attacks between the countries, and their position in this circle. All you would see is a bunch of green lines scattered all over the place like that so you can see how these positions really do pick up the structure that is there.

To summarize, my point is that singular value decomposition is a well-known tool in multivariate analysis that has been used for a long time, and there is a huge literature based on it, sort of in the additive case in the psychometric literature in lots of fields. It is something that is

well understood, although apparently the problem of identifying the dimension of the mean matrix has been left unsolved, at least in a Bayesian context, and actually in a non-Bayesian context it is also unsolved.

Using this framework, the idea is that we can use this natural singular value decomposition model that we are familiar with, and incorporate it into different types of data structures, along with regression effects and other things that we use when we are doing statistical modeling. That is the main point of the talk.

QUESTIONS AND ANSWERS

QUESTION: One quick question. Could you just say a couple of words on what you are thinking about here with the time series extensions, the dynamic extensions, because that is something that is sort of natural to the topics we are going to be covering.

DR. HOFF: I have always known that I should spend more time in the dynamic aspect of networks, but every time I gave a talk on latent variable models, everybody always asked, what is the dimension? So, I had to spend time doing the dimension first.

Now that that is done, there are a number of ways that you can do dynamic network inference for these things, or incorporate these ideas. One way is to have a time-series model for the latent characteristics, as a Markov model or something like that, and actually David Banks and Eric Vance are working on stuff like that. Again, I have some other ideas about how to set these things up in a general autoregressive structure. Their ideas are kind of interesting, they have a behavioral slant on things. They are using ideas from social networks like, I am at this location in the social space, and I form a tie with this person, and I should move at the next time step, in this other direction. There are sort of dynamical models, people starting to work on these things with these ideas here.

DR. BLEI: Your latent variables seem to model the connections between rows and columns that happen outside of the covariates. I can imagine, say, with countries, that two countries that don't get along, that the covariates explain that, that their latent positions would be far apart. Is there some way to deal with that?

DR. HOFF: In one sense, you could use this sort of approach with a slight modification to actually let the latent variables model the variation that is unexplained by those covariates. In a sense, that might be the goal.

DR. BLEI: You could actually model those jointly with the connection.

DR. HOFF: Yes, depending on what your goal is, you could do it in a variety of different ways. For example, you might want to leave that variable out of it. I mean, if you know what the variable is, you might leave the variable out of the model, fit the latent characteristics, and then plot, say, the latent characteristics versus that variable. It depends on what you are trying to get out. One issue with including things in a model is, you are saying what the scale is, or saying it is additive. In a lot of cases it is not. It is like multiplicative. There can be an advantage to leaving it out and then plotting it versus the latent characteristics later. You could do what I just said, try to—the way I have the model set up, you can constrain the latent characteristics to be orthogonal to things, and you could actually restrict them to be orthogonal to sort of the design structure of the regression and say, yes, it is the left-over variation. It depends on what issue you are trying to address at the moment.

DR. BLEI: One last question. A lot of these networks like the Internet are very sparsely connected, and I imagine that if you tried to fit a model that was just overrun with zeroes, say a logistic regression, you would end up just fitting the zeroes and you might not get anything interesting out of it. Have you had that issue?

DR. HOFF: Yes and no. Actually, the networks that I have been modeling, these conflict networks that my political science colleagues fit, are extremely sparse. So, actually we do that. One thing he likes to show to political science crowds is that he fits this logistic regression without any sort of latent variables, and then you make a prediction on where the conflicts are. Well, your best bet is to just predict no conflict and that is it, and you just predict all zeroes. So, the r^2 is very bad for these types of models. The benefit from doing something like this is, yes, you do predict mostly zeroes, but you can identify sort of those actors in the system that are active, and they get sort of big, latent vectors, and then you can sort of see what is going on there. So, you still predict mostly zeroes for most of the countries, but then you actually do pick up these things that are not sparse, you pick up the activity, and the fit is a lot better.

DR. BANKS: Do you have any thoughts on how to robustify this kind of analysis?

DR. HOFF: Against what?

DR. BANKS: In this business you can wind up with one or two atypical actors that don't quite follow the model that you specified for everybody else but, because of the interactivity, they influence so many other things that it gets really hard.

DR. HOFF: This type of approach actually can pick those things up pretty well, because it is fitting characteristics for each node. If it is true that there is a node that is sort of an outlier, it will have characteristics that are estimated as being way out there and they are very large and are

very active, without influencing too much the rest of the network. Actually, that sort of outlying stuff fit well with this type of structure. What is harder is when you have it not as a node that is very active. It is that a couple of pairs are outliers and you might get into trouble. Because of the nature of the model and where it fits, it is fitting these parameters for each node. You are very robust against outlying nodes.

REFERENCES

Handcock, Mark S., P.D. Hoff, and A.E. Raftery. 2002. "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association* (97).

Hoff, P.D. 2005. "Bilinear Mixed-Effects Models for Dyadic Data." *Journal of the American Statistical Association*, vol. 100.

Hoff, P.D., K. Nowicki, and T. Snijders. 2001. "Estimation and Prediction for Stochastic Blockstructures." *Journal of the American Statistical Association* (96).

Dynamic Networks

Embedded Networked Sensing (Redux?)

Deborah Estrin, University of California at Los Angeles

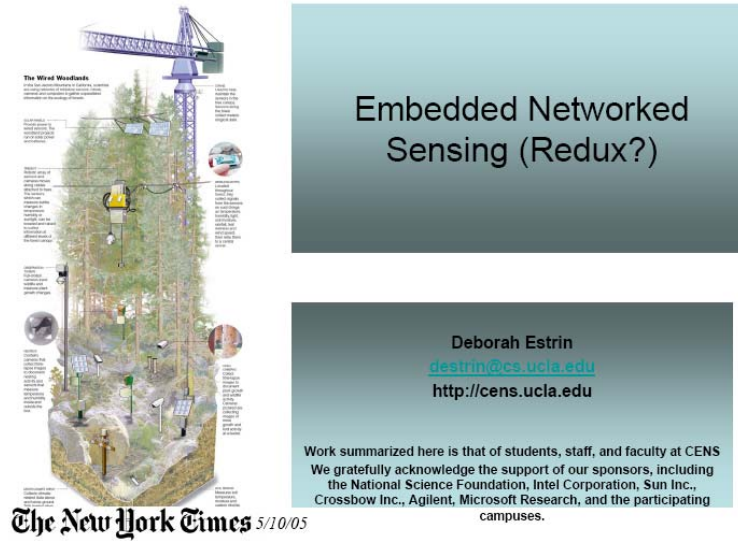


FIGURE 1

DR. ESTRIN: When I looked at the audience and I didn't recognize most of you, I added a question mark in my title, because maybe this isn't something that you have heard discussed, or perhaps not by me.

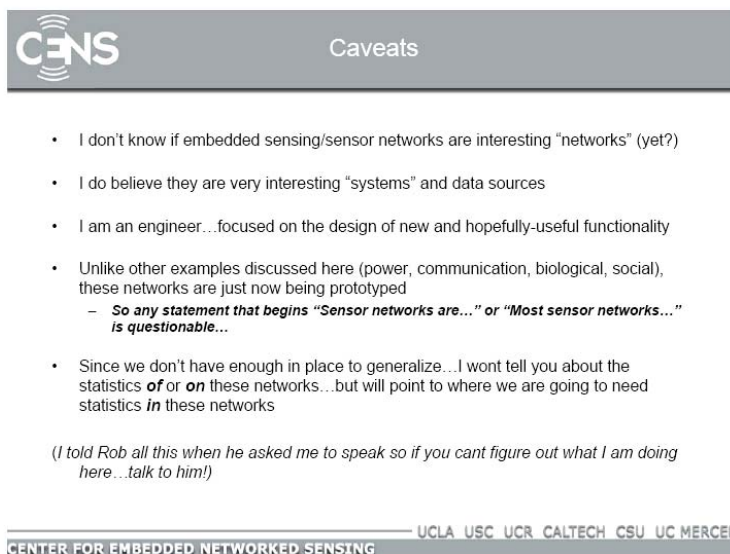


FIGURE 2

Let me start off with some caveats. The first of this is that I don't know if embedded network sensing or sensor networks are that interesting as networks. I do know that they are very interesting systems and data sources, and through this talk, recognize that I am an engineer so, most of what I focus on is the design of new and, hopefully, useful functionality and unlike the other examples here—power networks, communication networks, social networks, even, biological networks, neural networks— these networks are just now beginning to be prototyped.

There are things that can retrospectively be called sensor networks that have been around for a long time, like our seismic grids. Sensor networks of the type I am talking about are relatively new. We have a science and technology center at the National Science Foundation Science and Technology Center. It started three years ago and the purpose was to develop this technology, and there are lots of active programs all over the country and the world of people developing this technology.

As I said here, statements that begin sensor networks are, or more sensor networks, should always be called somewhat to question, because there aren't very many sensor networks, which makes it difficult to talk about statistics on or of these things. One thing is that there clearly want to be statistics in these networks, as I hope will become clearer, in the sense that there are statistics in image processing, and there are statistics in a lot of visualization and data analyses and things that we do. This is a little bit of a mind shift from some of what I heard this morning, and I hope it will be of some use. If not, Rob knew all this when he invited me. So,

you can talk to him about it.

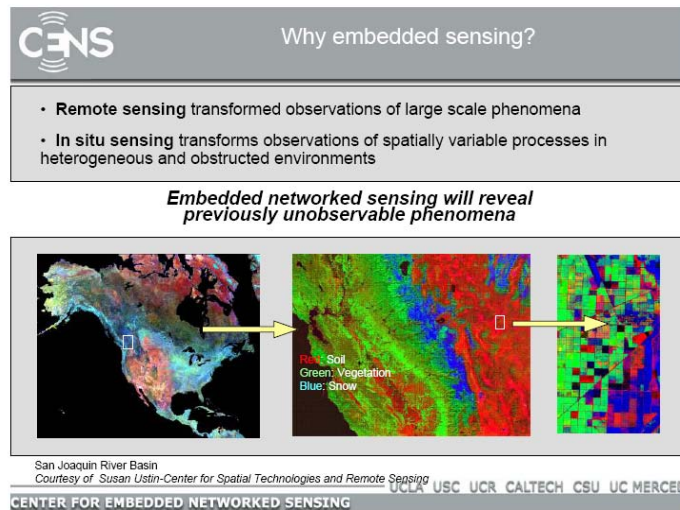


FIGURE 3

Why embedded sensing? Remote sensing has been around for a long time, and generates lots of very interesting data that continues to be a real interesting source of information for scientists, and a source of interesting algorithmic challenges. In situ sensing is a complement to that, where, as everyone knows, a pixel in a remote sensing image represents an average over a very large area.

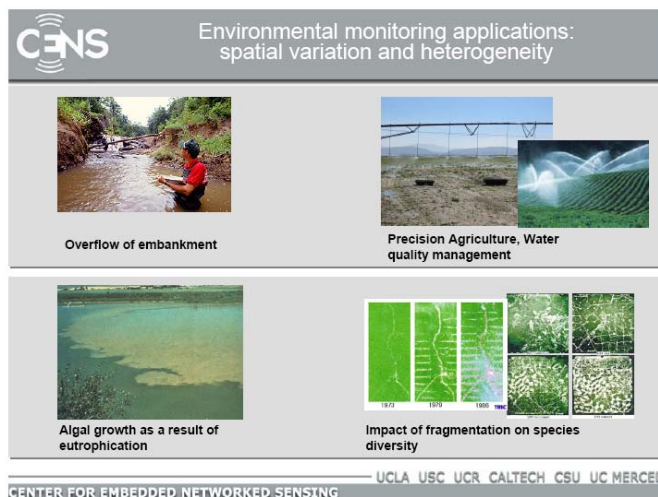


FIGURE 4

There are many phenomena, particularly biological phenomena, that simply the average over a large area isn't what you are interested in. What you are interested in is the particulars at the variations within that larger region. When that is the case you actually want to be able to embed sensing up close to the phenomenon, with the hope that it will reveal things that we were previously unable to observe. For us that means that embedded network sensing or sensor networks—I use that term, embedded network sensing—you already get a little bit of a clue that I don't think it is so much about the network.

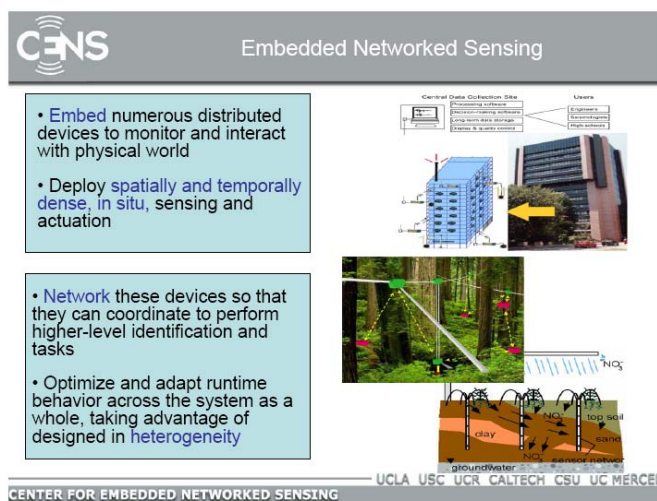


FIGURE 5

It is important that it is networked, if that is actually a verb or an adjective or whatever. It is important that you have collections of these things that are networked. That is an important part of the system. What is really important about it is that this is a sensing system and a network system, more so than the details of the communications network. In any case, where this technology has been most relevant is where you actually have a lot of spatial variability and heterogeneity. If you don't have variability, you don't need to have embedded sensing because you can take a measurement at a single point, or take an average, and you learn a fair amount from that. If you don't have heterogeneity, you can develop fairly good models that will allow you to estimate a value at a point where you are not actually measuring. So, embedded sensing is important where you can't do a good job of that estimation until you understand the system better.

In many contexts where we are building this apparatus, this instrumentation for people

now, it is for contexts in which they are trying to develop their science. It might not be that they end up deploying long lived permanent system places, rather, they are trying to develop a model for the physical phenomenon that they are studying. They need an instrument that gives them this level of spatial resolution and then, from that, they will develop appropriate models and won't necessarily need to monitor continuously in all places. In many places early on these systems are of interest to scientists to develop better models and, as the technology matures, we expect it to be increasingly used for then the more engineering side of that problem. Initially, you have scientists who need to study what is going on with the run off that is increasing the nitrate levels in urban streams, that is leading to larger amounts of algal formation and things like that, understanding that whole dynamic, because it is actually a fairly complex problem, the same kind of story in the soils. Right now we are building instruments for scientists to be able to understand and model those processes.

For the longer term the regulatory agencies will have models and know what levels of run off you are allowed to have from agriculture and from urban, and they will want to be able to put up systems that monitor for those threshold levels. When we talk about center networks, we are talking about both the design of those instruments initially to develop very detailed data sets, and in the longer term the ability to put out systems that last for long periods of time.

Embedded network sensing is the embedding of large numbers of distributed devices, and what we mean by large has changed over time. These get deployed spatially, in a spatially dense manner, and in a temporally dense manner, meaning you are able to take measurements continuously, although there are many interesting systems where you might do this over a short period of time, go out and do a survey to develop a model, and you don't necessarily need the system to be there and live for a very long period of time. Very important to these systems, as I will be describing is that the devices are networked, and you are not simply putting out a lot of wireless sensors and streaming all the data back to a single location.

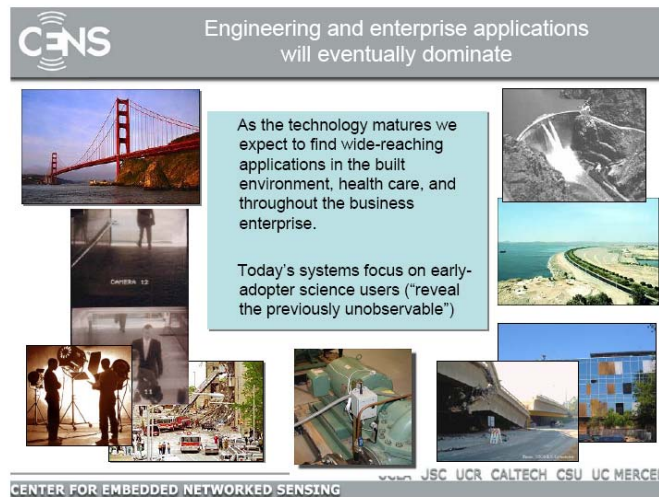


FIGURE 6

The fact is that there is a tremendous amount of potential data, and there are many modalities that you end up deploying that causes these systems to be fairly interesting, even when we are not talking about tens of thousands, or necessarily even thousands of individual devices. Most of the examples I will be describing are from things we are doing now with scientists, although the expectation is that, over time, the engineering enterprise, and even health-related applications will end up dominating.

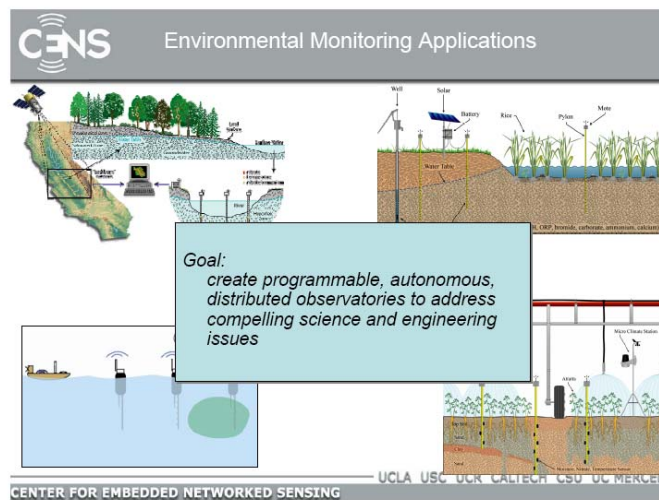


FIGURE 7

The systems that we are building are ones that—our goal is to have them be programmable, autonomous—although I will talk about interactive systems as well—distributed observatories that, for now, address largely compelling science and environmental engineering issues. Where we are right now, sort of three, five, 10, depending on how many years into this, when you think it all began, is that we have really first generation technology, basic hardware and software, that lets you go out and deploy a demonstration sensor network.

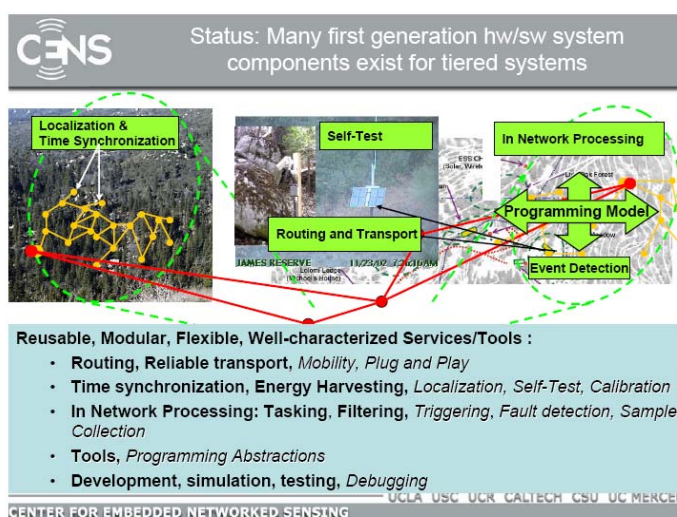


FIGURE 8

In particular, we have mature first generation examples of basic routing, of reliable transport, time synchronization, energy harvesting. Many of these nodes are operating based on batteries without any form of continuous recharge of that battery although, as I will be describing, in all the systems we deploy, we have some nodes that are larger and more capable, but energy harvesting largely refers to solar. We have basic forms of in-network processing, tasking and filtering. To some extent the systems are programmable in the sense that you can pose different queries to them, and have them trigger based on different thresholds and temporal patterns. We have basic tools for putting these systems together and doing development simulation testing. The things that aren't in bold are things that are still not very mature, things like localization. What we mean by that is, when you are doing your data collection on a node, you actually want to time stamp each of those data points. You also want to know where it is collected in three

spaces.

One of the problems that I remember in the early days, one of the things I would say is that, obviously, we are not going to be able to deploy these systems without some form of automated localization of the nodes, meaning a node has to know where it is in three space, so that it can appropriately stamp its data. Fortunately, I was wrong in the sense of identifying that as something that would be a non-starter if we didn't achieve it, because doing automated localization is as hard as all the others. You have to solve all the sensor network problems in order to do automated localization.

Although I was wrong about that, sometimes two wrongs make a right, because I was also wrong about the numbers and the rate at which we would very rapidly be deploying huge numbers of these things. Two wrongs make a right in the sense that the numbers of these things are still in the hundreds when you are doing a deployment. Going out with the GPS device, and identifying where this thing is, and configuring its location, the localization problem isn't what is stopping us from going to bigger numbers. That is not the biggest thing in our way, but it is a very interesting problem to work on. That is why I don't have that in bold because localization, while one of the most interesting problems, still isn't one that has an off the shelf solution, although we are getting nice results from MIT and other places. It is getting closer to that.

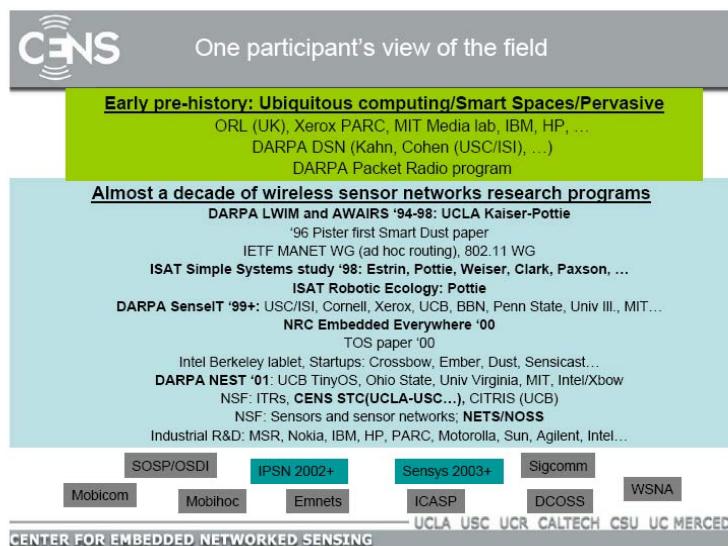


FIGURE 9

This is where we are, and a little slide on how we got there—Lots of work over time that was pretty small and scattered at the beginning, and then started to build up steam. One of the contexts in which it did that actually was with an NRC report. It seems appropriate to mention here, in about 2000, our *Embedded Everywhere* report. Then, various DARPA programs, while DARPA was still functional in this arena, and now lots of NSF programs in the area, and interesting conferences to go to and venues for people to publish, possibly too many venues for people to publish.

I said before that our price to earnings ratio in terms of numbers of papers to numbers of deployment is a little frightening at times, so that is where we are. It is a field that seems to have captured people's imaginations. People are doing prototyping of a small amount of industry activity, actually growing in the area. When I look back at what our early themes were, and what the themes were in terms of the problems to solve, this summarizes what I think are the primary changes. First of all, early themes, we talked a lot about the thousands—actually, I think I even had slides early on that said tens of thousands—of small devices, and the focus was absolutely on minimizing what every individual node, what each individual node had to do, and exploiting those large numbers, and making sure that the system was completely self configuring because, obviously, at that scale, you have to make the system self configuring. Hopefully, all of those will come back as being key problems to solve. They just aren't necessarily the first problems to solve and, since I am in the business of actually building and deploying these things, it doesn't help me to solve future problems before the current ones.

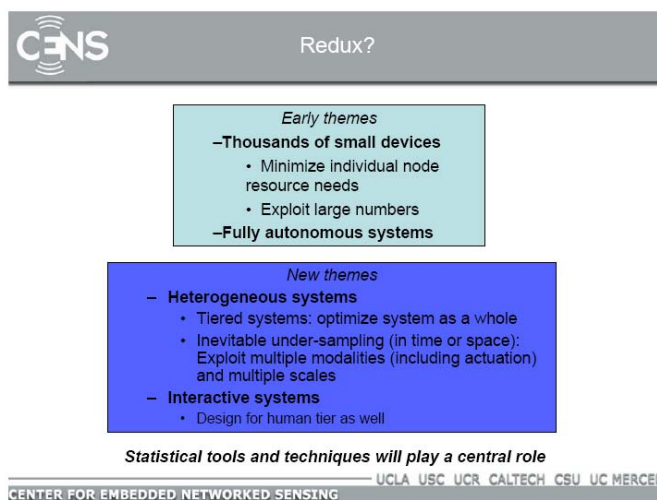


FIGURE 10

Our focus has been, and has turned to, systems that are much more heterogeneous, and I will talk about a few types of heterogeneity, and particularly the capability of the nodes, having mobility in the system, and then the types of sensing that we do. Also, our systems that we are finding very interesting to design and use early on are not fully autonomous in many of the cases. They are interactive systems where you are providing a surround sound three dimension or N dimensional, when you think of all the different sensor modalities, view to the scientists. It is the ability of the scientists to actually be able to be in the field and combine their human observations, but also their ability to collect physical samples, or go around with more detailed analytical instruments, that is providing a very rich problem domain for us. I mention it here in particular because there are lots of interesting statistical tools that scientists want to be carrying around with them in the field. That is probably when I come back to why I put that in the title. It is because there is this shift from very large numbers of the smallest devices, in our experience, to a more heterogeneous collection of devices, and a shift from a focus on the fully autonomous to interactive systems.

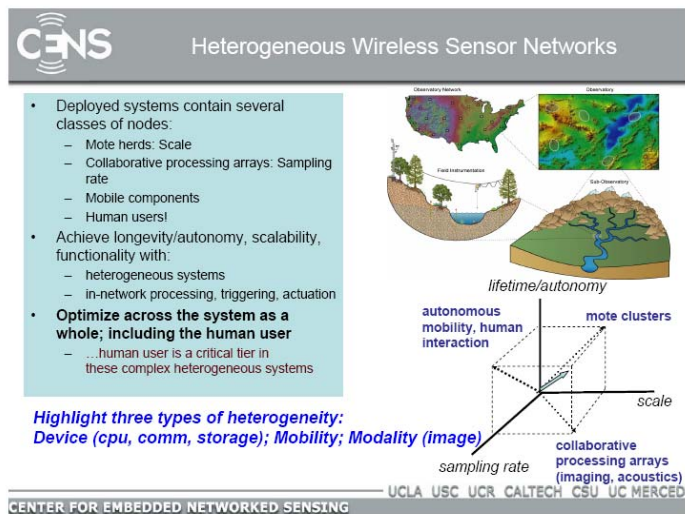


FIGURE 11

Let me say a little bit more about each of those things. First of all, what is important about this heterogeneity, and what are we doing to design for it, and what, if anything, would it have to do with, if you are somebody in statistics interested in this technology. All of our deployed systems and all the systems I know of contain several classes of nodes. At the smallest level you have little micro controller based devices that fit micro controllers that operate off of

coin cell batteries, and have on them a little bit of processing, memory, a low band width wireless and an interface to sensors, things like basic micro climate sensors and, for soil, a whole array of relatively low sampling rate chemical and physical sensors.

That is one class of device we put out there but, with every such collection of nodes, we always put out what we refer to as a micro server, or a master node, and usually several of those. These devices are very similar to what you have in your PDA or old lap tops. It is an embedded PC, it is a 32-bit device, runs linux, has your WiFi, has got relatively high band width communications on it. As a result of all of that, currently at least, it needs to be connected to some sort of energy supply, be it a solar panel, which is the case for us most of the time. You can think of that as just being a gate way. So, many systems that do exist, of the systems that do exist, you tend to have those 32-bit devices acting as gateways.

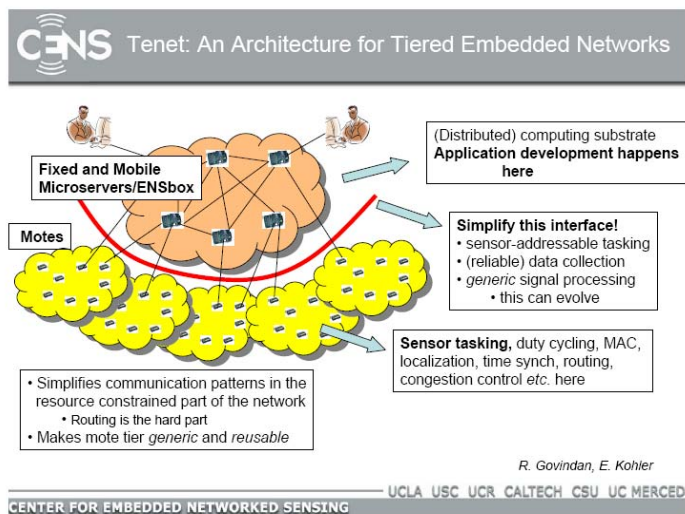


FIGURE 12

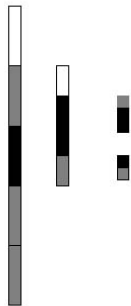
It turns out increasingly as we are doing more interesting things on the motes themselves, that those micro servers serve a more interesting role than that. Let me give you an example. In the old day we used to talk about doing a lot of in-network and aggregation among this large collection of motes, about the data that they were collecting as they were hopping it back to the edge of the network. Well, there are a few things. First of all, not surprisingly, you can go to P.R. Kumar's paper and other things, about the scaling limitations of wireless networks. It turns out that we don't build really deep multi-hop wireless networks. We build networks that the data hops maybe three or four of these low band-width wireless hops, before they get to an Internet

connected point. Secondly, it turns out that putting in any communications and having these devices listen to one another interferes with them doing aggressive duty cycling, because you want these small battery operated devices to be able to go to sleep as much as they can and have pretty deterministic schedules. If you add onto that job, you know, route things for your neighbor or aggregate your neighbor's data coming from any direction and know where your neighbors are, and choose a leader and then be the aggregation point, when you add on all of that stuff, it ends up interfering with their ability of doing the simple job of collect the data, do some local processing in terms of doing temporal compression, look for patterns or something like that. I send that data upstream when I see something, or if it matches a query. We found over time that it became easier for us to design systems where what the little nodes are doing is very simple. It is programmable. You can adapt the thresholds and adapt the little local analysis that it is doing, but it is not a node that is trying to have a view of what is going on across the network. Rather, the micro server that you have embedded out there is the natural point that is seeing the data coming from the different distributed points, and it is a natural place where you can adjust the ambient levels, where you can adjust the thresholds, and where you can effect what is more of a collective behavior. The micro servers in the network form that is more of a peer to peer, sort of any to any network, because they have the resources to do that, but in these clouds of motes, they tend to be doing very simple tree forwarding, localized processing, just based on an individual node and passing that data forward, with the rich interface in terms of what a micro server can tell a mote to do. That is interesting because we are starting to put more interesting sensors on our motes themselves—and I will describe one in a little bit—such as image sensors or acoustic sensors—so the image and acoustic sensors, we are not actually wanting to get an image feed or an acoustic feed. We are actually programming those nodes to look for particular image or acoustic patterns. Those patterns that we want them to look for, change in color histogram, significant change in geometry or size are things that are very sensitive to ambient conditions.

It is hard to have a simple algorithm running on this eight bit micro controller that works under all conditions. You need that larger device there that is seeing what ambient levels are, and what the general conditions are, to address that simple local rule that the device is doing. So, that is something that we are just beginning to do, both in the acoustic and image realm, and it is an interesting design problem and an interesting context in which to design algorithms, things that are like image processing problems, but in this interesting sort of heterogeneous, distributed architecture.

CENS Whole-System Optimization

- Goal is optimization of the hierarchical system
 - Not merely optimization of devices or any given layer
 - Models, devices, algorithms require co-design
- Context-driven algorithms
 - No single algorithm/device is best in all situations
 - Context is the state of the next level in the hierarchy; choose resources to apply when drilling down to next level according to this state
 - Probabilistic constraints and algorithms lead to more new optimizations
- Examples
 - simplified communication patterns and programming interfaces in more resource-constrained parts of system
 - adaptive and multi-scale sampling




UCLA USC UCR CALTECH CSU UC MERCED
CENTER FOR EMBEDDED NETWORKED SENSING

FIGURE 13

I am going to go back for a moment. One of the shifts here is that it is when you take this heterogeneous collection of devices and you decide, both at design time and at run time what should be done where, who should be doing the processing and the adaptation, we take this perspective now that it is really about this whole system optimization, not just looking at an individual node and seeing how to optimize its particular energy usage.

CENS Mobility/Actuation is an important dimension of heterogeneity

- Mobile and actuated (robotic) nodes
 - Articulation magnifies effective sensor range
 - Infrastructure-supported mobility enables sensor diversity
 - Enable adaptive, fidelity-driven, 3-D sampling
- Resulting data
 - Curves and surfaces
 - nitrate concentration in a vertical plane/slice of an urban stream
 - thermal mapping on a horizontal plane/slice of a plant species habitat
 - Adaptive and triggered sampling leads to irregular data sets



W. Kaiser

UCLA USC UCR CALTECH CSU UC MERCED
CENTER FOR EMBEDDED NETWORKED SENSING

FIGURE 14

The second form of heterogeneity that I want to mention that has become very important to us is related to mobility. This is part of a project that is headed up by Bill Kaiser. Something that I neglected to mention earlier is that Bill Kaiser and Greg Potti are really the folks who I consider as being the inventors of wireless internet working technology. Back around 1996 they had a first DARPA project. I wasn't at UCLA or anyway involved, but they had a first DARPA project where they had this effort of combining computation, sensing and communication, and having the devices coordinate to do in-network processing, to make the system long lived. These same folks, with a collection of other collaborators, had another excellent insight a couple of years ago, which is that we always wanted to be able to include mobility in these systems.

Static sensors are static. You are stuck on this manifold on which you place the sensors. Those are the only points at which you are able to do your measurements. In some sense, you are always under-sampling. You are just destined to always under sample, particularly in this mode in which we are trying to create models of phenomena that the scientists don't yet understand, so they don't even know at what point—they don't have a characterization to be able to say whether they are adequately sampling the system. Being able to move things around has always been very attractive but, just like the localization problem; robotics is every bit as hard as sensor networking. Looking to robotics to solve our problems, regular robotics where you navigate around on a surface, looking to robotics to solve our problem in the near term wasn't a very effective approach. Theoretically it was a good idea, but practically, it wasn't.

The insight that Greg Potti and Bill Kaiser had was that the way you move around more easily and the way you navigate more easily is by using infrastructure, be it roads or rails or what have you. They put the robots up on aerial cables, and the robotic devices are still robots and they are autonomously moving around, but they are elevated above these complex environments in which we tend to deploy things, and they can also lower and elevate sensing and sampling devices.

If you put up multiple cables, you start to get the ability to actually sense and sample in a 3-D volume, and you are not just stuck on the manifold on which you have placed your sensors. We actually use mobility and articulation at many levels. At this relatively coarse grained level, we use it in aquatic environments with a robotic boat, a little robo-duck thing that we have that goes around and does automated sample collection. In fact, there are things that you still can't sense in situ, and it will be a long time before you can actually do in situ biological sensing. What you have to do is do triggered sample collection where you are able to carefully identify the exact location, time and physical conditions under which a sample was collected, but then you

have to go and do the analysis back in the lab to identify what organisms are actually there.

We use articulation in many contexts, this one of mobility is key. What is interesting about this particular capability—it might sound a gimmick, but aside from being a gimmick—is it has that quality to it, the news stations like to pick it up as Tarzan robot or something like that. It has that appeal but it turns out to be far more than a gimmick than I had expected it to be. What has happened with our scientists is that this has allowed them not to have to leave behind their higher end imaging, spectroscopy and other instrumentation that otherwise you are not going to attach to mote any time. You also don't want to stick it in one place. The ability to combine these higher end imaging and data collection tools with a statically deployed, simpler sensor ray, has turned out to be of great interest to them, and this is being used by folks who are studying, as I used the example before, of nitrate run off into streams. This is being used not just above ground, but actually to study what is going on inside in the stream as well. It is suspended over the stream, and this hydrolab sensor does a traversal along the base of the stream, and at various elevations. So, we have several fielded systems, and the difference between the top and the bottom is that we have pictures of the real ones at the, and the bottom is depictions of bigger things that we want to do, but they are Power Point because they are not real yet.

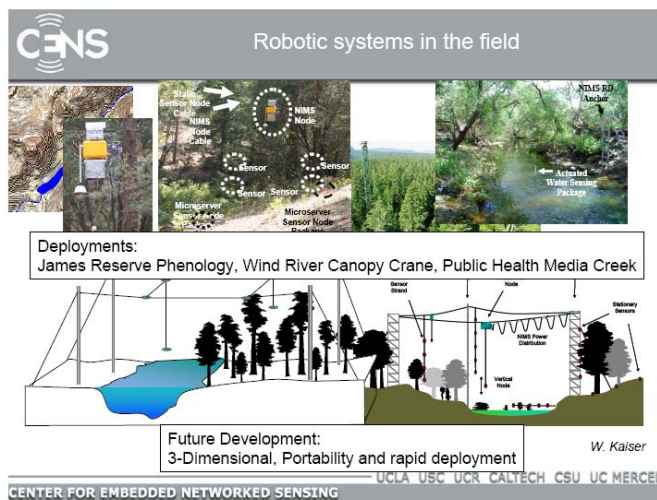


FIGURE 15

This is an example—this is Media Creek out in Thousand Oaks, not far from Los Angeles. You can't see it very well, but there is an NMS deployment that is traversing the stream, and that cylindrical object is a hydrolab which contains a number of different sensors in it.

We started to do some very interesting things such as trying to deal with the problem of calibration. The Achilles heel of sensor networks is and will be calibration, and we are not going to be able to solve it with gross over-deployment, as we thought initially, because it is very hard to over-deploy in this context, both due to cost and practicality. As you get into these applications, all of these sensors have a tremendous amount of drift to them. You are largely taking sensors that were used for analytical instruments, where a human being would go out and do measurements, but would therefore be regularly servicing the sensor. We now have a regime in place, because this is being actively used. This isn't just a bunch of computer scientists and electrical engineers and mechanical engineers who are trying to do this. We do this with a public health faculty member and a graduate student who is trying to get a thesis out of this and actually needs usable data.

One of the things that we have come up with is a calibration procedure, whereby this robotic device comes up and dips itself into some fresh water to clean itself off, dips itself into a known solution of nitrate level to do a calibration level, dips itself back, and then continues on its run. That is a low level mechanism, but it is something that it is hard to figure out how you are going to solve this problem without some form of automated actuation.

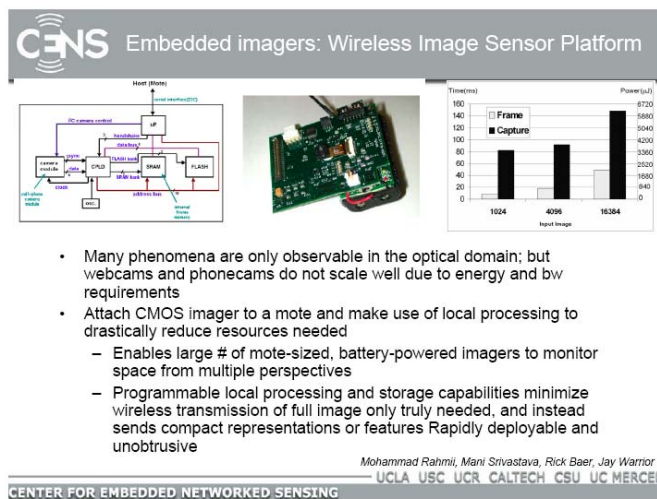


FIGURE 16

Moving on, these systems actually being more fully three dimensional is one important point. The other very important point is that it is through these systems that we discovered just how many interesting problems there are to address through portable and rapid deployment

systems. Initially, when we talked about these sorts of systems, we were always talking about very long lived permanent systems, and how you could put them out there, have them be fully autonomous and last the longest period of time. That is still of great interest, but it turns out that we skipped this and have come back to a capability that is of tremendous interest to scientists and engineers, which is the ability to go out and do a three hour, three day or three week detailed survey and study on some location. If you are focused just on that very long lived system, you leave behind a lot of higher end capability that is very powerful.

The last form of heterogeneity that I wanted to mention, and again, I chose these things—I forgot to say something. In this robotic context, one of the reasons I chose to say this, aside from its importance in center networks, is because there is a lot of opportunity for doing statistics in the network here. You don't want to just—if you have a phenomenon that have temporal variability to it, that is faster than your ability to do just a standard raster scan, what you want to do is adaptor sampling. So, there are very interesting algorithms to be designed here, where the system itself, based on what it is observing, spends more time in places where there is more spatial variability. We do both adaptive sampling and we triggered sampling where you have some static nodes that are temporally continuous, doing their measurements, and that can then give indications to the robotic node as to where there might be interesting phenomena to come and do more observations of. There are all kinds of interesting resource contention problems and lots of algorithms to be designed there.

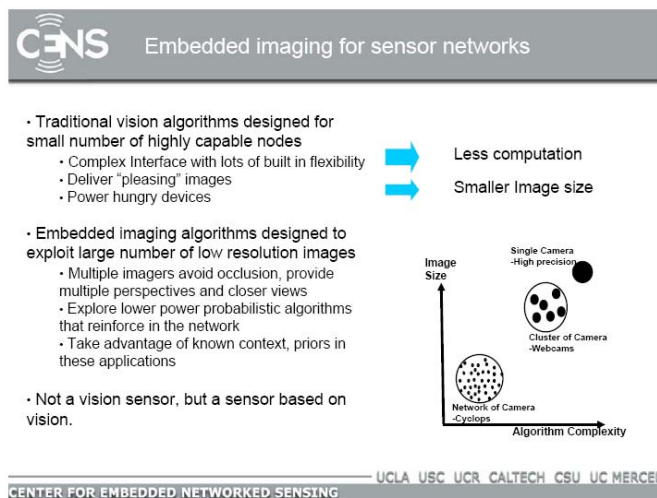


FIGURE 17

The third example, and last example of heterogeneity I bring up both because I think it is cool and interesting, and is going to expand the utility of these networks and because I think there are interesting statistical techniques and algorithms to be designed here, and that is when we begin to make use of imagers as sensors in these networks. Occasionally, you might pull back a full frame to look at an image, but the main point here is to enable us to do observations that can currently be observed in the optical domain. Many of my customers or my colleagues on this are biologists and, in general, all I can give them off the shelf are physical and chemical sensors. In some sense, these imagers can act as biological sensors, because you can observe phenomenology events, blooming events, major changes in color histograms, and in size and shape of objects.

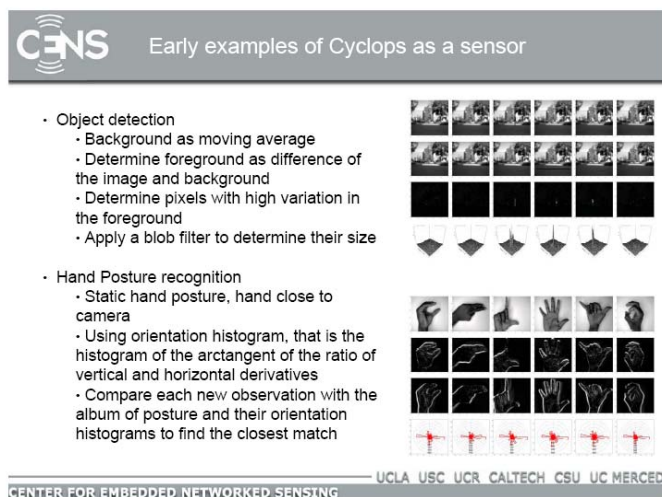


FIGURE 18

Yet web cams and phone cameras, which are all over the place are not embeddable, because they largely collect rather high data rate images and continuously send them back. What you need is to take that little camera off your cell phone and put it on a small low powered device and program into it simple computations that can allow it to quickly, or not so quickly, locally analyze those images. We are not talking about taking images continuously we are talking about many times a day because these are phenomena that are not that rapid. The same sort of thing applies in the acoustic realm, and this is actually one of those technologies, both on the image and the acoustic side, where we start to see or can think about at least fun, if I am not sure how serious applications are in the consumer realm. You can imagine leaving behind, in places or

with people that you care about, something that is going to grab an optical or an acoustic snapshot, not necessarily privacy invasive, because it is not able to—the band width of that radio is small because the battery is small and you have to be low energy, but you can leave behind, near your kid or your pet or your parents, depending on where you are in the life cycle,

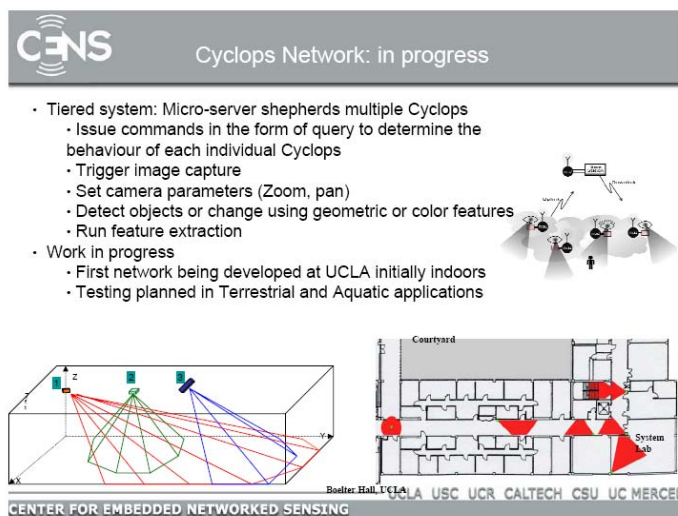


FIGURE 19

something that gives you some indication, or your favorite restaurant or cafe where you happen to work or, in my case, in my office, where there is construction outside, and I want to know how noisy is it in my office. This is a frivolous thing, but the ability to velcro onto the wall little acoustic and image activity level indicators starts to give us some interesting things with which to play. It certainly gives us something that we want to be able to program with very tightly crafted analyses. If you put these on a graph where you have high position cameras, or clusters of web cams, we are talking about down here where we have really small image sizes, really low algorithmic complexity, but the potential of putting out multiple of these things you can actually deal with things like occlusion and multiple perspectives pretty easily. One of the things that we have to our advantage here is that we are talking about putting these things in well defined environments where you know what you are looking for. They don't have to solve high end vision problems, and you can have a lot of the information about all sorts of priors and things you need for these things to be able to work well.

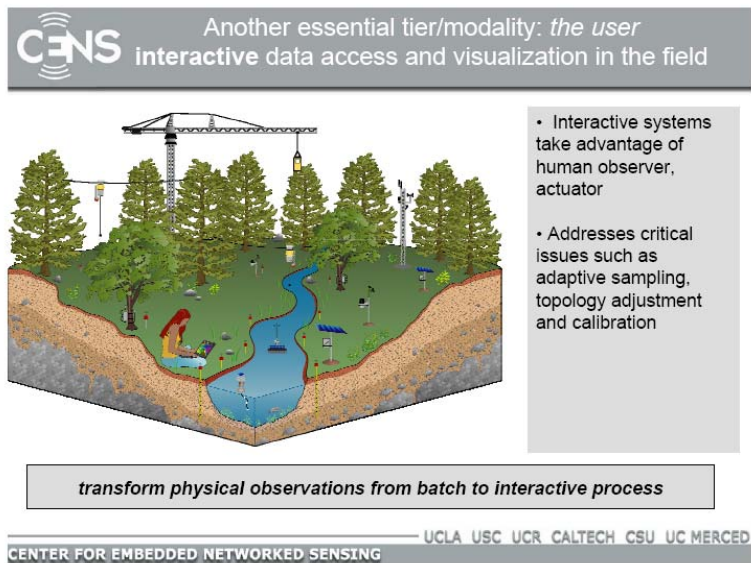


FIGURE 20

The last thing that I wanted to mention, which could just be called another form of heterogeneity in these systems, is what I referred to before separately as this issue of moving from completely autonomous to interactive systems, is that last sort of tier in our system which is actually the user. In retrospect, this is completely obvious. You know, artificial intelligence went through this as well. Looking for complete autonomous intelligence isn't what produced the useful technology that AI has produced. It is starting to look at expert systems and all sorts of interactive systems that act as decision making aids to human beings. The same thing applies here. We have no excuse for having made the same mistakes. That being said, again we are

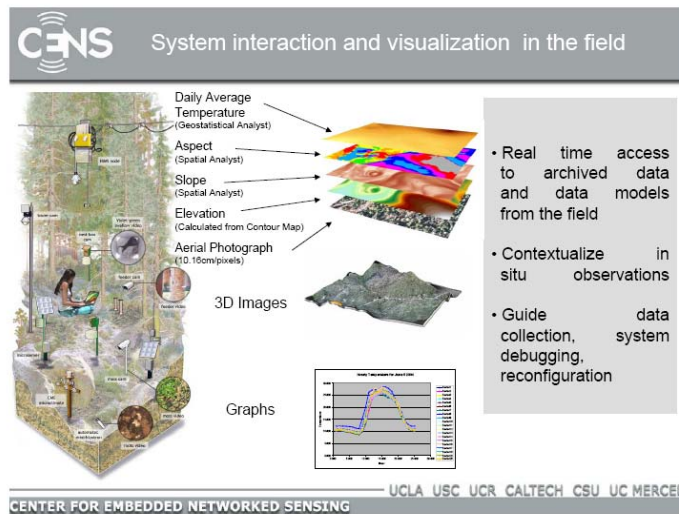


FIGURE 21

finding that our scientist customers are finding it very useful to look at these systems as systems that allow them to interact with their experiments and their set ups and their measurements interactively, sort of converting this, if you will, their experimental modes from being batched to interactive. Now what we are trying to put them with is the ability to go out there in the field, have access to their various GIS models and data, to remote sensing data, to their statistical models, to their statistical models. What they get back are these points of measurement

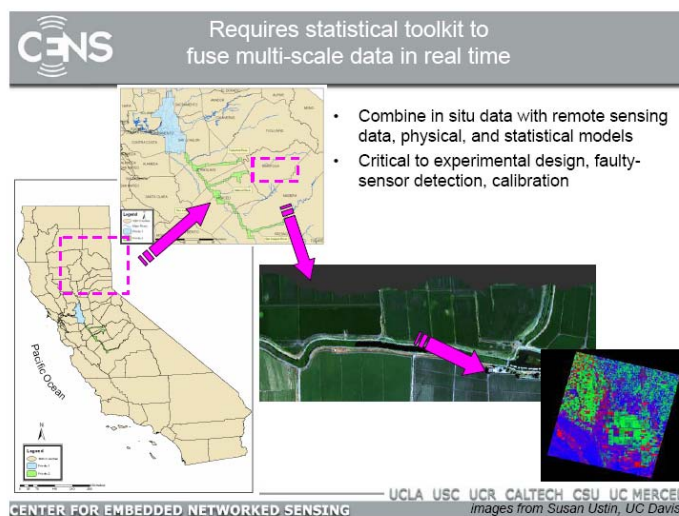


FIGURE 22

that they can start to combine into their analysis, what they are going to ultimately do with the data anyway, and identify where else might they need to deploy, what physical samples might they need to collect to actually be able to say anything with any certainty about the phenomenon they are studying. This turns out to be an interesting requirement for them. I don't know how much research there is to do, but there is definitely system building they need for equipping them with pretty nice statistical tool kits that they can take out into the field. Obviously, this in situ data by itself, these points that I say we are always under-sampling, are not most valuable by themselves. They are most valuable in the context of other data and models that the scientists have.

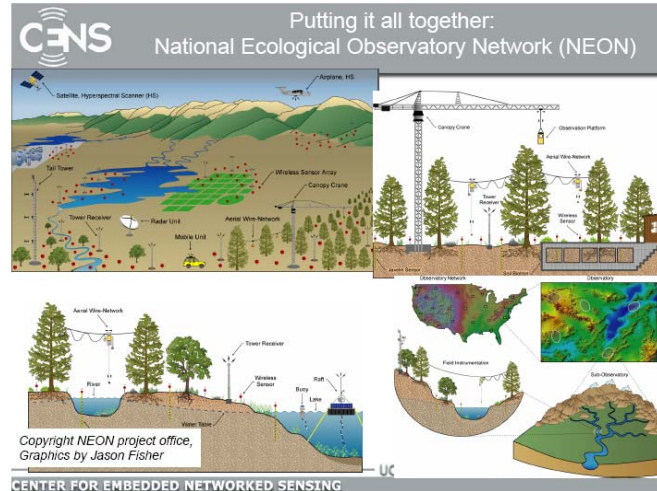


FIGURE 23

In conclusion, we have been working with lots of individual and small groups of scientists and then, over this past year, things have been getting a little more interesting in the context of planning for what are really continental scale observatories and, in particular, the national ecological observatory network, whose planning is well underway. Cleaner is another one for environmental engineers, where it really is a multi-scale, multi-modal sensor network that is being planned.



FIGURE 24

Again, as you can see, we are back to cartoons; therefore you know these things aren't real but you have a very large community of serious scientists defining how this technology needs to play out to serve them. Obviously this has lots of relevance to other issues, and there are lots of interesting problems to be solved in here, and lots of statistics gathered throughout it, at least from an engineer's perspective. There are lots of places to follow up such as conferences in the field.

How important are statistics to sensor networks? My answer is, Mark Hansen is a co-PI on our SENS center renewal, and that speaks to the point, because statistics, and Mark in particular, have become quite central. There are lots of planned observatories. You can Google ("NEON, Cleaner, GEOSS") and there is lots of work going on around the country.



Follow up

- ACM Sensys (San Diego Nov 05), IPSN
- Mark Hansen, UCLA Statistics Dept (and co-PI on CENS renewal)
- Planned Observatories: Google[NEON, Cleaner, GEOSS...]
- Excellent work at Berkeley, Harvard, MIT, Ohio state, U Illinois, U Mass, U Virginia, U Washington, U Wisconsin, Intel, MSR, PARC, Sun, and others...

UCLA USC UCR CALTECH CSU UC MERCED
CENTER FOR EMBEDDED NETWORKED SENSING

FIGURE 25

QUESTIONS AND ANSWERS

DR. BANKS: Thank you very much, Deborah. We have time for a few questions while Ravi gets set up.

DR. KLEINFELD: Part way through the talk you mentioned static sensors, the problem with undersampling, and the need for mobile sensors. You talked about a lot of things you were sensing—temperature, flora, fauna—but what are the scales? At what point do you know you are no longer undersampling? What are the physical scales for temperature, and the physical scales for observing?

DR. ESTRIN: Obviously, that all depends on the question. I can give you specific answers. For the ecologist it depends on who was asking that question about micro climates and how changes in forest canopy will end up changing the life structure and the micro climates that the ground plant species will end up experiencing, there the relevant scales are cubic meters. Am I answering the question?

DR. KLEINFELD: Yes.

DR. ESTRIN: If you are talking about, for example, our aquatic system, there, for example, these algal blooms, one of the things they are trying to understand is how, with different thermal climates, what it is the combination of light, temperature and current conditions that causes this base transition from small numbers of these algae to one of these algal blooms in the ocean or

in streams. There, you are actually talking about what could be smaller scale than even a cubic meter. So, it is very important to send this robotic sample collector around that can get to sampling on even smaller spatial scales than that. For the most part, a cubic meter is about as small we get to in terms of being willing to accept a similar measurement within that volume. Of course we are not going to have uniformity with cubic meters. You are talking about some experimental design that is not uniform.

The Functional Organization of Mammalian Cells

Ravi Iyengar, Mount Sinai School of Medicine

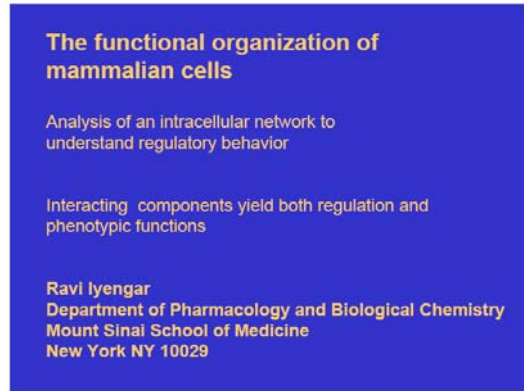


FIGURE 1

DR. IYENGAR: I should tell you a little bit about my background because much of what I am going to tell you probably won't sound like the language that we all heard this morning. I majored in chemistry and am a biochemist by training. I have spent most of my research career purifying and characterizing proteins that work on cell signaling pathways. Since many of these pathways interact with one another, the interacting pathways are called signaling networks. I come from a biology background, and I am trying to use network analysis to understand cellular functions such as those shown in Figures 1 and 2.

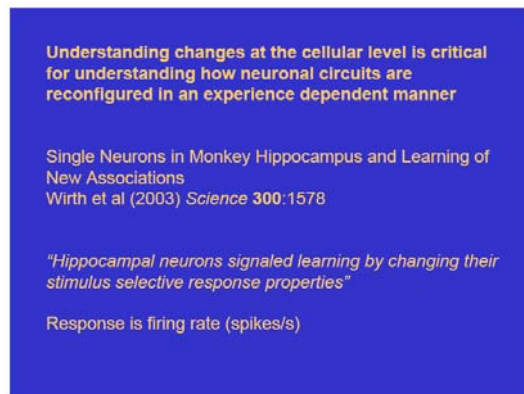


FIGURE 2

What I hope to convince you of is that graph theory-based statistical analysis of networks turns out to be a very powerful tool to understand how things are put together within a cell. We are still at the beginning stages of developing such an understanding. Our long-term goal is the

functional organization of mammalian cells. The system we are focused on is the neuron, and we want to understand the function of biochemical networks that control how the behavior of neurons changes in an activity dependent manner. Function is manifested in a very tightly regulated manner. There is continuous regulation of the intracellular processes, and this regulation, in turn, allows for many different events to occur in a coordinated way. I have been studying signaling systems for 25 years. Over this time, the numbers of signaling molecules have greatly increased. Proteins of the extracellular matrix which, when I was in graduate school, were thought to be proteinaceous glue that held cells together, have now been shown to be signaling molecules that tell cells about their immediate environment on an ongoing basis.

Almost no component in living organisms is really inert. They exist because they communicate constantly with other components, and this constant communications allow the components to work together. This constant communication is an important feature of the living cell. Underlying all of these analyses is the notion that all live processes arise from interacting components. These components are mostly proteins. But cell regulatory networks are not made up of just proteins, because within all our cells, as Dr. Kopell talked about it, ions are important constituents. Along with proteins and ions, nucleotides, sugars and lipids, all come together to form a multifunctional network. So we can approach the problem with the hypothesis that phenotypic behavior that arises from interactions between cellular components. Here my focus is on single neurons. Although networks between neurons are very important for information processing, quite a bit of information processing actually goes on within single neurons. One of the nicest examples of physiological consequences of information processing in single neurons comes from the work of Wendy Suzuki and Emery Brown and others, who published a paper in *Science* showing that, in live monkeys, during the learning process, there are changes in spike frequency of individual neurons that correlate with learning.

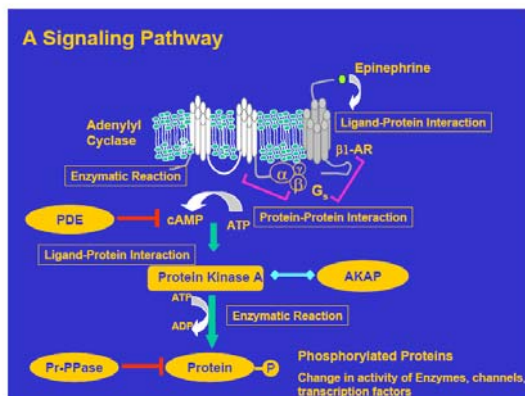


FIGURE 3

Signaling pathways has been studied for 50 years. This particular signaling pathway, the cAMP pathway which I have studied for a very long time, is a well characterized one. The most famous example of this pathway is adrenaline—“fight or flight” response. When adrenaline binds to its receptor, it initiates a series of steps. The steps eventually lead to change in activity of the enzyme that mobilizes glucose. The glucose released into the blood provides energy to run or to fight. The cAMP pathway is shown in Figure 3. As you can see, the cartoon in Figure 3 encompasses two sorts of information. One is the underlying chemical reactions, which are protein-protein interactions and the ligand-protein interactions. There are also arrows (activating interactions) and plungers (inhibitory interactions) and dumbbells (neutral interactions) that can be used for defining the edges that allow the nodes to come together to form a network. Such networks with direction-defined edges are called directed graphs. The overall organization of the hippocampal neuron is summarized in Figure 4. The organization indicates that that receptors that receive extracellular signals regulate the levels and/or activity of the upstream signaling molecules such as Calcium, cyclic AMP, and small GTPases that in turn regulate the activity of the key protein kinases. The change in activity of the protein kinases modulate the activity of the various cellular machines in a coordinated manner. This change in activity of the cellular machines results in changes in phenotypic behavior.

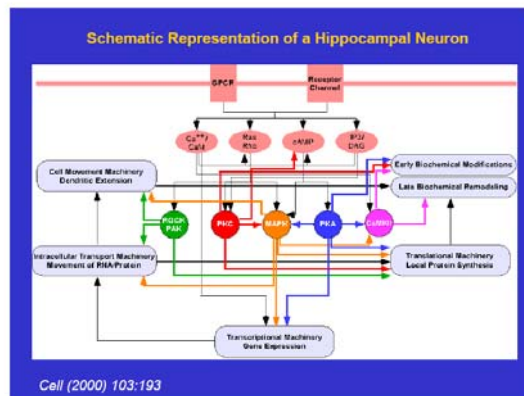


FIGURE 4

This minimal view of cellular organization has been developed from a large number of studies from many investigators. The NIH funding for basic research over the past 50 years has allowed for the collection of a lot of information on binary interactions in many systems. It has been possible, in the last 10 years or so to put this information together and this synthesis gives the high-level schematic representation of what a hippocampal neuron might look like as is shown in Figure 4.

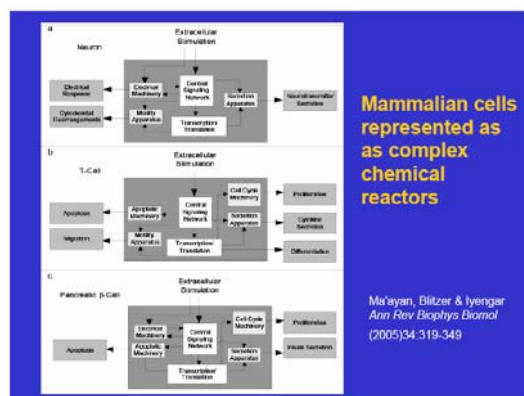


FIGURE 5

Such an abstracted representation can be used to depict many different mammalian cell types. Three examples are shown in Figure 5. Neurons, T cells and pancreatic beta-cells. If we are studying electrical properties of neurons and how they change in an activity dependent manner, one might focus on early biochemical modifications such as phosphorylation of channels. With signals flowing through the network one can modify a channel, regulate gene expression, and move components around the cell and so on. This minimal view can be expanded to depict many cellular components and interactions as shown in Figure 6.

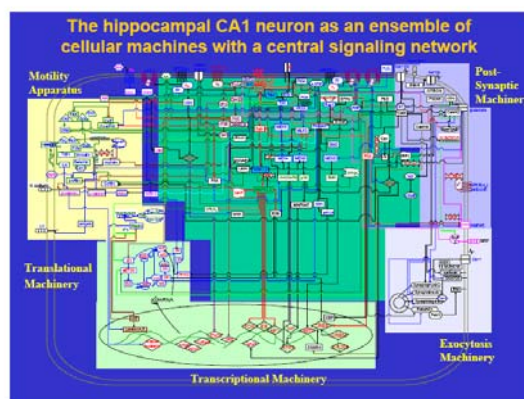


FIGURE 6

This brings one to a complex picture. Often these complex diagrams are considered to have minimal value because people question just what can be learnt from these diagrams. Actually, the map in Figure 6 has just a few hundred components. It has about 250 nodes but one can make this even bigger. People are struggling to understand how such complex systems can be tackled to make some sense of the functionally relevant organization. This question and

approach we use are shown in Figure 7.

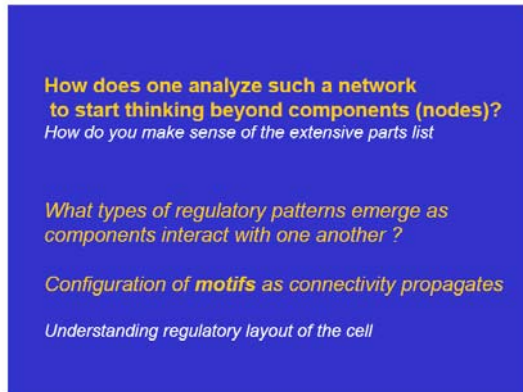


FIGURE 7

John Doyle in his talk also asked this question: how does one make sense of the extensive parts list? The reasoning we use to answer this question is based on the assumption that you can make sense of the function of complex systems by understanding the underlying regulatory patterns that these systems possess. These patterns are likely to be dynamic. I will explain our use of the term dynamics that in a minute. First, let us focus on what does the word pattern means within the context of cellular networks. This is where I found the papers of Uri Alon and his colleagues on motifs as building blocks of networks relevant. We have used the approach of Milo, Alon and colleagues to define the configuration of motifs as the connectivity propagates through the cellular network to identify patterns. This approach turns out to be a very useful way of understanding the regulatory capability of the cellular network. The occurrence of motifs does not mean that regulation occurs, but that the capability for regulation exists. In an experimental sense these analyses gives us a set of initial ideas and hypotheses that we can go and test in the laboratory. All of what I am going to describe is the work of one graduate student. We have 15 co-authors on the 2005 paper in *Science* that is cited in Figure 8. Most of the others have made valuable contributions in curating and validating the interaction data set for the construction of the network. All of the analysis has been done by one graduate student, Avi Maayan. The sort of biology area that I come from, everybody keeps saying, the days of individual laboratories are finished, you need this massive network biology. In reality, when one gets one smart graduate student who challenges your way of thinking, brings something new to the lab it is possible to make real progress, as has been the case here.

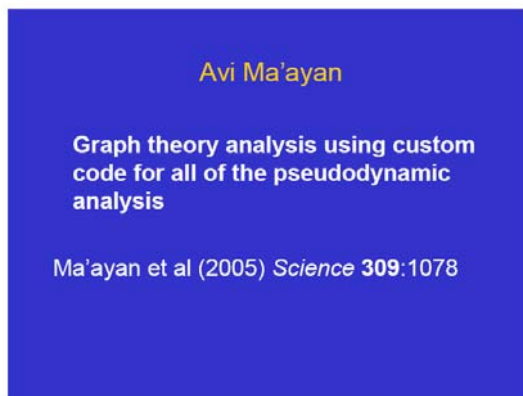


FIGURE 8

Cellular networks can be represented as graphs at varying levels of detail. The various representations are shown in Figure 9.

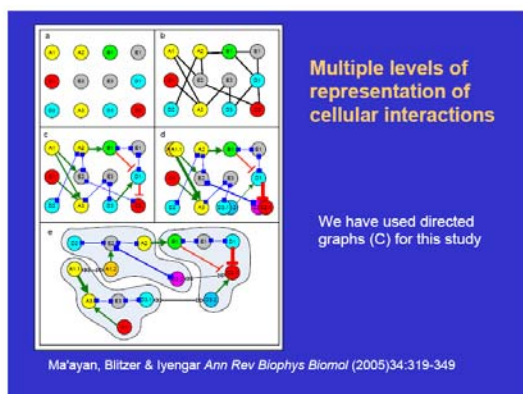


FIGURE 9

Graphs consisting of nodes that are cellular components such as proteins, small molecules such as metabolites and ions are connected by edges that are chemical interactions between nodes. These graphs can be undirected (B), directed (C), directed with weighted nodes and conditional edges (D) and directed with spatially specified nodes and edges (E). In the analysis described here we have used directed graphs (Type C, Figure 9) which are these directed graphs. They are directed in two ways. There are edges that are positive—that means that A stimulates B—and negative—that means A inhibits B. Then there are these neutral edges, that important for cellular networks because they represent interactions with anchors and scaffolds. In this representation we have not actually dealt with space in an explicit way, however these neutral

edges give the initial incorporation of spatial information into these networks. In this representation we have made a simplification, which I should state up front. We treated these three classes of edges as independent interactions. In reality, what happens in cells is, when two components come together in a spatially restricted manner they interact, that means there may be an active scaffold protein C that binds both A and B allowing for the $A \rightarrow B$ interaction. A to B edge exists because the A to and B to C edges are operational. We have not incorporated such conditionality between the edges themselves and hence type C graphs are simplified representation of biological systems. Type E graph would be the case where one would actually incorporate spatial constraints for the same components when they are in different dynamic compartments, interacting different sets of components. There is also the issue of weighting the links. Such weighting is needed to represent the levels of changes in activity of the components leading to differing interactions. These complexities have to be dealt with the underlying assumption that all of biology is a continuum. It is a question of when one needs to move to weighted nodes and edges to capture the behavior of the system. For the initial analysis we have assumed that Type C graphs will suffice.

We constructed a network *in silico* using a function-based approach. For this we have focused on binary interactions since the validity of these interactions is clear from the biological literature. From lots of studies in biochemistry, cell biology, and physiology people have shown that A interacts with B, and there is a consequence to that interaction. These are well-characterized interactions and there is a vast literature on such binary interactions from the literature we have identified that A talks to B, B talks to C and C talks to D to build the network. The approach we have used is shown below in Figure 10.

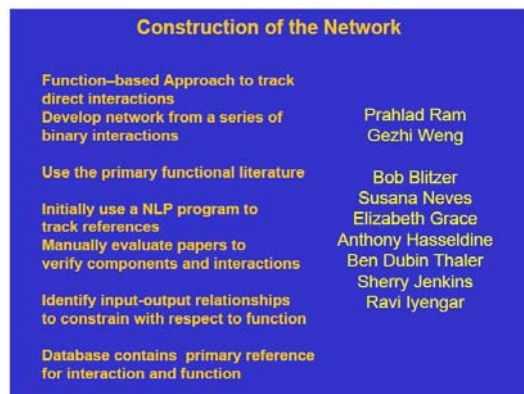


FIGURE 10

In many cases, because groups have studied input-output relationships, it is also possible

to constrain the network from the published data. Starting from A, one can get to D, or starting from B one can get to F. Such distal relationships are not always known and there is some ambiguity about pathways within networks. What we did was to use a function based approach to identify direct interactions and develop a network as a series of binary interactions. For this we searched the primary literature. That is a daunting task, so initially we developed a natural language processing program to pull out the papers. We then decided that we could not use the papers directly so, I recruited everybody else in my laboratory to actually read, verify and sort the literature. So we are reasonably confident about the validity of our network. In almost all cases, we searched for input-output relationships to constrain the connectivity relationships with respect to function and we have a database with all the primary papers and the references that go to make up these interactions. What one can do by this process is to generate a large series of subgraphs as is shown in Figure 11. This is one for calcenurin an important cellular enzyme that is a phosphatase.

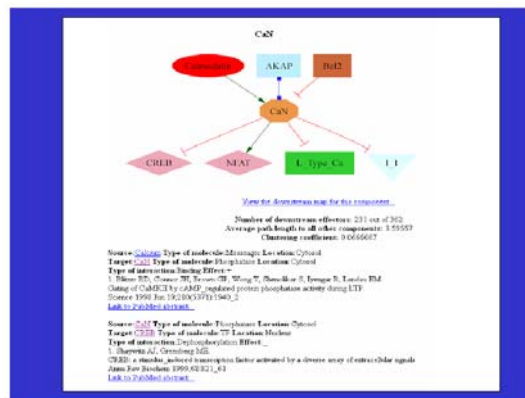


FIGURE 11

In this subgraph, calcenurin, is activated by calmodulin. Activation is represented by the green arrow. Calcenurin is anchored by AKAP. This means calcenurin is held in a certain location because of this protein, and that is that shown by the blue dumbbell. Activated calcenurin in turn can either activate (green arrow) the transcription factor NFAT, or inhibit (red plunger) another transcription factor CREB. You can construct a huge number of these subgraphs and put them all together, and use them to define how a CA1 neuron might work.

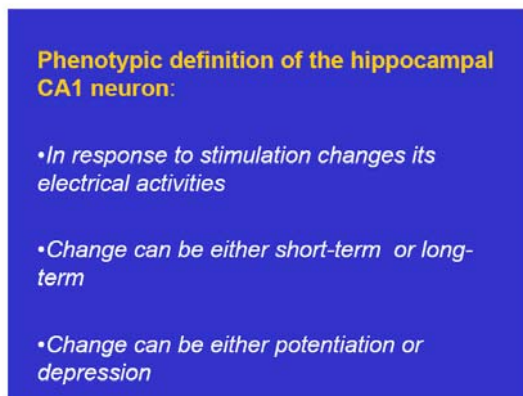


FIGURE 12

Functionally, as shown in Figure 12, a CA1 neuron changes activity in response to stimulation, and this change can be either long or short term. These changes in activity arise from changes in behavior of its components. This is now well documented. A list of changes in activity and the cellular components associated with these changes are shown in Figure 13.

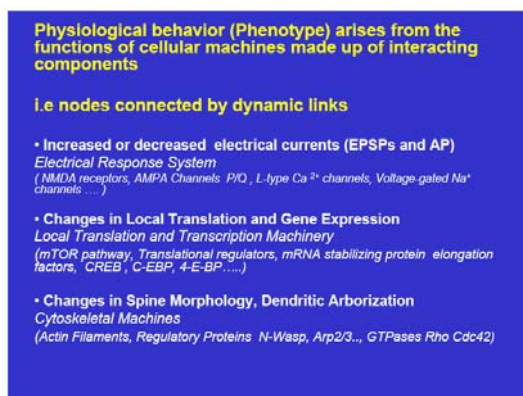


FIGURE 13

There are groups who study changes in the activity of channels, and the NMDA receptors. Others study changes in gene expression and translation, and actually even the way the spines are formed. Karel Svoboda from Cold Spring Harbor Laboratory, has been studying how the morphology of the connections change as the stimulus propagates. There are other groups who have characterized in detail the components that underlie these functional changes. One can put the data from these two groups of investigators together and make a functional network such one shown in Figure 14.

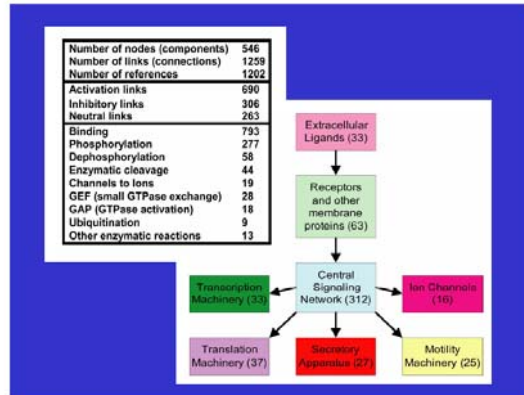


FIGURE 14

For this network we know the underlying biochemistry of all of reactions we don't use this information in the graph theory analysis. Overall we have a system of some 546 components and nearly 1,300 edges. You can parse these nodes out in these functions, as shown in Figures 14 and 15.

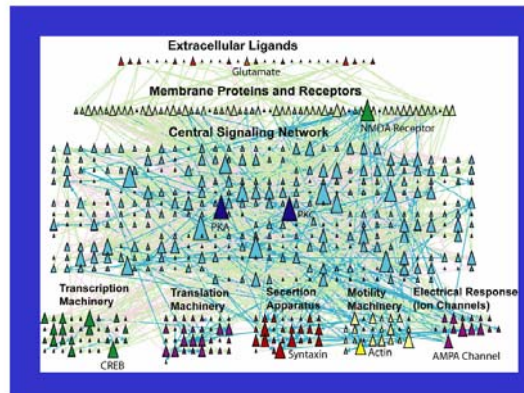


FIGURE 15

Figure 15 is a modified Pajek diagram of the CA1 neuron network with triangles as nodes and the size of the triangles indicative of the density of associated edges. The characteristics of the network are summarized below in Figure 16.

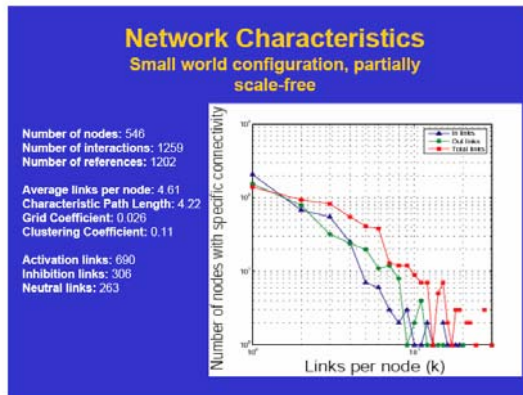


FIGURE 16

This is largely a scale-free, small-world network. The characteristic path length is 4.22. The clustering coefficient is 0.1, indicating that the network is quite highly clustered as compared to a randomized network. The clustering coefficient has turned out to be one of the most informative parameters in understanding the cellular space, because it becomes an estimate for things are put together by anchors and scaffolds. Neurons, especially, have substantial geometry and components are not evenly distributed within the cell. Understanding how things are in close proximity to one another that becomes very useful for understanding local function.

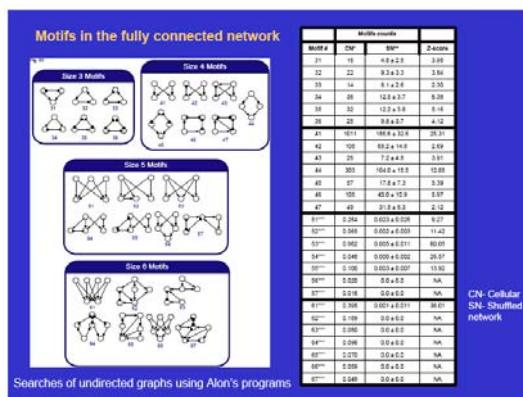


FIGURE 17

We used the Uri Alon's Pathfinder program to characterize the various types of motifs that are present in this network. Motifs are groups of nodes that act as a unit and have the ability to process information to alter input-output relationships. In cell signaling systems information transfer occurs through chemical reactions. Typically information processing by motifs involves changing the input/output relationship such that there is a change in the amplitude of the output

signal. There can be a change in the duration of the output signal or the signal can move to a different location within the cell. In many cases, signaling networks produce each of these effects to varying degrees.

There are many of these motifs that when organized together gives rise to signal processing. The various configurations of the motif are given in Figure 17. I want to draw your attention to the feedforward motif, which is motif number 44. I really like this motif since it gives you two for the price of one. One, it allows for redundancy of pathways, which is very important in these systems to ensure reliability of signal flow, and two, it essentially works as a positive feedback loop that allows for persistence of the output signal, which almost always alters the interpretation of the signal for the mounting of functional responses. Another noteworthy motif is the bifan motif, which give rise to local interconnectivity and signal integration. The minimal size of these motifs is three or four nodes. We have found that motifs with five and six nodes arise from juxtapositions of the smaller ones.

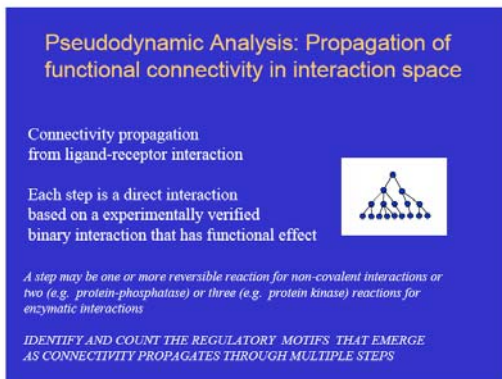


FIGURE 18

To understand how signal flows through this network we conducted a type of Boolean dynamic analysis which we have termed pseudodynamic analysis, as outlined in Figure 18. We used the adjective pseudo to signify that although the analysis represents the dynamics of the underlying coupled chemical reactions, the values of the links are all equal and hence do not capture the different reactions rates for the different equations. This simplifying assumption is largely valid because inside the cell, past the membrane receptors, and not quite at the level of gene expression, the reactions rates are largely similar for 80-85 percent of the reactions. The term pseudodynamics is similar to the term “apparent K_d ” in biochemistry and pharmacology. Although rigorous measurements for K_d determination require equilibrium dialysis, most of us

often used steady state methods to determine K_i from which apparent K_d s can be calculated. Such a simplified approach can also be used to study propagation of signal from the receptor into the cellular network.

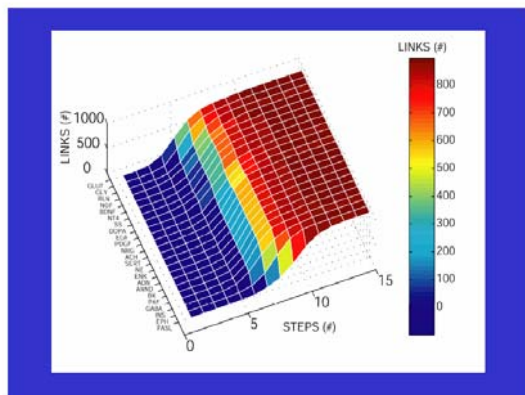


FIGURE 19

We looked at connectivity propagation going forward from the receptor. Each step signifies the formation of a link (edge) and represents a direct interaction. Such direct interactions may represent a single chemical reaction for noncovalent reversible binding interactions or 2-4 chemical reactions when the interaction is enzymatic. The numbers of links engaged as signal propagates from a many ligands that affect the hippocampal neuron is shown in Figure 19. This is a complex plot with many ligands that affect the hippocampal neurons. At one end is the major neurotransmitter, glutamate. Signals from glutamate rapidly branch out and by 8 steps engage most of the network. At the other end is the fas ligand that causes apoptosis in neurons. Fas takes nearly 12 steps to engage most of the network.

By the time we reach 10 or more steps, we can get about 1,000 links engaged, indicating that most of the network becomes interconnected. We then counted the number of links it takes going from ligand binding to a receptor to get to a component that produces a functional effects such as a channel or a transcription factor that turns on a gene. Eight is the average number for going from receptor to effector. When we are tracking paths from ligand-receptor interactions to channels or transcription factors we can identify the regulatory motifs that emerge as connectivity propagates through the network. This is shown for three important ligands for the hippocampal neuron: glutamate, norepinephrine and BDNF in Figure 20.

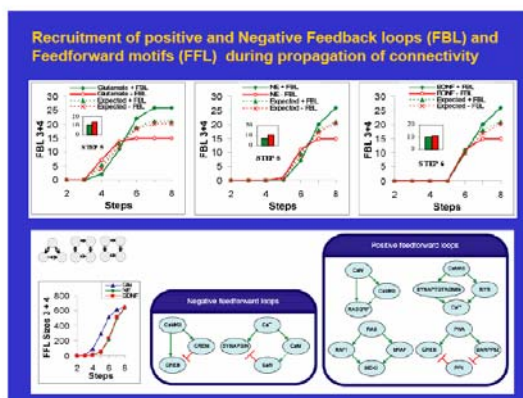


FIGURE 20

We counted feedback loops, sizes three and four, we then counted size three and four feedforward motifs. One thing we see is that, there are a lot more feed forward motifs than feedback loops in the cells. I think this represents the molecular basis for redundancy. Feedforward loops can arise from the presence of isoforms. The higher we go up in the evolutionary isoform hierarchy, there are more isoforms for many signaling proteins. The same protein comes in three, four or five different forms. They often have some what different connections, so we think of these proteins not quite as full siblings, but more like half brothers and sisters with some common connections and some unique connections.

What we found most interesting, was non-homogeneous organization of the positive and negative motifs. As we start at the receptor, at the outside of the cell, the first few steps yield many more negative feedback and feed-forward loops, which would tend to limit the progression of information transfer. As you go deeper into the system, in each of these cases—with glutamate, norepinephrine, we pick up the positive feedback loops and positive feedforward motifs. It appears that, if the signal penetrates deep into the cell, it has much more chance of the signal being consolidated within the cell. This consolidation may trigger many of the memory processes by changing the cell state. This was our first big breakthrough that we got in understanding the configuration of the network and may satisfy my biology friends.

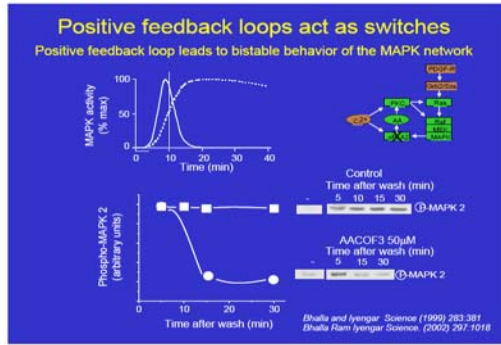


FIGURE 21

Two brief detours into the standard differential equation based modeling to illustrate the capabilities of the feedback and feed-forward regulatory motifs. Positive feedback loops work as switches. We have shown this a while ago for the MAP-kinase 1,2, system both initially from modeling analysis and subsequently by experiments in a model cell-culture system NIH-3T3 fibroblasts. A comparison of a model and experiments showing the input output relationship due to the presence of a feedback loop is shown in Figure 21.

Positive feed forward motifs also give you extended output, and this is shown below in Figure 22. Although what is shown is a toy model, we can see that over a range of rates the presence of a feed-forward motif affects input-output relationships. So, the presence of these regulatory motifs can have real functional consequences.

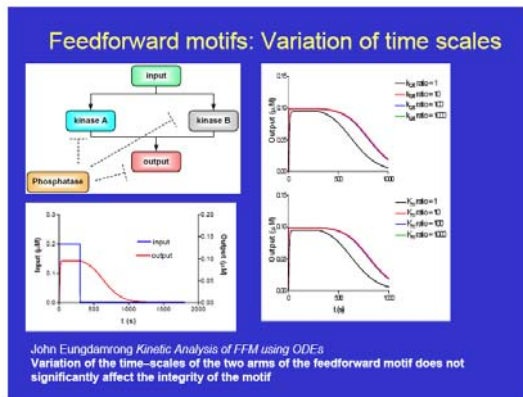


FIGURE 22

The next analysis was to simulate the formation of motifs as signals propagated from the receptor to an effector protein. Generally, in our initial analyses we started at the receptor and

allowed connectivity to propagate all over the cell. That does happen to some extent, but in most situations information flow is constrained by the input and output nodes. When we stimulate the hippocampal neurons, it results in changes in the activity of the channels or changes the activity of the transcription factors like CREB that results in altered gene expression. So, we decided to look at the system using a breadth-first of algorithm to go from receptor to effector with progressively increasing number of steps. This is shown in Figure 23.

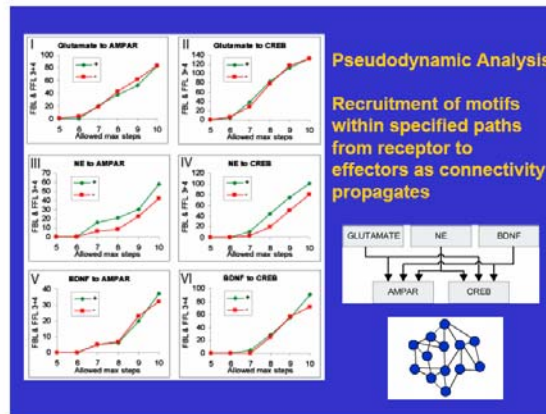


FIGURE 23

This analysis actually yielded the most satisfying part of our observations. It had been known for a long time that glutamate by itself would not allow this neuron to change state and get potentiated for an extended period without engaging the cAMP pathway. When we think about the cyclic AMP pathway, we think of the neurotransmitter norepinephrine, that in the hippocampus facilitates the glutamate dependent potentiation. When neurons become potentiated, they behave differently in response to stimuli.

The pseudodynamic analyses showed that for norepinephrine, when the number of feed forward and feedback loops were counted for whether they were positive or negative, far more positive loops were engaged with increasing numbers of steps. The preponderance of positive feedback and feedforward loops from norepinephrine to CREB provides an explanation of why this route is so critical for the formation of memory processes. CREB is often called the memory molecule since its activity is crucial for the formation of memory in animal experiments.

In contrast, for glutamate by itself, the numbers of positive and negative motifs are equal, and BDNF actually turned out to be a bonus. For neuronal communication there are always two cells, the presynaptic neuron and the postsynaptic neuron. Neurons can be potentiated by the actions that go on within themselves, and they also can be potentiated by changes in the

presynaptic neuron. It has been shown that that BDNF actually works in the pre-synaptic CA3 neuron, and in the postsynaptic CA1 neuron that we have modeled for the network analyses and the regulatory motifs for BDNF induced network evenly balance out. This statistical analysis gives us insight into how the configuration of motifs can affect state change in cells. If we engage more positive feedback and feed forward loops than negative loops we can induce state change (i.e plasticity). If the positive loops and negative loops are balanced then although signals propagate and acute effects are observed there is no state change. This is summarized in Figure 24.

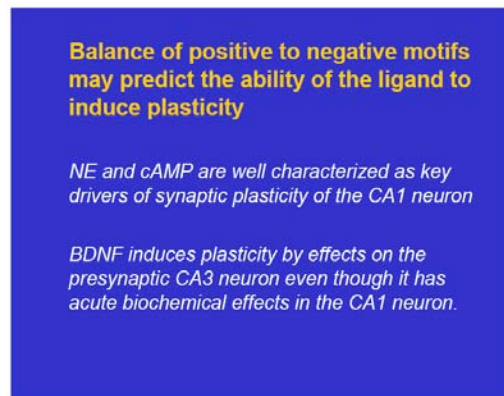


FIGURE 24

I want to make two further points. First, when we conducted this pseudodynamic analysis, we actually sampled the system for dynamic modularity. Modularity means different things for scientists in different fields. For those of us who come from a cell biology background, the word module actually means either a functional module, like components of one linear pathway, or it means a group of components in an organelle such as the proteins in the cell membranes, or the nucleus. In our analyses we used a functional approach going from receptor to effector protein. This analysis is shown in Figure 25.

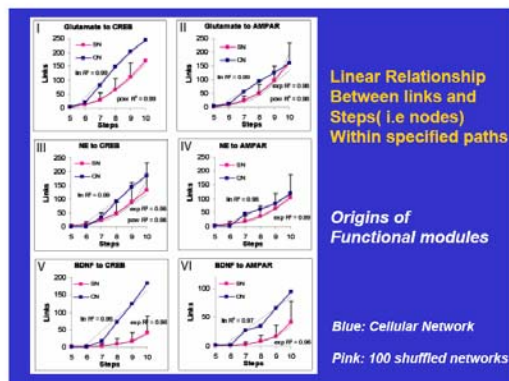


FIGURE 25

As we follow the number of links that we engaged when we go from, the NMDA receptor an initial glutamate target, to the AMPA receptor, which is also a glutamate target and the final effector in the system, or from the NMDA receptor to the transcription factor we found that these lines were relatively linear. To validate this profile Avi came up with a clever trick in creating shuffled networks that could be used as controls for this biological networks. What he did was maintain the biological specificity of the first uppermost connection, which is from the ligand to the receptor, and maintained the biological specificity of the last connection that is the links to come into the AMPA channels or CREB. So, there may be 10 components that feed into CREB with seven or eight protein kinases that regulate it and those links did not change. He then randomized everything in the middle. When he tracked paths in these shuffled networks we either got paths that yielded plots that fit either to a power law or an exponential function.

Two features of these paths are noteworthy. One is that the path is linear and, two, there are many more links engaged in the CA1 neuron network than in the shuffled networks even though only 10-20 percent of the links are engaged. I point that out to you because if you look at the scale, the number of links here is either 100 or 200 and if you go 10 steps without output constraint, without outward constraint, you engage about 1,000 links. So the input-output constraint and the number of steps (and if we use the number of steps as a surrogate for time) allows us to constrain the number of interactions that can occur starting from the point of signal entry to the final effector target. This boundary defines the module, and everything else on the outside becomes “separate” because one cannot engage these links within the time period defined by the number of steps. This type of analyses indicates that we are likely to have a series of dynamic functional modules. The properties of these functional modules are summarized in Figure 26.

Origins of Functional Modules

- Subnetworks from defined receptors to effectors show linear progression through multiple steps
- Biological binary specification of link between nodes leads to defining the boundaries of functional modules
- Functional modules have many more links per step suggesting a propensity for information processing


FIGURE 26

Second, we were interested in figuring out what these highly connected nodes do as part of the network. Avi started out with a system of mostly unconnected nodes and then asked the question, what happens to the system if we add nodes with four, five, six links, iteratively. He then determined both the number of islands as a measure of networking and the motifs that are formed as the network coalesces. This approach is described below in Figure 27.

**Pseudodynamic Analysis:
Propagation of connectivity in interaction space**

What role do the highly connected nodes play in the regulatory layout ?

Gradual inclusion of nodes based on connectivity



Look at network structure (count # of islands)

and count the emergent motifs of various types

FIGURE 27

Initially, at four links per node he had around sixty islands, and by the time he reaches 21 links per node, he was able to form one large island (i.e., a fully connected network). This is shown in Figure 28.

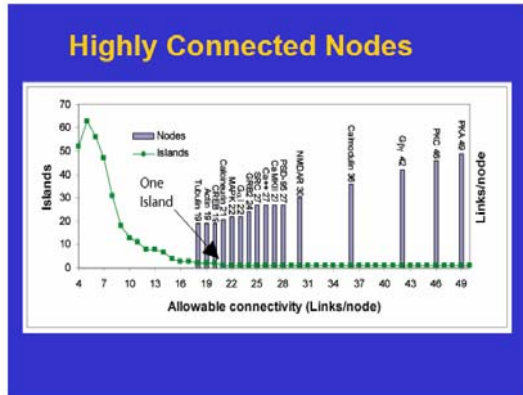


FIGURE 28

The surprise from this analysis was, none of the major players, the highly connected nodes, and ones that we know are biologically important were not needed to form the network. So what might the role of these biologically important highly connected nodes be? To answer this question we decided to determine what types of motifs are formed as these highly connected links come into play. The results from this analysis are shown in Figure 29.

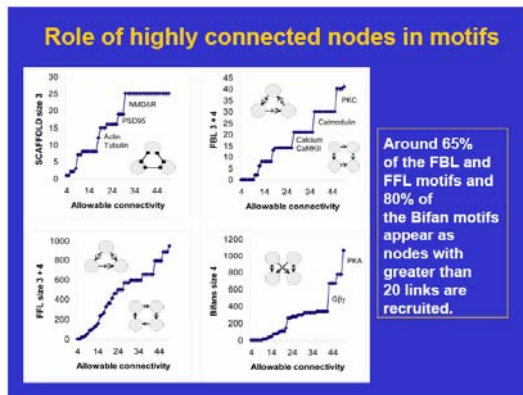


FIGURE 29

What we found is that the highly connected nodes disproportionately contribute to the formation of regulatory motifs. Eighty percent of the feedback loops and feed forward motifs occurred as these highly connected nodes come into play. For this network, it appears that the highly connected nodes are not needed for the structural integrity of the network, rather the highly connected nodes are required for the formation of the regulatory motifs that process information.

Thus the pseudodynamic analysis has allowed us to move from thinking about individual

components to groups of components within these coupled chemical reaction networks. The location of these regulatory motifs within the network allows us to define areas within the networks that are capable of information processing. Maps specifying the density of motifs at specific locations and their relative positions with respect to receptors and the effector proteins are shown. A heat map representing the density of motifs as a function of steps from the receptor is shown in Figure 30. A detailed distribution of the various types of motifs in the interaction space between receptor and effector proteins is shown in Figure 31.

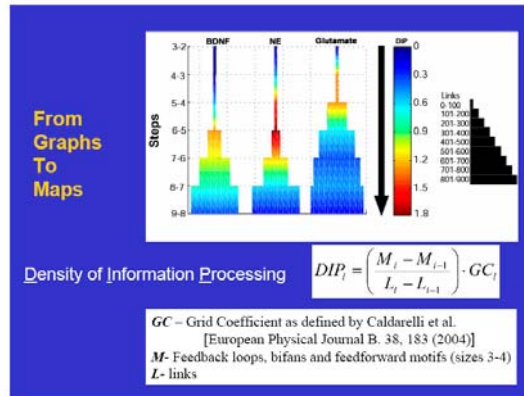


FIGURE 30

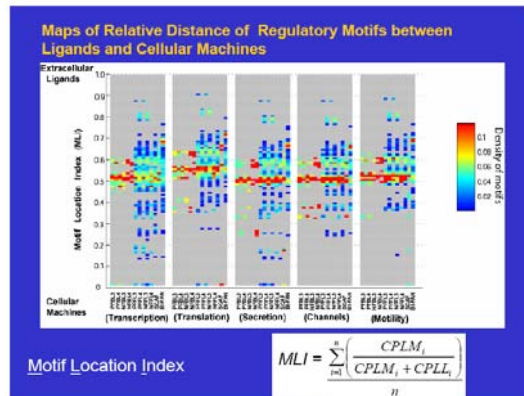


FIGURE 31

Detailed analyses for the location of the various motifs indicate that the motifs are densely clustered around the functional center of the network as shown in Figure 31. The map in Figure 31 would suggest that there are no regulatory motifs between channels. Channels pose an interesting representation problem for interactions between each other. Often channels use membrane voltage and membrane resistance to interact with each other. But voltage and resistance are not represented as entities within this network and hence these motifs are not

“seen” in our network. So, there are some corrections we need to make for biological networks that use electrical and physical forces as entities. In spite of these limitations we can make a numbers of conclusions from the type of analyses we have conducted. These are summarized in Figure 32.

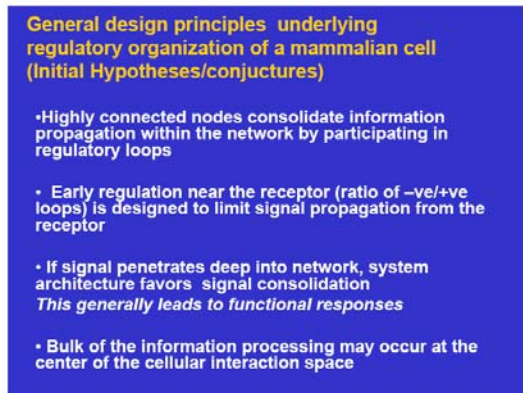


FIGURE 32

The major features of the cellular network within hippocampal cells are 1) highly connected nodes can consolidate information by participating in regulatory loops 2) Early regulation is designed to limit signals and presumably filter spurious signals. As signals penetrate deep into the network the positive loops that are formed favor signal consolidation that leads to state change. Thus description of the statistics of networks for somebody like me who works at a cellular level, is very useful in providing an initial picture of the regulatory capabilities of the cellular network.

It allows me to design experiments to test which of these regulatory motifs are operative and develop an overall picture that I would never get from a bottom up approach, if I was just studying one feedback loop or two feedback loops at a time. That is my experience with the statistics of networks. Thank you very much.

[Applause.]

QUESTIONS AND ANSWERS

DR. JENSEN: I am interested in the motif-finding aspect. Usually, with these algorithms you find all frequently occurring patterns, and then at some point you go on threshold. You say, you know things about the threshold, those are unexpected things, but because of the nature of them it is often difficult to do good hypotheses tests and say where we should draw that threshold. So, what approach did you use for saying these are motifs that seem big and interesting?

DR. IYENGAR: Actually, I did not have a real initial statistical threshold, because I was going from a biological point of view. The protein that participated was known to have important biology.

REFERENCES

Bhalla, U.S., and R. Iyengar. 1999. "Emergent properties of networks of biological signaling pathways." *Science* 283:5400.

Bhalla, U.S., P.T. Ram, and R. Iyengar. 2002. "MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network." *Science* 297:5583.

Eungdamrong, N.J., and R. Iyengar. 2007. "Compartment specific feedback loop and regulated trafficking can result in sustained activation of ras at the golgi." *Journal of Biophysics* 92:808-815.

Jordan J.D., E.M. Landau, and R. Iyengar. 2000. "Signaling Networks: Origins of Cellular Multitasking" *Cell* 103:193-200.

Ma'ayan, R., R.D. Blitzer, and R. Iyengar. 2005. "Toward Predictive Models of Mammalian Cells." *Annual Review of Biophysics and Biomolecular Structure* 34:319-349.

Dynamic Network Analysis in Counterterrorism Research

Kathleen Carley, Carnegie Mellon University

I am going to talk about some applied work that we have been doing, some of the issues that have arisen, and some of the challenges that the work drives for both network analysis in general or statistical approaches to network analysis.

To set the context, let me present two real-world examples that we have come across and had to deal with using our models. The first one occurred over a year ago when we happened to be out in Hawaii. A group of individuals associated with the Philippine terrorist group Jemaah Islamiyah (JI) had just been arrested in a move that foiled a major bomb plot. The question posed to us, was what was likely to happen to JI following that arrest. When a question like this comes to the analyst it often takes several months to deal with it; however, we were told they needed the answer yesterday. So, one of the issues is, “How can we answer such questions, rapidly and accurately?”

The other motivating example comes from analyzing the effects of an administrative change in a nation state. On the surface, the impact of a change in leadership seems like a very different question. We’re thinking of situations where, as a region of the world becomes more progressive, they become less antagonistic to the United States. Looking at the political elite in those countries to see how they are related to each other and to the military can reveal strategies by which small subtle changes in those groups can alter the lines of disagreement, possibly moving the groups more in alignment with U.S. interests. In contemplating such changes, we might ask questions such as what are the main lines of disagreement, can we influence them, would changing them likely alter the country’s attitude toward the United States. An issue here is, “How can we explore the impact of changes in a network of people on beliefs and attitudes?”

On the surface the problems of JI and of state change are very different. They certainly bring to the forefront a lot of different kinds of data. However, there are analysts who face both of these issues using the same kind of data and in both cases using data that has strong network characteristics. That is, in both cases the data includes who is connected to whom, who has what resources, skills or beliefs, and who is doing what. In both cases, the question arises, what would happen if a particular node were to be removed; e.g., as when a member of JI is arrested or a political elite is removed from power. These and other questions are often addressed—things like: How vulnerable is the overall system? What groups or individuals stand out? Who are the key actors, key groups, key ideas, and key things? How can we influence them? What are the

important connections and so on, or what is the health of the organization, how has it been changing over time? Can we infer where there is missing data, which helps to focus intelligence gathering ideas? How different are groups? And so on. There is a whole slew of questions like these that need to be addressed, some very theoretical while others are of near-term, pragmatic interest, such as “What is the immediate impact of a particular course of action?” These are the kinds of questions that need to be addressed using network inspired tools. The tools that I will talk about today are the first tiny steps on the long road to helping people address these kinds of questions and meeting this very real and practical need.

From a technical perspective what we want to do is be able to provide a system of evaluation for looking at change in multi-mode, multi-plex, dynamic networks. Maybe they are terror networks today, maybe they are drug networks tomorrow, but there is a whole set of these kinds of networks, and we want to analyze them under conditions of uncertainty. Finally, we want to place predictions in a risk context: that is, given sparse and erroneous data we want to estimate the probabilities of events and the likelihood of other inferences and the confidence interval around these estimates.

From a user’s point of view it is imperative that we provide the tools and the ability to think about analysis and policy issues from an end-to-end perspective. We all know that network tools are extremely data greedy, so we need to embed them in a larger context of bringing in the data automatically, analyzing it automatically, and using different kinds of prediction capabilities to make forecasts and so basically free up human time to do real live analysis and interpretation. At CMU we have developed a few tools as shown in Figure 1, but these are just examples of a lot of the tools that are out there. For every tool I will mention there are dozens more that more or less meet a similar purpose. Overall, our tool chain serves to bring in a set of raw text, like newspaper reports, and, using various entity extraction and language technology, identifies various networks. These are networks of people to people, people to ideas, people to events, et cetera. We then take those networks and analyze them. We are using a tool called ORA¹ for doing that, which lets us do things like identify key actors, groups, and so on. Once we have done that, the tools give us some courses of action that we might want to analyze. We take those, put them into a simulation framework, and evolve the system forward at time. All of this sits on top of data bases, etc. Basically, the set of tools help you build a network so you may find points of influence, and then help you assess strategic intervention. The important thing from a technology standpoint is that all of these things have to be built by lots of people, they have to be

¹ ORA is a statistical toolkit for meta-matrices that identifies vulnerabilities, key actors (including emergent leaders), and network characteristics of groups, teams, and organizations.

made interoperable, and we have to start thinking about things that we don't normally think about as network arrows.

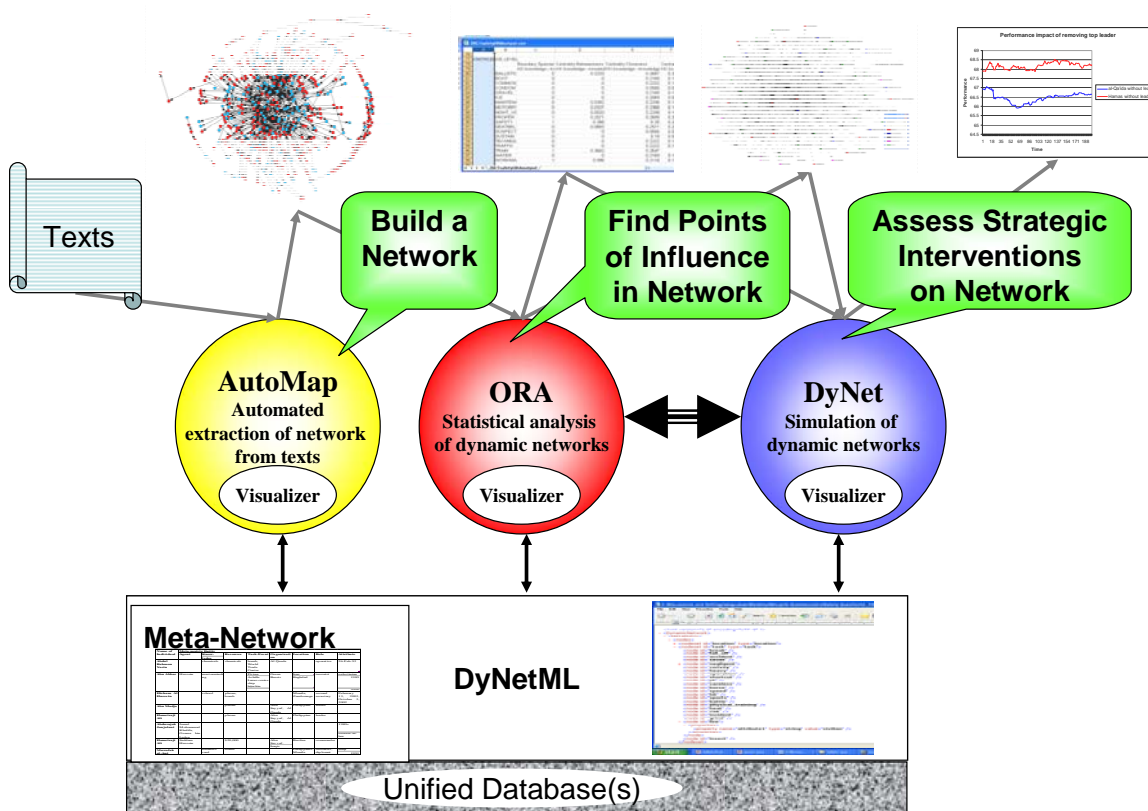


FIGURE 1 Integrated Tool Chain for Dynamic Network Analysis

From a network assessment perspective I am only going to concentrate on the middle block of Figure 1, which deals with network analysis. Basically, we ask four fundamental types of questions. One is who are the key actors? We want to do this from a multi-criteria perspective, not just worrying about who is key in terms of the social network but also other kinds of power, such as that stemming from access to money, resources, and so on. Second, we want to know about the emergent groups. What are the groups and who are there key members and leaders. Third, we want to know how we can influence someone or some group. And fourth, we want to characterize the network. Knowing what kind of network it is, is critical because the intervention options depend on the type of network.

Our tool for network analysis, ORA, is a dynamic network analysis (DNA) tool for

locating patterns and identifying vulnerabilities in networks. It lets you run a whole series of statistical tool kits on the networks and pull out a whole set of measures. It then organizes these into various reports, such as reports for intelligence, management, risks, etc. Importantly, ORA lets you utilize multiple kinds of data, not just who-talks-to-whom type data but also who has access to what resources, who has been involved in what events, who has been seen at what location. It uses these different kinds of multi-mode multi-plex data to help predict, think about, and infer actions from one network to another.

	People or Agents	Knowledge or Resources	Tasks or Events	Groups or Organizations
People or Agents	Standard Social Network	Knowledge Network Resource Network	Assignment or Attendance Network	Membership Network
Knowledge or Resources	Knowledge Network Resource Network	Information Network/ Substitution Network	Needs or Utilization Network	Core Capability Network
Tasks or Events	Assignment or Attendance Network	Needs or Utilization Network	Precedence Network	Institutional Relation or Sponsorship Network
Groups or Organizations	Membership Network	Core Capability Network	Institutional Relation or Sponsorship Network	Inter-organizational Network

FIGURE 2 Illustrative Meta-Matrix for Dynamic Network Analysis

In other words, ORA uses something we call the meta-matrix approach to evaluate connections among multiple entities at varying strength, as illustrated in Figure 2. Traditional social network analysis tends to focus on just a single entity class, such as people to people. In contrast, DNA uses networks connecting not just people to people, but people to knowledge or resources, attacks or events, and organizations, and there are other fields as well. This approach is multi-mode and multi-plex, and thus more powerful than a single mode technique. It takes us beyond the traditional social networks analysis by enabling the simultaneous analysis of relations among many types of entities.

Before telling you about some of the results from using these tools, I want to highlight some of the major issues we have come across, basically as caveats as to why it is difficult to get useful insights about dynamic networks if one only relies on simple social network techniques.

The first caveat is that you really need multi-mode, multi-plex data to reveal relations. This is certainly true when people are trying to hide, when groups operate covertly. In covert cases, we often have to infer network relations from other kinds of data, such as co-presence at a large number of events. Second, we need to collect data from multiple sources, implying different collections methods and biases, and so we need protocols that help us triangulate from those multiple sources to infer the actual relations. Finally, we have also found the need to consider networks over time; specifically looking at data from multiple time periods reveals what relations remain constant and which change.

As an example of the first of these challenges, I point to an analysis we conducted about one particular Middle Eastern country. We identified a social network for the political elite based on open source data. This observed network, Figure 3, suggests that the society is not strongly connected. (See Figure 3, noting the lack of connectivity and the isolated subgraphs, the labels are not important here.)

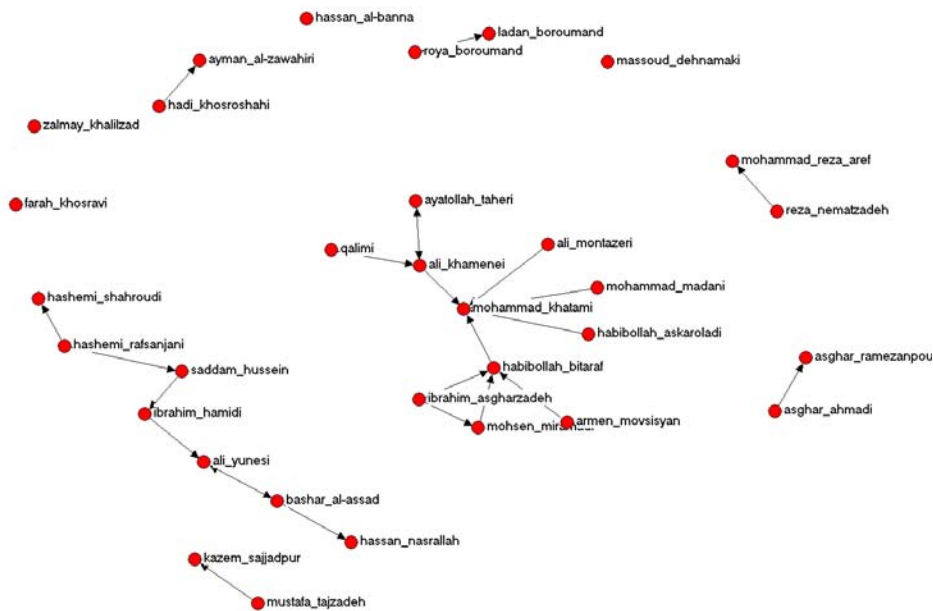


FIGURE 3 Political Elite Network Mideast

But this impression is wrong. When we dug deeper and examined the knowledge and resources that these various people had access to, we in fact found a lot of connectivity. The nature of the society is such that members who have shared a resource have probably actually met, and so we were able to infer many other social connections that we wouldn't have been able to infer without resource access data. The resulting network, shown in Figure 4, is a very connected group. This

structure reveals that the society has a dual core with two competing cores. One such core tends to be more reformist than the other. It is important to note that multi-mode data gives us a very different picture of the society and the connections among the elite than we would have inferred from just the social network in the open-source data.

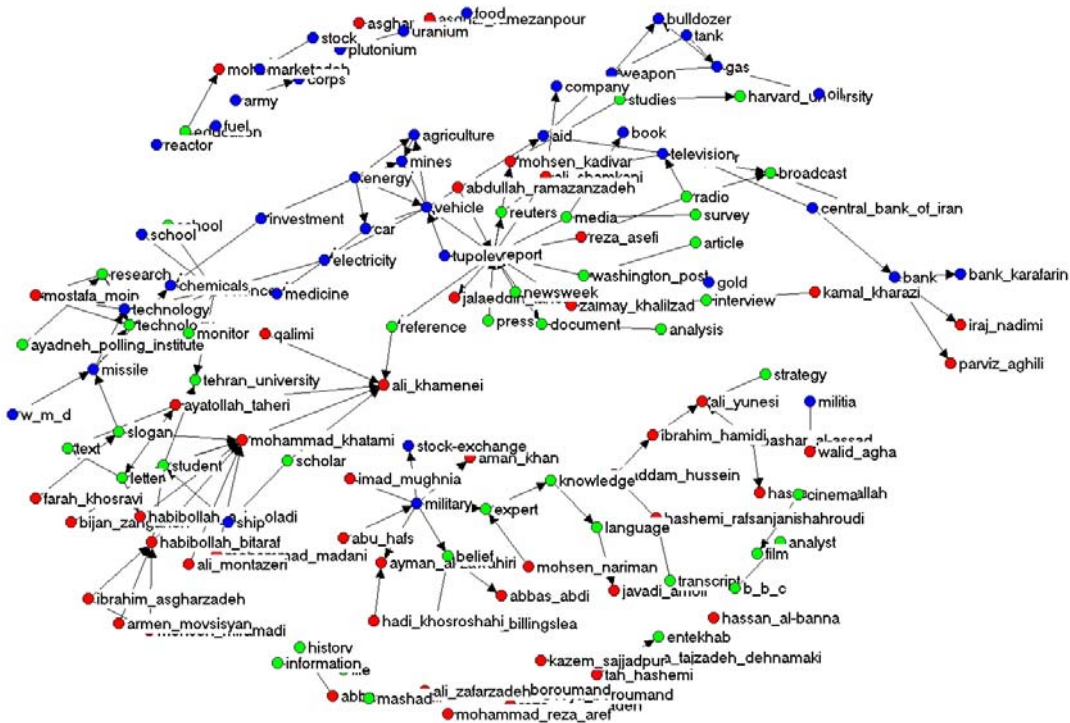


FIGURE 4 Political Elite with Connections Through Knowledge and Resources—MidEast

The next example is about collection methods. In order to understand better the insurgency in Iraq, two sets of data have been collected. But the data was collected by two separate subgroups that refused to talk to each other, and they collected the data in very different ways. The first group collected data on incidents (left in Figure 5); for each incident they generated a report that said who was involved in that incident and what was going on. Analyzing the incident data results in a network that is basically a set of isolated individuals or groups, because few of the people involved in the individual incidents had the same name. Looking at just this data, we might conclude that the insurgency is really a bunch of disconnected groups copying one another’s methods. In other words, we might conclude that fighting it, would be like fighting fires, and very difficult to counter or stop in the long run. Even when you codify location it doesn’t help; that is, the conclusion still appears the same.

The second group that collected data focused more on the resistance movement, and they included information on who attended resistance movement meetings. Based on this attendance data we get the picture of the resistance network shown on the right in Figure 5. It's a completely different structure than in Figure 4, showing more of a core periphery structure. Based on just this data, we might conclude that there is a central controlling unit that was calling the shots. In this case, fighting the resistance could be successful if we could identify and alter that core. Now many of the same people appear in both data sets. If we combined them we find that the little networks that were identified through incidence data are basically on the fringe of the core of the resistance. This would alter further the operational conclusion. The point of this example is that when dealing with covert groups and working to set operational direction, it is generally useful to collect data in multiple ways and then combine them before drawing any operational conclusions.

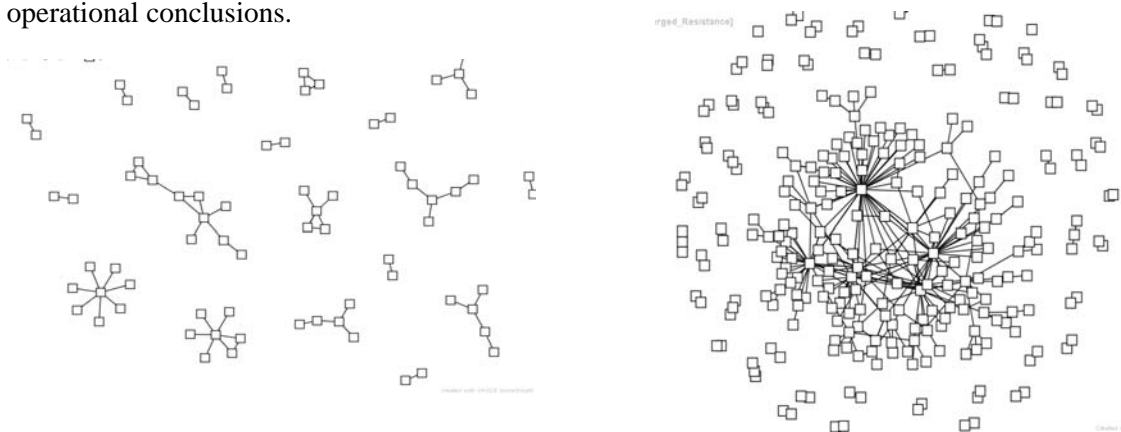


FIGURE 5 Social Network Based on Incidents (left) and Resistance (right)

What about the case where there is data from multiple time periods? For Iraq, we had the opportunity to analyze newspaper data prior to the election. In a very quick proof-of-concept exercise, we collected data on both the incidents and resistance movement. We had a set of people, and we tracked them over a one-month period from November 19 to December 25. At that time, the prevailing view was that the Iraqi insurgents were a bunch of disconnected groups, that didn't have anything to do with each other. When we actually started taking those people over time using newspaper reports we found that there was a core group of about 50 people who kept showing up repeatedly. As you tracked them over time you saw that they did have very strong connections with each other and they did have a very strong group structure. In fact, we were able to actually assess who the leaders were once we started capturing data over time. Thus, the third issue that is critical if network techniques are to be used in meet applied needs, is that

data must be captured and assessed over multiple time periods.

Once we got all that data we asked questions like who do we target? Where are the vulnerabilities in these kinds of networks, who are the leaders, who stands out, and so on? Using traditional social network measures, based on a single mode data of just who talks to whom, we would typically reveal node attributes such as the centralities: who is the most connected or which of two figures is on the most paths. In the 9-11 hijacker network constructed by Valdez Krebs, four individuals stand out—Alhazami, Hanjour, Atta, and Al-Shehhi. If you actually track the full network and are able to pull out the centralities you have a good understanding of different ways of affecting the group, because then you can affect the flow of communications, the transmission of illness, and so on. However, as was certainly the case with the hijacker network, often you only see part of the network. As such, these centralities can be misleading. Moreover, from an intervention perspective, you sometimes are more concerned with issues of power, control or expertise. As such, you might be better served by using exclusivity measures, like who has exclusive access to particular resources and so forth. Thus, another issue is that you need, since for many applied concerns, you need to move beyond single-mode single plex data, you also need new metrics for identifying core vulnerabilities.

In building the dynamic network tools we have tried to address the issues described above, and others. We assess not only the social network but also go on to ask what resources do the members of the network have access to, and how does that access relate to their position in the social network. With event and task data we go still further, and ask who has been at what events, who is doing what tasks, and so on, and use that information to infer missing social linkages and identify different social roles. It is still possible to calculate things that like centralities, but they are on multi-mode data and as such begin to reveal individuals' roles. Individuals can stand out on a number of dimensions, not just by virtue of their communication. For example, one metric we have developed measures who is likely to be the emergent leader, using a multi-mode multi-plex metric called cognitive demand. Additional roles can be defined in terms of work load, subsequent event attendance, and exclusive access to resources. All the things that we want to talk about with respect to individuals and why they are important in networks can begin to be pulled out when we place networks in this kind of dynamic multi-mode multi-plex context.

This dynamic approach affords a much richer understanding of how we might impact the underlying network. The person who is connected to a lot of other people but doesn't have special expertise or doesn't have high cognitive demand is probably a good target for going to and getting information. However, if I want to break the system, I might want to go in and pull out

the individuals who have exclusive access to particular resources, so I want to use an exclusivity metric. If I want to impact not just the formal but the informal leadership then I would want to use cognitive demand.

Once I've identified a possible individual, I might want to ask how I can influence that person. To answer that I would ask who are they connected to, what groups are they in, what do they know, what resources do they control, what events have they been at, and so on. Figure 6 shows the sphere of influence for one actor, showing their focus—their ego network, basically—but in this multi-mode multi-plex space. In addition, as part of the sphere of influence analysis, we also calculate how the actor fares relative to everyone else on a number of metrics. And, we assess who are the closest others to them in that network; i.e., who are most like them structurally. Now Figure 6 implies that Khamenei is close to Khatami in this data set. On the one hand, this is funny because they hold completely opposite political views. On the other hand, this makes sense, because their political positions ensured that they would be connected to the same others.

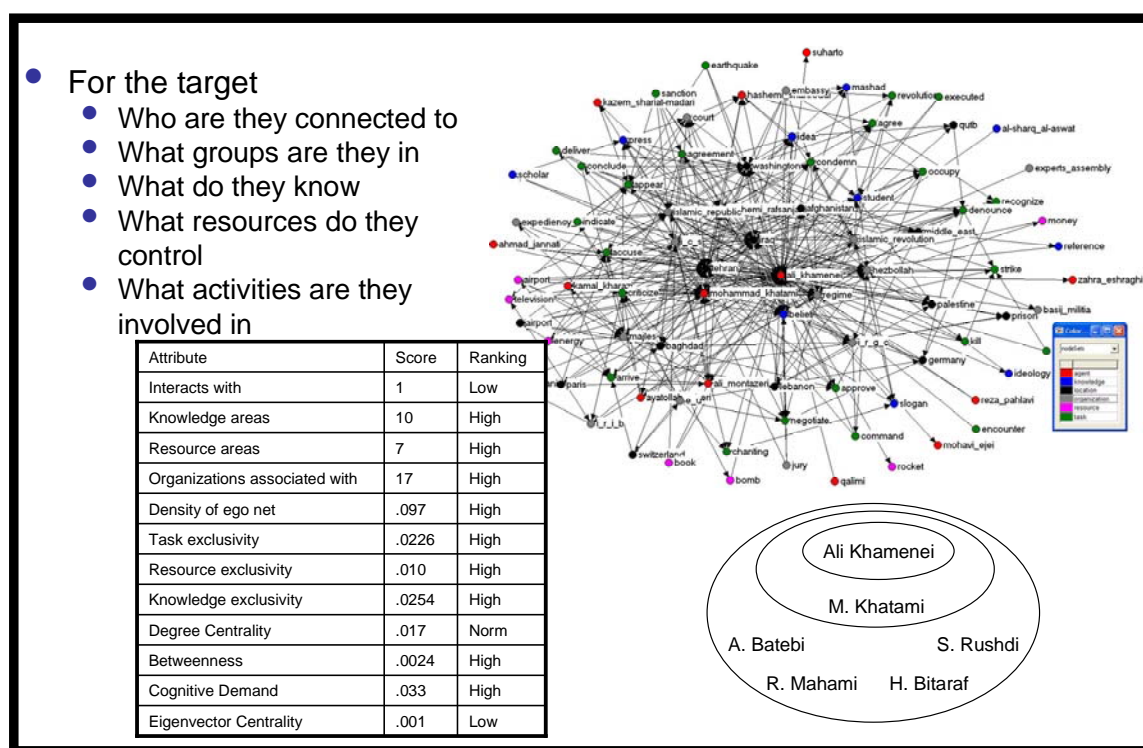


FIGURE 6 Sphere of Influence Analysis

We next want to ask how such a network can be broken into groups. In Mark Newman's paper, he introduced some grouping algorithms. Grouping algorithms are extremely important in this area. We use grouping algorithms in a dynamic sense to look at the evolution of the structure of a network. Figure 7 shows the structure of al Qaeda in 2000 as based on open source data. Figure 8 shows some of the groups that were extracted from that data using one of Newman's algorithms. The block model reveals a structure that basically shows a bunch of individuals who are all tightly connected along the diagonal to a bunch of separate groups, and then a few individuals who serve as liaisons connecting multiple groups. The upper red horizontal rectangle in Figure 8 is the row of group connections for Zawahiri, while the lower red horizontal one is for bin Laden. They cross connect between groups. This is what is now referred to as the classic cellular structure.

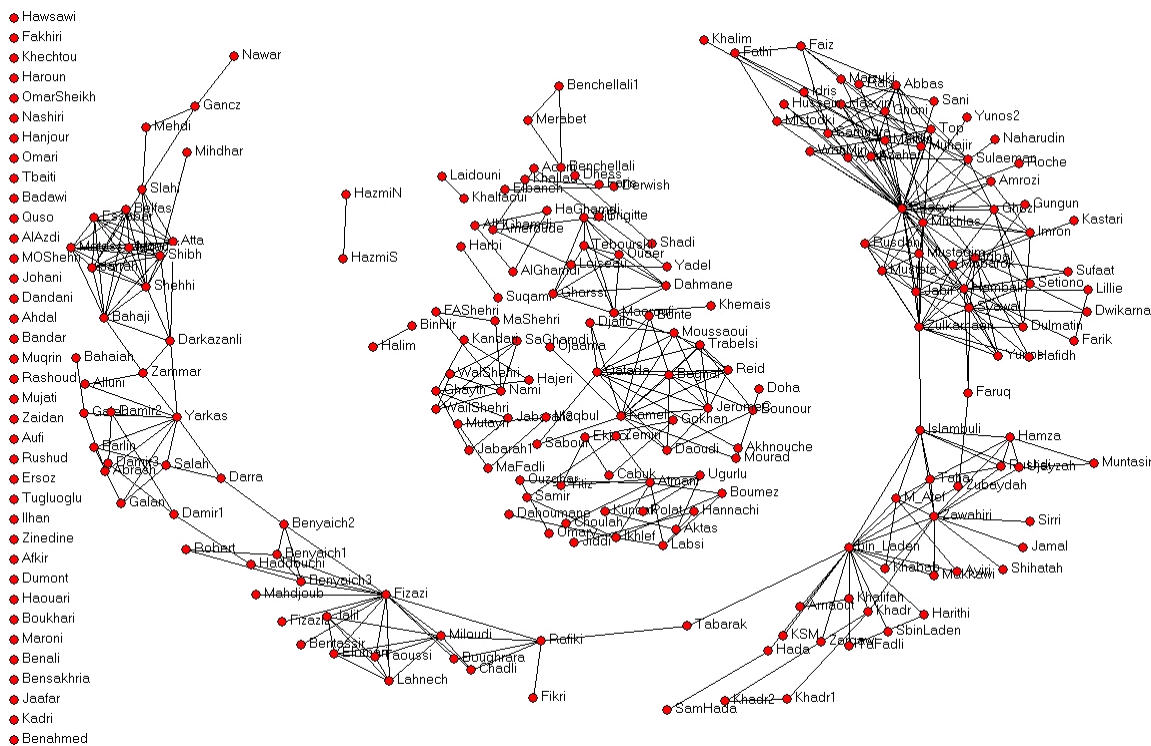


FIGURE 7 Al Qaeda 2000—Open Source Information

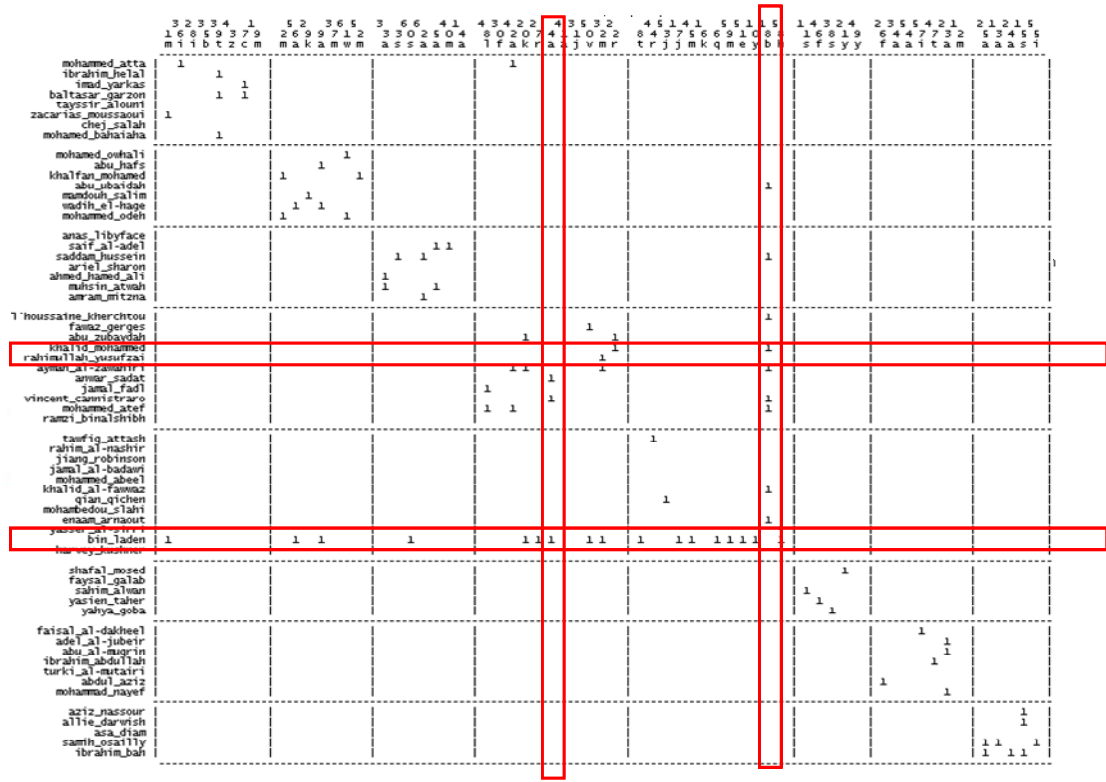


FIGURE 8 Al Qaeda 2000—Cellular Structure

Using those techniques, we can now look at change over time in network. By simply running these grouping algorithms on the network at multiple points in time, and tracking which critical individuals move between them, we begin to track the extent to which these networks are evolving and changing their basic structure. I am going to show you some networks and talk about this in the context of some work we have been doing on Al Qaeda using data collected from open sources on what it looked like in 2000, 2001, 2002, and 2003. In the Figure 9 the meta-network of al Qaeda circa 2000 (left) and 2003 (right) is shown; with different symbols for people, resources, locations, etc. The red circles are people. All the other things represent known locations where operatives were, resources, knowledge at their disposal, roles that people were playing, and critical events. The point that I want to make about this is that there is a lot more information here than just who was in contact with whom. We know a lot more about al Qaeda. This means that, if we are going to be smart about this, we have to use that other information to be able to think about how to evolve these networks and how they are changing and to

infer what the social network looks like.

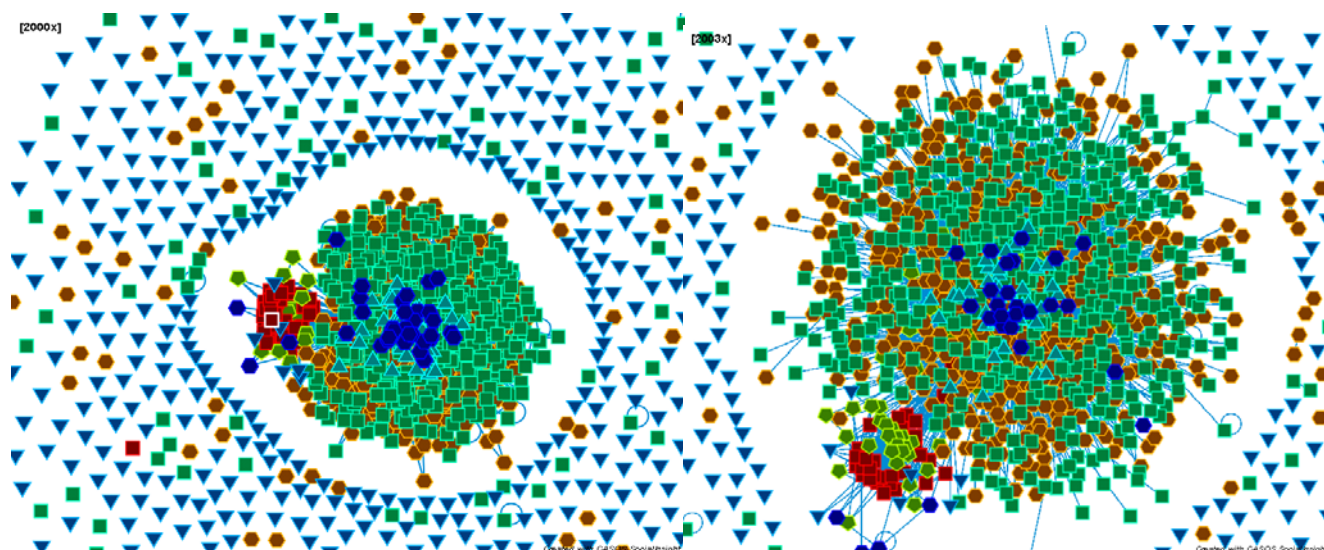


FIGURE 9 Meta-Matrix for Al Qaeda—2000 (left) and 2003 (right)

Has al Qaeda remained the same over time? In short, no; in fact, it has changed quite a lot. Although you can't tell it from Figure 10, if we were to actually zoom in on the graph for 2003, the red circles, which are the people, actually form two completely separate, broken apart, sub-networks. Whereas, it was one completely connected group back in 2000—except for the isolates—but in 2003 we have got two big separate cores with no connections between them. They are only connected through having been from the same locations, which may be dead letter drops, or who knows.

Let's examine this data in more detail. So far, I am just showing you two visual images and asserting that these pictures are different. Behind this are a lot of statistics for looking at how the networks change. For example, in Figure 10 we look at the movement of al Qaeda over these four years on a number of dimensions. Over time, the density of the overall network has gone down. Basically, this is attrition effect, and it has gone down not just in terms of who is talking or who is communicating with whom, but it has gone down in access to resources, access to knowledge, and involvement in various tasks. At the same time, the network was become increasingly adaptive as a group from 2001 to 2002, but it had suffered so much attrition by 2003 that it kind of stabilized in a new organizational form. In terms of other factors, we know that, over time, the communication structure of the group has changed, such that the average shortest path among actors, for example, has actually increased. So, on average, communication should

take longer. The communication congruence, which is a mapping of who needs to communicate with whom to do the kind of tasks that they are supposedly doing, has actually improved. This suggests the very different organizational structure that evolved between 2000 and 2003. A leaner, more tightly organized structure appears to have evolved.

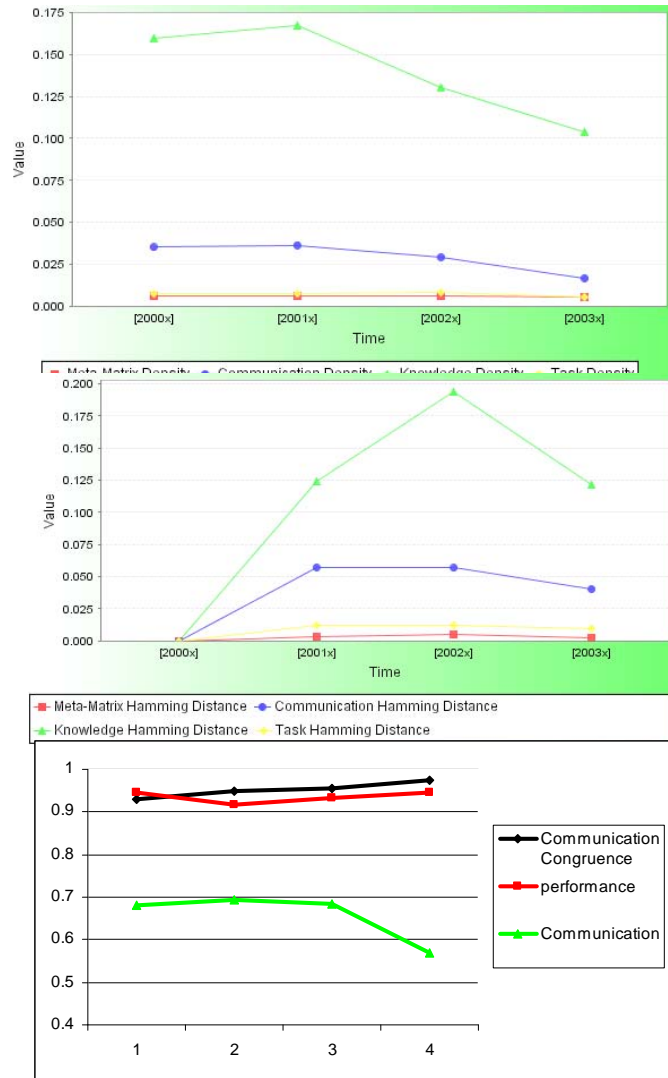


FIGURE 10 Change in Al Qaeda Over Time

What about performance? Learning, recruitment and attrition lead to changes in the network which in turn impact performance. Some of these changes, e.g., learning are natural, and others, e.g., attrition are the result of strategic interventions. We have a whole series of tools for looking at dynamic change in networks, for looking at the effects of both natural evolution and interventions. Basically these tools—we have used ORA to estimate immediate impact using

comparative statics and DyNet to estimate near-term impact using multi-agent simulation—let you look at change in networks immediately or in the near term, and to look at change hypothetically or for a specific network.

You can use these tools to ask questions about network evolution. Fundamental principles from cognitive psychology, sociology, and anthropology about why people interact, the tendency of people to interact with those who are similar, tendency of people to go to others for expertise, and so on, form the basis for the evolutionary mechanisms in DyNet a multi-agent technique for evolving networks. DyNet can be used to address both natural and strategic change. For example, we took the al Qaeda network, shown in figure 7 and evolved it over time, naturally. The result was that the overall network, in the absence of interventions such as arresting key members, oscillated back and forth between a two core and single core structure. The structure at 300 steps is similar to that at 100 steps, with a totally connected structure intermediate at 200 and 400 steps. Without intervention, it just oscillates between these two forms over time, according to these basic fundamental principles.

We also used DyNet to examine the effects of hypothetical changes to stylized networks composed by blending actual network data with network structure extracted from more qualitative assessments. For example, Figure 11 shows an expected change over time for Al Qaeda, which is in blue, and Hamas, which is in red, based on a combination of real and stylized data drawn from qualitative assessments. On the left, we see the state if we just left them alone, in which al Qaeda out-performs Hamas and so on, and on the right what would happen if the top leader were removed from each. We did this analysis a while back, when Yassin was the Hamas leader and, in fact, Hamas' performance did improve once Yassin was removed.

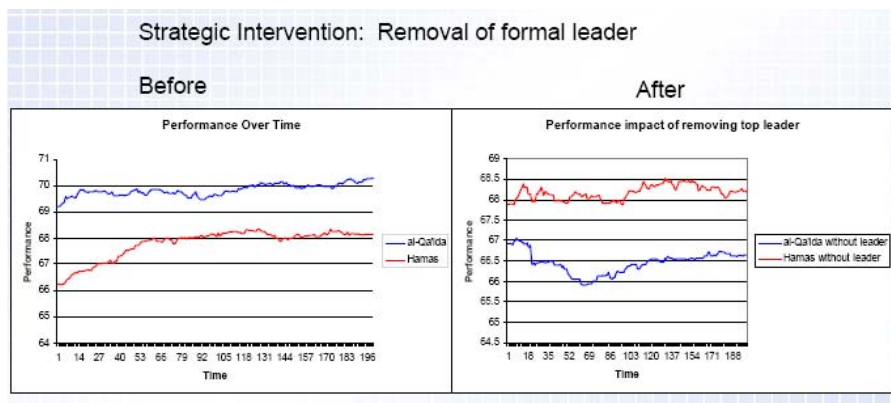


FIGURE 11 Impact of Removal of Top Formal Leader

Finally, we can use DyNet to do course of action analysis and ask “How do these networks evolve, when there are specific strategic interventions?” We might, take a network, for example the al Qaeda 2000 network from Figure 7, and isolate the individuals who stand out on some of the various measures. For example, the individuals who were highest in degree centrality or betweenness or cognitive demand might be removed; i.e., bin Laden or Baasyir. We could, alternatively, remove a whole group of people, such as the top 25 in cognitive demand. Each of these what-if scenarios represents various courses of action. Running this series of virtual experiments results in comparative impact statistics like those in Figure 12. In this example, any of the actions considered would lead the system to perform worse than it is now. Caveat: these results were generated using a multi-agent simulation on a network extracted from open source. The relative value of this analysis is not to make a point prediction of what actually happens to these groups, it is not to predict specific reductions such as that pulling out all the high-cognitive-demand people will result in 40-percent lower performance. Rather, the value of this analysis is to show relative impact; e.g., that the relative impact will be stronger were you to remove all those people versus removing just bin Laden and that any intervention is more crippling than none.

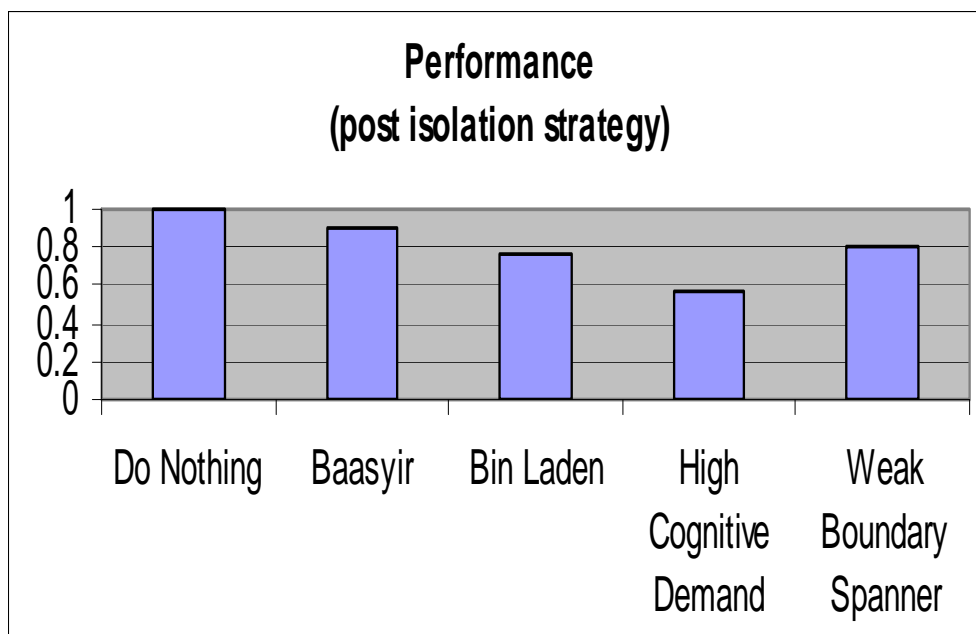
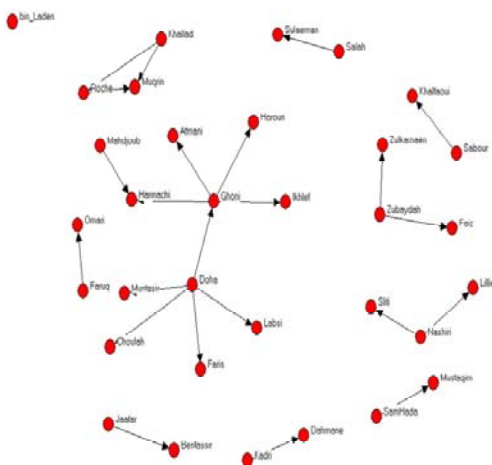


FIGURE 12 Relative Impact of Different Courses of Action

You can also use such an analysis to go further and ask, if we were to remove particular individuals, then—since people are learning and changing anyway—who is going to start interacting. That is, where should we see the impact of change? In fact, we might find many people starting to interact in the simulation. An example is shown in Figure 13.

Bin Laden Isolation → New Network Relationships



New Relationships based on bin Laden Isolation			
From	To	Value	NormValue
Salah	Sulaeman	0.07378	1
Nashiri	Sliti	0.016457	0.223052
Ghoni	Haroun	0.012465	0.168949
Zubaydah	Faiz	0.012083	0.163765
Sabour	Khalfaoui	0.011504	0.155918
Ghoni	Ikhlef	0.011457	0.155283
Nashiri	Lillie	0.01078	0.146111
Ghoni	Atmani	0.01067	0.144618
Doha	Ghoni	0.009806	0.132915
Mahdjoub	Hannachi	0.009584	0.129906
Roche	Muqrin	0.009504	0.128819
Ghoni	Hannachi	0.009374	0.12705
Doha	Faris	0.009175	0.124359

FIGURE 13 Emergence of New Relations After Removal of bin Laden

If I were the analyst, I would look at these results in Figure 13 and ask “What does it mean that the simulation is predicting that Salah and Sulaeman will start interacting with a probability of one?” Well, it is actually .999999999. Even so, is the probability really that high? What that means is that there is a good chance that these two individuals are already interacting and additional information gathering efforts should be directed to confirming this, if it is critical. The point here, is that suggestions for information gathering and about missing data are some of the side benefits from doing predictive analysis of network change.

In summary, from an applied perspective there is a need to move beyond simple network analysis to look at dynamic multi-mode, multi-plex approaches. Doing so will require us to put network statistical analysis as a component within a larger tool chain moving from network extraction, to analysis, to simulation, and employing more text mining, data mining and machine learning techniques. The tools I have shown you are a step in this direction, and their use has raised a number of key issues that the network community needs to deal with. At this point, I will open it to questions.

QUESTIONS AND ANSWERS

DR. HOFF: This is a very sophisticated model that you have for the behavior of the networks under changes. Even when we look at observational studies of the impact of, say, beta carotene on people's health, we get that wrong. So, you need to do a clinical trial to figure out what the causal effect is. Is there anything you can do, or what sort of methods do you have for sort of checking—you talk about what will happen if we modify X. Is there any way of checking that or diagnosing that?

DR. CARLEY: This is going to be kind of a long answer. There were three parts to that question. The first thing is that, in the model, there are three fundamental mechanisms for why individuals change their level of interaction. One is based on this notion called relative similarity or homophily based interaction. I interact with you because it is comfortable; you have stuff in common with me. Another is expertise, you have something I need to know. And the third is task based, we are working on the same task, so we start interacting. Those three basic principles all came out of different branches of science, and there is lots of evidence underlying all three of them. So, we began by taking things that had been empirically validated, and put them in the network context.

Then what we did was, we took this basic model and we have applied it in a variety of contexts. For example, we applied it to NASA's team X out in California and to other groups at NASA. We applied it to three different companies that don't like to be talked about. In all cases we used it to predict changes in interaction, and in general we got between a .8 and .9 prediction of changes and who started to interact with whom. It was less good at who stopped interacting. So, we know that there is a weakness there.

In terms of the covert networks themselves, we actually used this model to try to predict who some of the emergent leaders were in some of the groups over time. I can tell you that, in the case of Hamas, we correctly predicted the next leader after Yassin using this model. Does it need more validation? Of course, but that is the beginning answer.

The other part of the answer, though, is that one of the ways that things like this are validated in the field, and one of the things that people like about models like this in the field, is that it helps them think systematically about data. Thinking systematically is really important. The second thing is that, even if the model gets it wrong, it suggests places to look to gather data and, when you get that data back in, you can then modify and upgrade and adapt things based on that new data, which is what I think we are in the process of doing.

DR. GROSS: When you said you used a variety of statistical tools, other than calculating measures, did you use any other statistical tools, cross validation or boot strapping? I don't know, I am just mentioning random—

DR. CARLEY: The tool itself basically has a lot of measure calculations. It also has an optimizer system for optimizing the fit between a desired network form and an actual network form. It also has a morphing procedure for morphing from one form to another that is based on facial morphing technology, and it has some clustering and other techniques for grouping. In terms of the simulation models, we have tried different types of validation techniques.

DR. BANKS: I have a quick question. It is sort of technical. You were using a Hamming network on the two networks as a measure of adaptability, and I am not quite sure what the intent of adaptability is there.

DR. CARLEY: Basically, one of the ways in which people think of organizations as being adaptive is that they change who is doing what and who has access to what resources, and who is working together. Rapid changes in those are often considered a key leading indicator of adaptivity. That doesn't mean improvement in performance. It just means it is adaptive from the organizational sense to change. So, what we are looking at is a Hamming metric between—thinking of this as an organization—what the connections were in this multimode state at time one versus two and that becomes our measure.

DR. MOODY: Could you speak about what you mean by performance? You have seen reductions in performance, but I could imagine numerous dimensions of performance.

DR. CARLEY: Right now, in the system, there are two things that are considered performance metrics that people are using the system on. One is basically a measure of information diffusion, which is basically based on shortest path, with the idea that systems tend to perform better if they are closer together and so on. A second measure of performance is what we call our performance-as-accuracy metric, which is a simulation technique that estimates, for any given organizational form, its likelihood of correctly classifying information, given a very trivial classification choice task. It is not a specific measure of how good are you at recruiting or how good are you at X, Y, and Z, but it is a generic thing of how good is a structure, is this structural topology good at doing classification choice tasks. We use that because there is a lot of evidence out of organization science that, in fact, a huge portion of tasks that any group does are classification choice tasks. So, this seems to get at it, and we have some validation in the sense that we have looked at 69 different organizations in a crisis, non-crisis condition, and you can correctly get their ensemble performance differences using this kind of a generic measure. It is a good indicator.

DR. HANDCOCK: Leading into the next session, could you make some comments on the reliability and quality issues and, in particular, about missing sample data?

DR. CARLEY: Well, the first thing I will say is that these metrics in here don't yet have confidence intervals about them, showing how confident we are that this metric is what it is given the high levels of missing data. So, that is an unsolved problem. The second thing is that we have been doing a series of analyses, as I know several other people here have been doing similar analyses, looking at networks where we know what the true answer is, and we then sample from it, and we then re-estimate the measures to see how robust they are.

Steve Borgatti and I have just finished some work in that area, which suggests that, for a lot of these measures, if you have 10 or 20 percent errors, that as long as you are not trying to say, this person is number one, but instead say, this person is in the top 10 percent, you are going to be about right some 80-90 percent of the time. So, it has that kind of fidelity. Is it worse or better with particular types of network structures? We are not sure yet. Some of our leading data suggests that for some types of network structures, like corporate free networks, you can in fact do better than that.

Data and Measurement

Current Developments in a Cortically Controlled Brain-Machine Interface

Nicho Hatsopoulos, University of Chicago

DR. HATSOPOULOS: My background is in neuroscience. I started out in physics, went to psychology, and now I'm in neuroscience. Today I am going to talk about some work we have been doing starting 10 years ago with my collaborators at Brown University, which is trying to make sense out of large data sets collected from the brain, particularly in the cortex of behaving monkeys, and then where we are taking that to the next level.

It is going to sound a little bit engineering in flavor, but I hope to convince you that it has not just applied applications but also has scientific interest. Let me start off by telling you what people in my line of business do and have done historically. For about 40 years, my field might properly be called behavioral electrophysiology. What we are trying to understand is electrical and physiological signals in the brain and how they correlate with behavior, whether it is sensory signals coming in, visual or auditory signals, cognitive process, or motor outputs. Since the late 1960s people have been able to record these electrical signals in behaving animals and, for the most part, what people have done is address this encoding problem. In our case, we have worked with a monkey that we trained for several months to play a video game by using a joy stick to move a cursor to particular targets. Historically, what people have done is insert an electrode into the an animal's brain to record the extracellular action potentials from individual neurons while the animal performs a task. Particularly in cortex, signals are believed to be highly noisy so you have to do some sort of multi-trial averaging, requiring the animal to perform the task over and over again.

Figure 1 shows an example of five trials. These are raster plots showing you the time occurrence of the spikes, which are action potentials. These are extracellular recordings. By averaging many of these rasters you can get an average response which is shown in the yellow graph, and this is a typical average response for a neuron in the motor cortex. What we are plotting on the y axis is the firing rate of the neurons versus time. This vertical bar is the onset of movement when the monkey first begins to move. Typical of the motor cortex, it starts firing maybe 200 or 300 milliseconds before the moving begins, and it is believed to be intimately involved in driving the arm, driving the motor neurons in the spinal cord that ultimately activate the muscles and move the arm. This same approach has been used in the sensory domain as well. You present a sensory stimulus multiple times, so as to wash out noise, and you get this sort of response.

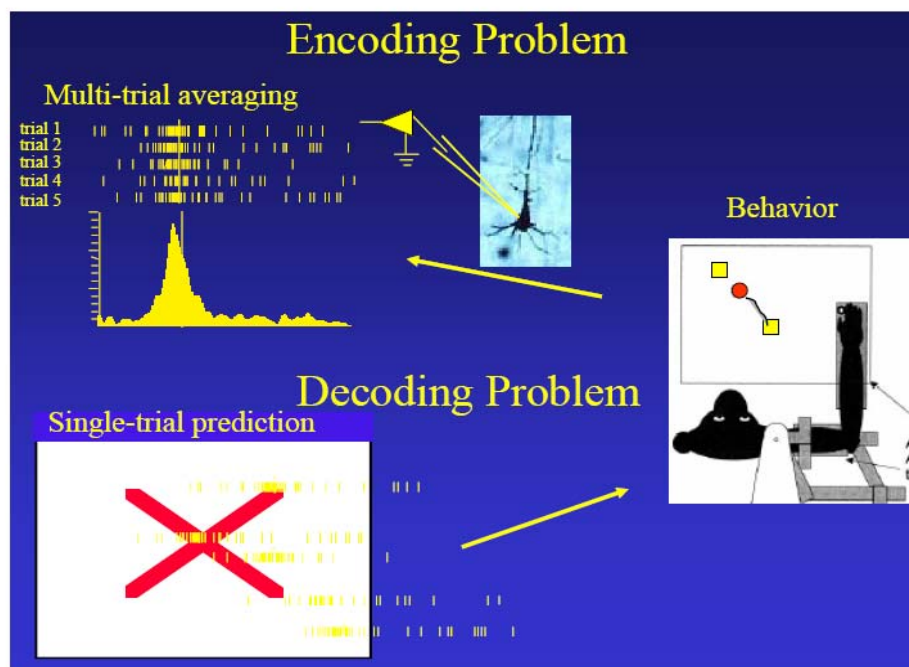


FIGURE 1 Unpublished.

What we have been doing, especially over the past five years, is taking the inverse approach, the so-called decoding problem, which in some sense is more realistic in terms of what the brain actually has to deal with. We put in multiple sensors, multiple electrodes in a particular area of the brain, the motor cortex, record multiple single neurons, their action potential, and based on the activity on a single trial, tried to predict what the animal was going to do. We don't have the luxury of trial averaging, we just have a snapshot of activity at a certain moment in time and try to predict what the animal is doing. This multi-electrode sensor, shown in Figure 2, is this so-called Utah array that was developed by Dick Normann, a biomedical engineer at the University of Utah. It is a silicon based array consisting of 100 electrodes arranged in a matrix 10 by 10. Each electrode is separated from its neighbors by 400 microns. The length of those electrodes is typically 1 or 1½ millimeters long. So, we are picking up the cortical layer where the cell bodies lie, and the tips of these electrodes are what actually pick up the electrical signals. The rest of the shaft is insulated. These tips are platinized and we pick up these extracellular signals. The picture on the right of Figure 2 gives you a sense of the scale. Oftentimes when I first gave these talks I would present this thing on a big screen and people would think this is like a bed of nails destroying the brain here. In fact, that is my finger to show you the scale. It is quite small, and for a neurosurgeon it is essentially a surface structure although it is penetrating in

about a millimeter.

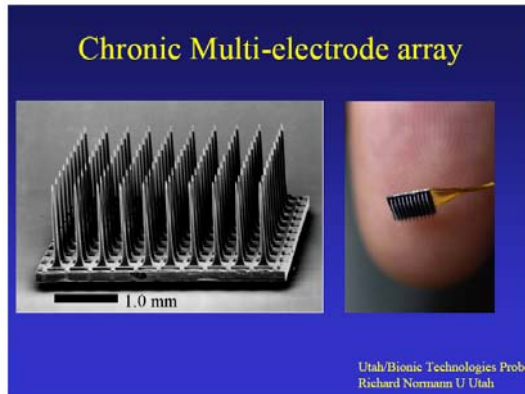


FIGURE 2 Left panel: electron-micrograph of Utah microelectrode array taken by Dr. Richard Normann. Right panel: photo taken by the Chicago Tribune.

I am going to work backwards. I am going to talk to you about where we have taken it to the next step then take you back into the lab with our monkeys, tell you where we are going for the future, and where this area of decoding is going to head in the next few years.

About three or four years ago we took it to the clinical setting. The idea was if we could predict, based on a collection of signals from the motor cortex, where a monkey was going to move, could we use that to help someone who was severely motor disabled like a spinal cord injured patient who can't move their limbs. Could we extract those signals while the patient is thinking about moving, and then control some sort of external device, whether it be a cursor on a computer screen, or a robotic device, or even their own arm by electrically stimulating the muscles. There are already devices out there for people who are partially paralyzed who can still move their shoulders but can't move their wrist or their fingers. Basically, these devices have little sensors in the shoulder and depending on the orientation of the shoulder joint, will generate different patterns of electrical stimulation, and generate different canonical grips like holding a cup of coffee or holding a fork or whatever. Ultimately we want to connect those devices with our cortical sensor and then provide a kind of practical device for a patient.

About three years ago we formed a company that took this to the clinical setting and initiated an FDA approved clinical trial involving five patients. We had three proofs of principle and milestones that we had to meet. First, by putting this in a spinal cord injury patient, could we extract human signals from a human cortex? That hadn't been shown before, at least with this particular device. Secondly, could the participant, the patient, modulate those neural signals by thinking about moving? Thirdly, could those modulated signals be used for a useful device? What we are going to talk about today is just moving a computer cursor.

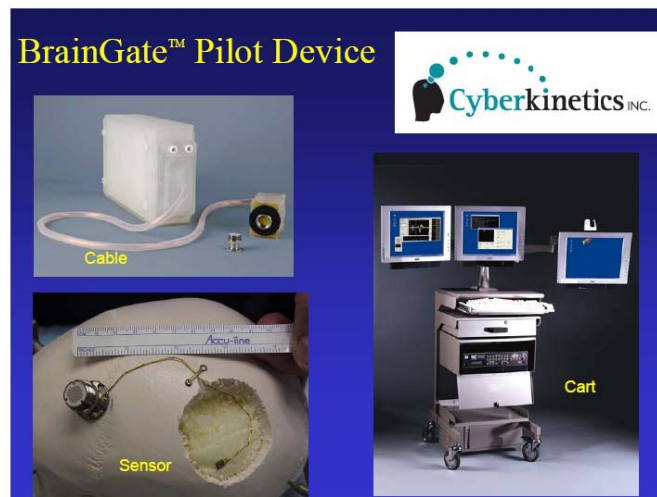


FIGURE 3 Photos taken by Cyberkinetics Neurotechnology Systems, Inc.

Figure 3 shows our so-called BrainGate pilot device which basically has this sensor which is the same sensor we use in our monkeys, connected to a connector that is secured to the skull. Going through the skin we can connect it to an external computer and acquisition system which is shown here. We collect the data, do our calibration, build our so-called decoding model to make sense out of those signals and hopefully, allow the patient to move a cursor.

Our first patient was a young man in Massachusetts who was basically tetraplegic and decided to participate. We implanted the sensor array in his motor cortex during a 3-hour surgery, he had an unremarkable recovery from the operation, and several months afterwards we began picking up electrical signals. Figure 4 shows examples of raster plots from three channels while we had the patient imagine opening and closing his hand. You can see all three of these neurons tend to fire when he imagines closing the hand, and then stop firing when he imagines opening the hand.

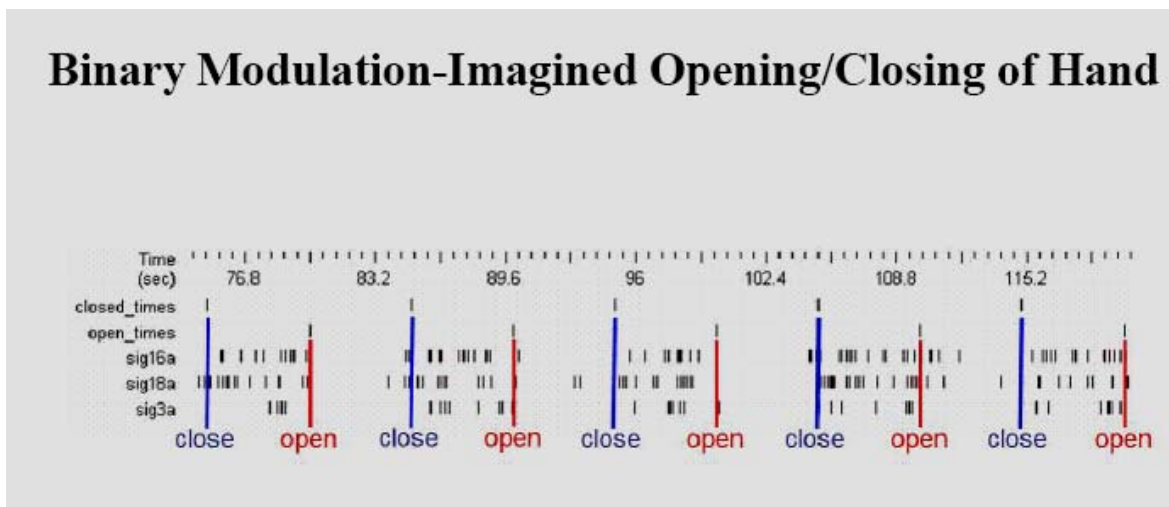


FIGURE 4 Figure generated by Cyberkinetics Neurotechnology Systems, Inc.

Here is a video clip of this voluntary modulation. [A video clip was shown at the workshop.] When the technician asks the patient to think about moving the whole arm either to the left or to the right or relax, the top raster plot displays a big burst of spiking as the patient imagines moving his arm to the left. Thus, we met the first two milestones of our clinical trial, demonstrating that the neural signals can be detected and that the participant can modulate that neural output.

We met the third milestone by actually showing that the patient could basically read his e-mail, turn on the television, and so forth. This is an example of what can be done. It is pretty crude, and there are already a lot of devices out there for spinal cord injured patients. This patient can speak, so we have to really demonstrate that what we are providing for this patient isn't something he could do for himself with a much cheaper solution, such as an existing voice-activated system. I think we have just begun to scratch the surface with this technology.

One of the interesting results from this study and also with our monkey studies is that it has always intrigued me how few signals we can extract and yet get reasonable control and reasonable accuracy and prediction of movement. Under that little chip, that sensor array, there are about 1.6 million neurons, and yet, by extracting signals from maybe 20 or 30 neurons or multi-unit activity, we can do a reasonably good job.

Why do we have all those neurons in the first place? It is intriguing, and there are all kinds of hypotheses. There is some redundancy in the system. I am not sure if that is the whole answer, but that has always been intriguing to me. Again, this sensor array is just put into the arm area of the motor cortex. We know it is activated when the arm is moved or planned to move.

Aside from that, we are not picking particular cells or particular areas.

Let me switch to the work that I have been doing in the lab with the monkeys, and taking this idea of a brain-machine interface to the next step. One of the ways we view interaction with the world is to consider two different general modes of interaction: a continuous (analog) mode, such as moving a mouse on a computer screen or moving one's limb in space and time, and a discrete (symbolic) mode, such as language or pressing keys on a keyboard, or even grasp configurations such as holding a cup and so forth, which can be considered discrete kinds of movement. Our feeling is that those two different modes of movement or selection processes might involve two different kinds of cell populations and different kinds of decoding algorithms.

To tackle this problem with our monkeys we trained them to perform two different kinds of tasks. To get at this continuous mode we had the monkey basically scribble to follow a set of randomly positioned targets as shown in the top left part of Figure 5. When they get to a target it disappears and a new target appears. They move to it and it disappears and so forth. Over multiple trials they generate the kind of mess shown in the lower left. To get at the discrete mode of operation we use a much more common task, which is called a center-out task, where basically the monkey reaches from a center target to one of eight targets positioned radically away from the center, and keep repeating, doing the same task over to one target. In the actual experiment each target is randomly selected on any given trial. By doing this they generate the relatively stereotyped movement to different targets shown in the lower right of Figure 5. The key here is that we break up the center-out task into a planning phase and an execution phase: in the planning phase, one of the squares on the outer ring is turned to yellow, signifying it is the target, but the monkey is now allowed to move. He has to wait there for about a second and presumably plan his movement. After about a second, the target starts blinking, indicating that he has to move. When the monkey waits for that blinking, we know we have trained the animals. It takes a lot of training to get them to do that because as soon as they see that target their natural inclination is to move to it because they know they are going to get their apple juice and be happy. Again, it takes some time to train them to do that, but they will wait for about a second. We are going to look at that early planning phase and see if we can then predict which target the animal has selected. That tackles two different kinds of modes of operation. What we were interested in is whether there were different cortical areas that might be better suited for these two different modes of operation.

Continuous versus Discrete Control

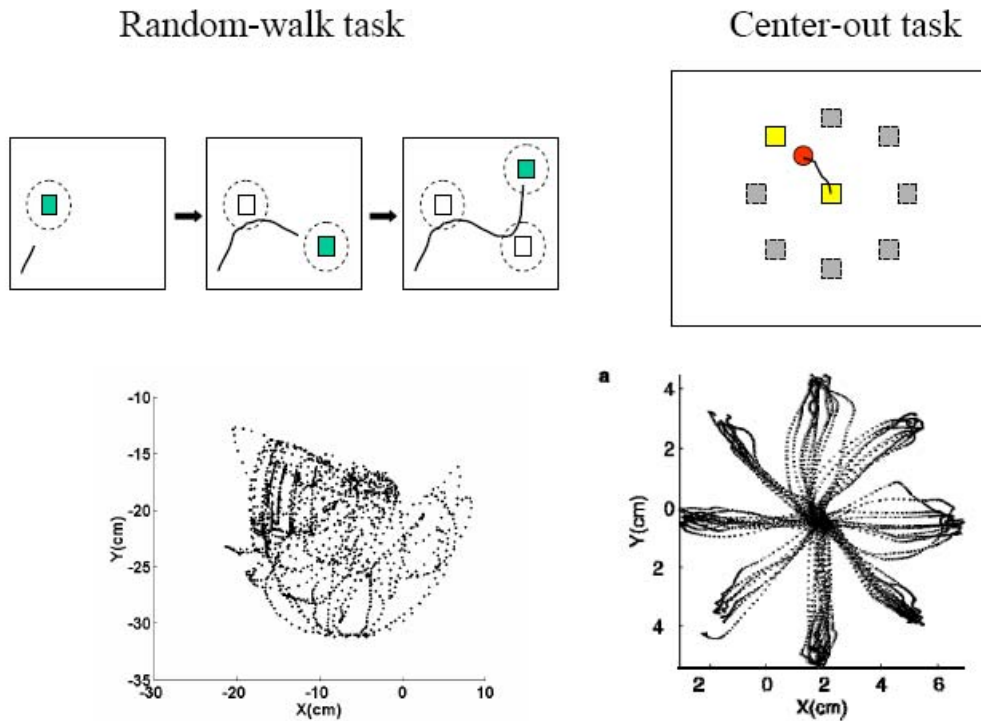


FIGURE 5 Top left and top right panels generated by myself, unpublished; bottom left and right: taken from Hatsopoulos et al. (2004). “Decoding continuous and discrete motor behavior from motor and premotor cortical ensembles.” *Journal of Neurophysiology* 92:1165-1174.

For the most part we have been implanting into the central motor strip of the monkey cortex. It has a rough topography—a leg area, an arm area, and the face area. For the most part we have implanted in the arm area of the primary motor cortex. Since I moved to Chicago I have been doing these dual array implants, and am now doing triple array implants, but I will be showing you these dual array implants where we have a second array implanted in the pre-motor cortex in the dorsal part of the pre-motor cortex, which is believed to be involved in this early planning or target selection process.

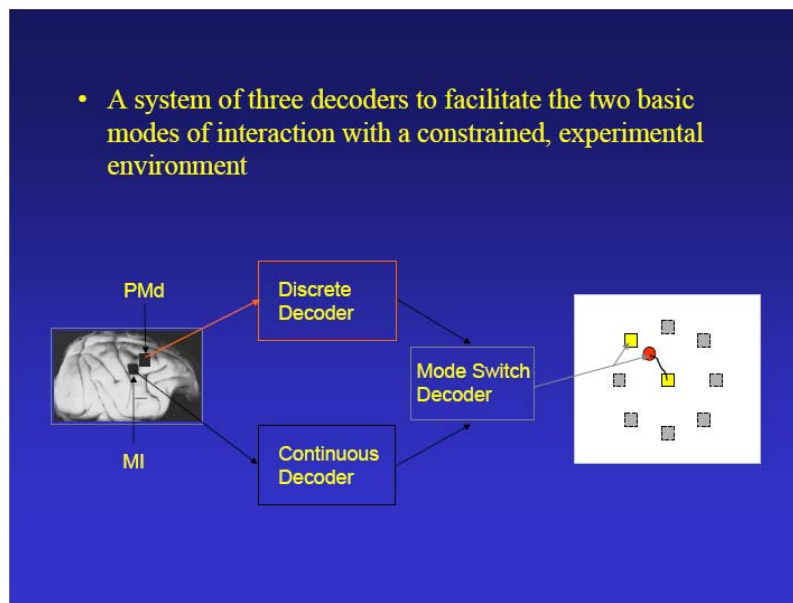


FIGURE 6 Figure generated by myself, unpublished.

The photograph on the left of Figure 6 shows you the two arrays implanted in surgery. What we are proposing here is a system of three decoders to facilitate these two different modes of interaction with the world within a constrained experimental environment. We have got these two arrays in two different cortical areas. The pre-motor cortex signals will be fed into a discrete decoder to predict which target the animal is going to go to. If the monkey incorrectly predicts, based on his brain activity, which target he is going to go to we then switch to a continuous mode based on signals in the motor cortex. We have a third decoder which allows us to switch between the two modes. We control the switch but ideally we want the switch to occur voluntarily by the patient or by the monkey, and we are going to use a different kind of signal to instantiate that switch.

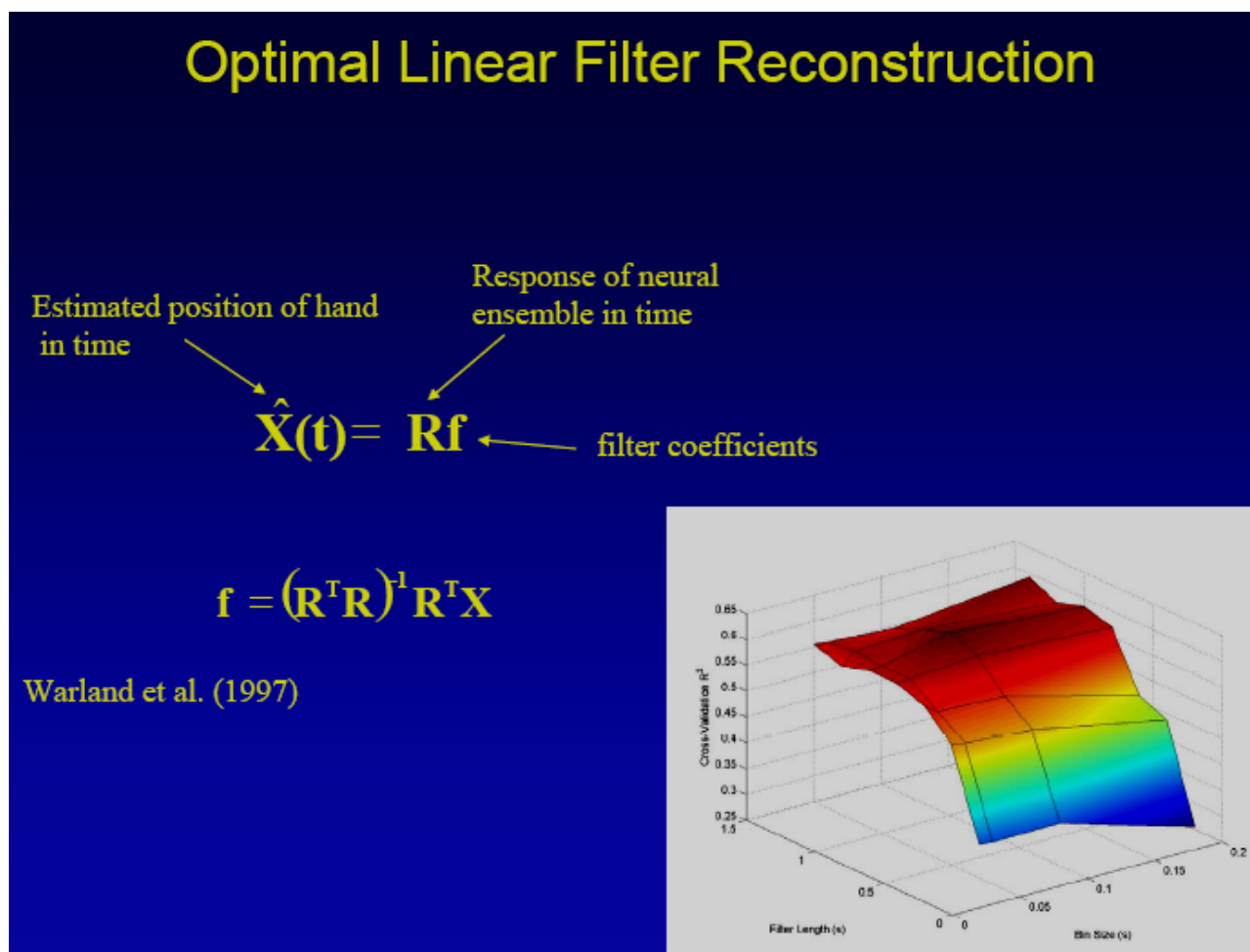


FIGURE 7 Taken from Hatsopoulos et al. (2004). “Decoding continuous and discrete motor behavior from motor and premotor cortical ensembles.” *Journal of Neurophysiology* 92:1165-1174.

Figure 7 shows some elementary statistics. Basically, we have tried a number of different approaches to predicting the position of the monkey’s hand based on the response of multiple neurons at multiple time lags. This is essentially a linear regression problem. We have tried nonlinear approaches and we have tried more sophisticated linear approaches. For the most part we can get a bit of improvement but not that much better. In fact, this very simple linear approach does remarkably well. The key to this is we are looking in this response matrix; we are looking at the neural activity not at just one instant at time but in multiple time points in the past, up to about a second in the past. It turns out that neural activity way back in history can actually participate in predicting what the current position of the monkey’s hand is. Figure 8 shows the output of just one simulation. This is an off line simulation where we are predicting the position of the monkey’s hand, and the blue is our best prediction based on the neural activity.

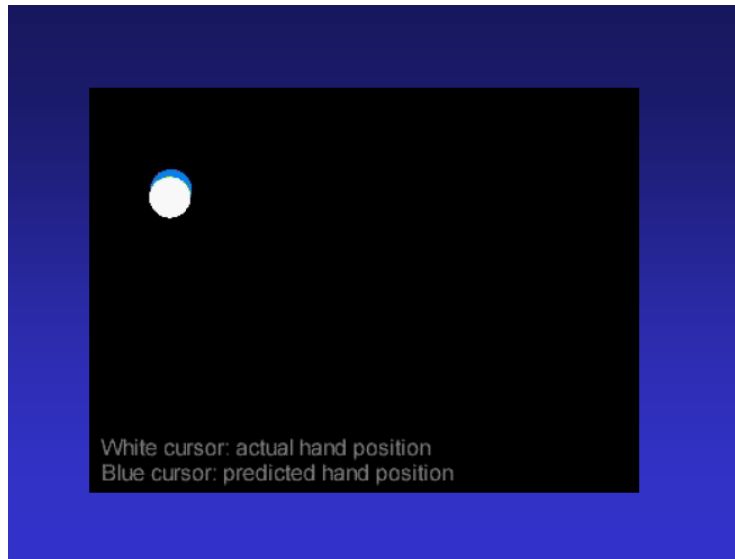


FIGURE 8 Figure generated by myself, unpublished.

What we found—this in retrospect isn't so surprising, although there was some data in the literature that suggested otherwise—was that cells from the motor cortex, primary motor cortex, M1—the red trace in Figure 9—did a much better job of predicting the actual position of the monkey's hand.

Figure 10 shows the results from a monkey playing that random walk task; he is jumping around and you can see he is oscillating; he is doing harmonic oscillation. M-1 activity predicts better than pre-motor cortex. In this little video [shown at the workshop] you can probably see how well M-1 does compared to pre-motor cortex, and pre-motor cortex is lagging behind; it is not doing such a good job. That is something we have seen repeatedly; it is a consistent finding and not very surprising.

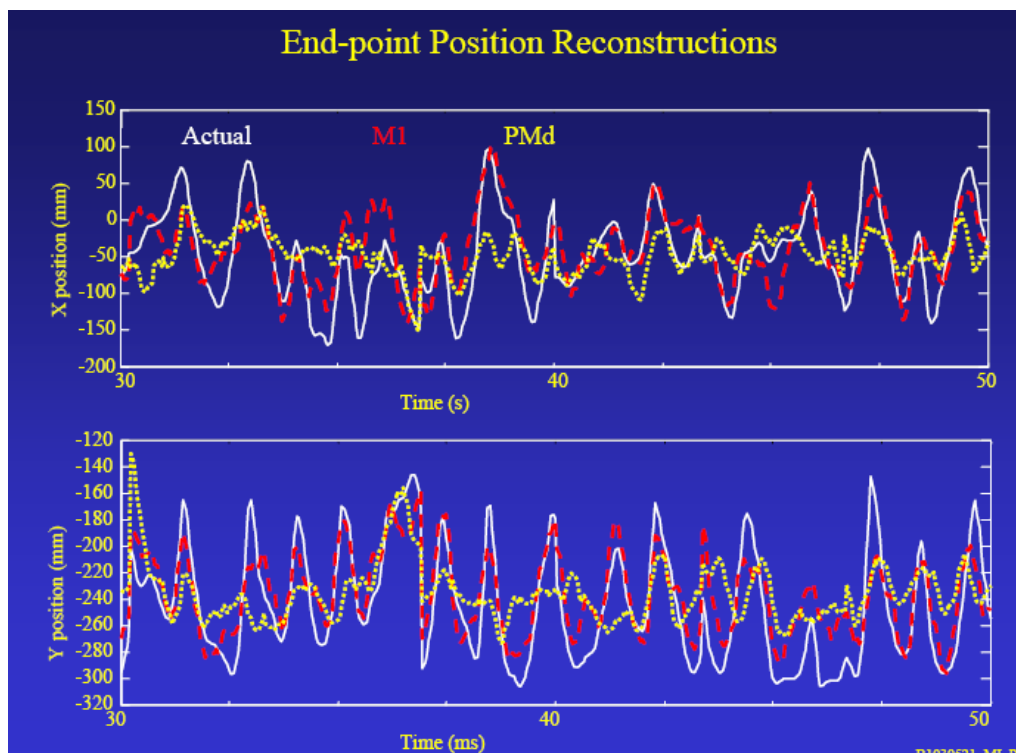


FIGURE 9 Taken from Hatsopoulos et al. (2004). "Decoding continuous and discrete motor behavior from motor and premotor cortical ensemble." *Journal of Neurophysiology* 92:1165-1174.

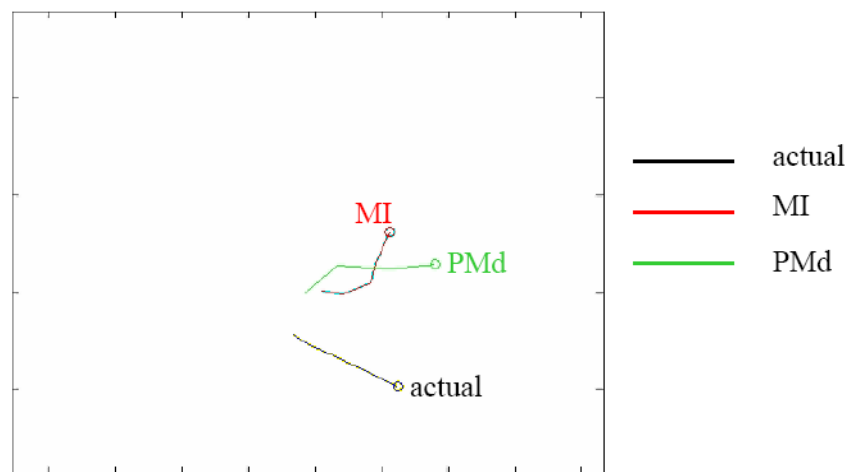


FIGURE 10 Figure generated by myself, unpublished.

QUESTION: It looked like the pre-motor signals actually had a higher frequency response than the motor signals. Is it true that some combination is better than one or the other?

DR. HATSOPOULOS: Certainly a combination is better than either one alone. We have demonstrated that as well, it is true. We were interested in, if you had a choice between one cortical area versus another, which one would you choose for this kind of continuous control, but you are absolutely right, and we have demonstrated that.

So, what we found in Figure 11 was what percentage of the variance we account for based on a selection of neurons. The x-axis shows the number of neurons used to predict the movement of the hand. So, we have different sized ensembles, ranging from one neuron to 20 in this example. As you can see, overall, M-1 is doing much better than pre-motor cortex, and this is the mean response as a function of the number of neurons.

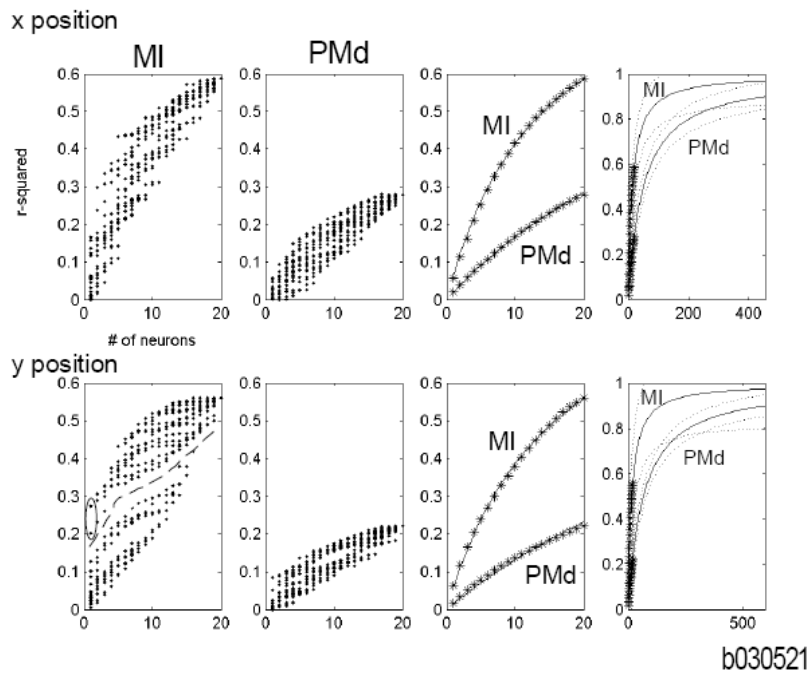


FIGURE 11 Taken from Hatsopoulos et al. (2004). “Decoding continuous and discrete motor behavior from motor and premotor cortical ensembles.” *Journal of Neurophysiology* 92:1165-1174.

There are a couple of things that are particularly interesting. One is, if you look at the lower-left graph, if you go back to one neuron—let’s say you have only one neuron to work with—you can see a big cluster of neurons that do poorly; they account for maybe 10 percent of the variance, and then you have two outliers higher up that account for a much larger percentage of the variance. I have termed these “super-neurons” because we consistently find these unique

neurons that fall outside the main cluster of neurons that do a much better job. In fact, if you look at a large ensemble, and you consider 10 neurons in your decoding you can see this pattern so you have got these big spaces. This cluster of 10 neurons always contains one of these two super-neurons. That is one interesting finding. The other interesting one is when you try to extrapolate based on 20 (or we have actually gone out to 40 neurons or 50 neurons), and try to extrapolate how many neurons would you need to get ideal performance. Our definition of ideal performance would be that 95 percent of the variance is accounted for. Now extrapolation is always fraught with problems, but if you were to do that you get the hypothetical curves in the two graphs on the right side of Figure 11. You might say, wait a minute, this is just make-believe. You are right, it is kind of make believe, although these are curves that have been proposed in the past. Let's say they were true, how many would you need? You would need about 200 neurons in the primary motor cortex.

Yet we have shown in real time applications with patients and with our monkeys that it is irrelevant whether or not we were restricting the hand. The fact is the animal and the patient can control the cursor remarkably well so this is the continuous mode. For the discrete mode we had this other task where it is basically a center-out task, and we have this instruction period where the monkey is asked to bring the cursor to the center of the target and wait anywhere from 600 to 1,500 milliseconds. This is shown schematically in Figure 12. We are going to look at this early activity to try to predict which of the eight targets he is going to move to prior to actually moving at all.

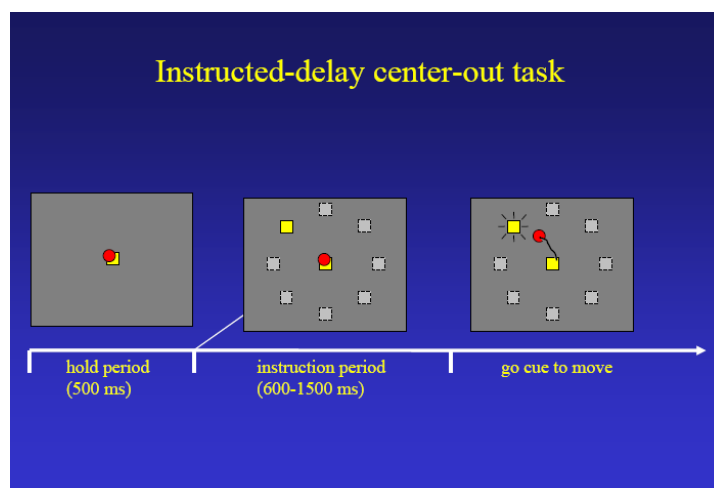


FIGURE 12 Figure generated by myself, unpublished.

Figure 13 shows the the average responses for this task—these are now peri-event histograms which are generated in the same old-fashioned way as they have been done for 50 years, you just have the monkey perform each of eight directions. Each column is now one of eight directions and each row is a different neuron. The top three are from pre-motor cortex, and the bottom three are from M-1, primary motor cortex. You just average the response over multiple trials of the same direction. What you see is this interesting increase in activity as soon as the instruction signal comes on. Zero represents the onset of the instruction signal comes on so zero represents the onset of the instruction signal. You see this increase in firing rate, then it goes down, and then about a second into the trial the movement begins. This is well before movement begins. You don't generally see this activity in the primary motor cortex, although it is not so cut and dried. Generally you see more of this early target selection activity in pre-motor cortex.

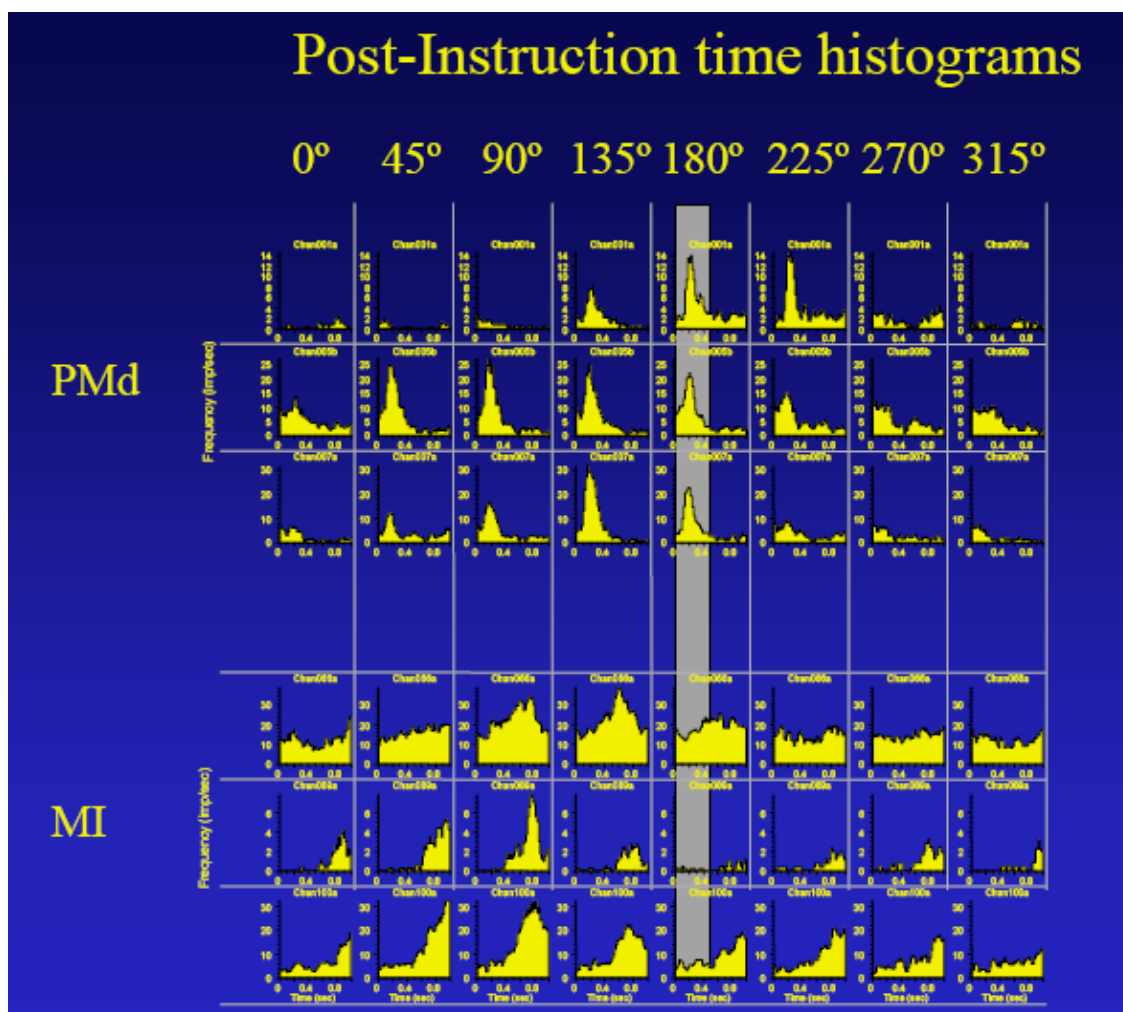


FIGURE 13 Taken from Hatsopoulos et al. (2004). “Decoding continuous and discrete motor behavior from motor and premotor cortical ensembles.” *Journal of Neurophysiology* 92:1165-1174.

When we tried to model the probability of the response of the whole ensemble—let’s say we have 20 neurons—using a maximum likelihood classifier, this R vector is a 20-dimensional vector. We model this probability conditioned on each of the targets, so we have eight of these models, and we have tried Gaussian or Poisson models for the probability. Poisson tends to do better than the Gaussian. We then maximize the likelihood of the data and use that to predict the target. Figure 14 shows examples of that, where we are now looking at our prediction starting from instruction onset. You can see chance is 12.5 percent, shown in the blue dotted line. Pre-motor cortex does considerably better than M-1 during this early period, which is not surprising based on those average responses and, in fact, we have gotten a lot better than even 42 percent. We have gotten up to about 80 to 90 percent based on activity in pre-motor cortex.

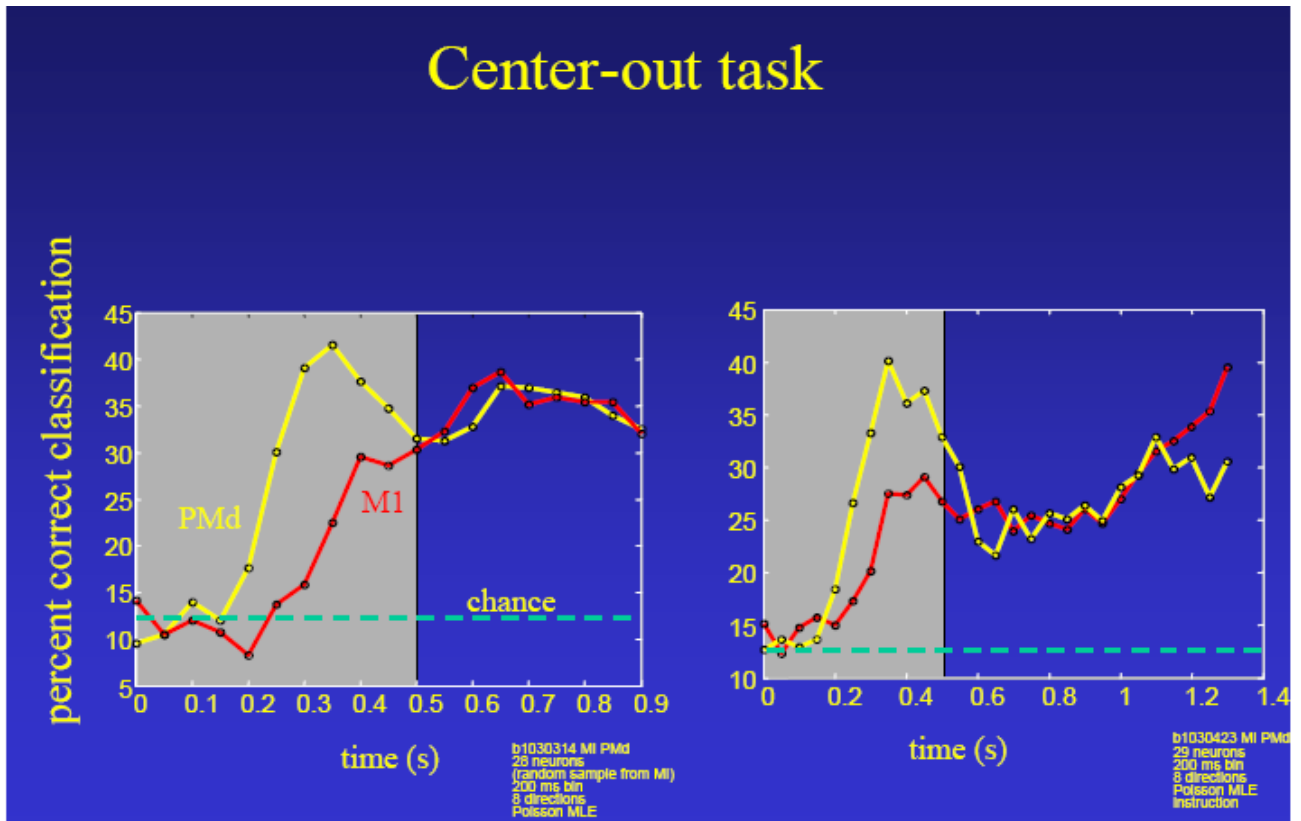


FIGURE 14 Taken from Hatsopoulos et al. (2004). “Decoding continuous and discrete motor behavior from motor and premotor cortical ensembles.” *Journal of Neurophysiology* 92:1165-1174.

Figure 15 shows plots where you examine the performance of the classifier as a function of neurons. As seen in the previous figure, pre-motor cortex does exactly the opposite of M-1.

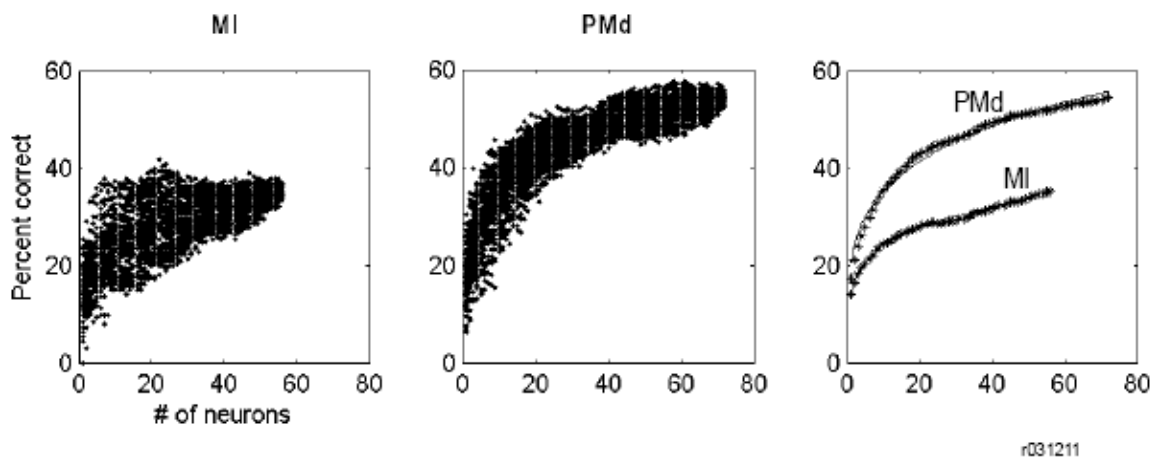


FIGURE 15 Taken from Hatsopoulos et al. (2004). “Decoding continuous and discrete motor behavior from motor and premotor cortical ensembles.” *Journal of Neurophysiology* 92:1165-1174.

I just want to thank members of my lab, in particular Jignesh Joshi and John O’Leary, and John Donohue at Cyberkinetics, as well. Thank you very much.



QUESTIONS AND ANSWERS

DR. MOODY: Perhaps one quick question and then another. In the discrete trial of the histograms, and each degree on the clock there was a great deal of variance, it seemed, in those histograms across the different directions. Could you explain why that is the case?

The second question is, can you explain or can you speak perhaps more about whether or not it really matters so much if you are tapping into what they are doing if, instead, they learn how to respond and manipulate, even subconsciously, what they are doing to get the outcome. Can you speak between the two of those, the observation and the learning?

DR. HATSOPOULOS: The first question was the variance in response over the eight directions. The whole point was, it not only modulates with that direction, it modulates during that planning phase, but it varies with direction, with target. That tells us, in fact, that it is providing target information. It is saying, this neuron likes target five and not target three, and that is the whole point. If it was common to all directions, we would have a very hard time decoding. The second point was, oh, yes, in some sense, you could in principle use any brain area. It is a form of biofeedback. Is that what you are getting at? Absolutely, that is true. In fact, there are EEG-based systems for doing this kind of thing in patients. The only thing I can say about that is that they are functional, but they take a long time to train the patient to use. They basically have to associate certain global brain signal brain patterns with certain movements or certain targets, and it takes a lot of training to get them to do that. With this device, it is invasive, that is the draw back, but the pro to this approach is we are extracting signals from the motor area, the area that was intentionally used in the normal intact system to control the arm, and this requires no training at all. Some of these videos were done during the first attempt, or almost the first attempt. There was no extensive training in the animal. You are right, in principle that is true. It is biofeedback, in some sense.

DR. KLEINFELD: I have a question on some of the noise between cells. Were they really independent units? Did they co-fluctuate at all?

DR. HATSOPOULOS: Yes.

DR. KLEINFELD: You had a very coarse scale. You had a .4 millimeter scale. Is there an issue if you get too close, that you expect to have sort of synchronous or noise sources?

DR. HATSOPOULOS: You are talking about cross talk.

DR. KLEINFELD: Common fluctuations to neural outputs so that they are no longer considered independent variables.

DR. HATSOPOULOS: In fact, that was the talk I wanted to give, but then I decided to give this, looking at neural synchrony in the motor cortex. Basically in about 10 to 15 percent of cell pairs, we find evidence of synchronization, anywhere from a millisecond to 20 milliseconds. If you plot a cross-correlation histogram, we find some widths that are very narrow, maybe three millisecond widths, and other ranging all the way up to 20 milliseconds.

Some Implications of Path-Based Sampling on the Internet

Eric D. Kolaczyk, Boston University

DR. KOLACZYK: To all the organizers, thank you for the invitation. I've enjoyed very much what I have seen so far and hope I can follow well here in the next half hour.

I have heard sampling and measurement talked about a few times throughout the day and, as Mark said, we now have a session trying to delve a little deeper into some of those issues. I want to look at the issue of sampling and its implications on certain inferences. What I am going to talk about are two projects I've done with collaborators on Internet network modeling, which both end up leveraging classical sorts of problems from the statistical sampling literature going back, say, to the middle of the 1900s in that sense. Given the context that we are in, they involve new methodology and new perspectives. Basically, working with the network is breeding a need for new thinking in this area.

The two statistical problems I will discuss both involve path-based sampling on the Internet: what we call `Traceroute` Internet species and Network kriging. There are lots and lots of ways you can think about statistical sampling on a network. In some sense you have sampling of links, sampling of subnetworks, and sampling of paths. Our two projects are both in the context of sampling paths. In both cases, the path-based sampling design necessitates innovative solutions to new versions of classical statistical problems, and the underlying network structure plays an important role. The Internet species problem tries to perform inference of the network and the network kriging project performs inference for measurements on a network. Species problem is a classic one. Kriging is a fancy name from the spatial literature for what is essentially linear prediction.

I am going to proceed first with the Internet species problem and then I am going to switch to the network kriging problem. I figure I have about a half hour long talk here, so what should I do, cut down one of these hour long talks? No, of course not, take a quarter of each and squeeze them into a half hour. That is what I am going to do. This is going to be a rather high level view of each of them. What I am hoping to do is get a point across messages.

The Internet species problem is how to infer global network topology characteristics from a sample subnetwork. This is work I did jointly with Fabien Viger, Luca Dall'Asta, Alain Barrat, and Cun-Hui Zhang. Our measurements are what I am going to call traceroute-like paths between source and target nodes, and in a minute I will show you a slide talking a little more about that. The approach here is to characterize the estimation problem of certain characteristics in the

network as a so-called species problem. A species problem refers to an old problem, the canonical example of which is a biologist sitting in the middle of a forest, watching the animals going by. It is all well and good to mark which animals, but the biologist is really interested when they see new species. They might see a tiger, a bear, another tiger, another tiger, or an ostrich. There are three species there that I mentioned. The question is, how many species are there, total? It turns out that some questions regarding the inference of certain network topology characteristics from particular types of measurement can be paraphrased as species problem, and we propose estimators based on subsampling principles. We have concentrated on accurate estimation of a so-called easy species, which is really just the size of the network in terms of nodes. I will also try to shed a little bit of light on why some of the harder species, such as the number of edges and the degree distributions, are a little harder to get at.

Figure 1 defines the basic problem. I am going to let \mathcal{G} be a network graph. It can be directed or undirected, which really doesn't matter here. So, we can just think of a simple undirected graph. We are going to suppose that a set of measurements yield a sampled subgroup, \mathcal{G}^* . At a very general level, the question I am interested in here—which is really what is underlying a lot of the types of studies that you see in the literature—is how to take a sample of the network and infer something about the network itself. I have certain characteristics of the sample network. I am going to infer that these are representative of the network I actually saw. The examples we are going to concentrate on are, for example, N , the number of vertices in the network, M , the number of edges, and the degrees of a given vertex.

Inferring Network Topology Characteristics

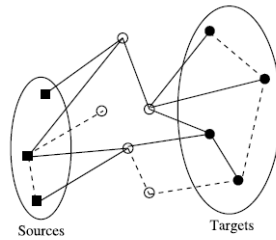
- Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network graph.
- Suppose a set of measurements yields a sampled subgraph $\mathcal{G}^* \subset \mathcal{G}$.
- **Question:** To what degree can we estimate a characteristic $\eta = \eta(\mathcal{G})$ of \mathcal{G} well using just the measurements underlying \mathcal{G}^* ?

Examples: $N = |\mathcal{V}|$, $M = |\mathcal{E}|$, and degrees $d_v, \forall v \in \mathcal{V}$.

FIGURE 1

Let me tell you a little bit about Traceroute sampling, and my apologies to the computer science community: this is going to be an extremely high level sketch of it. Basically, what we have is a network, as depicted schematically in Figure 2. Of course, the Internet and the nodes are referring to sources and routers, which are your standard machines on the Internet. There is a standard experiment, and there are various versions of it, but there is a certain tool called Traceroute that I am now labeling all of it broadly as. These are Traceroute-like experiments. Basically, the idea is that you have ownership of certain machines and, from there you can send probes into the network. These are going to be sources within the overall network. The nodes within the network are characterized by their addresses, their so-called IP addresses. There are various ways that you can pull off addresses. You might generate them randomly. You might have large lists, but you have some sense of the overall address space. You make up some list of target addresses to which you want to send probes. The details of the particular Traceroute mechanism aren't really important for the moment, for what I want to talk about; the end effect that results from that is all that I am really interested in. I take a probe and I equip it with an address of a certain target. It gets sent into the network and is routed so it goes from one node and then to another and then to another. At each stage information is being accessed in that packet to find out where do you want to go and then saying, you need to go down that street. You get there and then they say you need to go left down here. The information that you get back at the source is being sent back to you from each of these saying, yes, I have received it, I am this far away and I have sent it on, at a very rough level. From that information, you can reconstruct paths taken by the probes. You can see that I have three sources in this little schematic and I have four targets. There have been all kinds of studies about the topology that is inferred, the characteristics that are inferred from these types of studies (and to some degree the dependability of the characteristics you infer), and so on.

Traceroute Sampling of the Internet



- n_S source nodes; n_T target nodes.
- 'Trace' shortest paths from each source to all targets.

NAS Wkshp, 09/26/05 - p. 5/1

FIGURE 2

Basically, this is enough of a paradigm to suffice for what I am going to talk about. Let me recap again and put more detail to the species problem, and in a moment I am going to take the Traceroute slide I just showed you and explain why these are essentially species problems.

Figure 3 defines the species problem. A population is composed of C classes, and you observe N members of that population. Overall, I might observe N animals in a forest and there are something like C species. Of those N members you observe, they are going to fall into some subset of all the available species, say C^* species, which is bounded above by C . One of the standard questions is what is the value of C ? What you are really trying to find out is how many species were seen. This is a very hard problem in many contexts. It is a classic problem in statistics, as I said, but it arises in biology in the example I gave you.

The Species Problem

A population is composed of C classes.
You observe n members of the population,
representing $C^* \leq C$ classes.

Question: What is the value of C ?
i.e., How many species did you not see?

- Classical problem in statistics.
- Arises in biology, numismatics, linguistics, etc.
- Problem ranges from challenging to ill-posed.

FIGURE 3

The same problem arises in numismatics, as in when an archeologist digs up a number of ancient troves of coins and, from comparing the different coins that you have and the mint dates, you would like to know how many coins were minted in a certain period of time. If you are talking thousands of years back, this is the data you have. You don't have much more than that unless you have records that you can also access.

Linguistics presents another one instance of the species problem. There is a famous paper by Brad Ephron and Ron Thisted in the early 1970s called, "How Many Words Did Shakespeare Know." There are lots of versions of this in the linguistics literature. It is all getting at the same type of idea. What is the effective vocabulary of an author? One of the cute games they can play is going into the words of Shakespeare and once in a while you see that there is a new sonnet that has been discovered, and it is attributed to Shakespeare. The linguists do various types of quantitative analyses attempting to compare word frequencies and such. A variant on that idea is to say, okay, there are some new words that were found in new sonnets that he could have known. He is a pretty bright guy but, in some sense, how many words might he have known for the purpose of writing sonnets? This problem is all over the place. Some versions are challenging but certainly do-able, others slightly ill posed in a way that can be made formal.

Regarding `Traceroute` species, I would argue that N , M , and the degrees are all essentially species under the `Traceroute`-like standpoint. Let me take the idea of the number

of vertices. You have a source and you have a target and, for each of those, you are getting what I am characterizing as a single path between them. While you are doing that, you are discovering nodes. There is one discovered, there is one discovered, there is one and there is one. The sources, of course, are known, and the targets are already known as well. This particular target has been seen once, going there, it has been twice going there, it has been seen three times and four times. It is a species, and you saw four members of that species. You can do the same thing with edges, and you can actually do the same thing with the degree of the node. Each of these is getting successively harder, so we are going to go after the issue of estimating the number of vertices. This is, in some sense, a toy problem. How many words did Shakespeare know, how big is the Internet? There are an enormous amount of issues under that if you really wanted to look at that question. There are other measurement devices that you would most likely want to bring to bear, etc. The point is that a number of these issues of inferring graph characteristics can be phrased in this paradigm, and this is the one that we are using to bring to bear some focus on the problem.

You could think about doing parametric inference. We have an argument, which I haven't attempted to sketch, that shows it would be quite hard given the current levels of knowledge. If you think about it, what you need to be able to do is have a model, a parametric model, for the particularly species that you don't see. The problem is that there is typically an arbitrary number of species that you don't see in an arbitrarily low frequency. Without some kind of control on those two quantities, how many you haven't seen and in what frequencies, there is very little you can say there, and that is what leads to the sort of ill-posedness and the worst conditions. Nonparametric techniques, in the sense of not being reducible to just a handful of parameters, have typically had more success in this area. So, that is what we choose to pursue here. As I said, pathway sampling makes this problem non-standard.

‘Leave-One-Out’ Estimator

Idea: Information on unseen nodes gained through rate of return per target node.

Assumptions: Low marginal rate of return from any single target node; simple random sampling of targets.

Formal argument leads to

$$\hat{N}_{L1O} \approx (n_S + n_T) + \frac{N^* - (n_S + n_T)}{1 - w^*},$$

where w^* is the fraction of target nodes not discovered by traces to any other target.

NAS Wkshp. 09/26/05 - p. 8/1

FIGURE 4

Figure 4 shows the essence of one of the estimators. Of all of the estimators we have looked at, we have essentially designed subsampling type estimators. You have a sample, now let's sample from that sample in certain randomly chosen ways, so that you can utilize the information in an intelligent fashion to give you more information on what you haven't seen. One version that we have looked at is the so-called leave-one-out estimator. This is the strategy that underlies cross validation. It is based on the idea that the information on unseen nodes can be gained through the rate of return per target node. One of the assumptions here—I won't show you the derivation—is that there is a low marginal rate of return from any single target node.

There is a paper by a couple of people in the audience here—John Byers and Mark Crovella, colleagues at Boston University—which, if I remember correctly, was called “The Marginal Utility of Trace Route-like Sampling.” One of the figures they have in that is taking a look at, for each target, if I have already done `Traceroute` to all of these targets from these sources, how many new nodes do I discover. If I recall correctly, the number was roughly on the order of a rate of 3-ish or so, per target. We have done simulation studies where we have found similar things. So the amount of bang for your buck of a target is not that much. What that means is if you eliminate any one of these from the study, you look at which ones were not discovered, those are unseen species.

Taking these out is giving you insight on the rate at which you have not seen certain species. We do that for every one of these, and we do a type of averaging over those. What we end up with is essentially an estimator of the form shown in Figure 4, where $n_S + n_T$ is the number of sources plus the number of targets. I want an estimate of how many nodes are there in the network. Number of sources plus targets, of course, plus the numerator in this estimator is the number of discovered nodes, and star being the number that we actually saw, minus sources and targets, divided by an adjustment factor. This adjustment, $1 - w^*$, is the fraction of the target nodes not discovered by traces to any of the other. So, they are essentially the fraction that, if you left them out, would have been unseen species.

Numerical Results

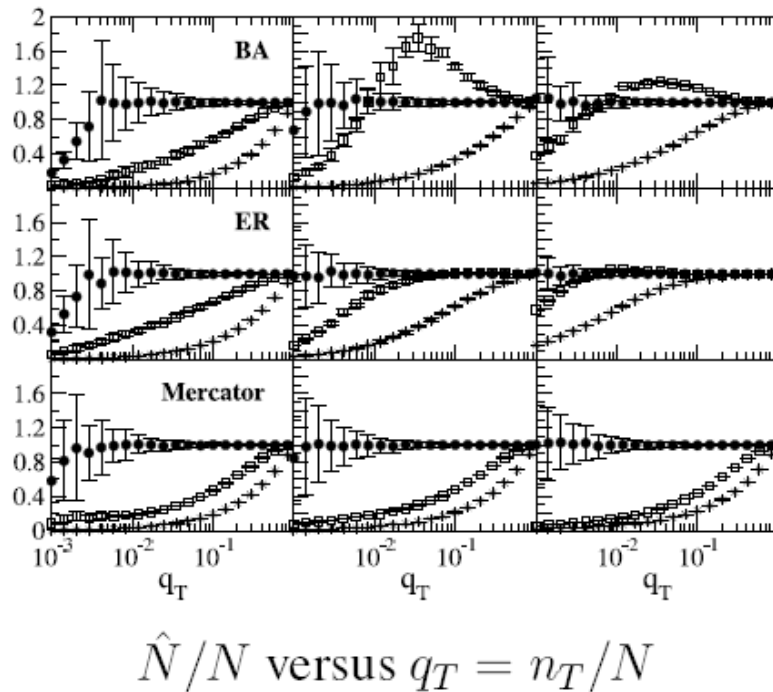


FIGURE 5

Figure 5 is an illustration of how this type of thing works. Around 1999 we took a couple of simulated networks, and then we took a measured topology from the Mercator, so this is rather

dated. The row labeled ER is taken from an Erdős-Renyi random graph model, and the row labeled BA is taken from a Barabasi-Albert random graph model. The stars here show the actual proportion of nodes discovered. So, you just do `Traceroute` and count how many do you see. Of course, that is an underestimate, but that is a fine place to take as a grounding point. We also had a re-sampling estimate where we said, in some sense, the quantity that you are sampling are actually paths. What if you took a source and, from each source you randomly sampled some of these paths? If there is a similarity between the networks at different sizes, then you might expect that that would give you some reasonable results. It turns out it gives you some improvement in that it lies above the actual sample values, but not nearly what you would like.

What works very well is the leave-one-out estimator, and these are the black circles in Figure 5. What you are looking for in this ratio is a ratio of 1. What you can see is that with low sampling rates, which is a fraction of targets compared to all the vertices, you very quickly capture what the size is. The reason is that within the species problem there are gradations of difficulty, and in between version of difficulty is a so-called coverage estimation problem. Coverage estimation is not quite actually trying to find out how many species there are; it is trying to find out what proportion of the overall population was represented by those you saw.

The easiest problem is not a species problem at all; it is where you know how many species there are, and you are estimating the relative frequency of those species. That is an easy one. The species problem is hard, and coverage estimation is in between the two. As it turns out, to a reasonable degree you can argue that estimation of N is actually closer to a coverage problem. The estimation we have shown is essentially a coverage-based estimate and these, in the literature, when you can use them, have tended to be better.

I am going to switch gears here. I had two problems for you. This past one was a `Traceroute` problem, and we tried to exploit sampling-type principles for inference of something regarding the network itself. What I want to do now is take a look using sampling principles to estimate traffic on a network, so let me give you some background. This is going to be a high-level summary before I go into some details.

The problem is global monitoring of network path traffic. Now we are talking about the same network, the Internet or some sub-network like a backbone or such thing, and we are talking about traffic in the order of, say, packets from origins to destinations.

Every time you send an email, or every time you click on a browser, information is being exchanged, and that can be measured at the level of routers. We are essentially talking about that type of traffic-volume information. The idea is that, if you had a network and you wanted to monitor, in a global sense, the overall network, between all path origins and destinations, this is

infeasible on any reasonable network sites, because it is going to scale like $O(N^2)$, where N is the number of vertices. You have to go down and start sampling and very quickly it becomes a sampling problem. Measurements, origin, destination, packet delay, for example, you can also talk about packet loss. I will use the word delay throughout here, and some of the data I show later is also delay data.

The approach that we use takes a different point from what had usually been done in the literature to date. Some work came out of Berkeley and there is some other work, I believe, that came out earlier from Israel in which they are taking linear algebraic views of the problem. In a moment I am going to show you a linear algebraic formulation, and you will see that it makes perfect sense to have done that. What that work said was you don't need to measure all the paths in a network; you only need to measure a certain lower bound. If you measure just that many then you would actually be able to recover all the information in the network but, if you lose even one of those paths, you lose the ability and you enter into the realm of statistics. You have to; it becomes a prediction problem. We have characterized the problem as a linear prediction problem, and we saw how the network demography literature has done so well with a sexy name.

Instead of calling this "network linear prediction," let's call it network kriging, because kriging is the word from the spatial literature for the same problem, and it is a much neater name. The problem is to get a global monitoring of network path traffic based on measurements of origin-to-destination packet delays on a limited number of network paths. We characterize this as a linear prediction (i.e., kriging) problem and exploit the sparsity of routing matrices. The results are that we have an accurate prediction of whole-network delay summaries based on only a fraction of the number of paths previously used.

Illustration: Abilene



FIGURE 6

Figure 6 shows an example, taken from basic backbone that all the university systems are pretty much on. The idea is that if we are looking at delay between New York and Sunnyvale, at a certain level that is essentially the delay going from Sunnyvale to Denver, if it is routed that way, and Denver to Kansas City, and Kansas City up to Indianapolis, and up to Chicago, and up to New York. So essentially you have an additive sort of structure. Secondly, if I have a sense of the path delay just how long from here to there, and I also measure something from Seattle and I measure something from, say, Sunnyvale, they are all sharing this part of the network. They all have a certain amount of redundancy of information in those same measurements. In principle, the idea behind sampling is that we shouldn't have to measure all of those. We ought to be able to get more bang for our buck if we tile the network, in a rough sense, with appropriately chosen paths.

So, a little bit of notation here. As shown in Figure 7, I will let \mathcal{G} be a representation of a network path, \mathcal{P} is a set of all paths, the numeracy here, and we are going to be confining ourselves to situations in which we have path measurements, y , which we will just call delays, associated with link measurements x , through a linear relationship like this, where the matrix G is the so-called routing matrix. I am taking a static routing matrix, assuming that there are no changes and there are no differences in terms of choice of path like load balancing and what not. $G_{i,j}$ here is saying that the matrix is a one if path i traverses link j , and it is a zero otherwise. So,

you can put packet loss modeling in this framework, you can put delay modeling in this framework at an appropriate level of granularity. In Figure 8 we squeeze what we have done into one slide. The goal is to predict global linear summaries. There are a lot of ways that you can do that. We are going to consider the path information that you have.

Some Notation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a strongly connected digraph, and let \mathcal{P} be the set of all paths on \mathcal{G} .

Define $N = |\mathcal{V}|$, $M = |\mathcal{E}|$, and $n_p = |\mathcal{P}|$.

Consider $y \in \mathbf{R}^{n_p}$ and $x \in \mathbf{R}^M$, related via $y = Gx$, where

$$G_{i,j} = \begin{cases} 1 & \text{if path } i \text{ traverses link } j \\ 0 & \text{otherwise} \end{cases} .$$

Examples: Delay, packet loss, etc.

NAS Wkshp, 09/26/05 – p. 12/1

FIGURE 7

On the Abilene, there are on the order of 120 different origin-destination pairs. You can picture watching this in time, so you would have 120-dimensional multivariate time series that you would like to have information on. One way to reduce that is to take summaries of it, and the simplest class to look at is just linear summaries. You could take a look at the average delay over the network. You could take a look at differences of delays on subnetworks, and that gets into interconnections with things like multi-homing and various other applications. You want to be able to monitor or predict this type of summary based on only measuring k paths. We are going to have some sort of evaluation of predictors, and we are going to choose mean square prediction error, which is essentially looking at the quantity we want to get, which is random, minus some function of the samples that we take, squared, expected value.

Kriging for Global Linear Summaries

- **Goal:** Predict *global* linear summaries $l^T y$ using measurements on just k paths i.e.,
 $y_s = (y_{i_1}, \dots, y_{i_k})$.
- Evaluate predictors $p(y_s)$ using MSPE
i.e., $\text{MSPE}(p(y_s)) = E[(l^T y - p(y_s))^2]$.
- Standard linear prediction ('kriging') arguments lead to
$$\hat{p}(y_s) = l_s^T y_s + l_r^T V_{rs} V_{ss}^{-1} y_s.$$
- Select the k paths to minimize (ideal) MSPE; equivalent to use of algebraic subset selection algorithms.

NAS Wkshp, 09/26/05 – p. 13/1

FIGURE 8

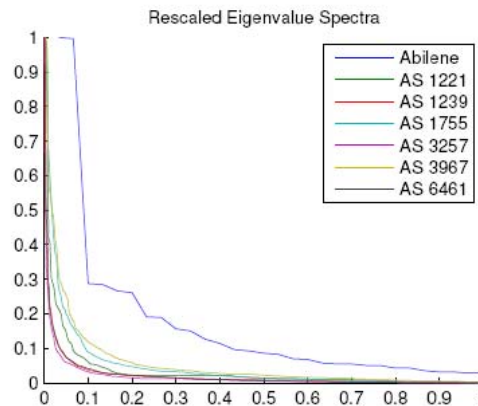
There is a series of arguments that are essentially standard up to some issues of rank, full rank, not full-rank issues. There are basically standard linear prediction arguments that can lead you up to an analogue of the so-called best linear unbiased predictor, except that this is not unbiased because of the rank issue I mentioned. You are essentially going after something in that same genre of arguments. It is saying take the sub-elements of this vector L , which might just be weights one on N , for all practical purposes—that is, taking the average of the whole network—take the average of the measurements you saw. For the measurements you didn't see, which is the weights are obtained in this vector, take the measurements you took, standardize them by their co-variance matrix, and then this is the cross co-variance matrix between the measurements r that remained in the network and the measurements s that you sampled. You then take that combination. There is a derivation behind it, but there is not time to actually be able to show this. Those of you who have seen these types of things should recognize the essence of this formula, and this is sort of a standard argument.

There is one thing that I didn't mention, which is how do you select the k paths that you want to use. Previously in the literature, the result that I mentioned earlier basically took the relation in Figure 8 and it said we are essentially looking at the row space of G . If we could find a set of rows of G that spanned it—because this matrix G is very tall and very thin, so you don't

need all those rows.

The idea in that set of papers is, if you take just enough rows to span the row space of G , then you can recover x and therefore you can cover all the rest of y that you didn't measure. They use a so-called subset selection algorithm from the linear algebraic literature to do this. If you cast things within this framework, you can actually show that algebraic subset selection algorithm is equivalent, in this framework, to minimizing the mean square prediction error over k , over choice of k . There is actually a statistical angle on what is being done when one tries to do that so there is still an issue of, should I really believe that this is actually going to work. There is an argument here that has much more of a story to it, but I am only going to show one slide, Figure 9.

Why Should it Work: Path Redundancy



Path Redundancy \iff Sparse G

NAS Wkshp, 09/26/05 - p. 14/1

FIGURE 9

Basically, this matrix G is, as I said, very tall, very thin: there are a lot of these paths that will be common. If a lot of the paths are common, then it means that a lot of the rows of G are very similar, which means that, for the most part, the dimensionality of G should be quite low. If you have a lot of path redundancy, you can then picture a lot of the rows being similar and, therefore, G should have a lower dimension. What I have here, standardized in the number of

rows up to the number of rows needed for full rank—that is, 1—and I have standardized the values here, I have item spectra of G^T , so essentially square of the singular values. What you can see in Figure 9—this is for the Abilene network I showed you plus six of the so-called rocket fuel networks that are measured by a rocket fuel project, which is considered a reasonably good benchmark for what these types of real networks would look like. You can see that all of these are saying there is a huge decay in these eigenvalues very fast, after which it sort of tails out very quickly. The dimension of these matrices is very small, uniformly, across these different sorts of networks, so we would hope we would be able to do well here.

Figure 10 presents a characterization of the theoretical performance, and all I am going to do is call your attention to the bottom here. What we were after was to bound the mean-score prediction error in a way that was meaningful in terms of the relevant quantities here. This interior part is all that is important for the story; it is the average link level delays. That, of course, has to play a role, that flights of certain lengths have a lot of information, less if not. The vector of combinations that you are using will have some role, but here what we have in equation 2 are two pieces. The term λ_{k+1} is the $k+1^{\text{st}}$ largest item value, and $f(k)$ is a function that is determining the rate, as a function of the number of paths you use, at which you can approximate the routing matrix by a submatrix. How well you can do in your prediction is going to be a function of these two things. In practice this thing ends up being much smaller than that, so the rate at which you can effectively do subset selection tends to dominate this overall performance.

Theoretical Performance Characterization

Theorem 1 Denoting the i^{th} row of G as $G_{(i)}$, let $p_i = \|G_{(i)}\|_2^2 / \|G\|_F^2$, where $\|\cdot\|_2$ and $\|\cdot\|_F$ are the Euclidean and Frobenius matrix norms, respectively. Let \tilde{G}_s be a rescaled version of G_s , under the operation $G_{(i)} \rightarrow G_{(i)} / \sqrt{k p_i}$ for each of the k rows in G_s . Then if

$$\left\| G^T G - \tilde{G}_s^T \tilde{G}_s \right\|_F \leq f(k) \|G\|_F^2, \quad (1)$$

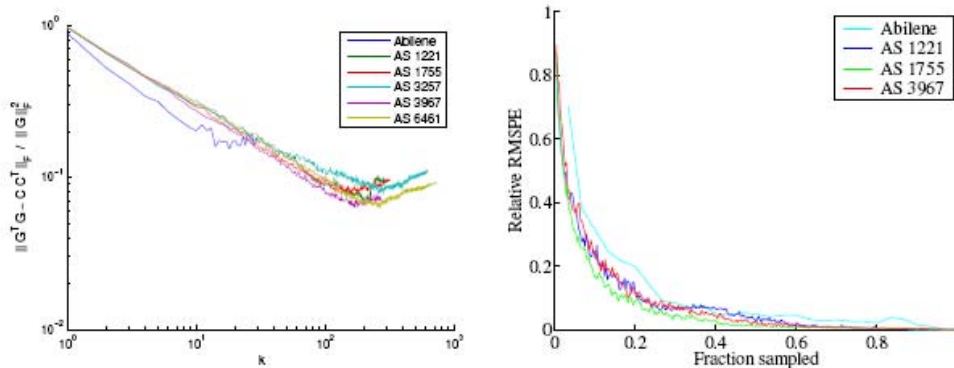
for some $f(k)$, the MSPE can be bounded as

$$\text{MSPE}(\hat{p}(y_s)) \leq (\|\mu\|_2^2 + 1) (\lambda_{k+1} + 2f(k) \|G\|_F^2) \|l\|_2^2. \quad (2)$$

FIGURE 10

Let me show you in numbers, for the same network I showed you before. The graph in Figure 11 is essentially this function $f(k)$, showing the rate at which we can approximate. What you see here is a slope roughly on the order of $k^{-1/2}$. This is almost the same for all these networks. There is a level of consistency here that all of those networks can essentially have a similar sort of rate. In the same networks, we did a simulation study where we simulated a very simple model of delay. We wanted to predict the average delay over the whole network, and we look at the relative route mean square as the proportion of the fraction sampled, up to full rank, where 1 means that you could actually recover all the paths. What you can find is that you get on the order of a 10 percent error with only 20 to 25 percent of the paths you originally needed.

Numerical Illustration



Left: Empirical calculations showing $f(k) \sim k^{-1/2}$.

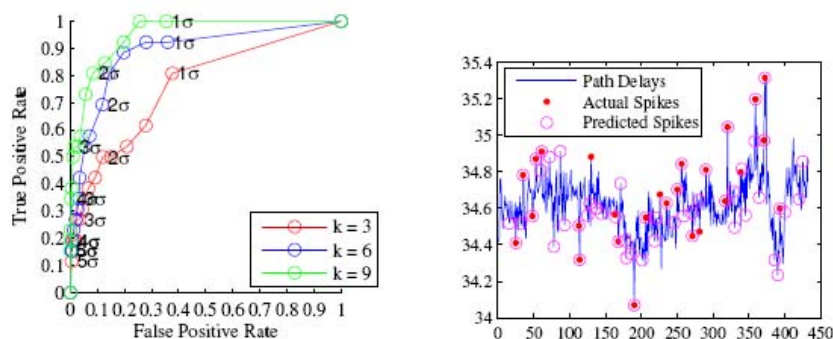
Right: Computed MSPE for simple packet-loss model.

FIGURE 11

Monitoring the average is the low-level test. Picture that you have a network of engineers. They want something that perhaps you put on the weather map for Abilene. Here is another something that you could put on—you watch it go up and down. High-level tasks would be something like anomaly detection. Can I infer high-level characteristics? What we have done in Figure 12 is take the average—blue is the average path delay on Abilene—for a period of three days. We generated the delays from that. There is a bunch of data processing that is involved in that, but we backed into the delays.

We then declared a criterion that, if the traffic at a certain point in time in average had spiked above three standard deviations from the previous hour, we declared an anomaly. Clearly you can substitute your own definitions of anomalies here. Then we asked what if we actually do our prediction based on just some small number of paths, such as 3, 6, or 9? In Abilene, if you do 30 path measurements in an independent fashion, then you can recover everything. We said fine, if we apply the same type of criteria with a certain threshold, how well can we do at recovering the anomalies in the original data?

Illustration: Anomaly Detection



- Potential 'anomaly' defined in true network average if value at time t exceeds 3 standard deviations of the past 6 epochs.
- ROC curve considers effect of both choice of k and choice of σ used in flagging anomalies in predicted time series.
- True Pos Rate = 0.81 and False Pos Rate = 0.08

FIGURE 12

On the left in Figure 12 is an ROC curve showing, as a function of the number of paths sampled—3, 6, and 9—and as a function of the threshold value we use for declaring anomalies, how well do we do. What you can see is that, based on even 9 paths, which is only about a third, you can do extremely well, roughly on the order of a true positive rate of a little over 80 percent. We can argue here that this is effective both for low-level and high-level sort of tasks.

So, let me close up here. Network sampling is something that, when I first started

working with my C.S. colleagues, and more recently with some biology colleagues, although I am definitely a neophyte in that area, I approach it at a certain level, and namely, here is the data, here is the network, blah, blah, blah, blah, and where can I plug myself in. As we delved into these, we started finding that we had to keep going back through the steps from which the things came. I think network sampling deserves much more attention than it has had to date, and I should follow up quickly—because I know that a number of people here in the audience are working this—that it is starting to happen. In the last few years, there have been, particularly in the area of heavy-tailed degree distributions, a number of people looking at issues of sampling and what is the effect of sampling. Could my sampling have made me infer certain distributions for degrees and what not? There is old literature and some more recent literature on social networks, and if I sample a network in a certain way, what are the inferences that I can draw? It is out there, but I think in the same way that our intro courses, for example, quickly brush aside sampling issues, so that we can get on to linear regression by the end of the semester, because we have to do that for the students, sampling experimental design gets shuffled under the table from the very day one of our training in statistics, and I think that percolates out then and comes back to bite us when we are talking about doing science questions. What is interesting, and should be a draw for statisticians, is that the problems are non-standard and non-trivial. Our work shows that network topology here can be an influence both on inference of network characteristics and on the inference of network and data structures.

QUESTIONS AND ANSWERS

DR. DOYLE: Have you thought at all about a slightly more Bayesian approach, which is, we know a lot about what routers can and can't do? You know, you have got the Cisco catalogue. Could you use that in any way to knock down that number quite a bit? I mean, obviously, you get a lot of redundancy, and that is a big win, but if you knew a lot about like core routers and their typical sizes and so on, would that be of any use in this?

DR. KOLACZYK: In which one?

DR. DOYLE: Either, but I was particularly thinking of the second problem.

DR. KOLACZYK: I think there is plenty of room to build on what we had there. We really have done the simplest thing you would think to do and it wasn't easy, but it was the place to start, and that is linear prediction, plain and simple. From there, there are sorts of stuff in the creating literature, for example, and the prediction problem, in which one of the first things you

might do is certainly build a Bayesian site onto that.

I am not sure where you build the information. It could depend a lot on what sort of information you had, and how well you could functionally map that to the traffic patterns themselves in a hypothesized fashion, to build your priors. Certainly the mechanism, what we put there is really a foundation for going in those directions.

REFERENCES

Barford, Paul, Azer Bestavros, John Byers, and Mark Crovella. 2001. "On the Marginal Utility of Network Topology Measurements." Proceedings of the ACM SIGCOMM Internet Measurement Workshop (IMW) San Francisco: CA. Pp. 5-17

Chua, D.B., E.D. Kolaczyk, and M. Crovella. 2006. "Network kriging." IEEE Journal on Selected Areas in Communications, Special Issue on Sampling the Internet (to appear).

Viger, F., A. Barrat, L. Dall'Asta, C.H. Zhang, and E.D. Kolaczyk. 2005. "Network Inference from TraceRoute Measurements: Internet Topology 'Species'". Submitted. Available online at <http://arxiv.org/abs/cs.NI/0510007/>.

Network Data and Models

Martina Morris, University of Washington

DR. MORRIS: As they say, I realize I am the only thing that stands between you and dinner. It is the end of the day. It is not an enviable spot, and all of these have become longer and longer talks, so I am going to see if I can work through this a little more quickly.

I am closer to the biologists since I put the people up front. We have a large group of people working on network sampling and network modeling—Steve Goodreau, Mark Handcock and myself at the University of Washington; Dave Hunter and Jim Moody, who are both here; Rich Rothenberg, who is an M.D., Ph.D.; Tom Snijders who some of you know, is a network statistician; Phillipa Pattison and Garry Robbins from Melbourne have also done a lot of work on networks over the year, and then a gaggle of grad students that come from lots of different disciplines as well.

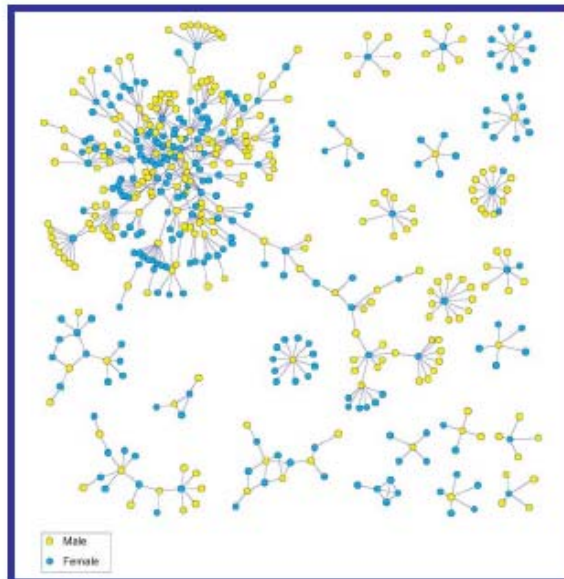
We are funded from NIH and we are primarily interested in how networks channel the spread of disease, so Figure 1 shows an example of a real network that comes from Colorado Springs, a visualization that Jim did. You can see that the network has a giant component, which is not surprising. It has this dendritic effect, which is what you tend to get with disassortative mixing. I think that is something that Mark Newman pointed out. After a while you get to look at these things and you can immediately pick that up. This is kind of a boy-girl disassortative mixing, and it generates these long loosely connected webs. This is also a fairly high-risk group of individuals, and it was sampled to be exactly that. John Potter thinks that he got about 85 percent of the high-risk folks in Colorado Springs. Every now and then you see a configuration like a ring of nodes connected to a central node, which represents a prostitute and a bunch of clients. Of course not all networks look like that.

Network Data and Models

Network Modeling Project:

Steve Goodreau, Mark Handcock, Martina Morris (UW),
Dave Hunter (PSU), Jim Moody (OSU),
Rich Rothenberg (Emory), Tom Snijders (Groningen)
Phillipa Pattison, Garry Robins (Melbourne),
Grad students: Krista Gile, Deven Hamilton, Dave Schruth

Funding from NICHD and NIDA



Colorado Springs "Project 90", John Potterat, PI
Visual by Jim Moody, Network Modeling Project

FIGURE 1

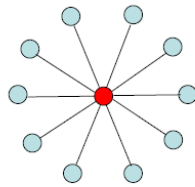
Just a little bit on motivation. Our application is infectious disease transmission, and in particular HIV, recognizing that the real mechanism of transmission for HIV is partnership networks. What we are interested in, in a general sense, is what creates the network connectivity, particularly in sparse networks. As most of you know, HIV has enormously different prevalence around the world. As low as it is here, which is certainly less than 1 percent, compared to close to 40 percent in places like Botswana. So, there is a really interesting question there about what kind of network connectivity would generate that difference and variation in prevalence.

There are clearly properties of the transmission system that you need to think about, which include biological properties; heterogeneity, which is the distribution of attributes of the nodes, the persons, but also infectivity and susceptibility in the host population. There are multiple time scales to consider, and time scales are something that we haven't talked about much here, but I think are very important in this context. You get the natural history and evolutionary dynamics of the pathogens, but you also have some interesting stuff going on with partnerships there. In addition, there is this real challenge of data collection. That is in contrast to Traceroutes—it would be great if we could collect Traceroutes of sexual behavior. Maybe we could do that with some of those little nodes that we stuck on the monkey heads, but at

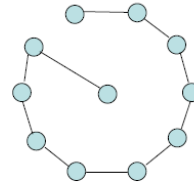
this point we can't really do that. So, for us, this means that the data are, in fact, very difficult to collect, and that is a real challenge. One of the things we are aiming for—basically, this is our lottery ticket—is to find methods that will help us define and estimate models for networks that use minimally sampled data. By minimally, I mean you can't even contact trace in this case, because contact tracing is, itself, a very problematic thing to do in the case of sexual behavior.

Connectivity in sparse networks

- High degree hubs



- Low degree linking



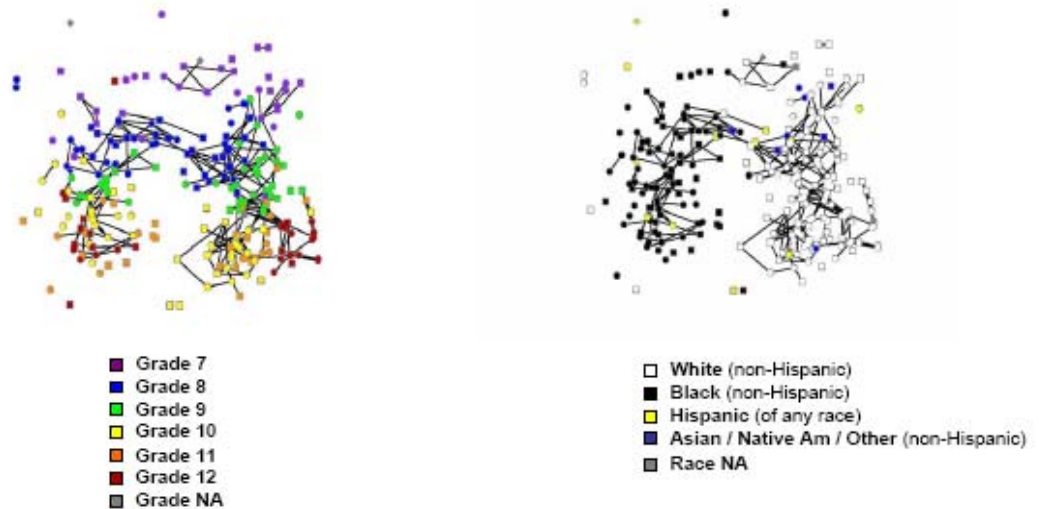
Both have mean degree = 1.9

FIGURE 2

So, how do you generate connectivity in sparse networks? Lots of people have focused on a scale of three steps: high-degree hubs, all you need is one person out there with millions of partners and your population is connected. In fact, you can generate connectivity in very-low-degree networks as well, as shown in Figure 2. Your simple square dance circle is a completely connected population where everybody only has two partners, so it is important to recognize that there are lots of other ways that you can generate connectivity, and that even if one person, for example, did act as a hub, and you figured that out somehow and you removed them, you would still have a connected population. I think it has some pretty strong implications for prevention.

Also, there is obviously clustering and heterogeneity in networks. Figure 3 shows data from a school. You can see in this case that there is a combination of clustering by grade that generates the different colors and also the very strong racial divide.

Clustering and heterogeneity



But why?

- Exogenous attributes (“birds of a feather”) or
- Endogenous process (“friend of a friend”)

FIGURE 3

You might ask yourself the question, why is this? What is the model behind this? What generates this? A number of people have hinted at this kind of stuff earlier.

Are these exogenous attributes at work—that is, birds of a feather stick together? If you are the same grade, you are the same race as me, that is why we are likely to be friends. Or is it some kind of endogenous process where, if two people share a friend, they are more likely to be friends themselves? That is a friend-of-a-friend process. It is interesting that in popular language we have both of those ideas already ensconced in a little aphorism. So, thinking about partnership dynamics and the timing and sequence and why that would matter, one of the things that we have started looking at in the field of networks and HIV is the role that concurrent partnerships play.

Partnership dynamics: timing and sequence

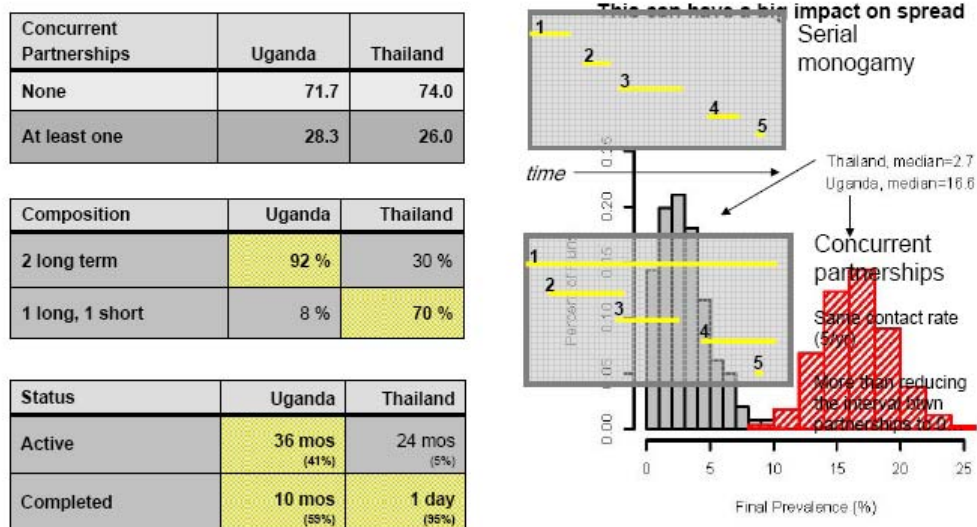


FIGURE 4

Concurrent partners are partnerships that don't obey the rule of serial monogamy. That is, each partnership does not have to end before the next one begins. You can see in Figure 4, in the long yellow horizontal line labeled "1", we have got somebody who has a partnership all the way across the time interval, and then a series of concurrent partnerships with that, including a little one night stand labeled "5" that makes three of them concurrent at some point in time. Now, this is the same number of partners as in the upper graph, so it is not about multiple partners, although you do need to have multiple partners to have concurrent partners. This is really a function of timing and sequence.

What we find in Uganda and Thailand is very interesting. Uganda at the time that we were doing this study had a prevalence of about 18 percent and Thailand had about 2 percent, which is 10 times less. In Uganda men reported about 8 partners in the lifetime on average and in Thailand it was 80. Now, 10 years later Thailand doesn't have an epidemic that comes from having 80 partners in your lifetime, so something else is obviously going on: it is not just the number of partners. We looked into this concurrency a little bit, and we found that in both cases you get concurrent partnerships. So, it is not the simple prevalence of concurrency that is the difference. The difference comes in the kinds of partnerships that are concurrent. In Uganda you tend to get two long-term partnerships that overlap, whereas in Thailand you have the typical sex partner pattern, which is one long term partnership, and then a short term on the side. The net result is that in Uganda the typical concurrent partnership is actually active on the day of the

interview—41 percent of them were, and they had been going for three years. The ones that are completed are also reasonably long, about 10 months. In Thailand you don't get anywhere near as many active on the day of the interview. They are about two months long. Ninety-five percent of them are a day long, so these concurrences happen and they are over.

That kind of thing can actually generate a very different pattern on an epidemic. The simulations that we have done of that suggest that if you take this into account you do, in fact, generate an epidemic in Uganda that has about 18 percent prevalence, whereas Thailand will typically have just about 2 percent. The approach that we take is thinking about generative models here, local rules that people use to select partners that cumulate up to global structures and a network. What we want is a generative stochastic model for this process, and that model is not going to look like you, for example, want to create clustering. A clustering coefficient, although it can be a great descriptive summary for a network, is not necessarily going to function well as a generative model for a network. It is also probably the case that when I select a partner I am not thinking I want to create the shortest path, the shortest geodesic to the east coast. That is probably also not going on. Again, it's a nice summary of a network but probably not a generative property. We want to be able to fit both exogenous and endogenous effects in these models, so that turns out to be an interesting and difficult problem. We also want this to work well for sample data. We want to be able to estimate based on observed data, and then make an inference to a complete network.

Figure 5 shows what this generative model is going to look like. We have adapted this from earlier work in the field. It is an exponential random graph model. It basically takes the probability of observing a network or a graph, a set of relationships, as a function of something that looks a little bit like an exponentiated version of a linear model, and then a normalizing constant down below that is all possible graphs of that size. This is the probability of this graph as a function of a model prediction, with this as the normalizing constant. The origins go back at least in the statistical literature of Bahadur during 1961; I talked a little bit about a multivariate binomial. Besag has adopted this for spatial statistics, and Frank and Strauss first proposed it for networks in 1986. The benefits of this approach are that it considers all possible relations jointly, and that is important because the relations here are going to be dependent, if there are any endogenous processes going on, like these friends of a friend. It is an exponential family, which means it has very well understood statistical properties, and it also turns out to be very flexible. It can represent a wide range of configurations and processes.

Generative Model: Exponential Random Graph (ERG)

Probability of observing a graph (set of relationships) \mathbf{X} :

$$P(\mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta' \mathbf{z}(\mathbf{x})\}}{c}$$

where: z = vector of network statistics
 θ = vector of model parameters
 $c = \sum_{\text{all } \mathbf{x}} \exp\{\theta' \mathbf{z}(\mathbf{x})\}$ normalizing constant

Origins: Bahadur (1961), Besag (1974), Frank & Strauss (1986)

Benefits: - Considers all possible relations jointly because of dependence
- Exponential family with well understood statistical properties
- Very flexible, can represent a wide range of configurations and processes

FIGURE 5

What does model specification mean in a setting like this? There are two things that we must choose, the set of network statistics $z(x)$ and whether or not to impose homogeneity constraints on the parameter θ . We are going to choose a set of network statistics that we think are part of the self-organizing property of this graph. A network statistic is just a configuration of dyads. Edges are a single dyad, that is the simplest network statistic, and that's used to describe the density of the graph. Others include k-stars, which are nested in the sense that a 4-star contains quite a few 3-stars in it, and the 3-star has 3 2-stars in it. So, that is a nested kind of parameterization that is common in the literature. We tend to prefer something like degree distributions instead in part because I think they are easier to understand. Degrees are not nested. A node simply has a degree, it has one degree only, and you count up the number of nodes that have that degree.

Triads or closed triangles are typically the basis for most of the clustering parameters that people are interested in. Almost anything you can think of in terms of a configuration of dyads can be represented as a network statistic. Then you have the parameter θ , and your choice there is whether you want to impose homogeneity constraints, and I believe Peter Hoff talked about this a little bit in his talk this morning.

Network statistics $z(x)$

The vector $z(x)$ can range from

- **Minimal** (# of edges – the Bernoulli model for a random graph) to
- **Saturated** (one term for every dyad)
- **In between** lie parsimonious summaries of the structural regularities in the network

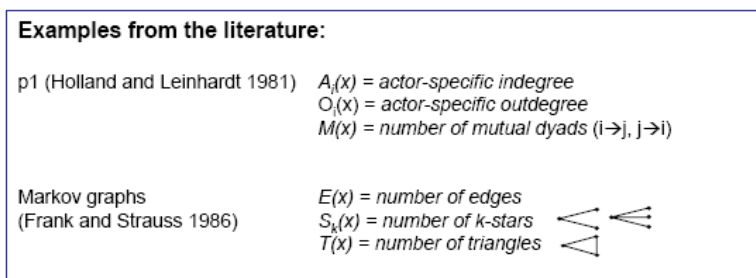


FIGURE 6

There is clearly a lot of heterogeneity in networks. Heterogeneity can be based on either nodal or dyadic attributes. People can be different because of their age, their race, their sex, those kinds of exogenous attributes, or different types of partnerships may be different. Let's think a little bit about how we create these network statistics. Referring to Figure 6, the vector here can range from a really minimal number one, such as the number of edges, which is the simple Bernoulli model for a random graph. But the vector and also be a saturated model, with one term for every dyad, which is a large number of terms. Obviously, you don't move immediately to a saturated model. It wouldn't give you any understanding of what was happening in the network anyway, so what we are trying to do is work somewhere in between these two, some parsimonious summary of the structural regularities in the network.

Figure 6 gives some examples from the literature that have been used in the past; the initial model was the p1 model by Holland and Leinhardt in 1981. Peter Hoff's talk this morning fit the framework of a p1 model: it had an actor-specific in-degree and an actor-specific out-degree—so, two parameters for every actor in the network—plus a single parameter for the number of mutual dyads, that is, when i nominates j and j nominates i in return. That is a dyadic independent model, which is to say, within a dyad, between two nodes, where there is the only dependence here for this mutual term, all other dyads are independent. That is a minimal model of what is going on in a network.

The first attempt to really begin to model dependent processes in networks—and that is, the edges are dependent—was the Markhov graph proposal by Frank and Strauss, which is also

shown in Figure 6. This model includes terms for the number of edges, the number of k-stars—again, those are those nested terms—and the number of triangles. In each of those cases you can impose homogeneity constraints on θ or not. So, for any network statistic that you have in your model, the θ parameters can range from minimal—that is, there is a single homogenous θ for all actors—to maximal, where there is a family of θ_i 's, each being actor- or configuration-specific. That was the case in Peter Hoff's model. Every actor would have two parameters there. You can say that every configuration has its own parameter, which is a function of the two actors, or multiple actors that are involved in it.

In the Bernoulli graph, the edges are the statistic, and there is a single θ that says every edge is as likely as every other edge. So, that is a homogeneity constraint. When we go with maximal θ parameters, you quickly max out and lose insight, but you can sure explain a lot of variance that way. In between are parameters that are specific to groups or classes of nodes—so you might have different levels of activity or different mixing propensities by age or by sex. In addition, there are parametric versus non-parametric representations of ordinal distribution. So, for a degree distribution, you can have a parameter for every single degree, or you could impose a parametric form of some kind, a Poisson, negative binomial, something like that.

The group parameterizations are typically used to represent attribute mixing in networks. We have heard a lot about that, but this is the statistical way to handle that. There is a lot of work that has been done on that over the last 20 years. The parametric forms are usually used to parsimoniously represent configuration distributions so degree distributions, shared partner distributions, and things like that.

ERG Model Estimation

$$P(\mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta' \mathbf{z}(\mathbf{x})\}}{c} \quad \text{This is the likelihood function to maximize}$$

Normalizing constant c makes direct ML estimation of the θ vector impossible

- e.g., with 50 nodes, there are 10^{369} graphs

Pseudolikelihood based on logistic regression approx. (Besag 1974, Strauss & Ikeda 1990)

- not good when dependence among ties is strong
(Geyer and Thompson 1992, Handcock 2003)

MCMC (Geyer and Thompson 1992, Crouch et al. 1998)

- theoretically guarantees estimation, but implementation for networks has been challenging

FIGURE 7

Estimating these models has turned out to be a little harder than we had hoped; otherwise, we would all be talking about statistical models for networks today. I don't think there would be anybody in here who would be talking about anything else.

The reason is that this thing is a little bit hard to estimate. Figure 7 shows the likelihood function $P(X = x)$ that we are going to be trying to maximize. The normalizing constant c makes direct maximum likelihood estimation of that θ vector impossible because, even with 50 nodes, there are an almost uncountable number of graphs. So, you are not going to be able to compute that directly. Typically, there have been two approaches to this. The first that dominated in the literature for the last 25 years is pseudolikelihood, which is essentially based on the logistic regression approximation. We now know, and we knew even then, that this isn't very good when the dependence among the ties is strong, because it makes some assumptions there about independence. MCMC is the alternative; it theoretically guarantees estimation.

The stumbling block: Model Degeneracy

Take a really simple model for a network with a specific density and endogenous clustering:

$$P(\mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta_1 d(\mathbf{x}) + \theta_2 c(\mathbf{x})\}}{c}$$

$d(\mathbf{x})$ is the density of edges
 $c(\mathbf{x})$ is the clustering parameter
(the fraction of 2 stars that are completed triangles).

What are the properties of this model? Examine through simulation.

For a network with 50 nodes:

- Set the density of the network to 4% (about 50 edges, for a Bernoulli graph the expected clustering would then be 3.8%)
- Set the network clustering 10 times higher than expected: 38%
- By construction, the networks produced by this simple model will have average density of 4% and average clustering of 38%

FIGURE 8

But that's only "theoretically." Implementation has turned out to be very challenging. The reason has been something called model degeneracy. Mark Handcock is going to talk a lot about this tomorrow, but I am just going to make a couple of notes about it today. Figure 8 shows a really simple model for a network, just density and clustering. Those are the only two things

that we think are going on. So, there is the density term, which is just the sum of the edges, and $c(x)$ is the clustering coefficient that people like to use to represent the fraction of 2-stars that are completed triangles. What are the properties of this model? Let's examine it through simulation. Start with a network of 50 nodes. We are going to set the density to be about 4 percent, which is about 50 edges for a Bernoulli graph. The expected clustering, if this were a simple random graph, would then just be 3.8 percent, but let's give it some more clustering. Let's bump it up to 10 times higher than expected and see how well this model does. By construction, the networks produced by this simple model will have an average density of 4 percent, and an average clustering of 38 percent. Figure 9 shows what the distribution of those networks looks like.

What degeneracy looks like

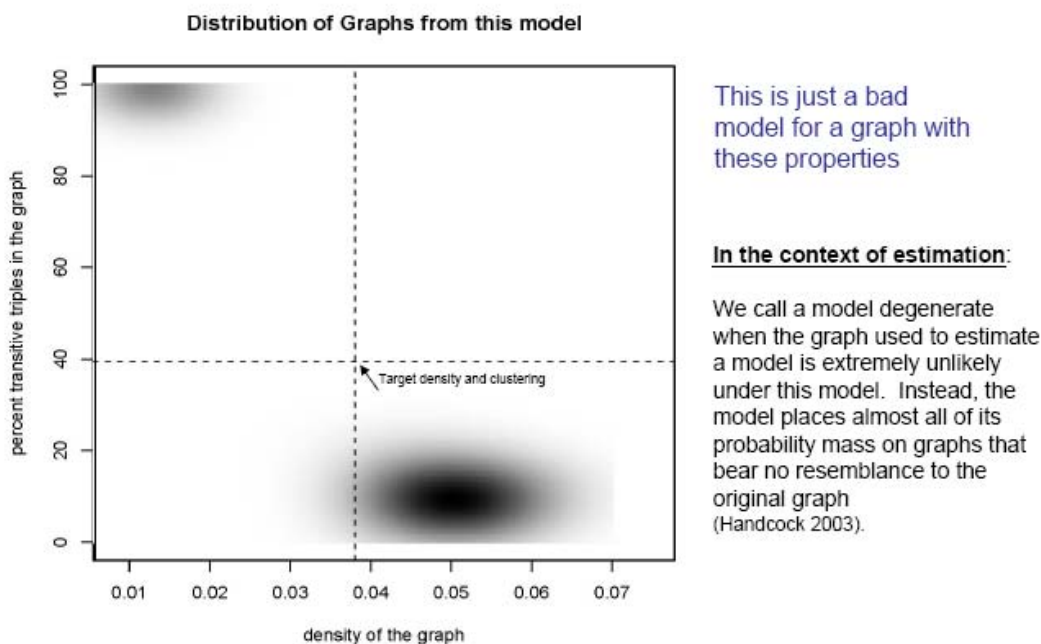


FIGURE 9

The target density and clustering would be where the dotted lines intersect, but virtually none of the networks look like that. Most of these networks, instead, are either fairly dense, but don't have enough triangles, or not so dense, but are almost entirely clustered in triangles. What does that mean? It means that is this is a bad model. For a graph with these properties, if you saw this kind of clustering and density, this is a bad model for it. This graph didn't come from that model, that is what it says. In the context of estimation, we call the model degenerate when the graph used to estimate the model is extremely unlikely under that model. Instead, the model

places almost all of its probability mass on graphs that bear no resemblance to the original graph.

We will never see anything like this in linear model settings, so this hung people up for about 10 years because they couldn't figure out why every time they tried to estimate using MCMC, the parameters they would get back would create either complete graphs or empty graphs, and nothing in between. What they did was make simpler and simpler models because they figured something had to be wrong with the algorithm or something else. It turns out that the simpler the model the worse the properties, and 10 years of science were down the drain.

The morals—and I think these really are morals, because they are a new way to think about data—are three. First, descriptive statistics may not produce good generative models, and I think that there is more truth in this setting than in the linear model setting. Second, starting simple and adding complexity often doesn't work—at least not the way that we expect it to work. Third, I think it is going to take a long time to develop new intuition, and we are going to have to develop new model diagnostics in order to help our intuition. One thing we can say now pretty clearly is that statistics for endogenous processes—so, like this friend-of-a-friend stuff, the clustering parameter—need to be handled differently, because they induce very large feedback effects. A little bit of that and all of a sudden the whole graph goes loopy. It has to create lots of whatever it is because there is nothing to prevent it.

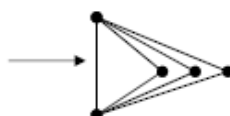
Thinking about new model terms

To capture endogenous clustering

Old : $t(x)$ = # of completed triangles in the graph
 $c(x)$ = triangles as a percent of all 2-stars

Every triangle here has the same impact, θ

New : Declining marginal impact, "weighted shared partner" statistic, where the weight is:



$$w = e^\delta \sum_{i=1}^{n-2} \left\{ 1 - (1 - e^\delta)^i \right\} p_i$$

p_i = # of linked nodes with i partners in common
 δ = dampening parameter

The more partners in common, the more likely two nodes are linked,
 but with a declining marginal return (Snijders et al. 2004, Hunter and Handcock 2004)

FIGURE 10

We have been thinking a little bit about new model terms to capture some of the endogenous clustering effects, because we think these are important, and so we do want to be able to capture them. As shown in Figure 10, the old forms of these were typically either the number of triangles—that is what you saw in the Markov random graphs—or a clustering percent. In those cases, every triangle has the same impact, θ . That is your parameter, and that is obviously a problem. What we have done is come up with new models where we parameterize a distribution of shared partners. You can think of those shared partners as each giving you an increasing probability of having a tie with the other person, but there is a declining marginal return to that. So, it doesn't blow up in the same way.

It looks ugly, but it turns out it is a fairly simple weight that creates an alternating k-star—that is one way of thinking about it. 2-stars will be positively weighted and 3-stars will be negatively weighted; 4-stars will be positively weighted, and 5-stars will be negatively weighted. That is what this thing does. It tends to create reasonably nice, smooth parametric distributions for shared partner statistics.

In the time I have left I want to give you a sense of how these models work in practice, and how friendly they look when you get them to actually estimate things well. Imagine a friendship network in which two different clustering processes are at work. There is homophily, which is the birds-of-a-feather exogenous attribute, in which people tend to choose friends (i.e., create a link) to people who are like them, in grade, race, etc. Then there is transitivity, which is that people who have friends in common tend to also become friends. This is a friend-of-a-friend endogenous process for creating links.

That can also generate a fair amount of assortative mixing, or can come from assortative mixing. Imagine three of our actors in the same grade, and two of them are tied to the same person. Then the question is, will they have a tie between themselves. If they do, it could either be because of transitivity, or it could be because they are in the same grade. So, the question is, how do we distinguish these? We have a number of different kinds of ties here. There are within-grade ties, which can come from both of these processes, and across-grade ties, which we can assume are due to transitivity. There are triangles that come from both, and there are ties not forming triangles, and we can assume that is homophily. So, we do have a little bit of information here that we are going to use to tease apart these things statistically, and we are going to work with data collected from a school.

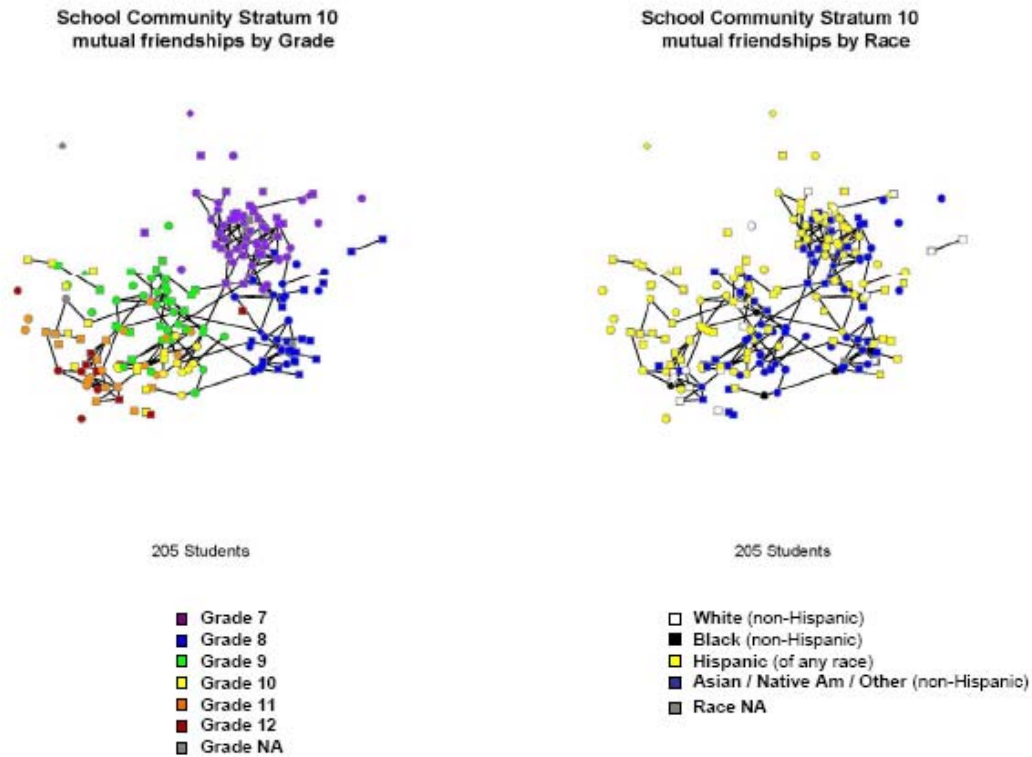


FIGURE 11

Figure 11 shows data from school 10, a school that has a junior high school and a high school. You can see there aren't a whole lot of links between those two. In addition, there is a fair amount of grade assortative mixing. Because we don't have a lot of black students in this school—it is mostly white and Asian—there is a little more mixing by race than you might get otherwise, so the race effects will not be quite as strong here.

We are going to try four models: edges alone (this is our simple random graph); edges plus attributes, which says the only thing going on in this model is that people are choosing others like themselves; edges plus this weighted edge-wise shared partner statistic, which is transitivity only, so just the friend-of-a-friend process; and then both of these things together. Figure 12 summarizes these four models.

Which model best captures the features of the graph?

Model	Statistics
Edges	# of edges
Edges + Attributes (homophily)	# of edges # of edges for each race, sex ,grade # of edges that are within-race, within-grade, within-sex
Edges + WESP (transitivity)	# of edges weighted shared partners
Edges + Attributes + WESP (both)	# of edges # of edges for each race, sex ,grade # of edges that are within-race, within-grade, within-sex weighted shared partners

FIGURE 12

Our approach is depicted schematically in Figure 13. We start with the data and our model, and we estimate our model coefficients. We can then simulate graphs with those properties, those particular coefficients, and those statistics, drawn from a probability distribution. We are going to compare the graph statistics from the simulated data to the graph statistics from our observed data, but the graph statistics that we are going to use are not actually statistics that were in the model. What we are trying to do is predict things like path links or geodesics that are not part of our original model, because that says we have got the local rules right. These are the local rules that generate those global properties. That is the approach we are going to take.

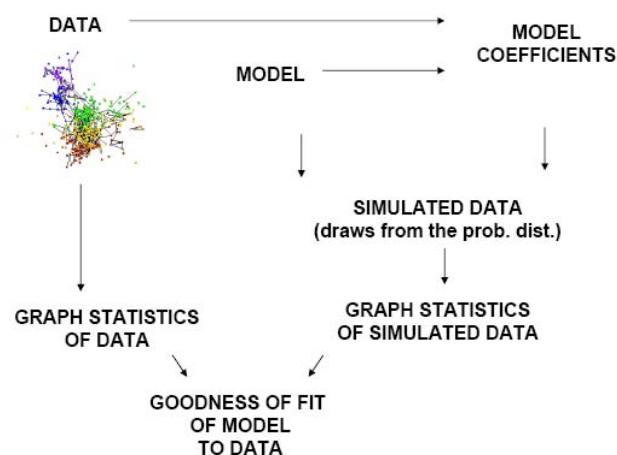


FIGURE 13

Figure 14 gives an example for a Bernoulli graph. You can see what the data looks like, and the chart shows the degree distribution from that graph. Students were only allowed to nominate their five best male and five best female friends. So, it is truncated at 10. Nobody

Goodness of fit measure 1: degree distribution

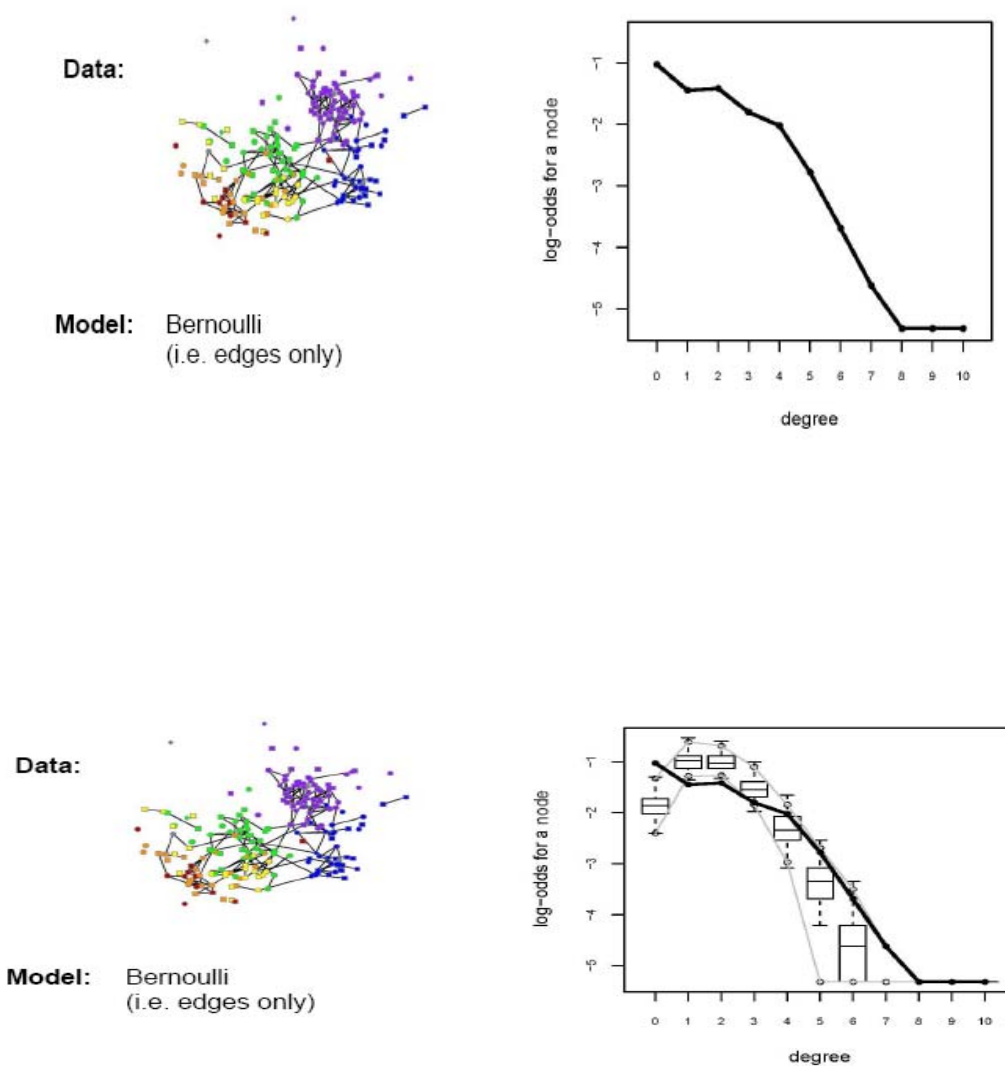


FIGURE 15

Figure 15 shows what the simulations look like from the Bernoulli model. You can see we don't get the degree distribution very well.

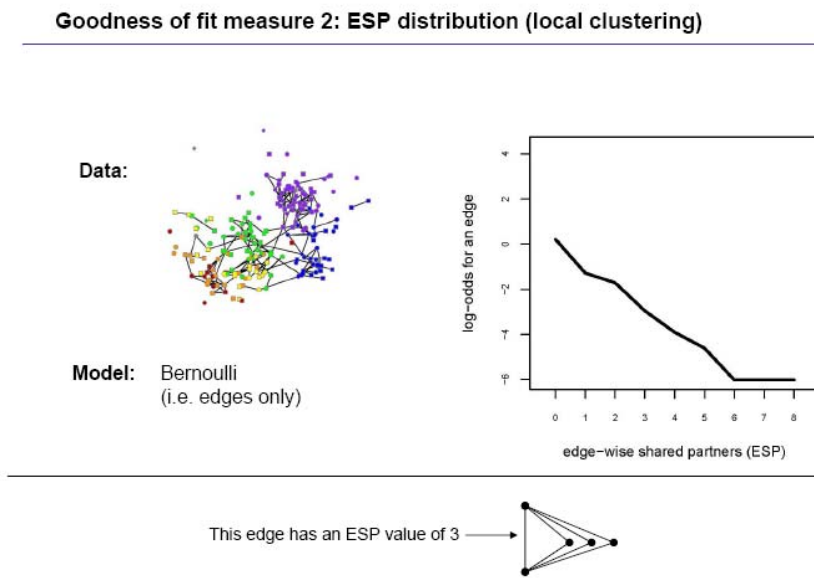


FIGURE 16

Figure 16 shows the edgewise shared-partner statistic. This is, for every edge in the graph, how many shared partners do they have. That is like a generalized degree distribution. That is what it looks like in the observed data. Figure 17 shows what it looks like from the Bernoulli model, so obviously the Bernoulli model isn't capturing any of that kind of clustering at all.

Goodness of fit measure 2: ESP distribution (local clustering)

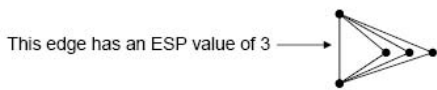
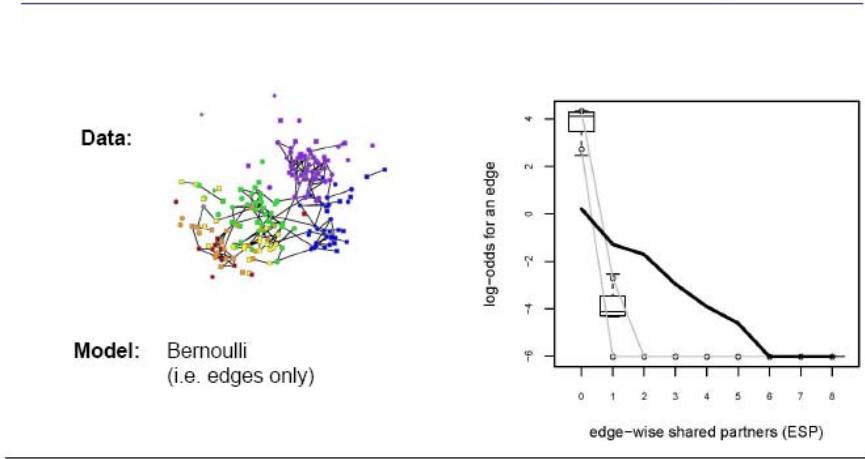


FIGURE 17

Finally, Figure 18 shows the minimum geodesic distance between all pairs, with a certain fraction of them here being unreachable.

Goodness of fit measure 3: geodesic distribution (global clustering)

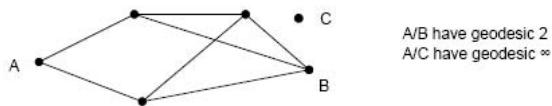
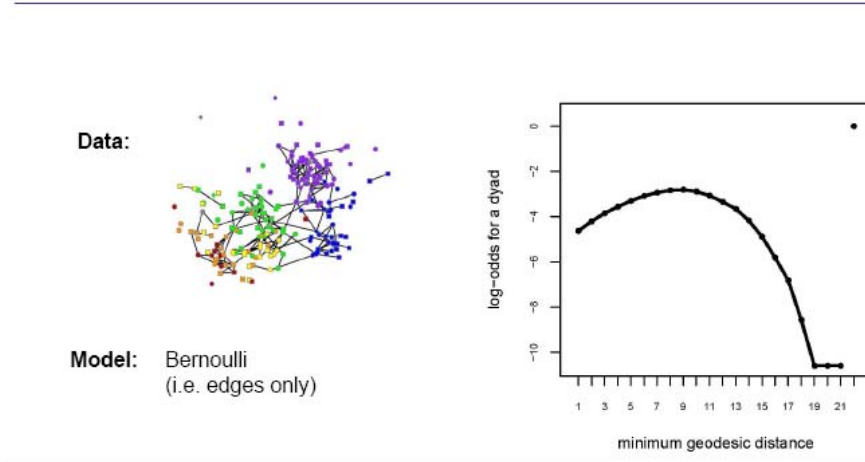


FIGURE 18

Figure 19 shows how the Bernoulli model does, and it doesn't do very well.

Goodness of fit measure 3: geodesic distribution (global clustering)

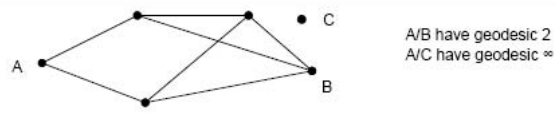
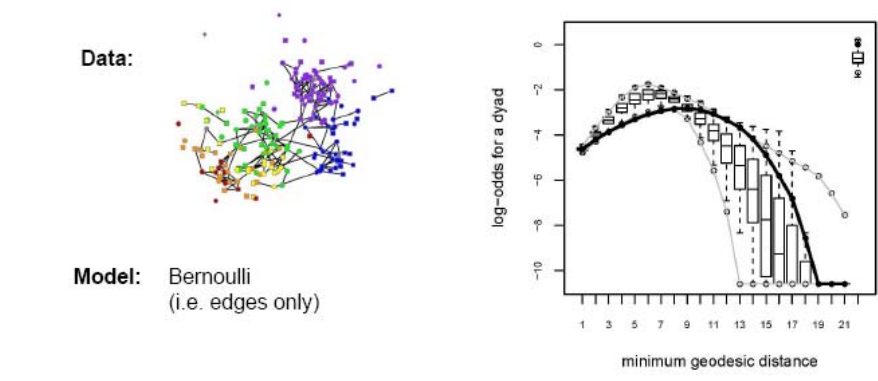


FIGURE 19

Putting these all together, Figure 20 shows your goodness of fit measures.

Goodness of fit measures assembled

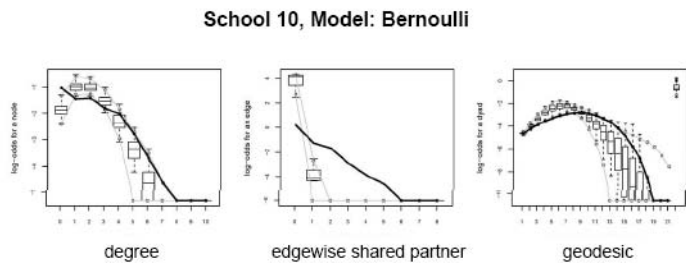


FIGURE 20

Figure 21 shows what it looks like for all the models. The first column shows the degree distribution; even a Bernoulli model is pretty good. Adding attributes doesn't get you much, but adding the shared partners, you get it exactly right on, and the same is true when you add both the shared partner and the attributes. For the local clustering, which is this edgewise shared-partner term, the Bernoulli model does terrible. I can't say the attribute model does a whole lot better. Of course, once you put the two-parameter weighted shared-partner distribution in, you capture that pretty well.

School 10

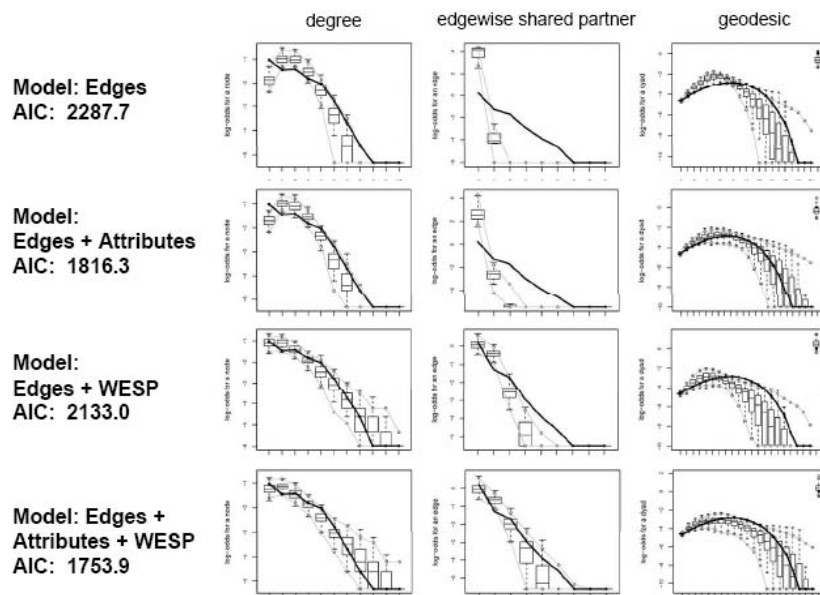


FIGURE 21

You don't capture the geodesic well with edges alone, but it is amazing how well you get the geodesic just from attribute mixing alone, just from that homophily. In fact, it actually doesn't do so well, the local clustering term for this edgewise shared-partner doesn't capture it anywhere near as well, and you actually don't do as good a job when you put both of them in.

So, Figure 22 is the eyeball test. That is a different approach to goodness of fit. One thing you want to make sure of with these models is that they aren't just crazy. Obviously, those degenerate models were crazy, and you could see that very quickly. They would either be empty or they would be complete. For this figure, it actually would be hard to tell which one was simulated and which one was real. They are actually getting the structure from the eyeball perspective pretty well.

School 10: the eyeball test

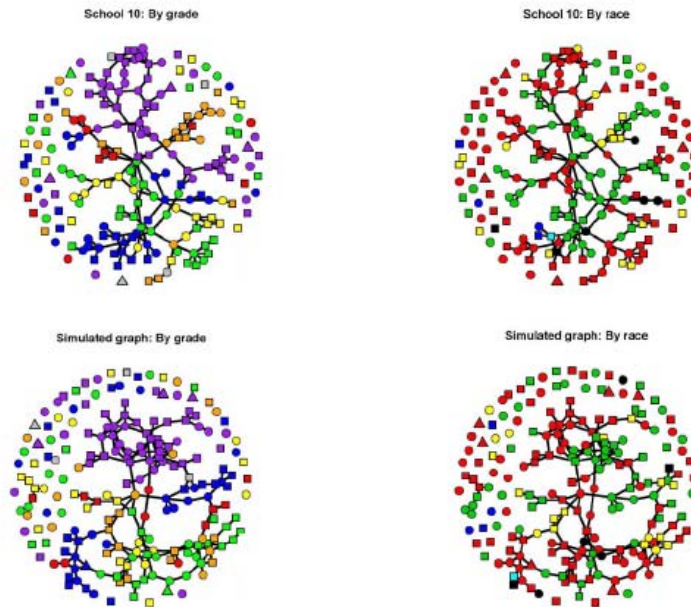


FIGURE 22

There are 50 schools from which we can draw information. There are actually more, but 50 that have good information. They range in size from fairly small—and this school 10 is one of the smaller mid-sized schools, with 71 students in the data set—but we can use these models all the way up to beyond 1,000. We have now used them on 3,000-node networks, and they are very good and very stable. Figure 23 compares some results for different network sizes, using the model that has both the friends-of-a-friend and the birds-of-a-feather processes in it. It does very well for the smaller networks, but as you start getting out into the bigger networks, the geodesics are longer than you would expect. Basically, I think that is telling you there is more clustering, there is more pulling apart in these networks, and less homogeneity than these models assume. So, there is something else that is generating some heterogeneity here.

Attributes + WESP Model as network size increases

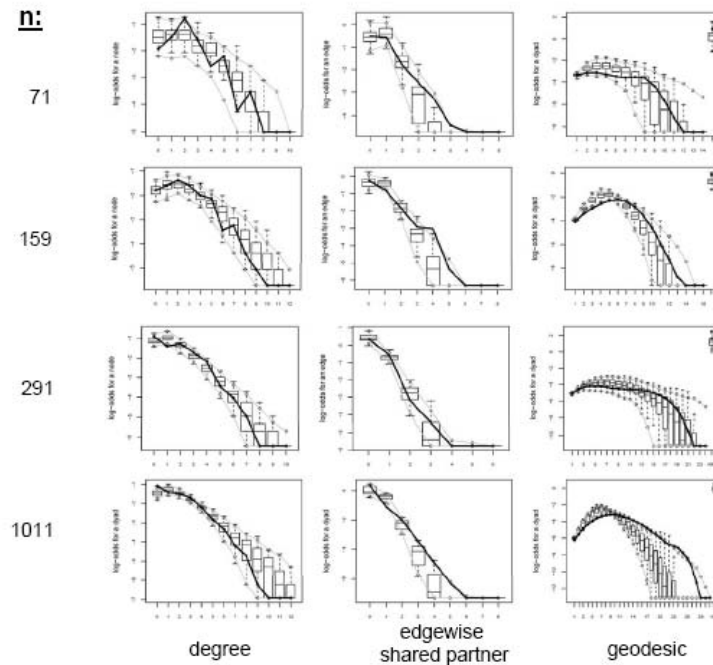


FIGURE 23

The other thing you can do is to compare the parameter estimates across the models, which is really nice. In Figure 24, we look at 59 schools, using the attribute-only model. You can see the differential effects for grades: the older students tend to have more friends.

Parameter values: Attribute only model, all 59 schools

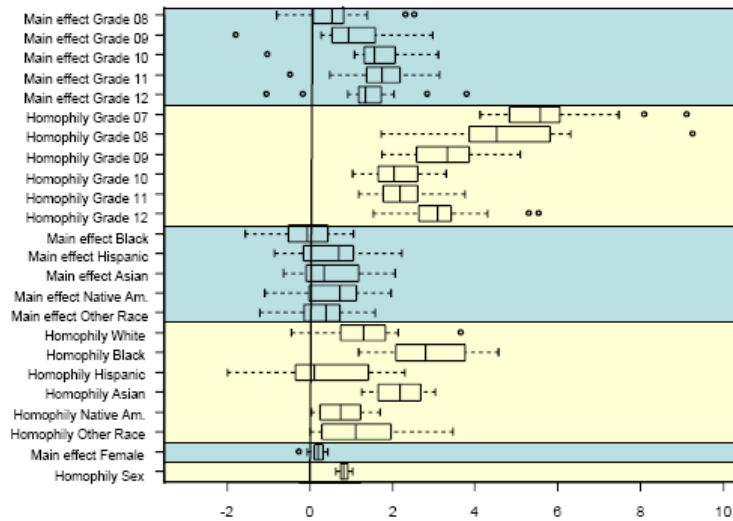
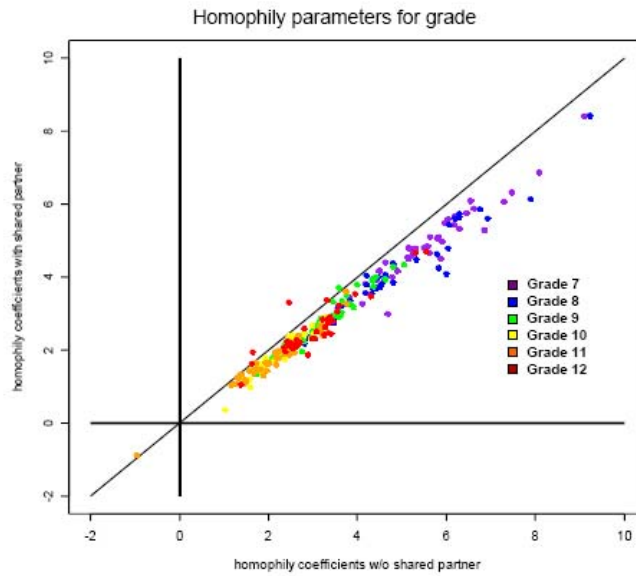


FIGURE 24

Figure 25 shows the homophily effect, the birds-of-a-feather effect, which is interesting. You see that it is strongest for the younger and the older, and those are probably floor and ceiling effects. Mean effects for race don't show up as being particularly important, but blacks are significantly more likely than any other group to form assortative ties. Interestingly, you can see that Hispanics really bridge the racial divide. So, there are all sorts of nice things you can do by looking at those parameters as well.

Finally, the other thing you can do is examine what is the effect of adding a transitivity term to a homophily model. I mean, how much of the effect that we attributed to homophily actually turns out to be this transitivity effect instead. It turns out the grade-based homophily estimates fall by about 14 percent once you control for this friend-of-a-friend effect. The race homophily usually falls, but actually sometimes rises, so once you account for transitivity you find that the race effects are actually even stronger than you would have expected with just the homophily model. This is shown in Figure 26.

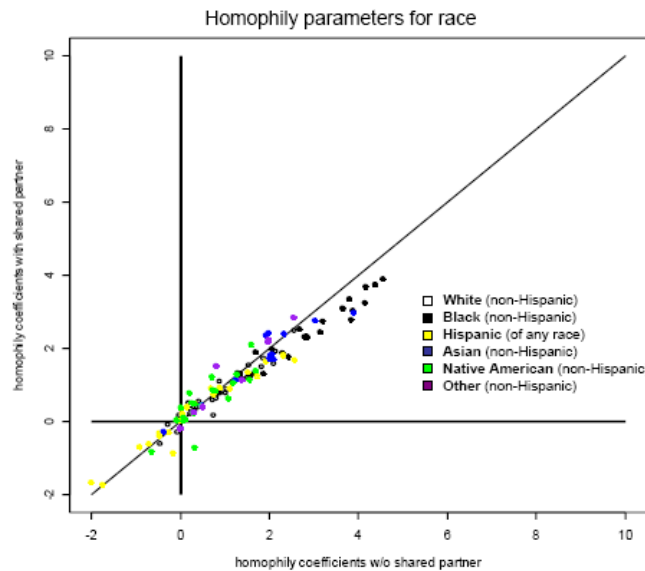
Grade Parameters: Homophily only vs. homophily+transitivity



Grade based homophily estimates fall by about 14% once you control for the transitivity effect.

FIGURE 25

Race Parameters: Homophily only vs. homophily+transitivity

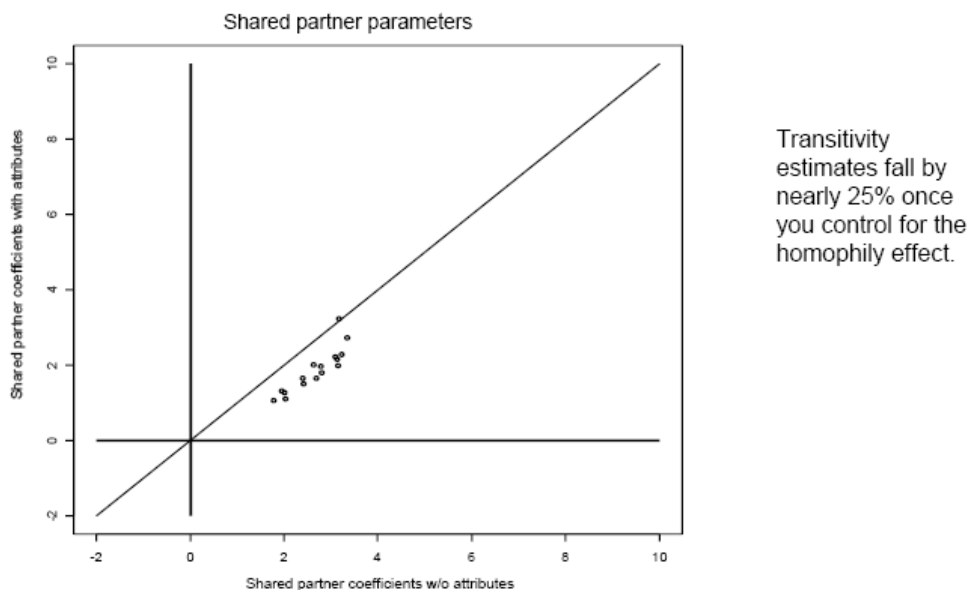


Race based homophily estimates generally fall, but sometimes rise once you control for the transitivity effect.

FIGURE 26

The transitivity estimates this friend-of-a-friend effect falls by nearly 25 percent once you control for the homophily term, as shown in Figure 27. This doesn't seem like much, but it is amazing to be able to do this kind of stuff on networks, because we have not been able to do this before. We have not been able to test these kinds of things before. What we have now is a

Transitivity Parameters: Transitivity only vs. transitivity+ homophily



What this approach offers is a principled method for theory-based network analysis where you have a framework for model specification, estimation, comparison and inference. These are generative models and they have tests for fit, so it is not as if you see there is clustering, but it is homogenous clustering how well that it fits. These give you the answers to those questions. We have methods for simulating networks that fall out of these things automatically, because we can reproduce the known, observed, or theoretically-specified structure just by using the MCMC algorithm. For the cross-sectional snapshots, this is a direct result of the fitting procedure. For dynamic stationary networks, it is based on a modified MCMC algorithm, and dynamic evolving networks means you have to have model terms for how that evolution proceeds. You can then simulate diffusion across these networks and, in particular, for us, disease transmission dynamics. It turns out the methods also lend themselves very easily to sampled networks and other missing network data.

With that I will say if you are interested in learning more about the package that we have, it is an R based package. We are going to make it available as soon as we get all the bugs out. If you want to be a guinea pig, we welcome guinea pigs, and all you need to do is take a look at <http://csde.washington.edu/statnet>. Thank you very much.

QUESTIONS AND ANSWERS

DR. BANKS: Martina that was a lovely talk and you are exactly right, you can do things that have never been done before, and I am deeply impressed by all of it. I do wonder if perhaps we have not tied ourselves into a straightjacket by continuing to use the analytic mathematical formulation of these models.

DR. MORRIS: The analytic what formulation?

DR. BANKS: An analytical mathematical formulation growing out of a p1 model and just evolving. An alternative might be to use agent-based simulation, to try and construct things from simple rules for each of the agents. For example, it is very unlikely that anybody could have more than 5 best friends or things like that.

DR. MORRIS: Actually, you would be surprised how many people report that. I am kidding. I agree, except that I think that, depending on how you want to define agent-based simulation, these are agent-based simulations. What I am doing is proposing certain rules about how people choose friends. So, I choose friends because I tend to like people the same age as me, the same race as me, the same socioeconomic status. Those are agent-based rules and, when other people are using those rules, then we are generating the system that results from those rules being in operation. I don't see a distinction between these two in quite the same way, but I do agree.

One thing that we did do was focus on the Markov model for far too long. It was edges, stars, and triangles. I think there is an intuitive baby in the bath water, and that is, edges are only dependent if they share a node. That is a very natural way to think about trying to model dependency, but I think it did kind of narrow our focus probably more than it should have.

DR. BANKS: You are exactly right. Your models do instantiate something that could be an agent-based system, but there are other types of rules that would be natural for friendship formation that would be very hard to capture in this framework, I think.

DR. MORRIS: I would be interested in talking to you about what those are.

DR. BANKS: For example, the rule that you can't have more than 5 friends might be hard to build in.

DR. MORRIS: No, in fact, that is very easy to build in. That is one of the first things we

had to do to handle the missing data here, because nobody could have more than 5 male or five female friends.

DR. HOFF: There was one middle part of your talk which I must have missed, because you started talking about the degeneracy problem with the exponentially parameterized graph models, and then at the end we saw how great they were. So, at some point there was the transition, by including certain statistics, or is it including certain statistics that makes them less degenerate, or is it the heterogeneity you talked about? I could see how adding heterogeneity to your models or to the parameters is going to drastically increase the amount of the sample space that you are going to cover. Could you give a little discussion about that?

DR. MORRIS: These models have very little heterogeneity in them relative to your models. So, every actor does not have both an in-degree and an out-degree. There is basically just a degree term for classes. So, grades are allowed to have different degrees, race is allowed to have different degrees. The real thing—that wasn't what made this work. What made this work was the edgewise shared partner. When we had originally tried using the local clustering term as either the clustering coefficient or the number of triangles with just a straight theta on it, those are degenerate models. The edgewise shared partner doesn't solve all problems either, but at least it was an ability bound, and that is essentially the effect that it has, is that it bounds the tail and it says that people can't have that many partners. So, that is what changed everything.

DR. JENSEN: I think the description of model degeneracy is wonderful. Fortunately, it wasn't 10 years of work in my case. It was more like 6 months of work that went down the drain and I didn't know why, and I think you have now explained it. Is that written up anywhere? Are there tests that we can run for degeneracy? What more can we do?

DR. MORRIS: That is a great question. Mark Handcock is really the wizard of model degeneracy, and I think he is going to give a talk tomorrow that can answer some of those questions. I don't think we have a test for it yet, although you can see whether your MCMC chain is mixing properly and, if it is always down here and then all of a sudden it goes up here, then you know you have got a problem. It is still a bit of an art. STATNET, this package, will have a lot of this stuff built into it.

DR. SZEWCZYK: My question is, is it model degeneracy, or is it model inadequacy? I look at a lot of these things and my question is, can we take some of these models and, rather than just fitting one universal model, can we go in there and, say, fit a mixture of these p-star models, or these Markov models or p1 models, rather than assuming that everyone acts the same within these groups?

DR. MORRIS: There are lots of ways to try to address the heterogeneity, I agree with you, and I think they need to be more focused on the substantive problem at hand.

So, just throwing in a mixture or throwing in a latent dimension, to me, kind of misses the point of why do people form partnerships with others. So, when I go into this, I go in saying, I want to add attributes to this. A lot of people who have worked in the network field don't think attributes matter because somehow it is still at the respondent level. We all know that, as good network analysts, we don't care about individuals. We only care about network properties and dyads.

DR. SZEWCZYK: We care about individuals.

DR. MORRIS: Attributes do a lot of work. They do a lot of heavy lifting in these models and they actually explain, I think, a fair amount. I would call it model degeneracy only in this case because you get an estimate and you might not even realize it was wrong. In fact, when people used the pseudolikelihood estimates, they had no idea that the cute little estimate they were getting with a confidence interval made no sense at all. It is degenerate because what it does, it performs the function. It actually gets the average right, but then it gets all the details wrong. So, you can call that inadequate, and it is. It is a failure. It is a model failure. That is very clear.

DR. WIGGINS: So, one thing to follow up on that I was wondering about, since each of these models defined a class, I wonder if you thought about treating this using not classifiers, large-margin classifiers, like support vector machines. Some anecdotal evidence is that sometimes you can tell if none of your network models is really good for a network that you are interested in. So, some of these techniques that measure everything at once, rather than measuring a couple of features you want to reproduce will show you how one network, if you look at it in terms of one attribute, it looks like model F, but if you look at it in terms of a different model, it turns out to be model G, and that might be one way of seeing whether or not you have heterogeneity or just none of your models is a good model. If you have a classifier, then all the different classifiers might choose, not me, as the class, in which case you can kind of see if none of your models is the right model.

DR. MORRIS: Yes, that is a nice idea.

The State of the Art in Social Network Analysis

The State of the Art in Social Network Analysis

Stephen P. Borgatti, Boston College

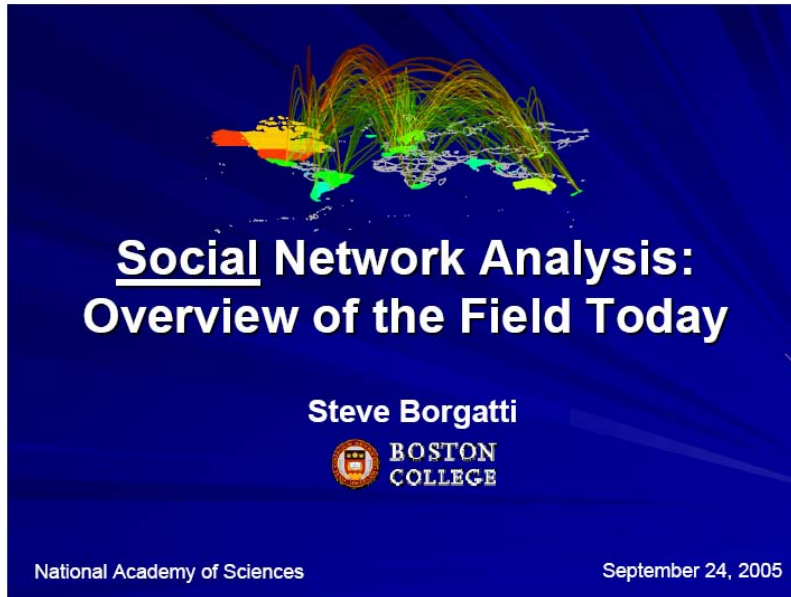


FIGURE 1

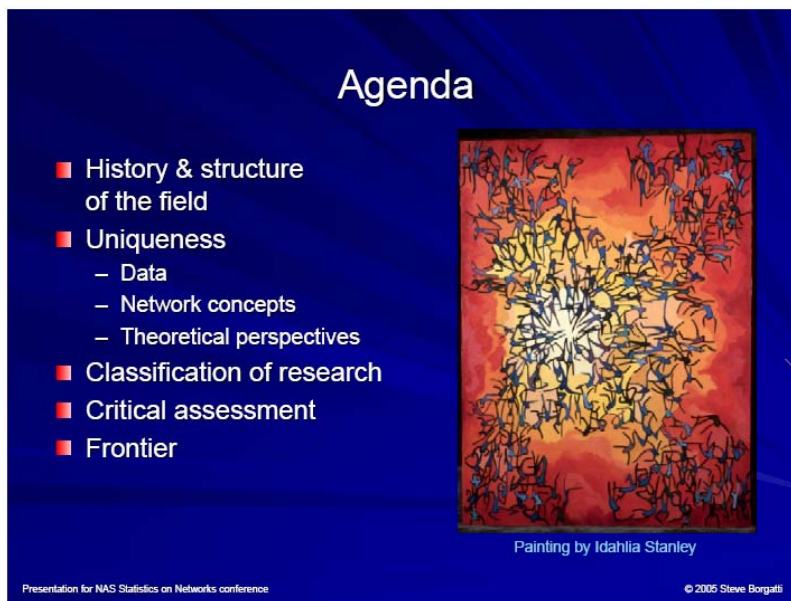


FIGURE 2



FIGURE 3



FIGURE 4

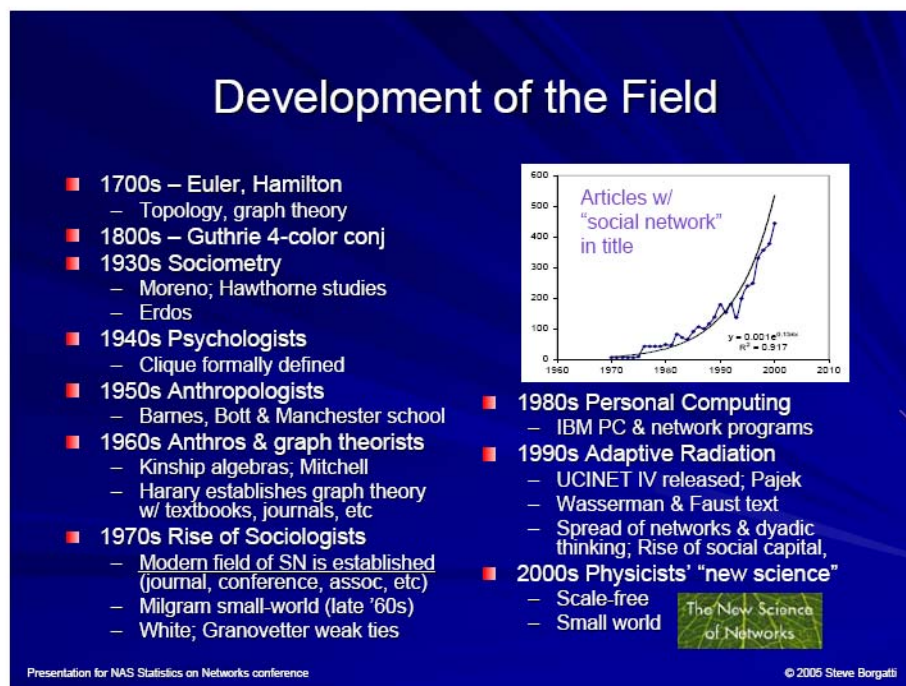


FIGURE 5

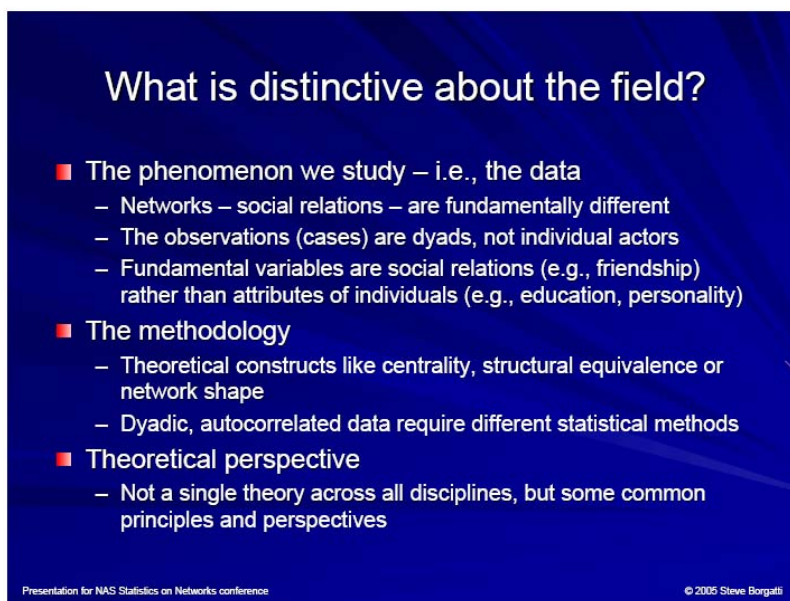


FIGURE 6

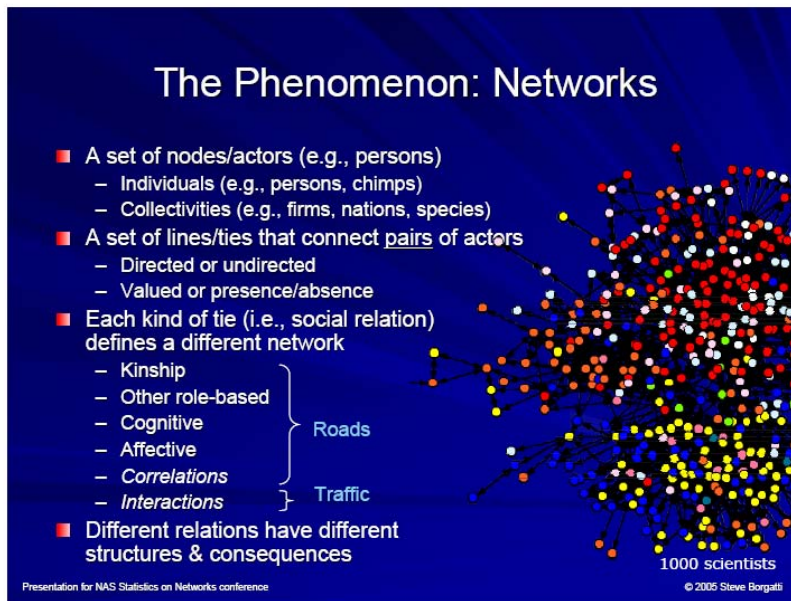


FIGURE 7

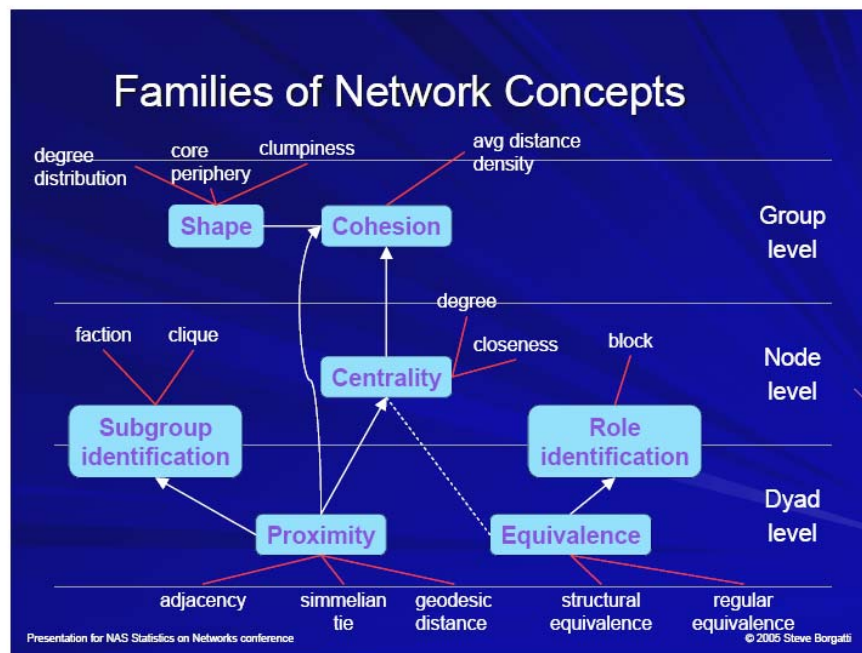


FIGURE 8

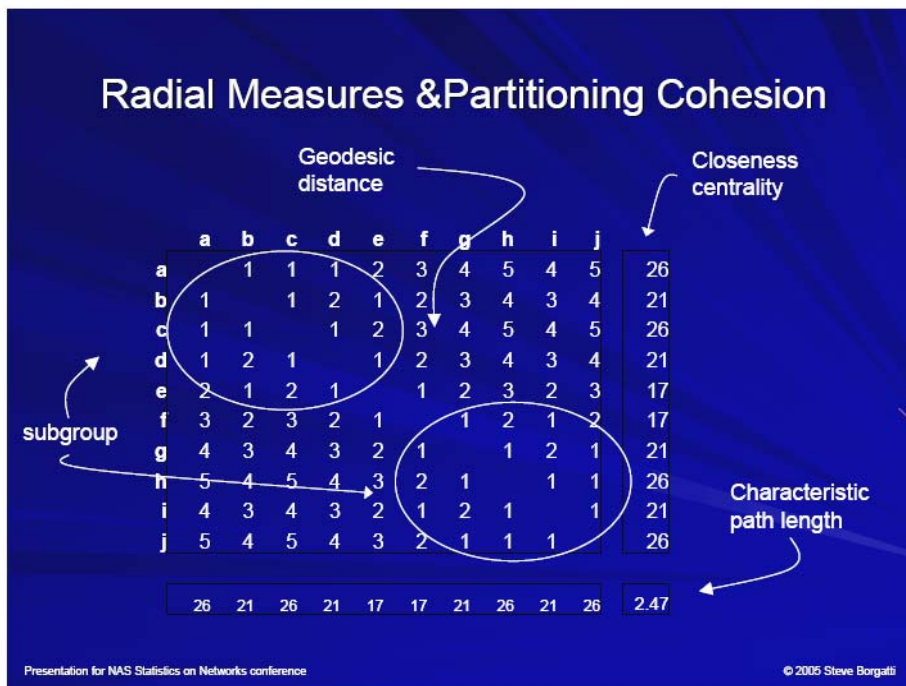


FIGURE 9

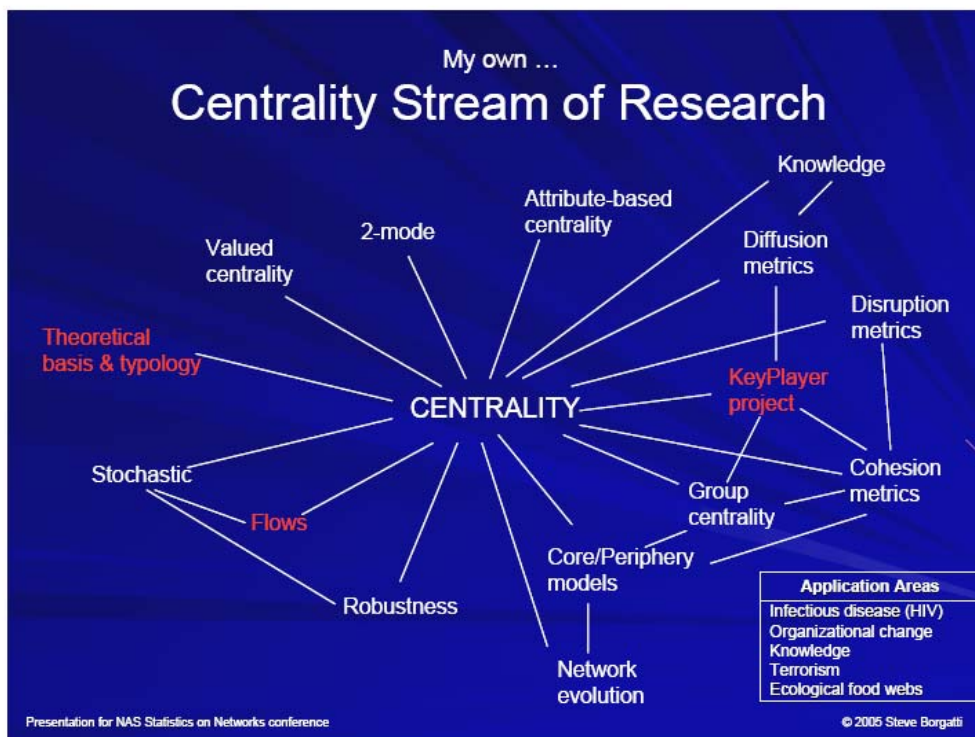


FIGURE 10

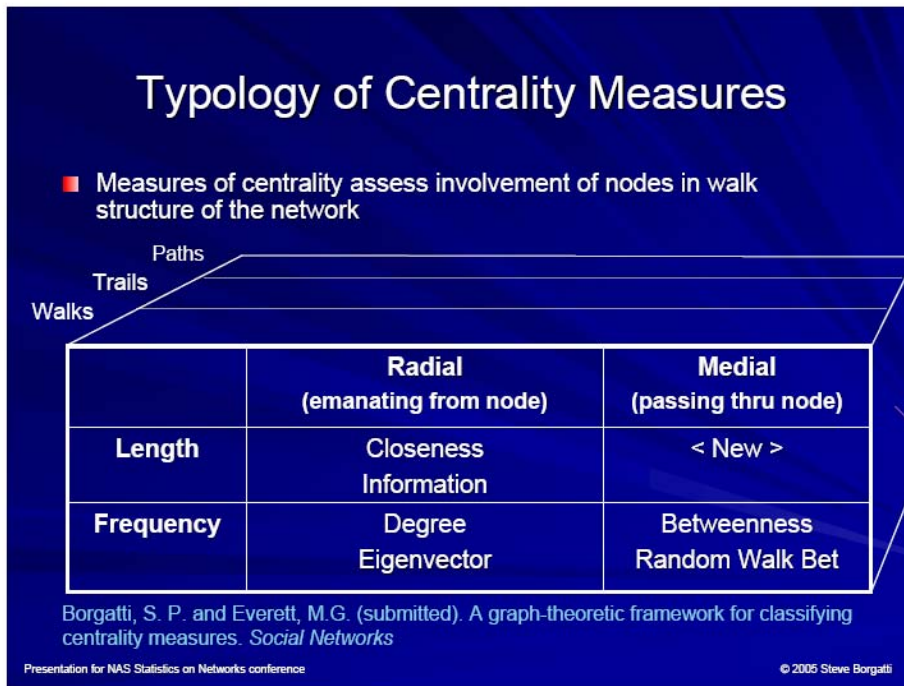


FIGURE 11

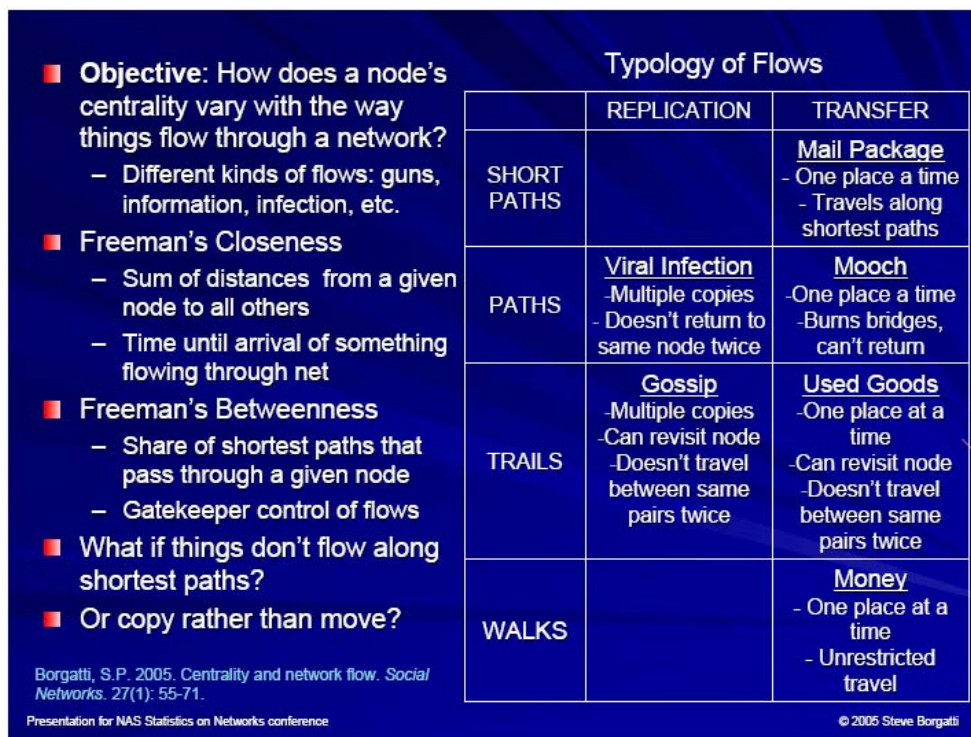


FIGURE 12

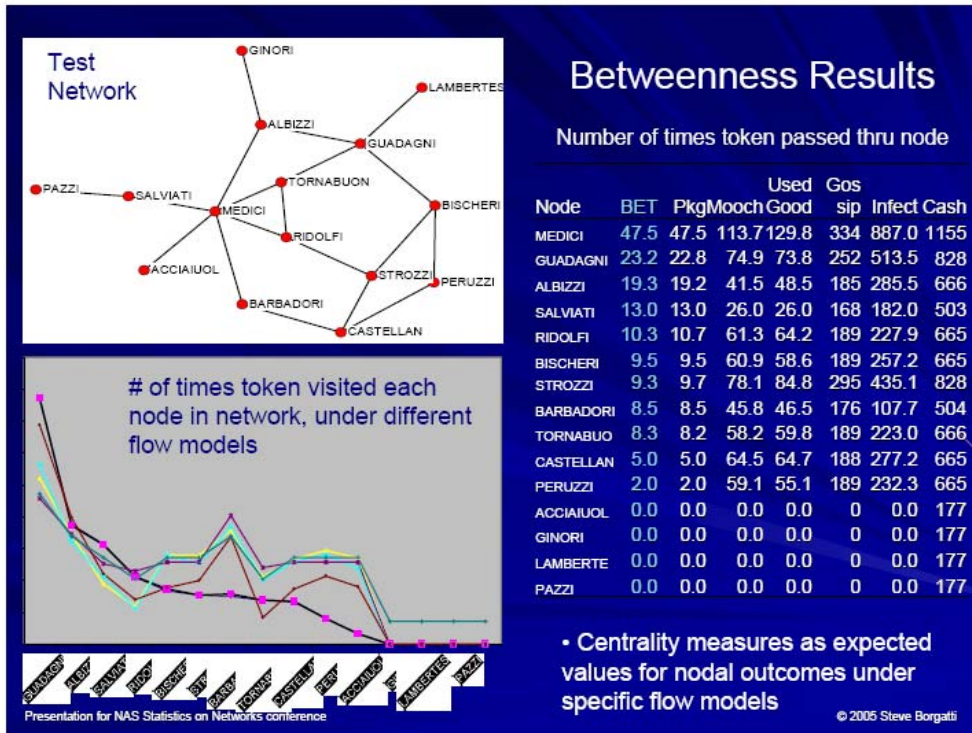


FIGURE 13

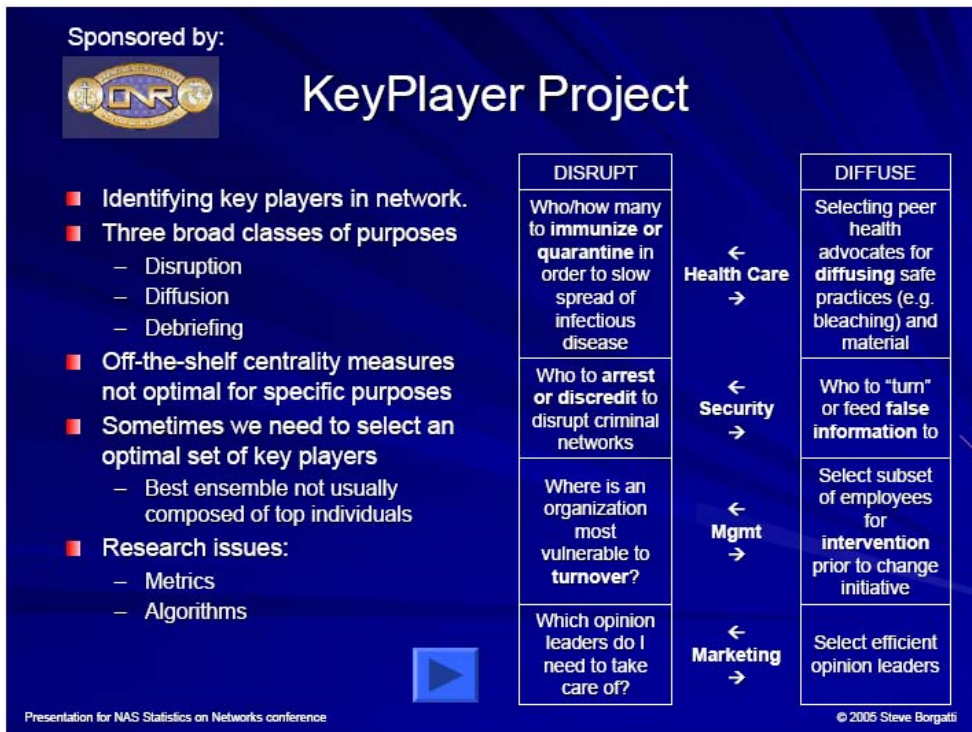


FIGURE 14

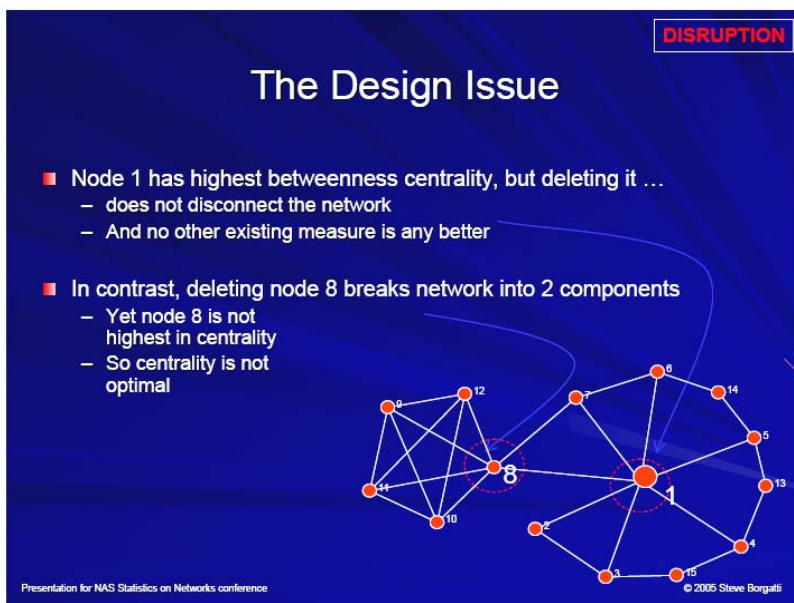


FIGURE 15

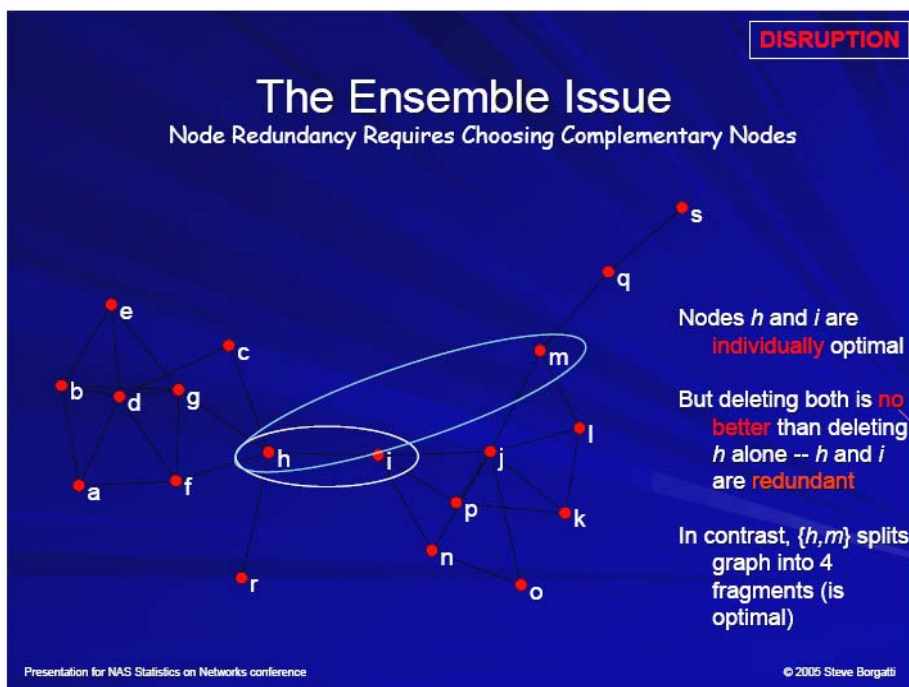


FIGURE 16

KeyPlayer DISRUPTION Metrics

■ **Strategy**

- Measure cohesiveness of network
- Remove key players
- Measure cohesiveness of new network
- Proportion reduction in cohesion (PRC) measures fragmentation potential of the key player set

■ **Optimization**

- Find sets of keyplayers with maximum fragmentation potential

■ **Cohesion measures**

- Fragmentation

$$F = 1 - \frac{2 \sum_{i < j} r_{ij}}{n(n-1)}$$

- Shortness

$$S = 1 - \frac{2 \sum_{i > j} \frac{1}{d_{ij}}}{n(n-1)}$$

Medial measures of centrality (e.g., betweenness) can be rewritten as PRC measures

Presentation for NAS Statistics on Networks conference © 2005 Steve Borgatti

FIGURE 17

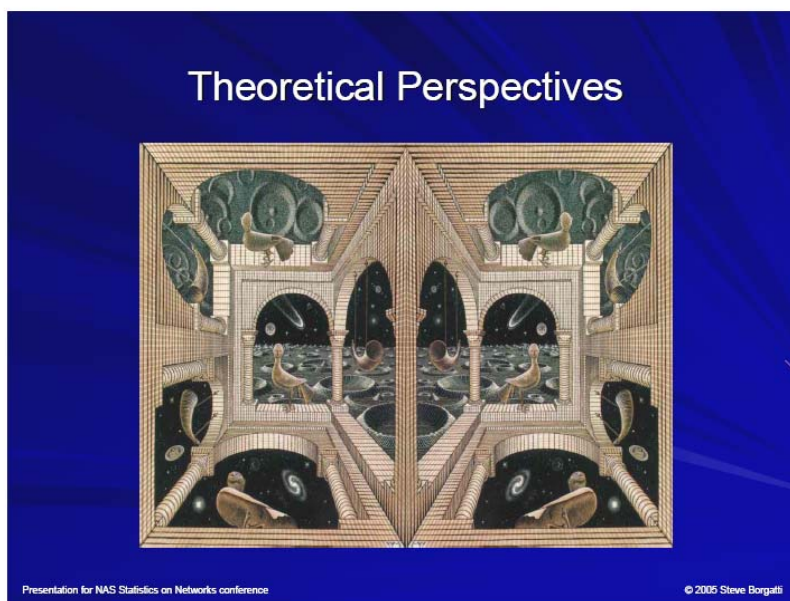


FIGURE 18

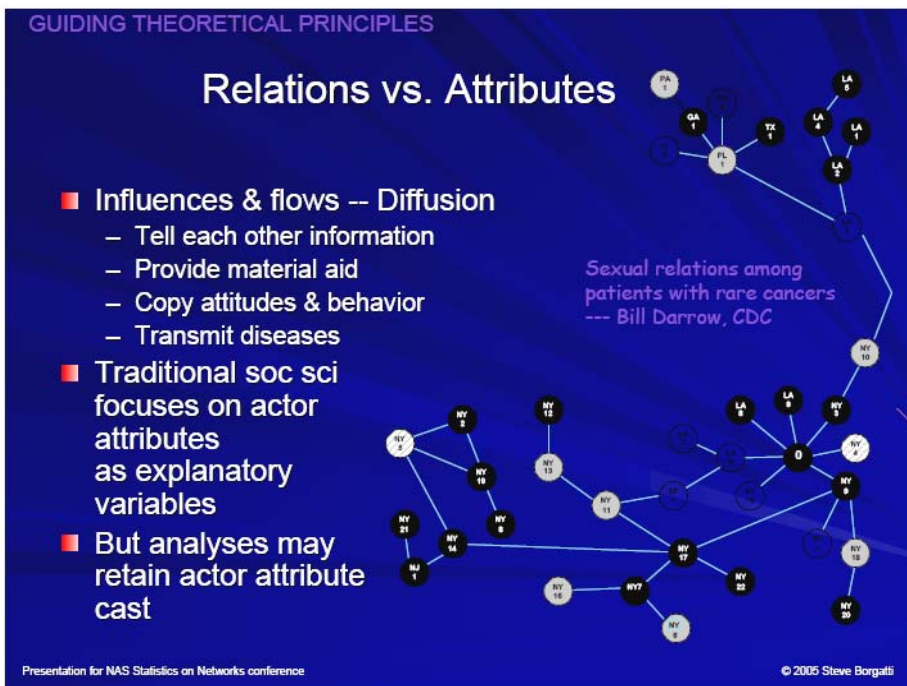


FIGURE 19

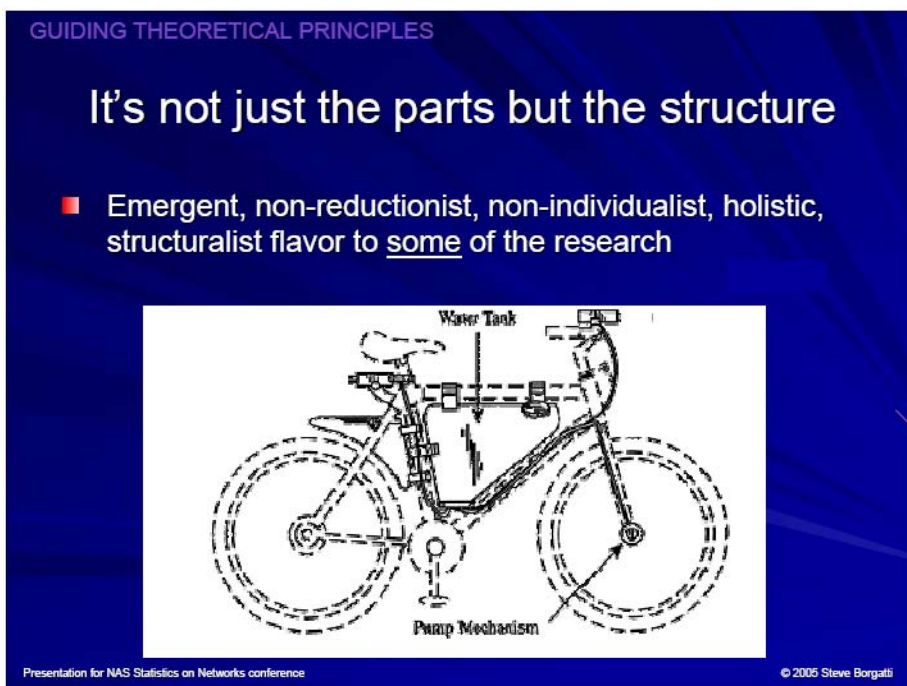



FIGURE 20

GUIDING THEORETICAL PRINCIPLES

Opportunities & Constraints

- A person's position in a social network (i.e., social capital) determines in part the set of opportunities and constraints they will encounter



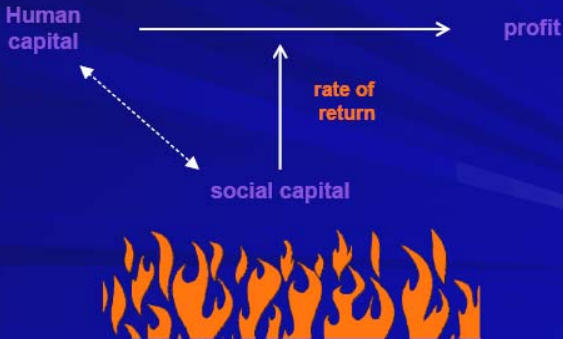
Presentation for NAS Statistics on Networks conference © 2005 Steve Borgatti

FIGURE 21

GUIDING THEORETICAL PRINCIPLES

Rate of return on human capital

- Burt: A person's connections determine the rate of return on human capital



Presentation for NAS Statistics on Networks conference © 2005 Steve Borgatti

FIGURE 22

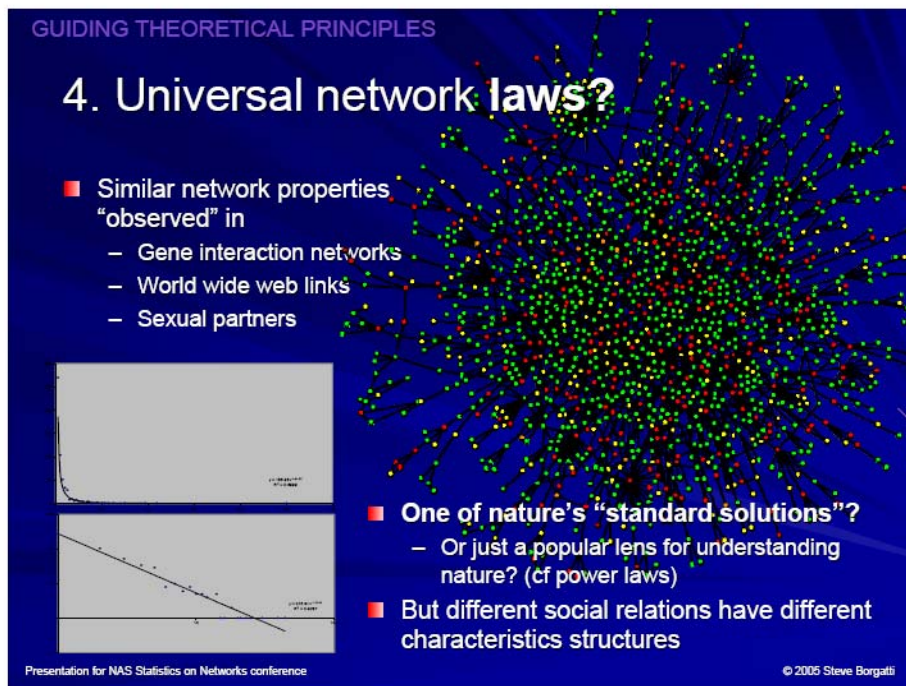


FIGURE 23

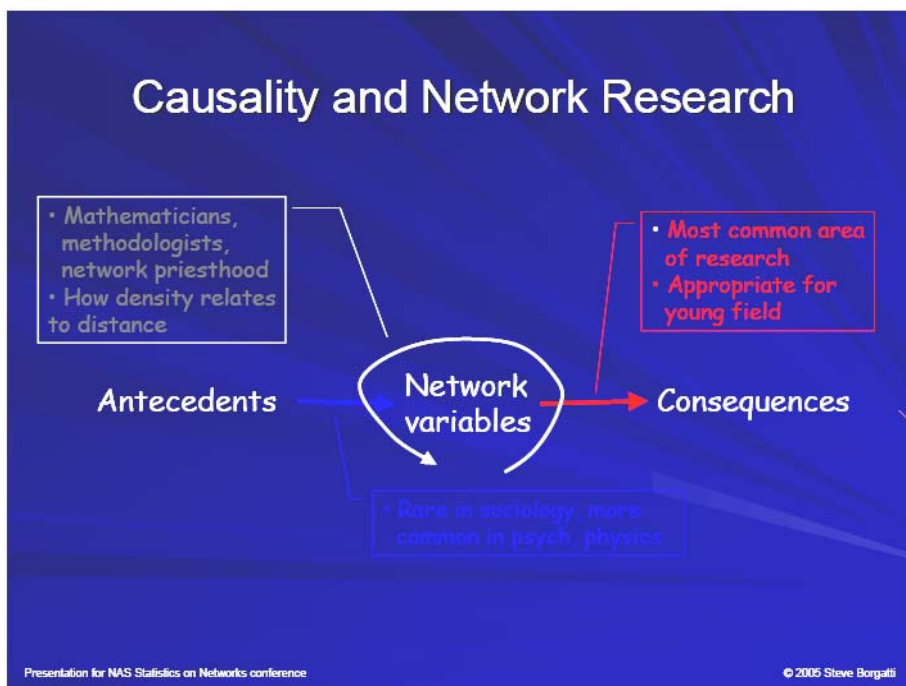


FIGURE 24

Consequences of Network Variables

Means \ Ends	Explaining Variance in Performance (social capital)	Explaining Social Homogeneity (adoption)
Connectionist mechanisms (flows thru ties)	Success comes from obtaining resources <u>through</u> social ties; It's who you know	People have same behavior because they directly influence each other & transmit ideas, beliefs, etc.
Structuralist mechanisms (emergent properties of topology)	Network positions /shapes provide opportunities for exploitation; It's how you know others	People have same behavior because their network positions are similar (and affect them similarly); same social environment

Borgatti, S.P. and Foster, P. 2003. The network paradigm in organizational research: A review and typology. *Journal of Management*. 29(6): 991-1013

Presentation for NAS Statistics on Networks conference © 2005 Steve Borgatti

FIGURE 25

Changes in the Field

<ul style="list-style-type: none"> ■ 25 years ago ... <ul style="list-style-type: none"> – Descriptive, methodological – Small datasets (< 100 nodes) – Structuralist cast – Focus on the consequences of network characteristics <ul style="list-style-type: none"> ■ Network is fixed ■ Cross-sectional data – Focus on the pattern of ties – Deterministic & analytical models – Inter-network comparisons 	<ul style="list-style-type: none"> ■ Now ... <ul style="list-style-type: none"> – Theory testing in soc sci – Large datasets 00s – 000s – Increasing attention to agency – Increasing attention to causes of network variables <ul style="list-style-type: none"> ■ Network change ■ Longitudinal data – Increasing interest in what flows through networks – Increasing interest in stochastic models & simulations – Comparison with theoretical baselines
--	--

Presentation for NAS Statistics on Networks conference © 2005 Steve Borgatti

FIGURE 26

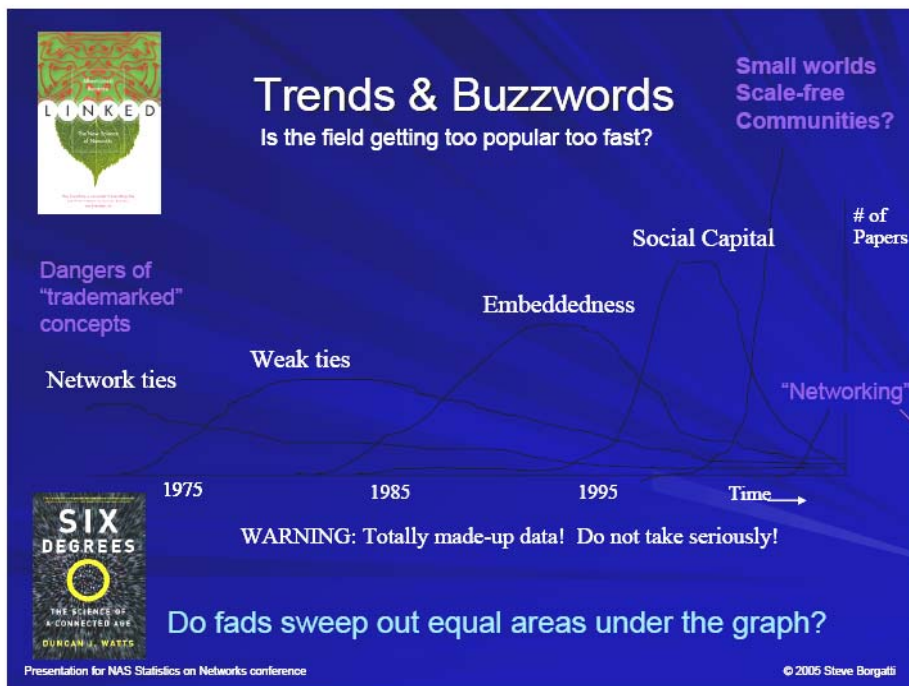


FIGURE 27

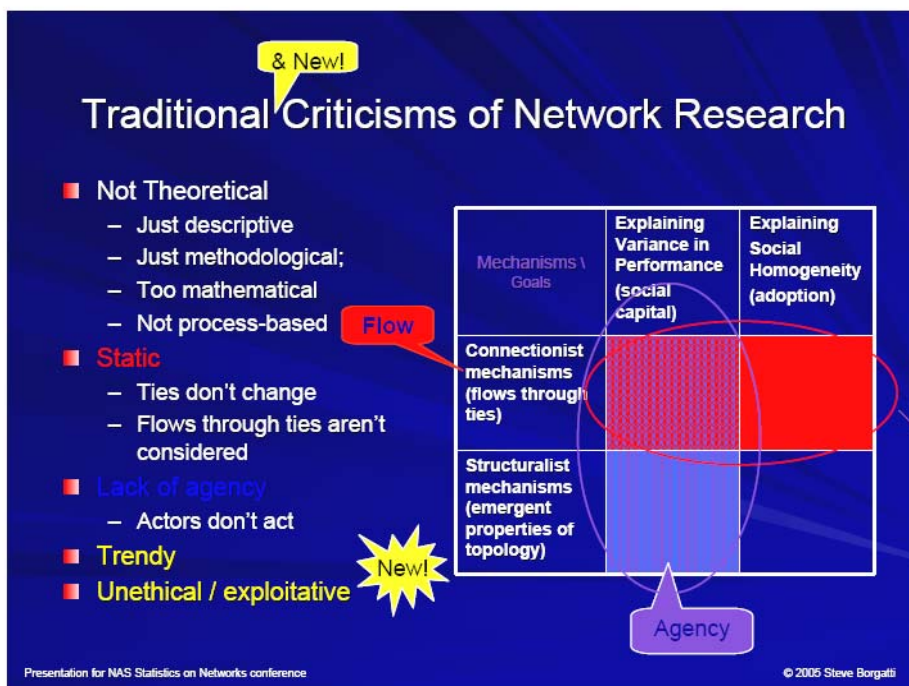


FIGURE 28

REFERENCES

Degenne, Alain and Michel Forse. 1999. *Introducing Social Networks*. London, U.K.: Sage Publications Ltd.

Kilduff, Martin, and Wenpin Tsai. 2004. *Social Networks and Organizations*. London, U.K.: Sage Publications Ltd.

Scott, John. 2000. *Social Network Analysis: A Handbook*. London, U.K.: Sage Publications Ltd.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, U.K.: Cambridge University Press.

Keynote Address, Day 2

Variability, Homeostasis per Contents and Compensation in Rhythmic Motor Networks

Eve Marder, Brandeis University

DR. MARDER: I'm a neuroscientist. I should say that I'm an experimental neuroscientist who has worked at various times and places with a variety of theorists, including my fellow speaker, Nancy Kopell. What I would like to do today is give you a sense of some problems and approaches where neuroscience and network science intersect. I realize that you'll think some of my take home messages are probably obvious, even though my neuroscience colleagues often get very disturbed and distressed and start foaming at the mouth.

The problem that we have been working on for the last 10 or 15 years, like all good problems, is both completely obvious and not. For this group it's basically the problem of how well-tuned do the parameters that underlie brain networks have to be for them to function correctly? If you were to have asked most experimental neuroscientists 20 years ago how well-tuned any given synapse would have to be, or how well-tuned the number of any kind of channel in the cell would have to be, they would have probably said 5 percent plus or minus. You should know that neuroscientists have traditionally had the perspective that things have to be very tightly regulated, and we can talk about why that is in the future.

About 10 or 15 years ago we started working on a problem that has since been renamed homeostasis and compensation. This comes back to something that is true in all biological systems, and not true in mechanical or engineering systems in the same way: almost all the neurons in your brain will have lived for your entire life, but every single membrane protein that gives rise to the ability of cells to signal is being replaced on a time scale of minutes, hours, days or weeks. The most long-lasting ones are sitting in the membrane for maybe a couple of weeks, which means that every single neuron and every single synapse is constantly rebuilding itself. This in turn means that you are faced with an incredible engineering task, which is how do you maintain stable brain function? I still remember how to name a tree and a daffodil and all the things I learned as a child, despite the fact that my brain is constantly, in the microstructure, rebuilding itself. When we started thinking about this, and this goes back a number of years, Larry Abbott and I worked on a series of models that basically used very simple negative feedback homeostatic mechanisms to try and understand how you could get stable neuronal function despite turnover and perturbations. What that led to was going back to step one and saying, how well do negative-feedback self-tuning stability mechanisms have to work? You have

to know how well-tuned do any of the parameters need to be in order to get the nervous system to do a given task. That's what I'm going to talk to you about today.¹

There has been a series of both theoretical and experimental investigations into the question of what really are the constraints. How well does a circuit have to tune all of its parameters in order to give an adequate or a good-enough output? By good enough, I mean good enough so that we can all presumably see a tree, name a tree, although we also all know that my brain and every one of your brains are different. (I would like to acknowledge the work I'll be showing you today of three post-docs in the lab—Dirk Bucher, Jean-Marc Goaillard, and Adam Taylor—who have done a lot of this work, along with two ex-post-docs who are now faculty, Astrid Prinz and Dave Schulz, and a colleague of mine, Tim Hickey, from the computer science department.)

I'm going to show you a little bit about the maintenance of central pattern generators (CPGs). Those are networks that produce rhythmic movements and that function despite growth. Among adult animals, we ask how much animal-to-animal variation there is in network mechanisms and underlying parameters, and then we will talk about the problem of parameter tuning: how tightly the parameters that determine neuronal activity and network dynamics have to be tuned and whether stable network function arises from tightly controlled elements or because elements can compensate for each other.

CPGs are groups of neurons in your spinal cord or brain stem. In this case I'm going to work in lobsters and crabs, and the particular pattern we will talk about runs a portion of a lobster or crab's stomach. As shown in Figure 1, if you drill holes in the carapace and put wires into the muscles of what is called the pylorus or the pyloric region of the stomach, you see one-two-three, one-two-three, one-two-three, which is a rhythmic discharge to those muscles that persists throughout the animal's life. Here you see it in these muscle recordings. When you dissect the nervous system that produces this motor pattern out, you can record intracellularly from the stomatogastric ganglion, which is responsible for producing this rhythmic discharge. In the lower left part of Figure 1 you see three intracellular electrodes from the cell bodies which show that one-two-three, one-two-three rhythmic discharge pattern. This discharge pattern really is a constrictor one phase, a constrictor two phase, and a dilator phase. You can see this either in extracellular recordings from the motor nerves, or intracellular nerves for the pattern.

¹ See also a recent review article on this same material, Marder, E., and J.M. Goaillard. 2006. Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience* 7:563:574.

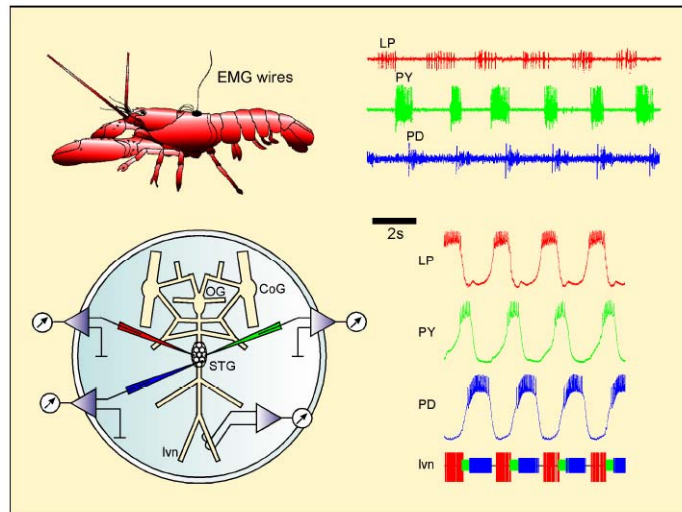


FIGURE 1

For many years people have tried to understand why this discharge exists and how the particular neurons and their connectivity give rise to things like the frequency and the phasing of all the neurons in the network that do this. The proof that this is a central pattern generator lies in the fact that the rhythm persists even when we have discarded most of the animal and are left with just the stomatogastric ganglion.

As neuroscientists we worry a lot about things like why the pattern has the frequency it has and why all these specific timing cues are there the way they are. In Figure 2 we see the circuit diagram that gives rise to this pattern. In these circuit diagrams, we use resistor symbols to mean that cells are electrically coupled by gap junctions, so current can flow in both directions. These symbols are chemical inhibitory synapses, which means that when one cell is depolarized or fires, it will inhibit the follower cell. This is the circuit diagram for the dilator neurons, the constrictor neurons. None of you could a priori predict what this circuit would do, because there is information missing in the connectivity diagram. I think that's really an important thing for those who worry about dynamics to think about. What is missing in this connectivity diagram is all the information about the strength and time course of the synaptic inputs. Additionally, and as importantly, all the information about the dynamics of all the voltage and time-dependent currents in each of these cells that give them their particular intrinsic firing properties will then be sculpted by their position in the network.

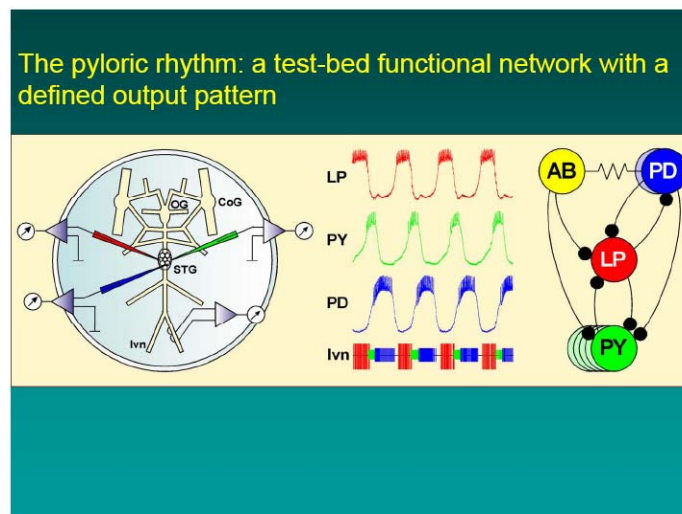


FIGURE 2

So, as neuroscientists, what we have to do if we want to understand how this connectivity diagram gives rise to this particular pattern is ask what kinds of voltage and time-dependent currents does each one of these cells have? Then try and understand how all of those currents act together to give rise to that cell's own intrinsic electrophysiological signature. Also to understand how they interact with the synaptic currents to end up producing this rhythmic pattern.

I can give you the short answer. The short answer is that the AB neuron shown in Figure 2 is an oscillatory neuron that behaves a lot like this in isolation. It's electrically coupled to the PD neuron, and it's because of that electrical coupling that the PD neuron shows this depolarization burst of action potentials: hyperpolarization, depolarization, et cetera. Together these two neurons rhythmically inhibit the LP and the PY neurons, and so force them to fire out of phase. The LP neuron recovers from that inhibition first, because of the timing cues or the inhibition that it receives, and because of its own membrane conductances. I can tell you as a first approximation why that occurs, but it took a long time to even get to that first approximation. That said, the question is, for this network, how tightly tuned do these synapses have to be in terms of strength, and how tightly tuned do all the membranes currently in each of these cells have to be? What I'm going to do is first show you some biological data that will constrain the question, and we'll talk a bit about some models that frame part of the answer. I'll go back to some recent biological experiments that go directly after the predictions of some of the models.

The first little piece of data I'm going to show you are from some juvenile and adult lobsters. As lobsters mature to adulthood, they grow considerably, and obviously their nervous

system changes size as well. Figure 3 shows a juvenile's PD neuron (on the left) and one from an adult (on the right). You can see that the cells have a certain look that has been maintained through that size change. Those of you who are neuroscientists realize that as the cell grows from small to large size, every feature of its electrical properties has to change because its cables are growing—that is to say, the distance signals will travel. Positions of synapses have to change, therefore, the number of channels in the membrane have to change. All of these things have to change just as when you go from a 2-year-old or a 14-month-old who is learning how to walk, the whole peripheral plant is changing as the 14-month-old grows into an adult. It's a challenge to say how this has happens if it has to maintain constant physiological function.

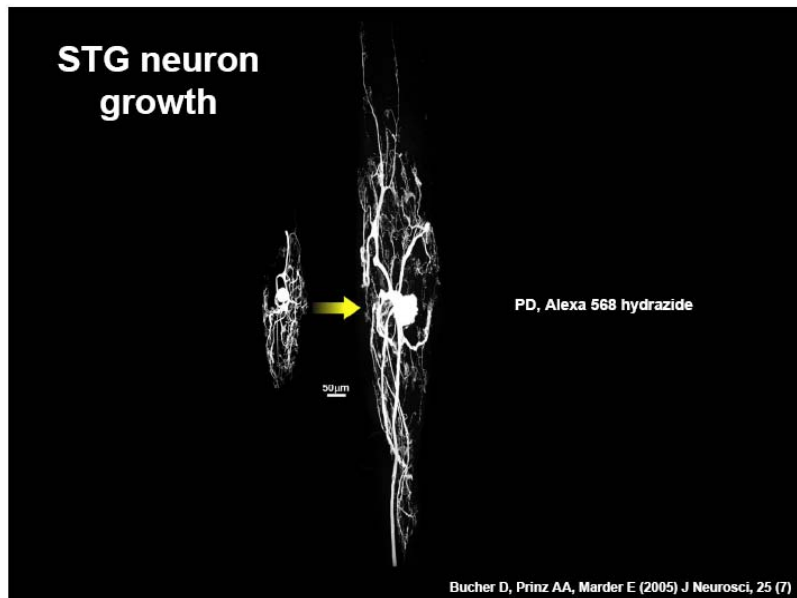


FIGURE 3

Here in Figure 4 are some intracellular recordings, some made from a juvenile, some from an adult. You can see by your eye that the waveforms and patterns look virtually identical between that baby lobster and the big lobster, which tells you that there has to be a mechanism that maintains the stable function despite a massive amount of growth, and therefore a massive change in any one of those properties.

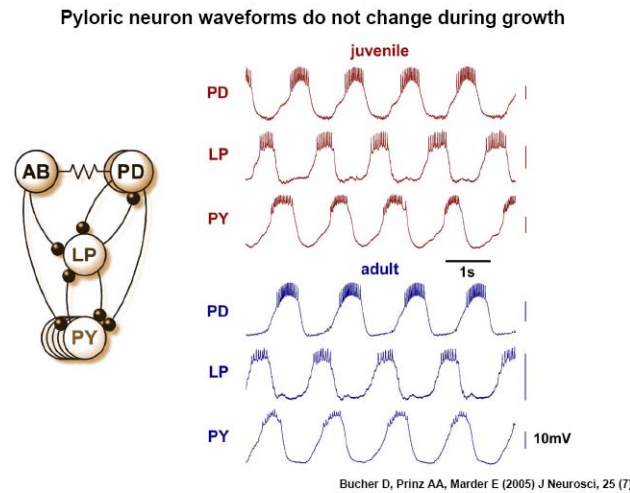
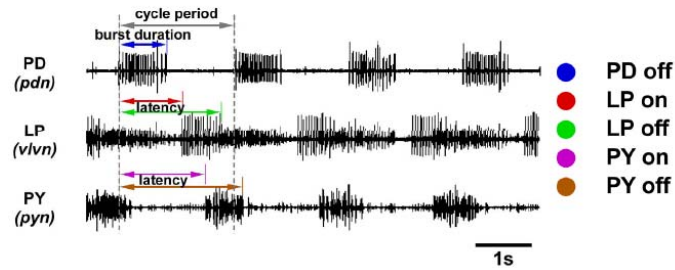


FIGURE 4

The question this poses is that there have to be clearly multiple solutions to producing the same pattern, or very similar patterns, because in the small animal and the big animal, many of the parameters are actually different, as are the number of channels. This tells you that at least during natural life, the animal manages to find its way from one solution and to grow its way continuously without making mistakes that shift it into other whole classes of solutions.

To go beyond making the simple assertion, I would like to show you what we realized we have to do, which was to ask to what's the variation among individual animals. Basically, this defines the range of normal motor patterns in the population, and what we are going to do to quantify the motor patterns is look at these extracellular recordings from the motor nerves showing one-two-three, the discharge pattern we have been looking at, that triphasic motor pattern. We can obviously measure the period. We can measure the duration of the bursts. We can measure latencies and the phase relationships of all the cells in the network. (See Figure 5.)

Quantifying the temporal structure of the pyloric rhythm

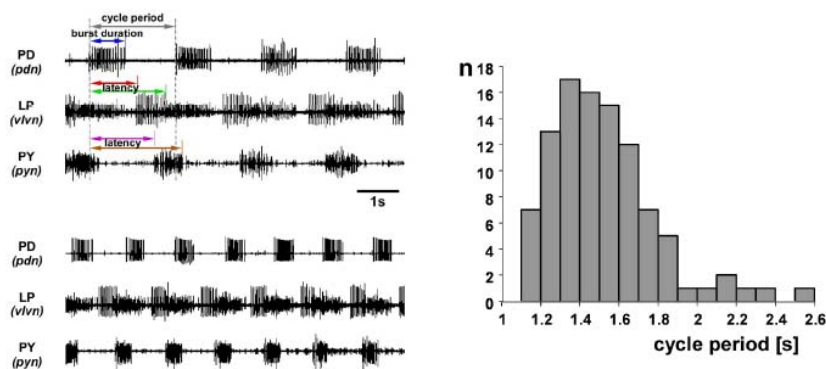


Bucher et al., 2005

FIGURE 5

The first thing I would like to show you is that if we look at different animals, the mean cycle periods vary about two-fold. Figure 6 shows a histogram of the cycle periods in 99 animals.

Mean cycle periods vary about 2-fold (n=99)



Bucher et al., 2005

FIGURE 6

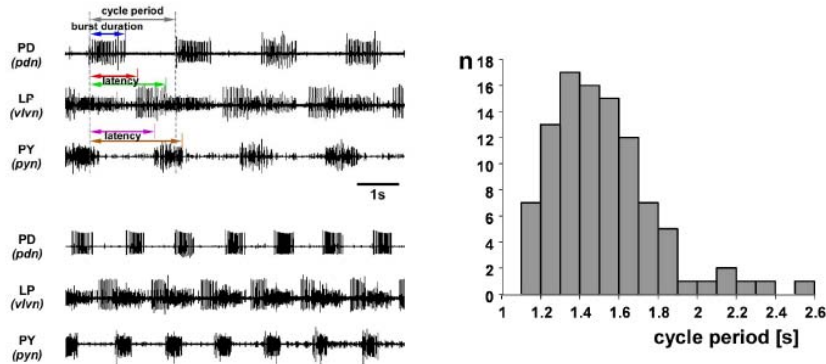
DR. BANKS: I could imagine that every lobster is more than its own sort of starting point cycle time, and with age that cycle time slowly increases or slowly decreases. I think that type of trend would be invisible if you do destructive testing.

DR. MARDER: We do destructive testing so we have actually done a great deal of work on embryonic animals. That is a whole other story. This system is built in the embryo. It has to be put in place before the animals hatch, because these motor patterns need to be ready for the animal before it starts to eat. In the babies, we tend to see more variability, and to some degree slower patterns, which then consolidate so there is a whole other story about how things change very early in development.

DR. BANKS: You don't know that this pattern is stable over the entire portion of its life. It could change.

DR. MARDER: There are two answers to why we think it's stable within the range of the population. I don't believe for a moment that an animal keeps an identical period over time. All I'm saying is it's going to be within that range, and presumably as things are tuning continuously, it may wander round within the tolerance level, and we can't monitor it long enough.

Mean cycle periods vary about 2-fold (n=99)



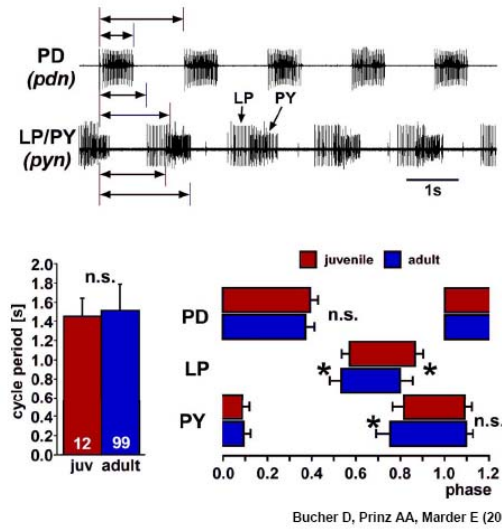
Bucher et al., 2005

FIGURE 7

Figure 7 shows the phase relationships, where we are looking at the time in a cycle where a cell starts firing and ends firing. And again, you see that over this very large range of periods, the phase relationships are remarkably constant. This in and of itself is actually a very challenging problem, because getting constant phase relationships over a large frequency range with constant time constant events, which is basically what you have, is tricky. Now, thanks to really beautiful work by one of my former postdocs, Farzan Nadim, I think we really do understand the mechanisms underlying this.

Now, if we look at these measures in the juvenile and the adult animals (Figure 8), we see the cycle periods are basically the same. Most of the phase relationships are very close, if not perfectly the same, and we just had fewer juvenile animals than adults, because they are very hard to come by. My guess is if we had 99 of both of them, they would all fit in totally the same parameter regime.

Similar bursting patterns in juvenile and adult lobsters



Bucher D, Prinz AA, Marder E (2005) J Neurosci, 25 (7)

FIGURE 8

That sets the stage for what we eventually want to account for, motor patterns with pretty tightly constrained, although not perfectly identical output patterns in different individuals. We now want to go after the structure of the underlying conductances.

To do so, I'd like to step back and spend some time on the single neuron. A lot of what I'm going to say is true of all neurons. Neurons, when you just record from them, can have a variety of different behaviors. Some of them can just be silent, and we'll call those silent neurons. Some of them will just fire single action potentials more or less randomly, and I will call those tonically firing neurons. Other neurons will fire what I'll call bursts, that is to say they will depolarize, fire bursts of action potentials, and then they will hyperpolarize. They will fire these clumped action potentials.

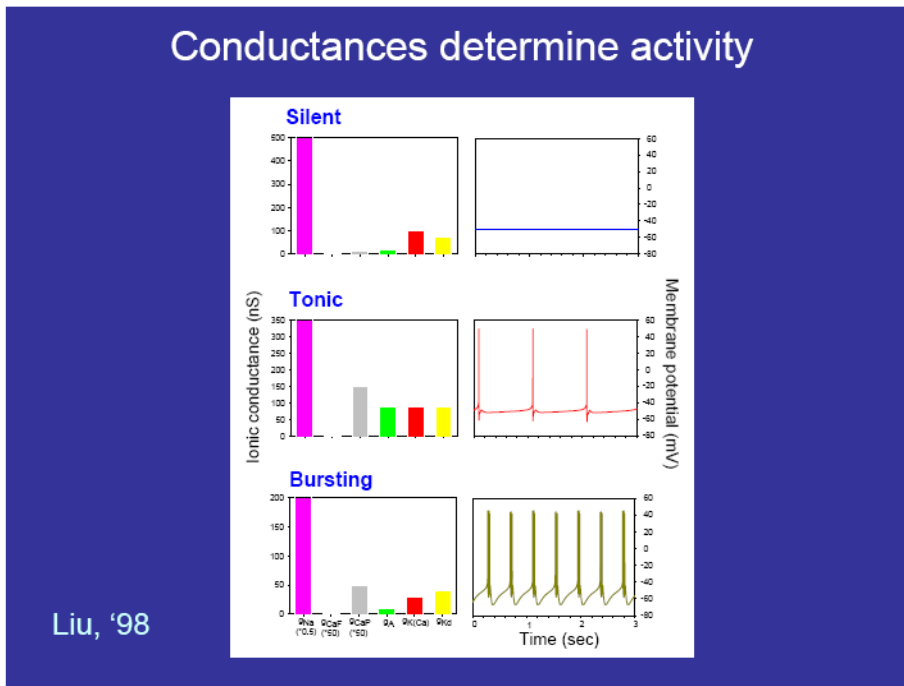
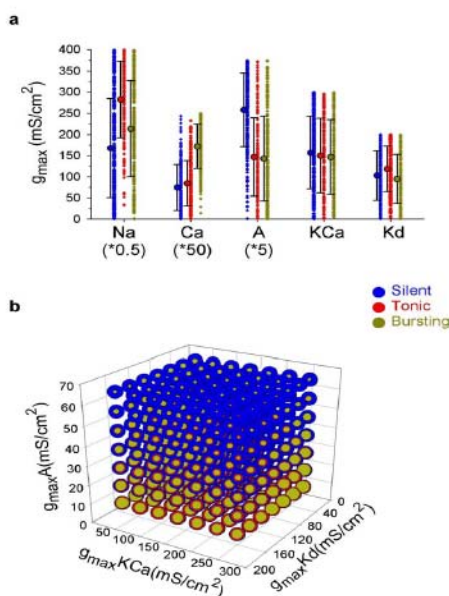


FIGURE 9

Figure 9, which was created by Zheng Liu when he was a graduate student of Larry Abbott, shows something that neuroscientists all take for granted, which is that as you change the conductance density or the channel density for all the different kinds of channels in the membrane, these behaviors can change. I should tell those of you who aren't neuroscientists that any given neuron in your brain might have—you probably all know about the classic sodium channel and the potassium channel in the Hodgkin-Huxley type axon where sodium carries current in, and potassium is partly responsible for the repolarization of the action potential. Real neurons have 6, 8, 10, 12, and 18 different kinds of voltage and time-dependent conductances. Some carry calcium, some of those carry chloride. Sometimes they are cationic conductances, sometimes they are gated by other things. Each one of these has its own characteristic voltage and time dependence that can be characterized in very much the same way by differential equations of the same form as Hodgkin-Huxley characterized the action potential equations. Zheng made a model to make these traces that contained a sodium current, a conventional one, two different kinds of calcium currents, and three different kinds of potassium currents. Then he varied the relative density of those currents, and that's what gave these different characteristic amounts of activity.

What we experimental neuroscientists do is use a method called voltage clamp to actually isolate, and then measure each of the currents. They are measured one by one in the cell. We can then fit those experimental data to differential equations that characterize those properties. Then we put all the currents from a cell back together again in a simulation and try to recover the properties of the cell. So, that's where the data in Figure 10 come from. These models can be very useful in giving you intuition about what goes on, which combinations of currents must be there to determine these properties, and so on.



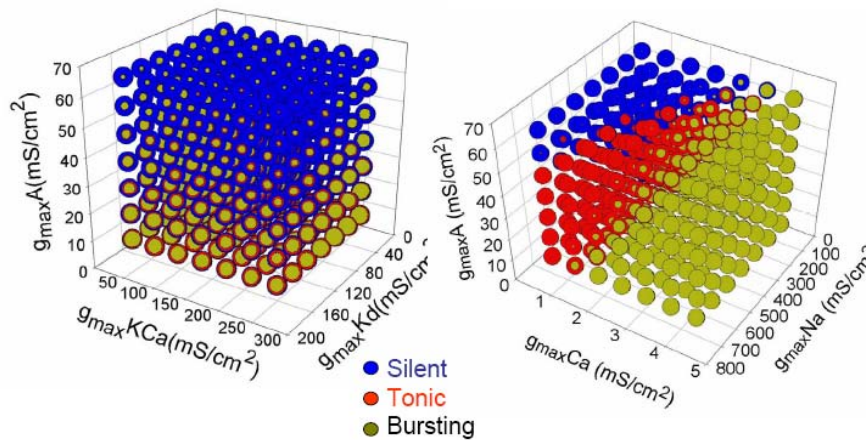
Goldman et al '01

FIGURE 10

The first thing that I would like to show you from these sorts of experiments is illustrated in Figure 10. These data come from Mark Goldman, who was Larry Abbott's graduate student and who is now on the faculty at Wellesley. They represent a model very much like the one I just showed you, one with a sodium current, a calcium current, and three different kinds of potassium currents. What Mark found, and this is really quite interesting, is he could have silent cells that have either low or high sodium currents. He had tonic firing cells that could either have low or high sodium currents, and bursting cells that could either be low or high. And the same thing was true for any one of the five currents in the model. What this tells you, contrary to what most neuroscientists believe, is that knowing any single current by itself will not be adequate for telling

you what the behavior of the cell will be.

Activity patterns are determined
by specific combinations of conductances



Goldman et al 2001

FIGURE 11

It turns out that if you plot the three potassium currents against each other on a 3-dimensional plot, as in Figure 11, you can see all these areas where you can have regions of silent, tonic firing, or bursting. So, even the values of those three potassium currents are insufficient to basically predict what the cell's behavior. On the other hand, the space partitions better if you plot one of those potassium currents against the calcium current and the sodium current. You can then see the space partitions nicely. This tells you that in this particular model, you have to know the correlated values of three of the five, and it has to be this subset of the three out of the five, which means it is actually a very inconvenient answer, because it makes it difficult to do many kinds of experiments.

We are going to take this approach and go one or two steps further. Figure 12 shows work done by Astrid Prinz. She wanted to build a really good model for the pyloric rhythm, because she wanted to build a nice model of that cell self-regulating or homeostatically regulating. She wanted good models for the PD, LP, and PY cells, so she began by hand tuning. She got really frustrated, because hand tuning is unsatisfying, because it's very hard when you have eight or so, as we have here, conductances in the cell.

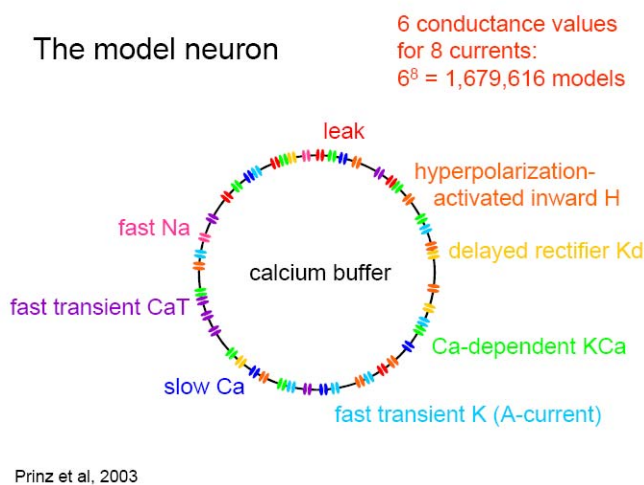


FIGURE 12

Instead of that, she decided to use brute force. There are eight different currents in the cell. We have a leak current, and a current I haven't spoken to you about before. It's a hyperpolarization activated inward current, three different potassium currents, two different calcium currents and sodium current. She decided to use brute force, and took six values for each of those currents, one of the six being zero in all cases. She then simulated combinations of all, and created a database of model neurons to span the full range of behaviors. So, that was the 6th to the 8th, or 1.7 million versions of the model. She ran them all and did this without telling me, because I would have screamed. (All good things that happen in my lab happen without me knowing about them, I kid you not.) She then wrote algorithms to search through the behavior of those cells, and to classify their behaviors in terms of whether they were tonic, firing, or bursting, and what they did.

She saved enough of the data so that we could search that database and find different neurons or different behaviors.

DR. BANKS: Is it absolutely clear that you distinguish a tonic from a bursting nerve?

DR. MARDER: You can make up criteria that will distinguish. We'll come back to that.

What do you see when you do that analysis? Obviously you see cells of all different kinds of behaviors, as shown in Figure 13. There is a single-spike burster, which has a single spike and then a long plateau phase. There is a more conventional burster, and there are just other bizarre things.

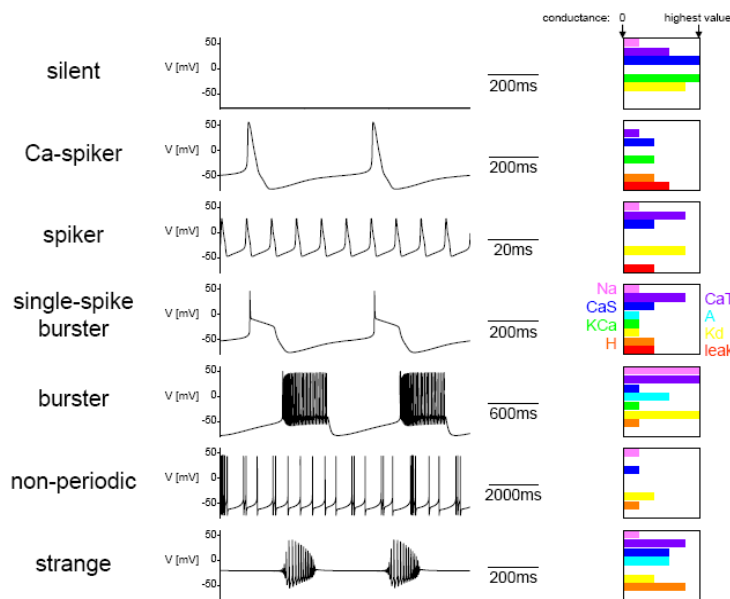


FIGURE 13

Astrid spent a fair bit of time classifying those neurons, but the real question which comes back to what we are asking is how do these partition in conductance space, what are the overall structures? What is of real importance to us, if we think about cells trying to homeostatically tune, is you want to know whether you have a structure like this. In two dimensions, it's real easy to see, or like this, where you have multiple regions of disconnected groups that can all be classified as bursters, but obviously this is going to be more problematic than if the space were better behaved. Obviously, this you all know.

What I'm going to do is very briefly show you a new way we can visualize what is going on in 8-dimensional spaces. It's useful to be able to bring eight dimensions down into 2-D, which we do through something called dimensional stacking. You start out for two conductances, everything else set to zero, which yields a plot. Then we take that grid and embed it in another grid which would have those full range of behaviors embedded for all these other values in two dimensions, and we would repeat that grid over and over again. Now we would have all of those four dimensions represented on the plot, and then you repeat that, and repeat it again and again until you have all the conductances plotted where you have one large and then smaller and smaller and smaller sets of repeating units, where everything has dropped down and repeated over and over again. This gives you all of those 1.7 million neurons in principle, represented on one plot. The real key is to use color to represent their activities.

Because there are 40,000 different variations in the way you can do the stacking—e.g.,

what axis you put as the smallest or the biggest—we asked ourselves what would be a most useful stacking order. Tim Hickey and Adam Taylor, who actually did a lot of this work, decided that maximizing large blocks of color would mean there were fewer transitions, and might allow you to see where different patterns of activity were best.

To go back to this issue of connectedness, in this particular model, what do these parameter spaces look like? What Adam Taylor did was to ask in the 8-dimensional space whether he could find all the cells that were nearest neighbors of one another by only varying one value of one current at a time. He then defined those as a population. It turns out that of all the tonically firing neurons, 99.96 percent of them can be found in one connected parameter space region. We were very surprised and pleased by that. 100 percent of the silent neurons are found in a connected region of parameter space. The bursters that have 6-10 spikes per burst, 99.3 percent of them are in one region of connected space, and there are all these little regions down where many of the others sit.

Surprisingly, almost all of the models in each of the three major activity types are in single islands, which means that if a cell is going to have to tune, it is actually going to have a much easier job, because it can stay moving around, making minor modifications of one or more currents, but it doesn't have to find 4 or 12 different isolated islands, and therefore cross into very different regions of activity in order to get there.

I would now like to show you one thing about our single spike bursters. Of all the single spike bursters, just 99-point-something percent of them are in a single island of connected space. The underlying conductances are in different regions of the map. Indeed, each one of these cells, which to an electrophysiologist would look virtually identical, actually vary considerably in their underlying conductances. So if I start looking at any one of these currents, you can see that all of these models, which have very similar behavior, have nonetheless very different underlying structure. That says that widely disparate solutions producing similar behavior at the single cell level may be found in a continuous region of parameter space, and therefore that relatively simple tuning rules and targeting activity levels might be used to tune cells.

At the level of the network, Astrid Prinz did the same sort of thing that we just did at the level of the single neuron, but she did it to try and form a pyloric rhythm. She took five candidate AB-PD cells from the single-cell database, and she wanted to use different ones to make sure that the results were not too idiosyncratic to any one model, and six versions of a PY neuron, 5 of LP neuron. She then took all the synaptic connections and simulated either five or six of each of these. She ended up simulating more than 20 million versions of the three-cell network. Now we can go in and see what those do.

The results include some answers where you get a tri-phasic rhythm in the right order. You get other answers where you don't get a tri-phasic rhythm at all, and then a whole variety of other kinds of mistakes. The tri-phasic rhythm is in the wrong order. It's almost a tri-phasic rhythm, but one of the cells doesn't follow. You can get really bad solutions and better solutions. There you have different model neurons of the same synaptic strengths, and again you get the full range of behaviors. This tells you that as you vary either synaptic strength or the intrinsic properties of the individual cells, you can get a variety of different network outputs, which I'm sure you all understand.

Now the question is how many of these meet the criteria for being a good model of pyloric rhythm using the biological criteria that we established in the very beginning of the talk? Of the full set of all networks, 19 percent of them were pyloric-like. That is to say, they were tri-phasic in the right order. If we go in and use the criteria from the biological database, and we use those to pick out all the model networks that meet all of these criteria, 2.4 percent of them were in the right frequency range, had the right duty cycles, the right phase relationships, and so on; there were something like 12 or 15 criteria that led us to that subset of networks.

We can now ask of that 2.4 percent that fit within a range that we would expect to be relevant to the biological system, what do they look like in their underlying structure? Analysis shows that there are large differences between the specifics of the conductances for the two model networks. For instance, the synapse conductance from PD to PY is much greater in model network 2 than for model network 1. Similarly, the AB-to-LP synapse is considerably different between the two model networks, as is the case for many of the other conductances. So, if you look at all of these, you say even though the behavior of the two networks is very similar, there are very disparate sets of underlying conductances giving rise to that very similar behavior. What this tells the biologist is that there are probably a whole series of compensating changes so that one parameter over here is compensated for by a correlated or a compensating change over there. The interesting question is to ask, what set of correlations and multi-dimensional correlations are there that give rise to these viable different solutions?

In the model we see parameters that can vary over very large ranges. Some of these parameters in these successful solutions may vary 5-fold, 10-fold, or 20-fold, and this is saying that if you have enough dimensionality, even if you have a very restricted output, you can come up with a large range of solutions. If we look now at 633 models with very similar outputs, we see that none of them have a small PY-to-LP synapse, and none of them have a large LP-to-PY synapse, so these strengths are obviously important to producing that particularly output. On the other hand, the AB-PY synapse is found either small or large in the population of networks, so

there are presumably some things that are much more important or more determinative than that conductance.

These data were very interesting to us. Now we have the responsibility to go back to the animal and measure the synapse from LP to PD in 20 different animals to see how variable it is in reality. Or if we measure the amount of potassium current in 20 animals, how variable is it? So, I'm just going to finish off with some new experiments that basically say given the conclusions of this sort of analysis, the prediction would be that each individual animal will have found a different solution set. Therefore, if we measure a given parameter, we should be able to see that variation in levels, so that's what we are going to do. We're going to measure some of those things, and we're going to do this in crabs.

We looked at two intracellular recordings from the LP neuron, and two ongoing rhythms. There is one LP neuron in each animal so it's very easy to compare them. Superimposing them shows that their wave forms and dynamics in the network are almost identical. But when we measured the amount of the potassium currents, one can see differences in the amount of calcium-activated potassium current. Looking at 8-9 animals, we found a range of values, about 2-fold to 4-fold. This range is very similar to what most biophysicists see in most cell types. Experimentalists have always attributed that range to experimental error, but if you ask biophysicists to go back and tell you, they will all say, oh yes, 1.5-3 or 2-4 or something like that. It's a very common range.

To conclude, I would just like to say that taken at face value, these data would say that it's very important for an animal to maintain stable physiological function throughout its lifetime, even though the nervous system has to constantly rebuild itself. Part of what we see is a tremendous heterogeneity in solutions found by individual animals, and probably by the same animal at a different time in its life. What we eventually have to understand is which parameters are very tightly controlled, which are co-regulated, which are coupled, what things are free to vary, what things don't matter. And then we have to understand the rules that allow cells to wander around in their protected regions of parameter space as they are constantly rebuilding themselves to maintain our networks functioning throughout our life.

QUESTIONS AND ANSWERS

DR. DING: I would like to ask two questions. One is, are there multiple PD cells? You said for each cell type there is only one?

DR. MARDER: There are two PD cells in every animal; there is one LP cell. There are four GM cells, et cetera.

DR. DING: Okay. Another question is, for each of the model neurons that you studied, over a million neurons studied, presumably these neurons are very highly nonlinear units. Couldn't they be capable of multiple behaviors?

DR. MARDER: Under different conditions. If you inject current, yes. But I just showed you their unperturbed behavior. Presumably, if you do a whole variety of different perturbations, they will change their activity patterns, and you could then select for all the cells that were bursting, unless you do X, in which case they might have made a transition to bi-stability or something like that.

DR. KLEINFELD: I had the same question. Because you cited your work and Ron Harris Warwick's work when you add small modulators to the animal, the animal has its own endogenous modulators, you change the firing properties of cells. Because the firing properties of the cells change with endogenous modulators, in fact you showed like 2.4 percent of all possible solutions were biologically relevant in these steady state conditions. Do you in fact get a much smaller range of solutions if the cells actually have to track the effective modulators?

DR. MARDER: David has asked that question much more politely than some people have asked it. There is a really interesting question about neuromodulation. With neuromodulation you apply a chemical, and you might alter one of the conductances 30 percent, and you may see a very big difference in the cell's properties. One piece of the question is, if the population has a 2-fold to 4-fold range, why do you see an effect at all with a 30 percent modulation? And that's because 30 percent from one of those solutions can still, with all the other compensating currents, give you an effect, even though that 30 percent variation would still be part of the expected range within the population.

I think the same thing is true of the network. What we know from our old data is that many neuromodulators have what we like to call state-dependent actions. That is to say they had much stronger actions on networks that were going slowly than they did on networks that were going quickly. Part of the answer is that every network is in a slightly different configuration. Modulators may always take them in the same direction, but the extent to which they will take them will depend on the underlying structure.

One of the things that we're trying to do now is address this question in the models to really see whether many modulators go after multiple currents and multiple sites of action. So, what we think is going on is that modulators may find correlated sets of currents and synapses to act on so that the whole system can go in the right direction, but it gets more or less of that

movement because of its action on multiple different currents in different sites.

DR. KLEINFELD: So, you're saying the modulator should cause things to co-vary in principle a little like you see in these analyses?

DR. MARDER: Or the modulators have been chosen to go after the right set of compensating or cooperative currents and parameters.

DR. KLEINFELD: So, the second question was does number one imply that if you added an interfering RNA, that you should be able to drive the from—

DR. MARDER: Yes, if one could get the interfering RNA to work in your lobster that is sitting at 12 degrees and has really large pools of message and protein, yes.

DR. DOYLE: So, I sort of just want to make a quick comment tying it up with my talk yesterday. If you were to do the same kind of analysis on a range of clocks, you would presumably have found that the parameter space that gave good behavior was very small. And so, early in the development of timepieces, it was very difficult to get clocks that were accurate. And if you take and do the same analysis on this, you would find that almost none of the parameters matter much at all, and that there are a few that have to be perfect more or less.

I have only studied bacteria, but bacteria appear to do the same thing that the advanced technologies do. In advanced technologies, like my digital watch, you use the network architecture to create extreme robustness to the things that are hard to regulate well. And you make extremely fragile architectures for things that are easy to regulate. So the reason digital watches are so cheap is there is a large amount of integrated circuitry in there that does all of the extra stuff, and it's cheap and sloppy. And there are a few little things that can be manufactured to very high tolerances also cheaply. By structuring the network in the right way, you can combine these very sloppy things that are cheap to do with a few things that are very precise, and the consequences are that the network then has incredible robustness. We know that everything that has been studied in bacteria seems to hold that. I don't know if it holds here. But it looks as though this is a standard thing in biology and advanced technologies, but not in primitive technologies. It's only in our most advanced technologies that we are doing that.

DR. MARDER: My guess is that anything useful from bacteria the nervous system has kept and used, because that's how brains were able to develop. I think that's really cool. The advantage of working in bacteria is you have fantastic genetics. It's going to be much harder for us neuroscientists to get some of those answers than it was for you.

DR. POULOS: I'm Steve Poulos from the NSA. What keeps these cells pyloric?

DR. MARDER: The way I like to think about it is that early in development these cells are determined to have a certain set of firing properties, and that is to say certain genes are turned

on that says I want to burst at approximately this frequency. Those then become set points for a series of homeostatic tuning.

REFERENCES

Bucher, D., A.A. Prinz, and E. Marder. 2005. "Animal-to-animal variability in motor pattern production in adults and during growth." *Journal of Neuroscience* 25:1611-1619.

Goldman, M.S., J. Golowasch, E. Marder, and L.F. Abbott. 2001. "Global structure, robustness, and modulation of neuronal models." *Journal of Neuroscience* 21:5229-5238.

Liu, Z., J. Golowasch, E. Marder, and L.F. Abbott. 1998. "A model neuron with activity-dependent conductances regulated by multiple calcium sensors." *Journal of Neuroscience* 18:2309-2320.

Prinz, A.A., D. Bucher, and E. Marder. 2004. "Similar network activity from disparate circuit parameters." *Nature Neuroscience* 7:1345-1352.

Dynamics and Resilience of Blood Flow in Cortical Microvessels

David Kleinfeld, University of California at San Diego

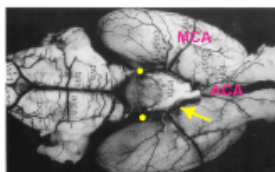
DR. KLEINFELD: Like Eve Marder, I'm also a neuroscientist, but I'm going to talk to you about blood flow. Some years ago we used some optical techniques invented by Larry Cohen to study the cortex, which Eve uses as well. That data came out with unexpectedly large variance as a result of blood flow. This gave rise to a series of studies—both measurement studies and perturbation studies on blood flow in cortex. What I'm going to talk about today is work that is aimed at understanding the relationship between the topology of the vasculature, which are graph-like structures, and the dynamics of blood flow within the cortical vasculature. It's interesting, because as you look at this problem from a physics or math point of view, you see highly interconnected networks; and the first thing you think about is percolation networks. You can add defects to these networks, and they should keep working until some critical junction occurs, and then they will stop flowing. If you talk to a neurologist or a stroke doctor, they also have this same notion, that you are constantly building up defects throughout your life in your cortical vasculature, and every so often you get these small, little microstrokes that occur throughout your life.

I broke the talk into three topics. First, I want to warm everybody up to just how violent the world of blood flow is in your brain, and also introduce you to a little bit of the technology that we use to measure flow in the cortex. In the spirit of using optics not just as a tool to visualize things but also as a tool to perturb, we will talk about two sets of experiments. One takes place on the vasculature and runs across the top of your neocortex and actually supplies blood to individual cortical columns. Another is about a separate network that exists in three dimensions and is within the bulk of the cortex itself, and how one could perturb flow in that region, and what the nature of that response is. So, the idea is measure and perturb and measure again.

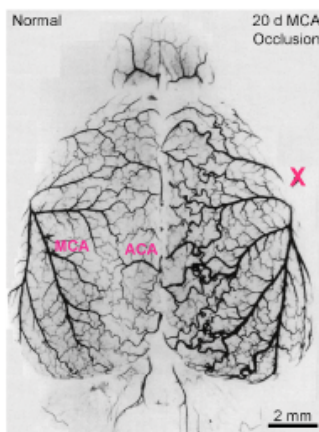
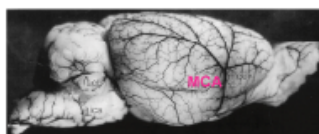
Just for a little bit of background, Figure 1 shows a rat's brain. Just to give you some flavor in terms of your own anatomy, there are four carotid arteries that come up your neck. Two of them feed directly into this thing called the Circle of Willis that is on the bottom of your brain. It makes your brain somewhat resilient. If somebody tries to choke you on one side, you'll get full flow of blood on the other side so there is a level of redundancy built in. Coming out of the Circle of Willis are another set of cerebral arteries. One in particular is called the middle cerebral

artery that is labeled in the middle image in Figure 1, and this comes around the side of the cortex and branches as it comes off. In fact, it meets up with another artery that comes through the midline called the anterior cerebral artery. The point is that the redundancy actually begins to break down, and if a block is made in the main trunk of the middle cerebral artery, which has been a favorite model system of the neurology community, large swathes of your brain will start to die off.

Architectonics of the Circle of Willis (arterial supply to the brain)



Architectonics of the Middle Cerebral Artery



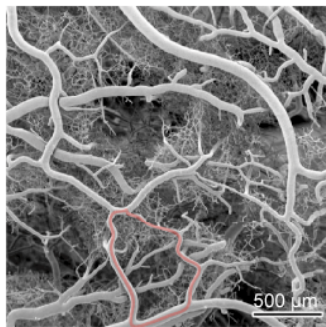
Data from the review by O. U. Scremin (1995)

FIGURE 1

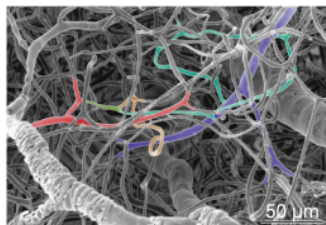
We are now going to look in finer scale at the region marked with the “X” in Figure 1, which is fed by the branches of these major cerebral arteries, and discuss what’s known from the past. Others besides me, notably Rob Harrison, also got intrigued by the idea of variability between neural signals and blood flow. He did a beautiful experiment. He had done imaging studies and saw a lot of variability across the cortex. He wanted to see if this variability could be

explained in terms of different vascularization. This is shown in the top image of Figure 2. These images are from rats in which the entire vasculature was filled with latex, then you chew away the tissue and cover the latex with gold and look with a scanning electronic microscope. They give you beautiful pictures, but they are quantitatively useless, because you can't see deep into the tissue. Nonetheless, what Rob found is that if you look at the surface architecture, you see a lot of loops. All of a sudden, this is the first serious indication that, rather than a tree-like structure that you see in textbooks, you actually have this loopy or graph-like structure.

Latex Casts of Cortical Vasculature Reveal Surface Loops



Latex Casts Reveal Arterial to Capillary to Venule Paths

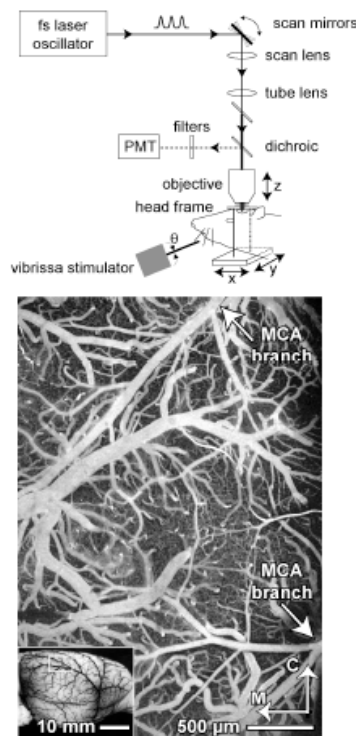


Harrison, Harel, Panesar and Mount (2002)

FIGURE 2

What Rob did next is look deep into the brain—the bottom image in Figure 2 shows a region in which a piece of the cast has chipped off—and follow from arterial to vein to venule, going from red to yellow to blue. At least qualitatively it looks like one sees fewer loops. This was very suggestive that the topology on the surface of the brain, maybe if only because it's in 2-D, is rather different from the topology below the brain's surface. Our question here is what in fact are the dynamic consequences of these differences in topology? So, we need a method.

TPLSM with Fluorescein-Dextran Labeled Blood Plasma to Monitor RBC Motion



Kleinfeld, Helmchen, Mitra and Denk (1998)

FIGURE 3

The first thing we need to do is make a nice map of the entire surface vasculature, so we use a technique invented by my friend Winfred Deale that is called two-photon microscopy (TPLSM). This allows us to see deep into scattering tissue. We are allowed to see about 500 or 600 microns into tissue where 100 microns of thickness is enough to scatter away about 90 percent of the light, which will allow us to make maps of the vasculature. Keep in mind during the whole time that there is a rat here; there is a living, breathing rat.

What we could do is home in, going down from the surface to look and make a little map of individual vessels. In the previous talk, Eve Marder showed these nice reconstructions of neurons that were made by making many individual planes and then reconstructed. We are doing basically the same thing with scanning in a Z-direction, and taking all these planes, and looking down at them from the top. That is shown in Figure 4 on the left. We can home in on the vessel that is inside the box. It's about six microns in diameter. That makes it a capillary, and we can then home in on the capillary per se.

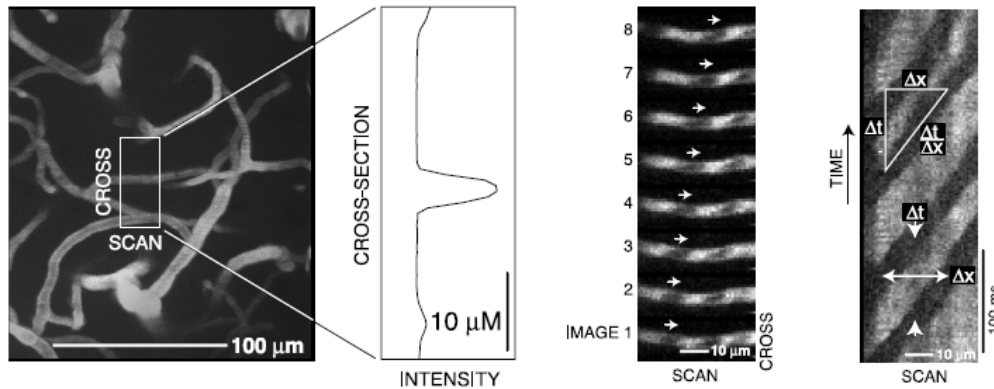


FIGURE 4

We want to get some contrast to see movements of individual blood cell components. In this case we can use a very simple mechanism. We put large dextrose in, big sugar molecules that are coated with dye, so the blood plasma glows bright and fluoresces. Particular red blood cells eschew the dye, so they are dark objects; we are imaging dark objects on a bright background. If you look at successive frames from 1 to 2 to 3 to 4 to 5 in the right half of Figure 4, these are roughly at video rate, there is a dark object progressively moving over to the right. That is the movement of a single red blood cell moving from a capillary, and this forms the notion of our basic data. We could look at individual red blood cells, or any cell for that matter, moving through vessels that are within, in our case, about the top millimeter of cortex, halfway through the thickness of rat cortex.

As a technical issue, the red blood cells move pretty quickly, maybe 1 millimeter per second or so. In order to quantify what the speed or the flux of these vessels is we use a technique called line scans. We scan the laser repeatedly across a single vessel. When there is just plasma, we get a bright image. When we hit a red blood cell, we get no light coming back, and when we get past the red blood cell it goes bright again. As we scan a little bit later in time, the red blood cell has moved, so it takes a while longer to hit the dark region. We then build up a succession of strips in a space/time plot, as shown in the right-hand image of Figure 4. The slope of this goes as one over the speed, and that's basically enough to quantify what is going on.

Basal Blood Flow in Capillaries Deep to the Pial Surface

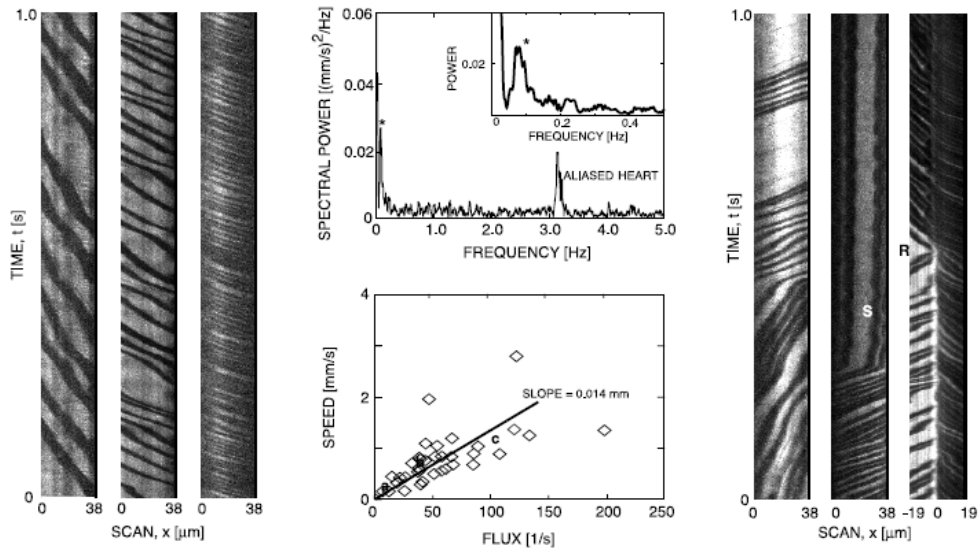


FIGURE 5

Let's see what happens just walking around the cortex in different columns, and looking at different capillaries. The images in Figure 5 are in some sense typical. You see these in a succession. These are slow speeds, medium speeds, high speeds, and different slopes. Not surprising, if you plot the speed versus the flux, the number of particles that pass per second is rather linear at low values of flux. In fact, as you might expect, it may begin to saturate at high flux. It's in 1-D and it's going to have a Pringle potato chip kind of transition.

The real thing that gets you is that you are sitting there, and all the sudden the speed jumps. It's very hard to find long swatches of time where the speed is uniform, or if this is a case where the red blood cells are moving, all of a sudden they stall and stop within the capillary. They will sit there for tens of seconds, or even more interesting, if nature very politely decided that capillaries should meet at Ts, what's a better present to somebody who is doing scanning. Here we are scanning two legs of a T. We have seen one come in from this side and one come in from that side. Particles are conserved, so we know what's coming out the third leg. Right in the middle of this, marked R for reversal on the right-most image of Figure 5, the direction flips; therefore, this business of things flipping in direction is just a signature of feedback loops. This data doesn't tell you what the spatial scale of the loop is, but this is the typical thing and it's wild.

Fritjof Helmchen and I gave a demonstration at Cold Spring Harbor Laboratories a

couple of summers ago at an imaging course, and it was like driving in Boston. We found a little round-about and you see the blood go this way, and then it goes that way, and then it goes this way, so it's like driving in the round-about. The other thing that is very interesting is you could ask where most of the variation lives. We could turn this kind of data into a time series of speed or velocity, because the direction is changing versus time. We can then compute the spectral power density of this time series. This is shown on the top plot of Figure 6, and what you see is that most of the variation lives at this very low frequency that comes in at about 0.1 Hertz. This is an old topic; it is actually a 110-year old topic. The noise is known as vasomotion. Its origin is unsolved and there is a real prize for this. This is the dominant noise in imaging techniques like functional magnetic resonance imaging and optical imaging. So, already at this capillary level you see this dominant noise source, and it is what got us into this game in the first place.

Vibrissa Stimulation Induces a Change in Blood Flow in Deep Capillaries

The change is comparable to the level of baseline fluctuations

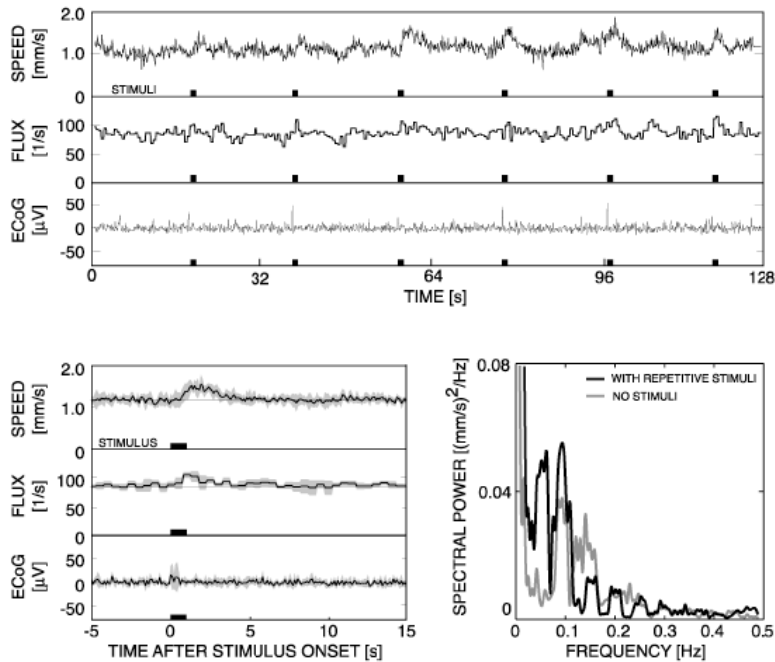


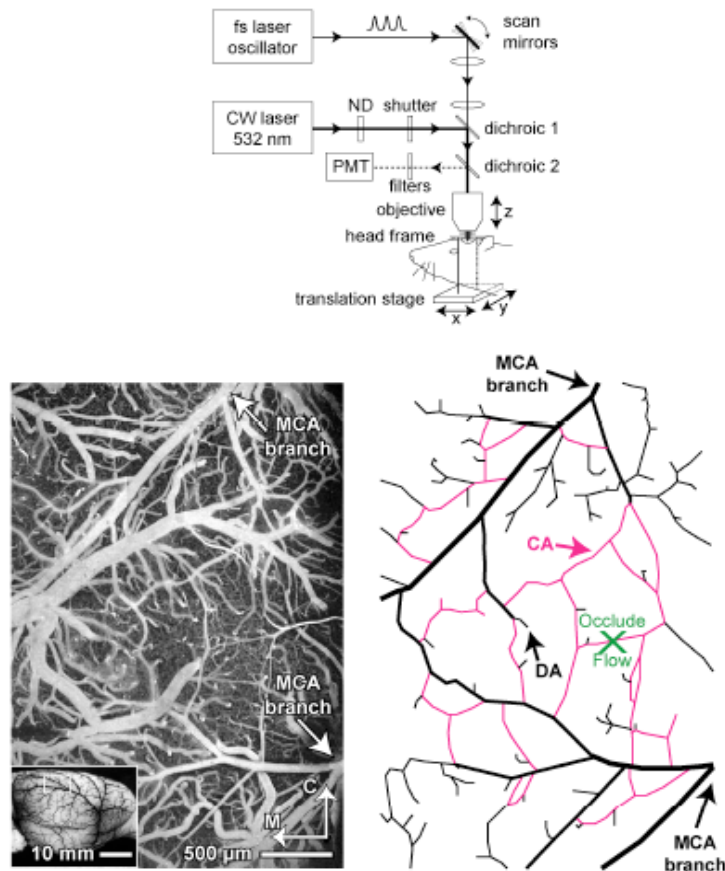
FIGURE 6

What is the scale of this variability? How can I put this into something realistic? One thing I could do is give a stimulus. I could ask if the speed of the blood changes as a function of stimulation. I believe this is true from various sorts of functional imaging experiments. At the

very top line in Figure 6, which is the time series of speed, you see that it's noisy or ratty—no pun intended. Towards the end of the trace you see a number of bumps. These little boxes mark the time stamp of when we actually went in and simulated the animal. In this case, we are recording from the vibursa area of the rat cortex, and we are tweaking the vibrissae at a level where we get about one spike per tweak, which is the normal physiological level of spiking. Sometimes you see a blood change and sometimes you don't. It's a statement of a signal-to-noise of one. If we take all those traces together and average them as shown in the lower left of Figure 6, there is clearly a response, an increase in the speed during the stimulus. The standard deviation is pretty much on the order of the mean itself, again, a statement of a signal-to-noise ratio of 1. To be more precise, if we take that time series on top and calculate its spectral power density, we have a stimulus peak that comes in at about 0.5 hertz, 1/20 seconds, and that's shown in the lower right. We then have our noise peak, our vasomotion peak, and they are about the same amplitude. You might argue that this is entrainment, so you look right before we put the stimulus on, and you see about the same amplitude. This is the bottom line, the variability in blood flow, and the scale for this is that the changes in blood flow in response to normal stimulation are on the same size as the noise. This could be why functional signals are weak, because they are right around the noise level. It could say something about regulation.

Things are so noisy, and there are all kind of loops, then you think this might almost be resilient to defects. We went ahead and tried this kind of experiment, and now I'm going back to our set-up, as shown in Figure 7. I'm imaging on the surface, and the first series of experiments we're looking at are changes in, or rearrangements of, flow on the surface. These branches of cerebral arteries constantly form these interconnections, and this has a name. It's the connecting or communicating arterials, and what we could do is measure the direction and speed and magnitude of flow in all of these different vessels. We will then go in and accrue data, make a blockage at this point, and look at the downstream flow.

TPLSM and Photothrombotic Occlusion to Targeted Surface Communicating Arterioles (CA) Loaded with Rose Bengal



Schaffer, Friedman, Nishimura, Schroeder, Tsai, Ebner, Lyden, Kleinfeld (submitted)

FIGURE 7

We may ask, does it just stall out? This is what you would expect for a tree structure. Is there going to be some rapid rearrangement to the flow? We need to introduce another new toy to do this. The trick is these terrible molecules called photosensitizers that make a free radical when you shine light on them. Within a couple of milliseconds these free radicals will actually damage the nearest piece of tissue. In our case they will cause a little bit of irritation to the wall of a vessel. This means we could inject these into an animal, do our observation at a wavelength that doesn't excite these molecules, and then come in with a second beam of light to excite the molecules of interest and then see what's going on.

The gist of it is the following, as illustrated in Figures 8 and 9. We map the surface vasculature to determine where we're going to target a particular point. In this example flow is

going this way, and then it actually comes down and it branches, first here, and then branches in this direction. Then I'm going to shine light. The deal is that I adjust this dye, this rose bengal, which is very fancy name for a molecule that does damage. It's everywhere in the bloodstream, but I want to make a block just at the surface, so I come in with just-above-threshold levels of intensity. I could actually activate the molecule on the surface, but once I get below the vessels, lensing due to the vessel and scattering through the issue will drop the intensity below threshold. I could then gradually build up a clot at this point and see that the flow, which was now all moving downward on the time scale of the block, would rapidly rearrange itself.

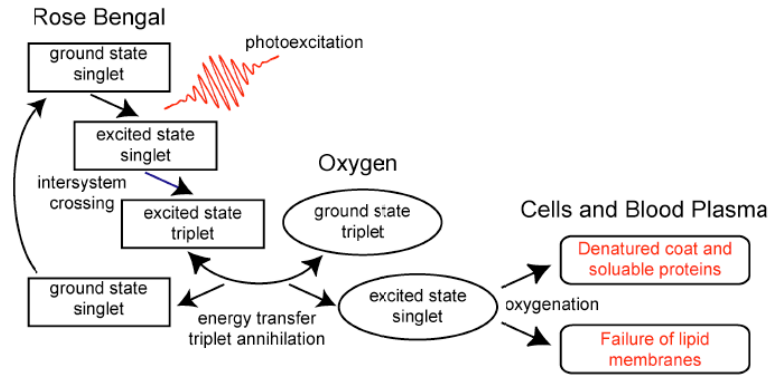
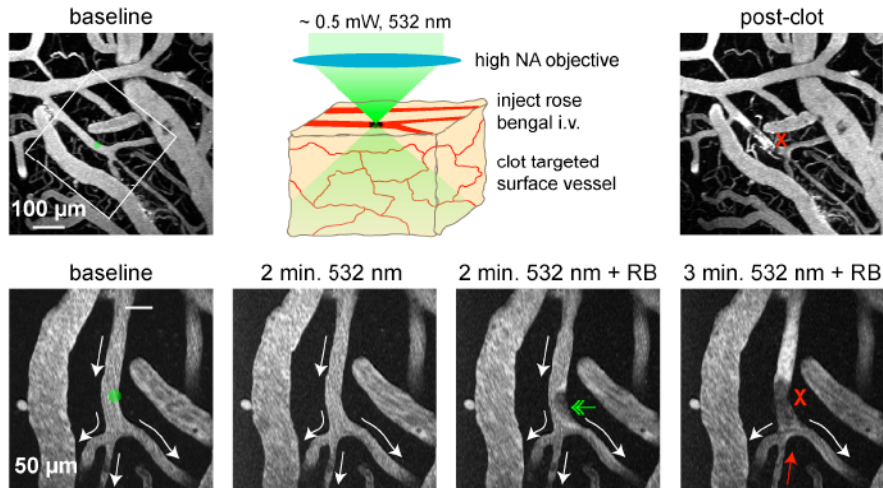


FIGURE 8

Lateral Localization of Optically Induced Rose Bengal Mediated Occlusion

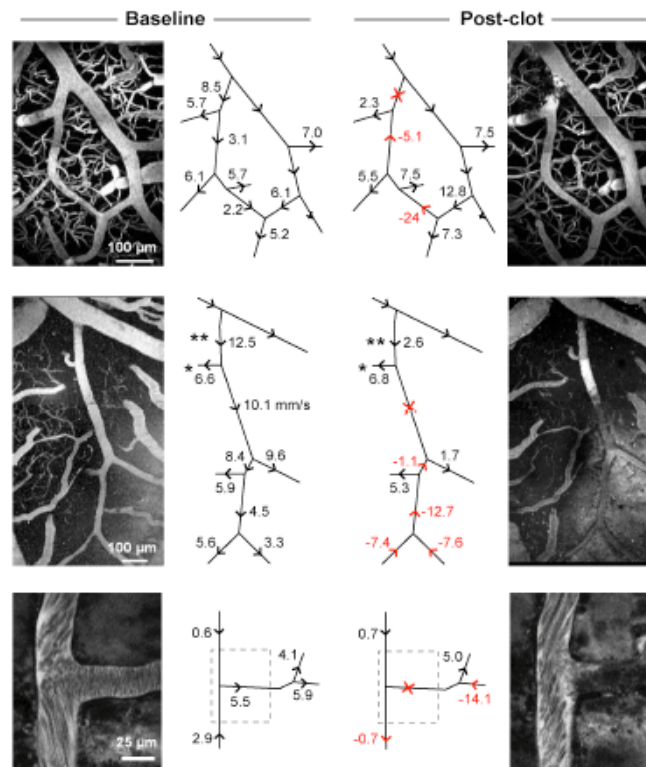


C. B. Schaffer, B. Friedman, N. Nishimura, L. F. Schroeder, F. F. Ebner, P. D. Lyden, D. Kleinfeld

FIGURE 9

Looking at Figure 10, I'm going to start irradiating at the point marked with a red X and then begin to build up a clot. The clot begins to form, and slowly the direction of these vessels is beginning to reverse, and gradually we'll make a stable clot. As a matter of fact, the flow is almost completely stopped at this point, so the key result is already shown here, that we make a block, and we immediately get this rapid reversal in flow. More so, the point is that the speed in the reverse vessel is very much on the same order, which means the rate of perfusion is on the same order as the initial flow.

Downstream Arteriole Flow Reversal after Targetted Occlusion



Schaffer, Friedman, Nishimura, Schroeder, Tsai, Ebner, Lyden, Kleinfeld (submitted)

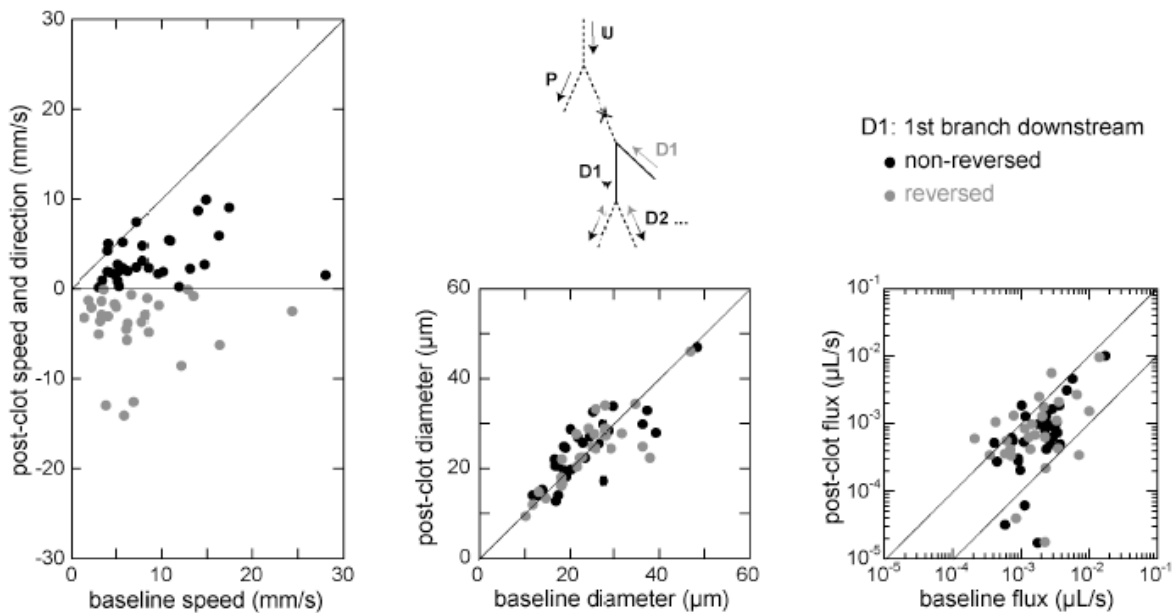
FIGURE 10

There are a number of examples here, and they illustrate the loopiness of what's going on. Let's look at the top example in Figure 10 for a moment. What we see is a blockage that is

shown on the right by the red X. Initially, the speed of the flow coming into the downstream vessels was on the order of 5.7, let's say 6, and 3 millimeters per second. After the blockage, it's down to 2 and 5 millimeters per second. So, it's the same order of speed.

You can just run through many animals. We have done this on 30+ animals and 60+ vessels, and you always immediately see a reversal in flow. Things don't stall. This magnitude of the speed is on the same order as the initial speed. In fact, if we could plot this, we could plot the speed after the clot as a function of the speed before the clot on these immediate downstream vessels which are marked D1 in Figure 11. The speed is slightly down, but not by much. The critical thing is that if we multiply the speed times the cross-section of the vessel, which we can also measure we get an idea of sort of the volume flux. That distribution is shown in the right-hand graph of Figure 11. The diagonal line there is the line of no change and the line parallel and to the right of it is about a reduction to 10 percent of the initial value. This is the point at which physiologically you begin to get in trouble. The point is that the rearrangement is well within physiological means to keep the neurons viable, and there is a set of histology that I'm not showing here that is also consistent in showing that neurons and glia stay viable.

Changes in Downstream Flow after Targetted Occlusion

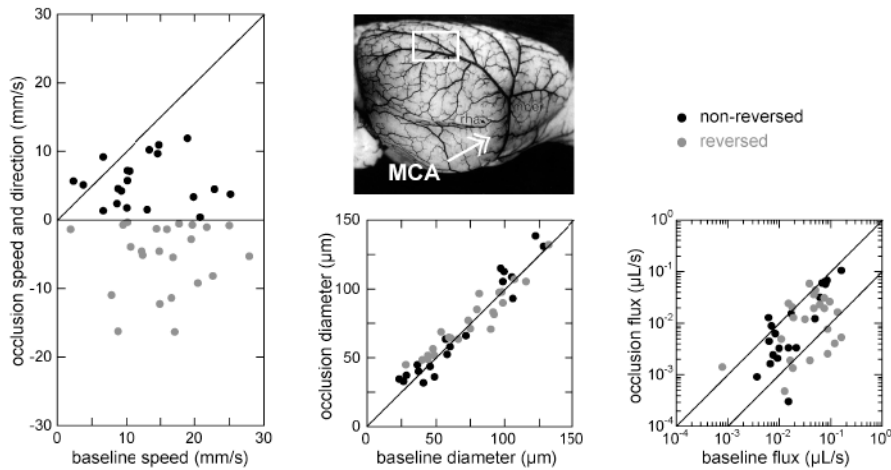


C. B. Schaffer, B. Friedman, N. Nishimura, L. F. Schroeder, F. F. Ebner, P. D. Lyden, D. Kleinfeld

FIGURE 11

The idea is that you have this mesh, and this mesh could rearrange its direction after a perturbation. This should also hold if we make a much grosser kind of perturbation to the system. Here we appeal to a technique that, as I mentioned earlier, has been in the neurology literature for a long time. That is putting a very fine filament through the carotids, and then up into the middle cerebral artery, and blocking this main artery itself. Therefore, it's not a complete blockage, and there is also a flow that comes in through the anterior cerebral artery. The point is that we changed the pressure balance in the system, so to speak. What we see in Figure 12 is again dominated completely by vessels changing their direction of flow when we go in and perturb even one of the main pipes. This kind of business of very delicate balance of flow and changes in the direction seems, at least experimentally or phenomenologically, to be somewhat general.

Changes in Downstream Flow During Filament Occlusion to the MCA to Reduce Overall Perfusion

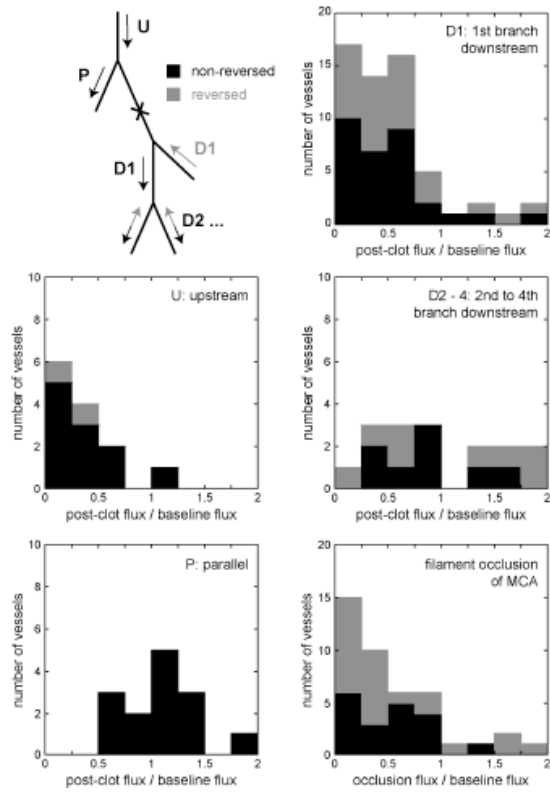


C. B. Schaffer, B. Friedman, N. Nishimura, L. F. Schroeder, P. S. Tsai, F. F. Ebner, P. D. Lyden, D. Kleinfeld (submitted)

FIGURE 12

One can summarize all of this as follows, as shown in Figure 13. If we look at the first downstream vessels I mentioned, the flow after the fact is on average about half its initial value. That is lower than physiological levels. If we make a blockage and look at vessels that fit parallel regions, there is virtually no change. If we look here at vessels further downstream, on average there is no change, but quite a lot of variability. It really is completely the loopiness of this system, and when you look in the textbooks they draw arterials as coming out as tree structures, and this is wrong. The images in Figure 14 are actually taken in vivo by staining the vessels with a lipid-like fluorescein dye that stains the surface. You can look at these regions and it's fed this way; it's fed this way and it's fed this way. It's just a fantastic amount of redundancy.

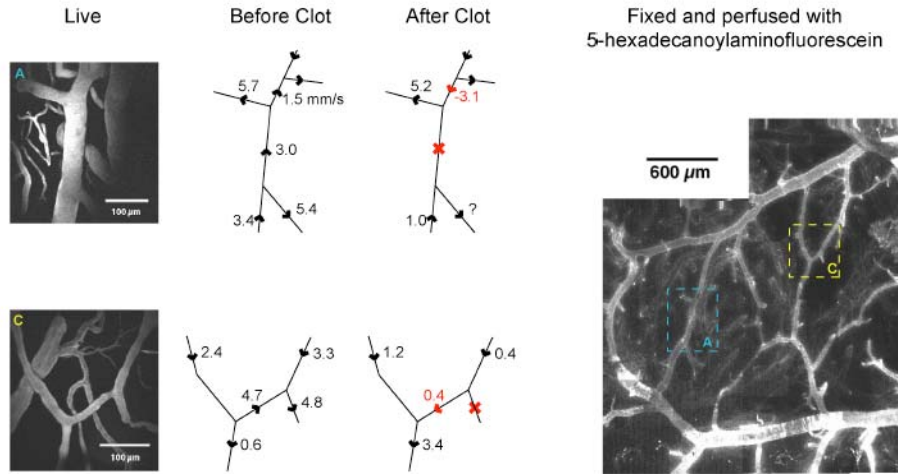
Summary of Flux Results on Reflow



C. B. Schaffer, B. Friedman, N. Nishimura, L. F. Schroeder, F. F. Ebner, P. D. Lyden, D. Kleinfeld

FIGURE 13

Reversal of Flow in Communicating Arterioles after Photothrombotic Ischemia May Reflect Collateral Circulation through Anastomoses



C. B. Schaffer, B. Friedman, N. Nishimura, L. F. Schroeder, P. S. Tsai, F. F. Ebner, P. D. Lyden & D. Kleinfeld (submitted)

FIGURE 14

DR. MARDER: How big is the section that is sensing the pressure? How far away from the inclusion do you see changes in flow?

DR. KLEINFELD: Immediately downstream.

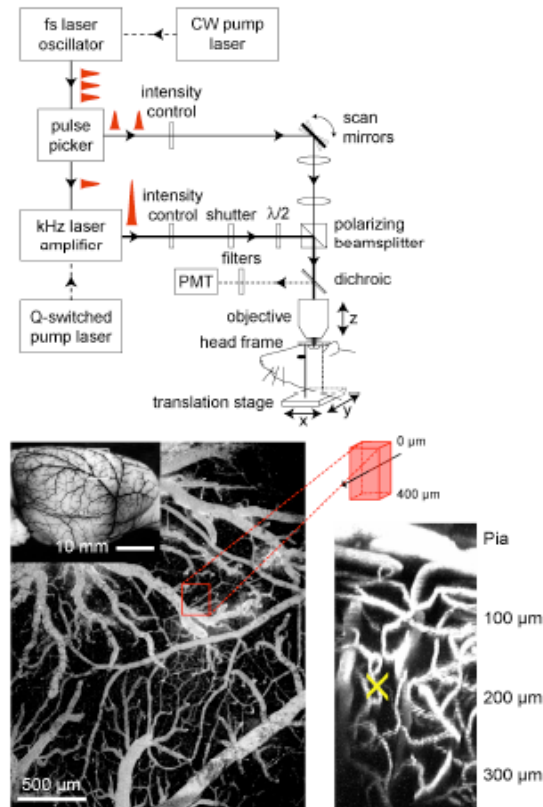
DR. MARDER: Two millimeters away?

DR. KLEINFELD: Let me give you a sense with the cartoon of vasculature in Figure 14. From here to here is a few hundred microns for the next sort of primary downstream vessels. By the time we get to secondary and tertiary vessels we are talking about 500 microns, and there, I think statistically, you would be even sensitive to this effect. We are talking scales of 100 microns, not millimeters, which may be, more importantly, of scale of cortical column in the rat. There is a deep issue, in fact, if this turns out to be a regulatory mechanism for blood flow in the rat. That will take an awake-behaving imaging experiment, which we are slowly moving towards.

The last point I want to get at is the noise in the system. We have talked about the redundancy on the surface, and now we actually want to move below the surface and look at the deep microvascular network, which at least superficially has less connectivity. Now we have to introduce a different toy, as illustrated schematically in Figure 15, because initially we used light

to make a blockage, but we relied on the fact that the intensity was just enough to cause a blockage right at the surface. What we need now is a means to actually cause a blockage below the surface. That means I have to go nonlinear, so to speak, and I need to have an interaction, which is only effective at the focus of the lens.

TPLSM and Nonlinear Photodisruption (Hemorrhage, Occlusion or Extravasation) to Targeted Deep Microvessels

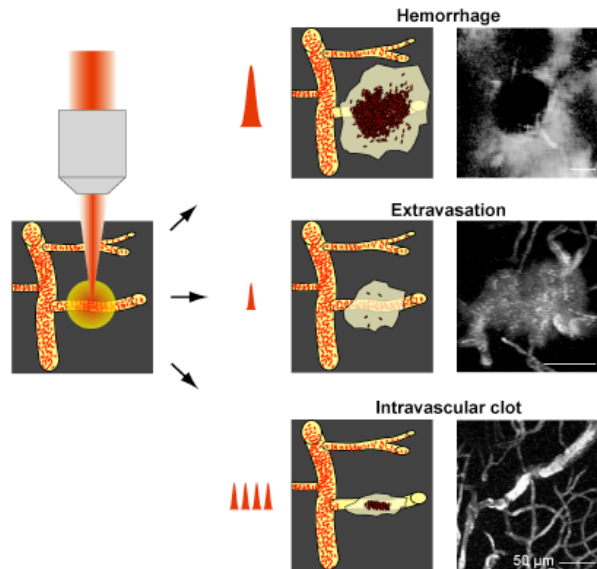


N. Nishimura, C. B. Schaffer, B. Friedman, P. S. Tsai, P. D. Lyden, D. Kleinfeld (submitted)

FIGURE 15

Just like we use nonlinear interactions to do our imaging, which means that we only get appreciable absorption at the focus, I can also come in with very, very short 100 femtosecond pulses, and I could only cause damage right at the focus of the lens. Again, we are going to make a map of the surface, and then I can also map down below the surface. I could find small arterials below the surface, and I could target these for damage.

Three Types of Vascular Disruption Induced in Targeted Microvessels with 0.1 to 1.0 μJ Ultrashort Laser Pulses



N. Nishimura, C. B. Schaffer, B. Friedman, P. S. Tsai, P. D. Lyden & D. Kleinfeld (submitted)

FIGURE 16

The ideas of lasers in biology to perturb things are not new. About 10 years ago people, particularly Eric Mazur at Harvard, started applying nonlinear interactions to actually perturb transparent objects. We picked up on these ideas of using very-high electric fields to perturb things only at the focus. I think it's a nice technique for all of neuroscience. What is going to happen is that we can target below, and there are three ranges of energies. I guess at some level of energy you longer have a rat, so that is not interesting. But at modest energies you actually cause an aneurysm to form. At more intermediate energies you get something very interesting: we cause a little cavitation bubble to form inside an arterial. We are picking a 10 micron object, and focusing light within the center of that object. We can cause a little cavitation bubble to form, and that will actually cause temporary leakage in the blood-brain barrier. It's basically pulling apart the endothelial cells and we inject dye into the surrounding media. It is similar energies, but by actually targeting the edge of the vessel, we can cause a blockage in a vessel that is below the surface. This will allow us to get access to a new topology, and this is what happens.

Rearrangement of Flow after Occlusion Induced with Ultrashort Laser Pulses

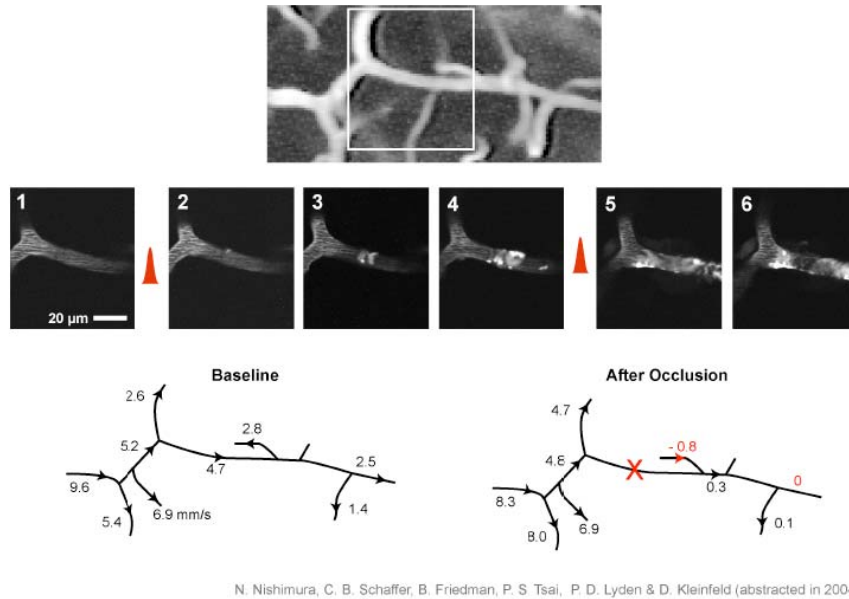


FIGURE 17

Figure 17 shows a vessel that is about 300 microns below the surface, and we are homing in on the small region that is covered by a box on top. There is flow going through this vessel, and I make a blockage at this point and map the speeds like I talked about before. The key difference here is that I made a blockage at this point in this less connected network that lies well below the surface.

When I look downstream, again I get a reversal of flow, but the magnitude of these changes is about an order of magnitude smaller than what I have gotten on the surface. This is not such a resilient network, so let's see what happens. On average we see the following. In this case, if I look upstream, there is not much change below the surface. If I look downstream I really do have pretty much a cessation of flow. This holds for a fairly large sample, about 16 animals and about 30 data points, as shown in Figure 18. This vasculature in a qualitative way is missing loopiness, and I can compare it directly to what happens on the surface. Again, upstream there is very little difference, but downstream on the surface you get this tremendous restoration of flow.

Rearrangement of Arteriole Flow after Occlusion Subsurface Microvascular Network versus Surface Communicating Arteriole Network

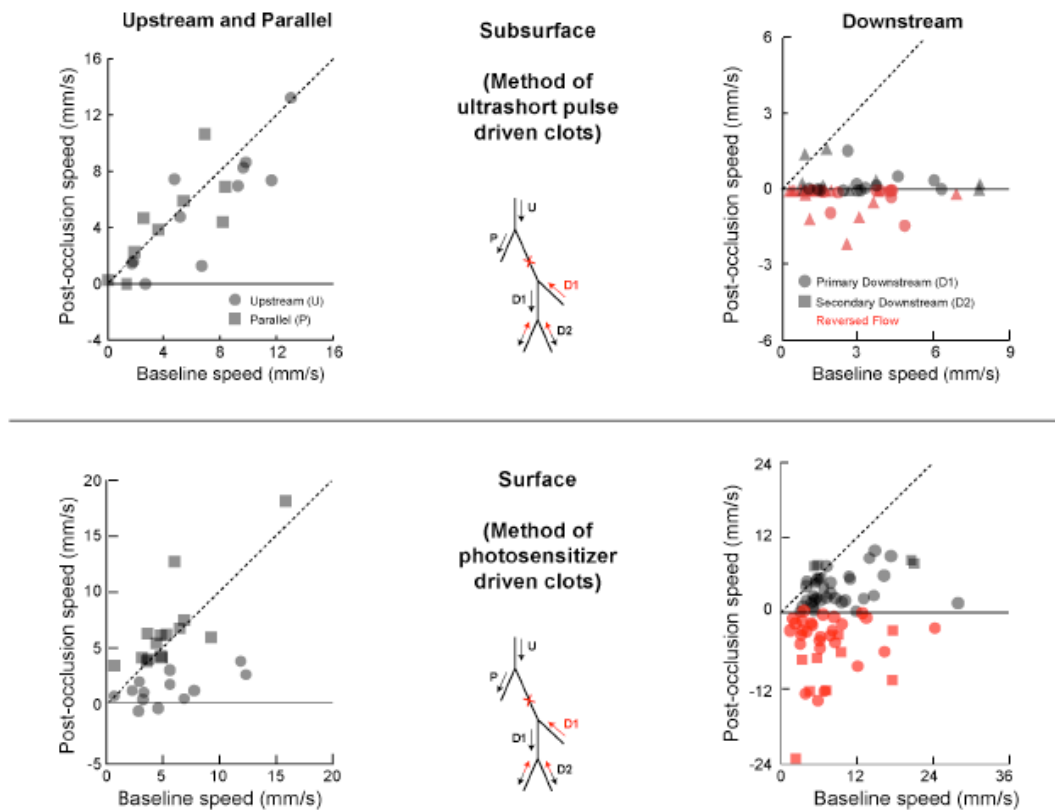


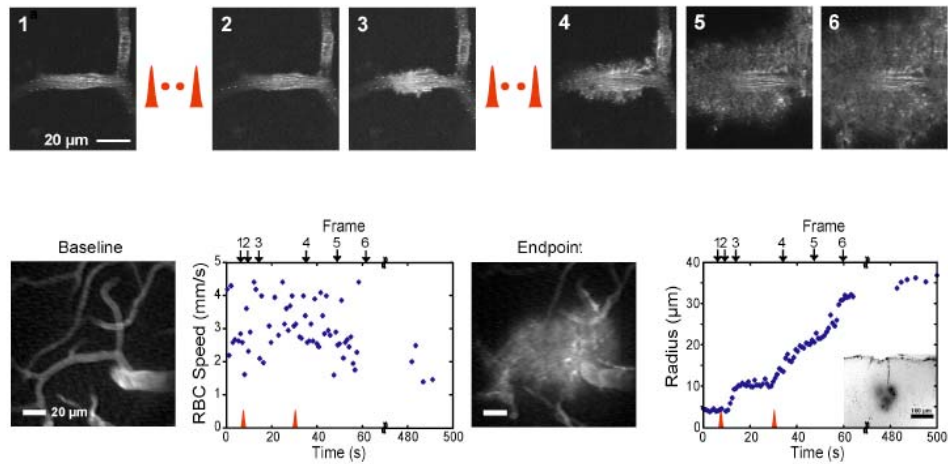
FIGURE 18

There are a lot of controlled experiments that we have done; there is a lot of histology I purposely left out. One thing you could always say is that these changes had nothing to do with making a blockage, because you have gone in artificially. You have made this transient mechanical disturbance and you have sent an electrical signal down the epithelial cells, down the vasculature. Maybe you have turned on all kinds of inflammation and damage mechanisms.

I could go back and look at this other case that takes place at the same energies, where we go in, target the middle of a vessel, and cause a disruption of the wall, which we sense because dye that is normally confined completely in the vessel sort of oozes out for a short period of time. The important point is during this entire procedure I have now hit the system with laser pulses. I have caused dye to leave, but I'm looking at the motion of the red blood cells at the same time, and that motion is unchanged as shown in Figure 19. So, at least on these types of timescales of

hundreds of seconds, whatever changes happened in flow appear to be due to the blockage, and not due to biochemistry.

Stable Plasma Extravasation, with Continuous RBC Flow, Induced in Deep Arterioles by Ultrashort Laser Pulses



N. Nishimura, C. B. Schaffer, B. Friedman, P. S. Tsai, P. D. Lyden & D. Kleinfeld (abstracted in 2004)

FIGURE 19

The last point has to do with what one could do to restore this flow, and the role that viscosity could possibly play in this mechanism. We did the same kind of experiment by making blockages at single points, and we asked if we looked further downstream, if some of the blood flow could be restored by actually dropping the viscosity of the blood. This is an old idea. What physicians used to do for stroke patients was just to dilute their blood. It sounds on the same order of leeches. Scary! The point is that we normally look at baseline levels, and then we cause a clot. Even if we try to dilute the blood, Figure 20 shows that there is virtually no change in the immediate downstream flow, and this region is stalled out. If I look a little bit downstream, things are also fairly stalled out. Yet, if you dilute the blood by 30 percent, you really do get this restoration of flow. This is just more data to say that we are limited by simple fluid dynamics as to how the flow rearranges itself. At least our neurology colleagues also think that this might be something that should be revisited in medical settings.

Hemodilution Partially Restores Perfusion Downstream from a Subsurface Clot

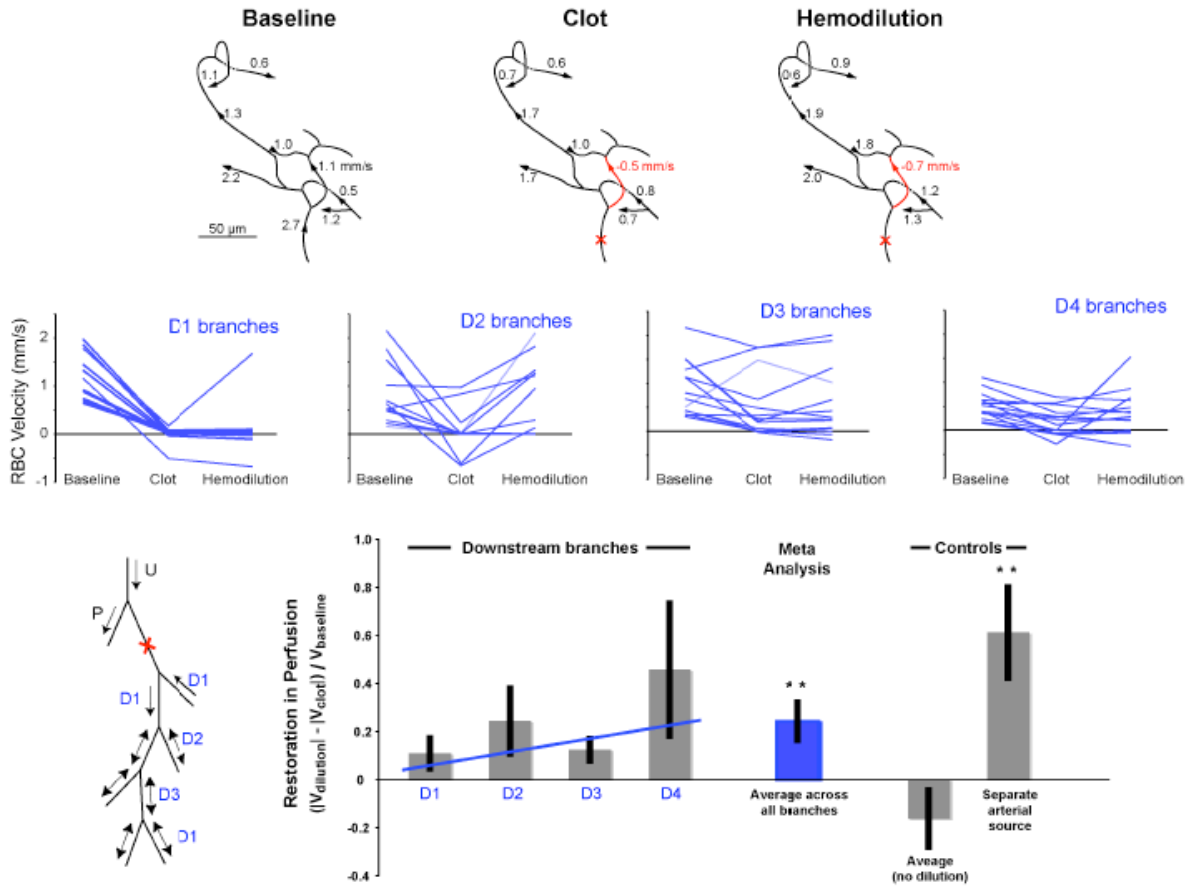


FIGURE 20

I have talked about ideas for trying to relate topology to blood flow, about how flow is very noisy in the vascular system, and about how we could go in and perturb flow on two different networks—a surface network and a deep network. This one is highly interconnected, this one is less connected, and how there are quite differences in the dynamics. Right now we are in the midst of quantifying the cortical angioarchitecture. I mentioned at the beginning that the casts are very difficult to reconstruct because you can't see into them. We actually have a way to reconstruct big swatches of the vascular.

**Is there a topological basis for resilient perfusion
in the surface network but fragile flow deep to the pia?**

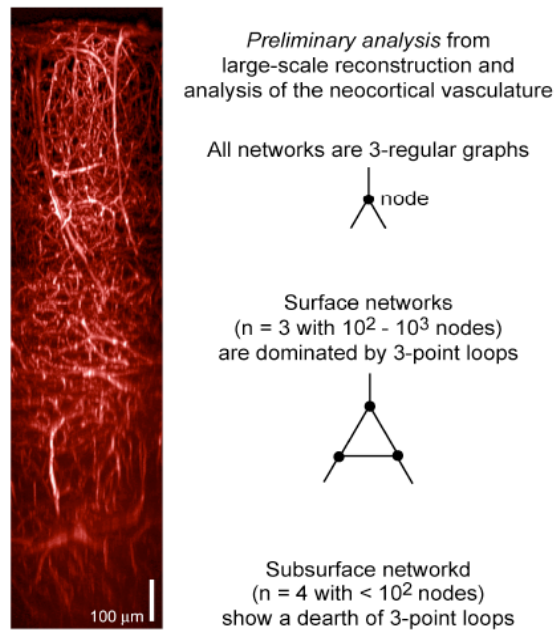


FIGURE 21

Figure 21 represents work in progress. From the point of view of networks though, all the vertices have three connections, so in some sense this should make the problem simpler, if not tractable analytically. If you look at the surface—just so far we have looked at a couple of networks—they are really dominated by these kind of little triads, the smallest possible loops that you could put together with these sort of three nodes. If you look at the subsurface networks you see very few of these threes, so what we hope to do over the next year is put together these dynamics and the topology in a hard way.

K. F. Ahrens
R. W. Berg
T. W. Caciatore
G. O. Clay
J. C. Curtis
M. S. Fee
K. Ganguly
D. N. Hill
S. B. Mehta
L. M. Merchant
B. Migliori
Q.-T. Nguyen
N. Nishimura
S. M. O'Connor
J. C. Prechtl
C. B. Schaffer
L. F. Schroeder
I. Stopkin
P. S. Tsai
R. Wessel
S. Venkatachalam
D. J. Whitmer

Neurophysics Laboratory at UCSD



Cumulative Government and Philanthropic Support
NIH - NCRR / NIBIB / NIDCD / NIMH / NINDS
NSF - DBI / IBN / IGERT / Physics
BEBRF Packard Foundation
HFSP US/Israel BSF
LJIS Whitehall Foundation

E. Ahissar
P. D. Brodfueher
T. H. Bulloch
L. B. Cohen
W. Denk
F. F. Ebner
G. B. Ermentrout
B. Friedman
D. Golomb
O. Griesbeck
A. Groisman
F. Helmchen
A. I. Farraguerri
M. R. Jarvis
J. P. Kaufhold
H. J. Karten
W. B. Kristan
H. Levine
V. Lev-Ram
P. D. Lyden
P. P. Mitra
B. Pesaran
R. N. S. Sachdev
J. A. Squier
H. Suhl
P. W. Taylor
R. Y. Tsien
H. P. Zeigler
M. Zochowski

FIGURE 22

I noted Eve reverted to math form, but I will revert to biology form and put the people at the end. The initial flow experiments involved my colleagues Winfried Denk, Fritjoff Helmchen and Partha Mitra. The National Academies more recent perturbation-based experiments were done in collaboration with Patrick Lyden's laboratory at UCSD and involved Beth Friedman, Nozomi Nishimura, Chris Schaffer and Phil Tsai. Beth and Phil, along with Pablo Blinder, Ben Migliori and Andy Shih, are continuing this collaborative effort.

QUESTIONS AND ANSWERS

DR. DING: My question is how long is the latency when you stimulate and measure the speed change?

DR. KLEINFELD: The latency is about 300 milliseconds.

DR. DING: So, it's very fast.

DR. KLEINFELD: Yes, but I think if you look at intrinsic imaging, which is where people measure in a gross sense by average the change from oxy to deoxy, they typically quote numbers like 400 milliseconds. I think the difference here is we are looking directly at individual vessels. I have to be straight, we could see no faster than 200 milliseconds, because we have to average over a window of time to detect the speed change.

DR. DING: A second question is that in the absence of stimulation you still see a lot of

activities in that flow. You used the noise to describe. Can you say a little more about what is underlying those noisy activities?

DR. KLEINFELD: I can and I can't. This is a deep intellectual embarrassment to the neuroimaging field. There are regions in the rat that are at least on the order of a couple of millimeters square in area, and in human patients they are probably a few centimeters square. This is based on Pertha Mitra's analysis of fMRI data. Basically, you get changes in the arterials and changes in the musculature in the arterial that regulate the flow. What controls this is not known. It's the question of neural control. There are interneurons and there are two populations of inhibitory interneurons. One puts out somatostatin, which causes things to get smaller, and one puts out vasoactive intestinal peptide, which causes it to get larger. The hypothesis is that there must be within some region of the cortex enough coupling among these interneuron networks that the blood flow is seeing some delicate balance of what's going on in these inhibitory networks, but that has not been tested. The one thing you can't do is put a catheter in the carotid artery to look at the fluctuations and predict what's going in the cortex. This would solve a lot of practical problems, but it has not worked.

REFERENCES

- Harrison, R.V., N. Harel, J. Panesar, and R.J. Mount. 2002. "Blood capillary distribution correlates with hemodynamic-based functional imaging in cerebral cortex." *Cereb Cortex*. 12(3):225-233.
- Kleinfeld, D., P.P. Mitra, F. Helmchen, and W. Denk. 1998. "Fluctuations and stimulus-induced changes in blood flow observed in individual capillaries in layers 2 through 4 of rat neocortex." *Proceedings of the National Academy of Sciences* 95(26):15741-15746.
- Nishimura, N., C.B. Schaffer, B. Friedman, P.S. Tsai, P.D. Lyden, and D. Kleinfeld. 2006. "Targeted insult to subsurface cortical blood vessels using ultrashort laser pulses: Three models of stroke." *Nature Methods* 3(2):99-108.
- Schaffer, C.B., B. Friedman, N. Nishimura, L.F. Schroeder, P.S. Tsai, F.F. Ebner, P.D. Lyden, and D. Kleinfeld. 2006. "Two-photon imaging of cortical surface microvessels reveals a robust redistribution in blood flow after vascular occlusion." *PLoS Biology* 4(2):258-270.

Robustness and Fragility

Robustness and Fragility

Jean M. Carlson, University of California at Santa Barbara

DR. CARLSON: I'm in the in the area of complex systems in the physics department. Over the last 17-18 years I have been looking at trying to come up with simple models and fundamental kinds of guidelines in thinking about complex systems. That has first led me into earthquakes and more recently into forest fires. Part of the goal has been to try to make connections between simple theory, detailed models and practical kinds of issues so I have benefited a lot from collaborations.

In light of everything that dominates the news these days my aim here is to step back and talk about natural disasters. I'm pulling from a lot of things that for me are less directly what I'm working on, but what are the impacts, what are the consequences of robust yet fragile behavior in terms of dynamics on the planet, and then also in terms of people. This is also an invitation and a question for all the people who work on problems in sociology, what can we do from the point of view of passing information from the scale of the modeling and the geophysical phenomena up to where it has a real impact, which is on sociological issues such as response as well as policy and planning and so on. I think we've got to do that better. Clearly, that is responsible for an enormous amount of the impact of these events, and the part that I have worked on is just the very beginnings of the geophysical phenomenon themselves.

Stepping back, I want to provide an overview of natural disasters. The question at the end is whether or not there are some interesting ways we can think about this. First of all, all these natural disasters are a natural part of the evolution of the planet. If we didn't have this sort of stuff we wouldn't have life, so there are many good things about natural disasters. When you go around and start collecting data though, you start to see that there are a lot of trends in terms of natural disasters such as costs as well as loss of life.

Figure 1 shows some statistics that have been drawn, and you can see that things are on the increase. Economic losses are on the increase and insurance on these losses is not keeping up with the actual losses themselves.

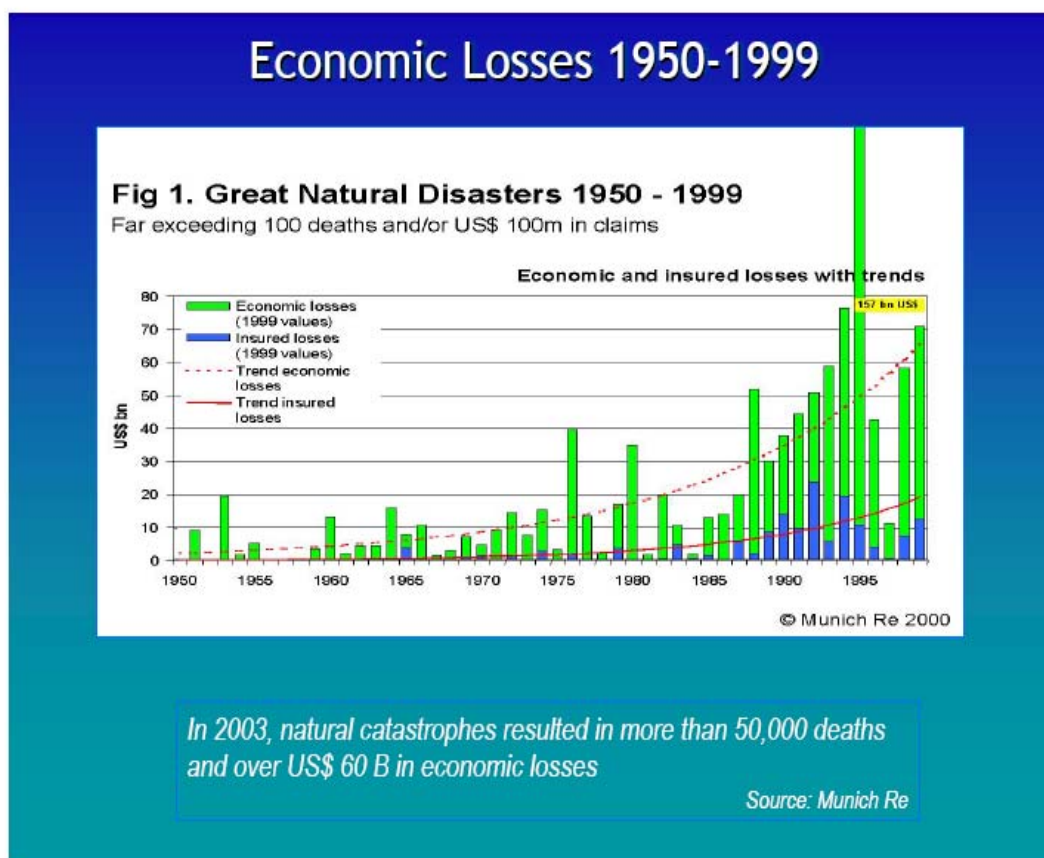


FIGURE 1

Figure 2 updates that picture through 2004. It doesn't include 2005, but it does reflect the tsunami in the Indian Ocean. You can see another big spike in terms of economic losses associated with the Kobe earthquake in Japan. And Figure 3 displays the trend in economic costs. Even correcting for inflation, you see that there is an overall increase in terms of global economic losses associated with natural disasters. This goes hand-in-hand with increased population in the world, and a lot of the population increase is associated with less developed countries.

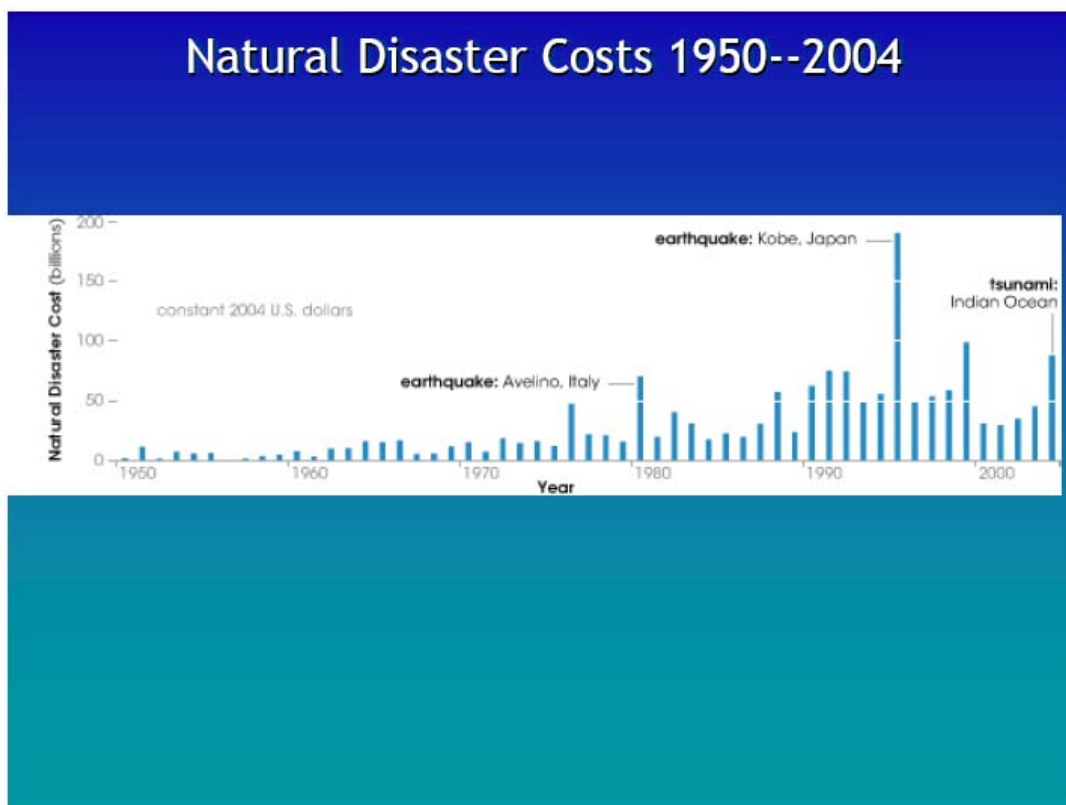


FIGURE 2

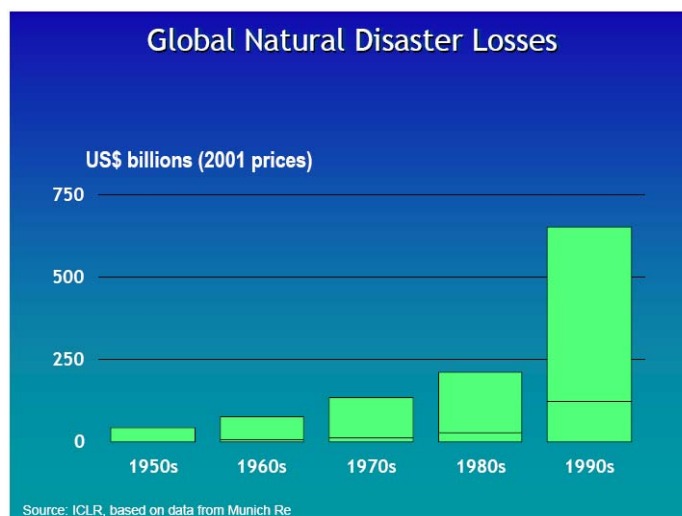


FIGURE 3

Natural disasters are also occurring increasingly in urban areas. The world is now approaching the point where the fraction of population in urban areas will equal the fraction in rural areas, and the total world rural population is expected to remain relatively flat or even decline in the 2020s. Figure 4 drives this home. Each set of bars represents a different large city's expected population growth over coming decades. In looking at some of the large cities you can see most of them are on the increase. Particularly, places like Nigeria and India are showing huge population increases in the large cities.

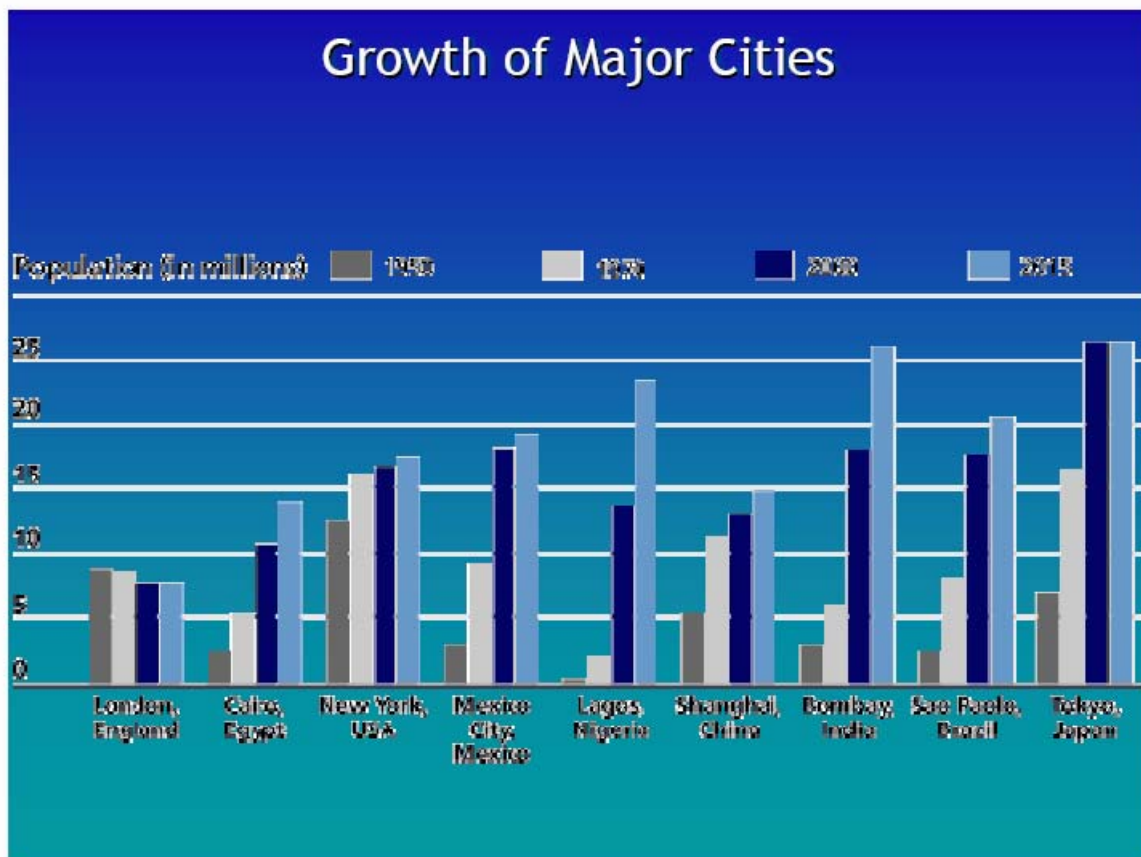


FIGURE 4

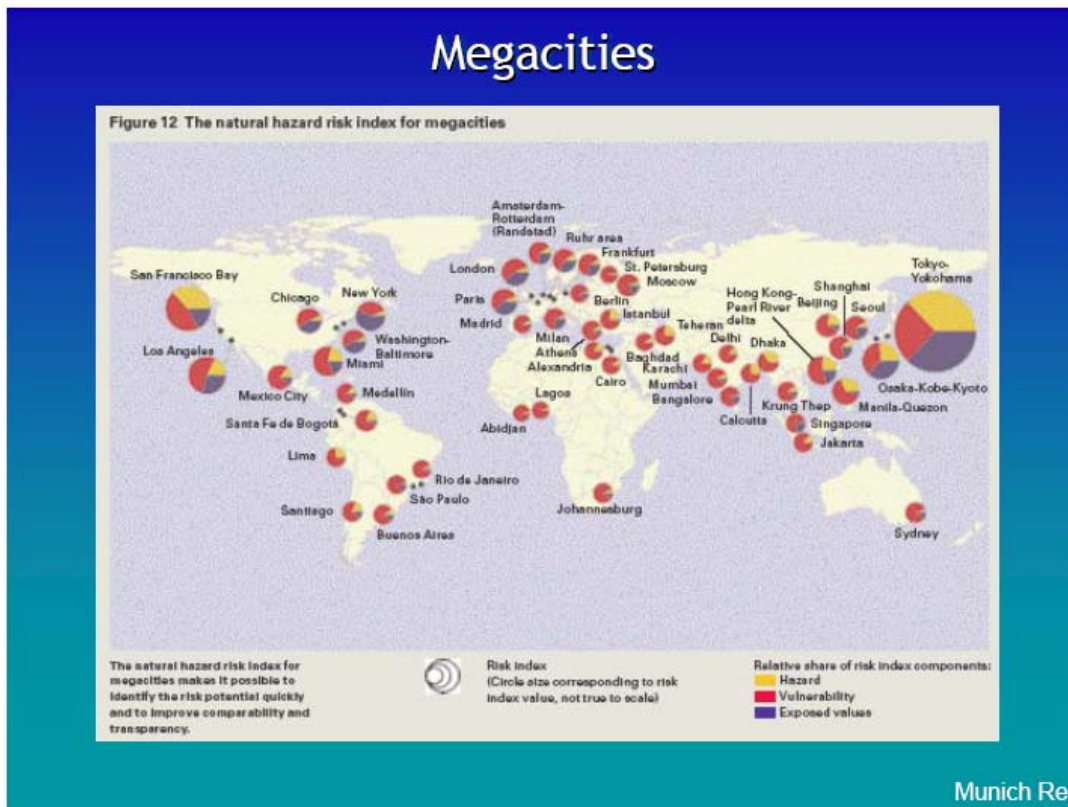


FIGURE 5

Figure 5 shows the class of cities that have more than 10 million people, called megacities. People look at these cities, which are often on coastlines, and classify them in terms of their hazard (yellow on the graphic), vulnerability (red), and exposed value (blue) in terms of hazard. People are evaluating these kinds of things. You can see that Tokyo is considered to be at high risk because of all of the geophysical kinds of phenomenon such as tsunamis, and so are the coastal cities in the United States. This is a natural hazard risk index.

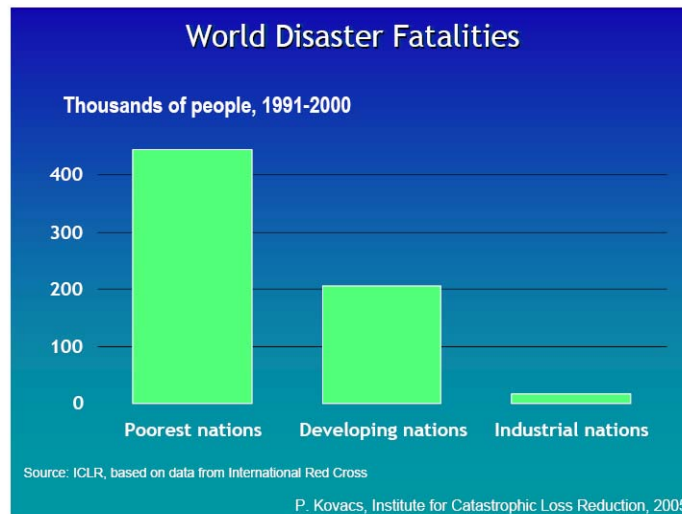


FIGURE 6

There are two things that people look at when they try to measure and plot this kind of data: fatalities and damage. Interestingly, the trends are opposite one another, as shown in Figures 6 and 7. The poorest nations dominate in terms of deaths from natural disasters, but industrialized nations dominate the economic costs. What are you trying to protect, how is a natural disaster measured if you try to think about where you are going to put your resources? You see that there are these two different measures to consider, loss of life and dollar impact.

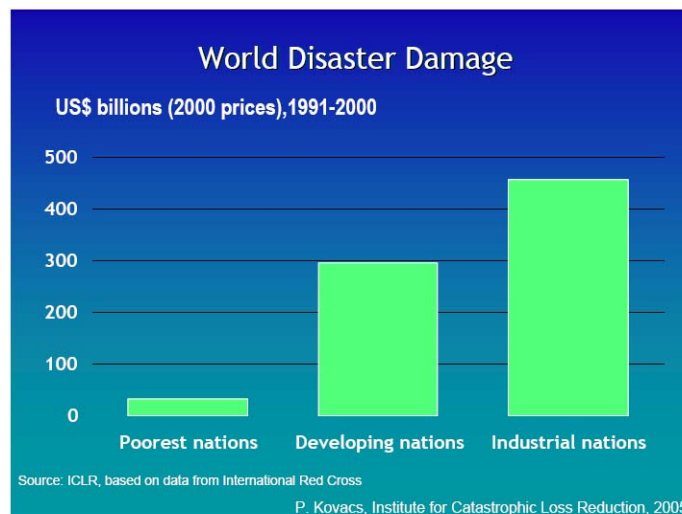


FIGURE 7

Returning to this issue of robust yet fragile and the talk that John Doyle gave yesterday, if you look at the distribution of sizes of events themselves (Figure 8) what you see are these power law distributions. Yesterday John talked about plotting the size statistics for natural disasters here. This figure is in terms of dollars for the natural disasters. They follow a power law distribution. That means that on a log-log scale they are very broadly distributed. Each dot represents one increase in the cumulative number, so does the largest event, second largest event and so on.

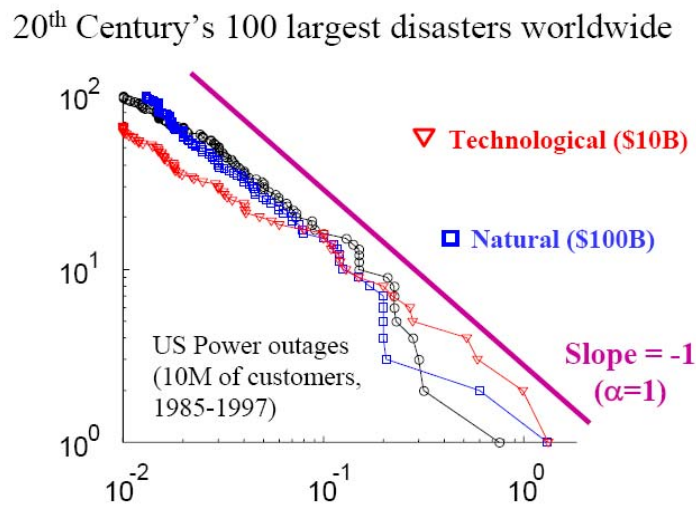


FIGURE 8

This is a power law distribution, and what that means in that the worst event, the events that dominate the losses, are much worse, orders of magnitude worse than a typical event. This is shown in Figure 9. Large events that we see, like Hurricane Katrina, are completely consistent with these statistical distributions, so they're not outliers or anything like that. They are consistent with the statistics and they are to be expected.

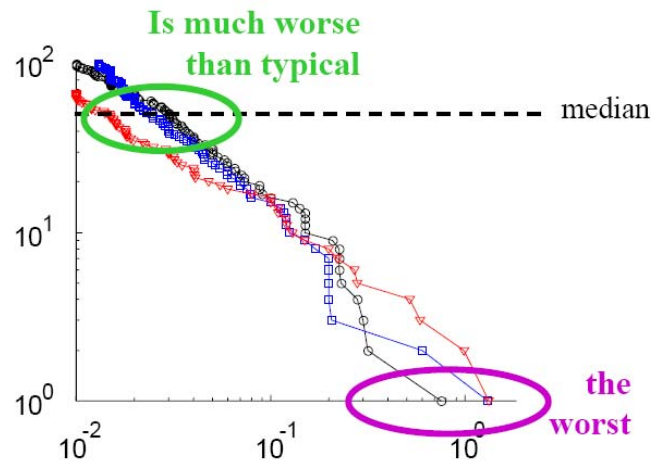


FIGURE 9

One of the things that John Doyle and I have been looking at in recent years has been trying to think about complex systems, cascading failure events, in the context of power law distributions in a fundamental way. This is a very quick summary of the work that we have been doing in that area. We try to bring in insights from biological and technological systems; systems which are highly optimized or have evolved through years and years of Darwinian mechanisms to be behave optimally in some sense, not like generic random systems but in an optimal way. We found that this demands a new theoretical framework on many different levels. We have used some of the simple kinds of models that arise in physics in order to try to show how robustness issues change these models, and robustness trade-offs lead to new fragilities and sensitivities. One hallmark of this is these kinds of power laws. Another statement is that sometimes heterogeneity and high variability can create opportunities for increased performance in some systems.

This HOT framework is one that talks about optimization or evolution of systems. It also fits into a broader framework for describing large network systems that have robust, yet fragile behaviors. Not necessarily all systems are obvious solutions of some kind of optimization problem like earthquakes, but they still have this kind of robust yet fragile behavior.

We are finding that feedback plays an important role. Interesting issues arise in terms of multi-scale and multi-resolution modeling and analysis. This really is a demand for a new approach to complex systems theory that goes all the way through. My hope is that we'll be able

to learn from what we study. Social systems will be helpful to understanding complex systems and natural systems. It doesn't mean they are the same, but it does mean that we are sharing tools. More and more we need to share those tools and we also need to be able to talk to each other, because our cascading failures pass from natural disasters into social systems and our network communication systems as well. In any case, this HOT fits within this general picture and a big, primary message from the HOT systems is to imagine having some sorts of resources that you are allocating to a spectrum of events. If you become, through design or evolution, robust to center kinds of perturbations, you introduce new fragilities as well, as a consequence of this architecture.

One of the kinds of pictures that John and I have worked on to describe this is a generalization of Shannon Coding Theory where you imagine that you have some kind of resources in your systems. We want to compress our data, which calls for optimizing $d-1$ dimensional cuts in d dimensional spaces. Fires, as shown in Figure 10, provide an example, where there is some template of trees in a forest and the resources might be associated with fire breaks or means to suppress fire. Given a constraint on the resources, you want to optimally assign the resources, given some spectrum of possible fires. We have also looked at this generalization of Shannon Coding Theory in the context of Web design.

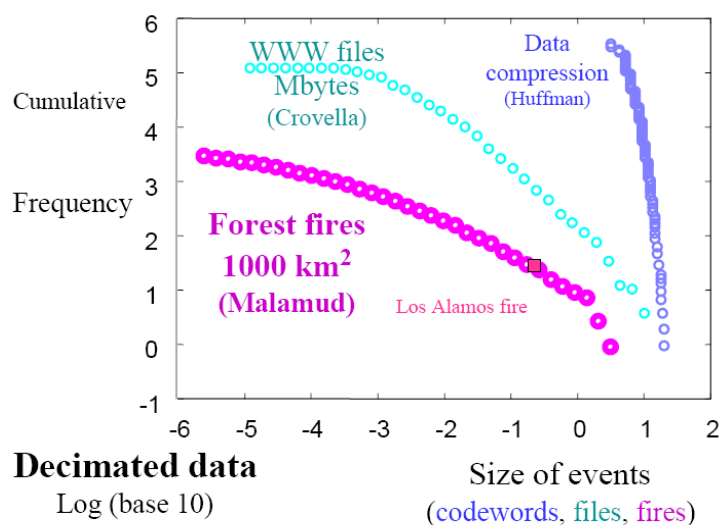


FIGURE 10

I'm not going to tell you the specifics of the model behind the curves in Figure 11, because I want to focus more on the robust and fragile features. We compared it with data, and a very simple model gave rise to very accurate fits that was much more accurate than we were expecting for statistics of forest fire sizes and Web file downloads. We then based it on data compression, so it has to fit there.

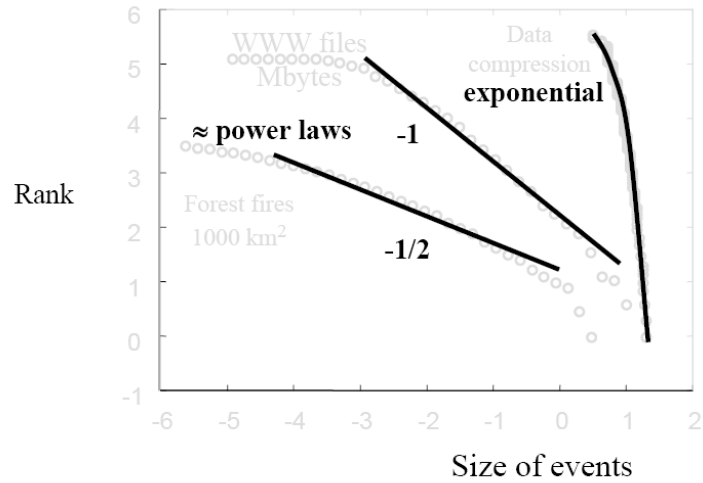


FIGURE 11

In this graph, “rank” is just frequency of possibility as a function of size, measured obviously differently in these two different cases, but the straight-line parts of the slope are really signatures of the power laws. In this particular case they have different exponents, because there is a different underlying dimensionality of the 2-dimensional forest and chopping 1-dimensional documents into files.

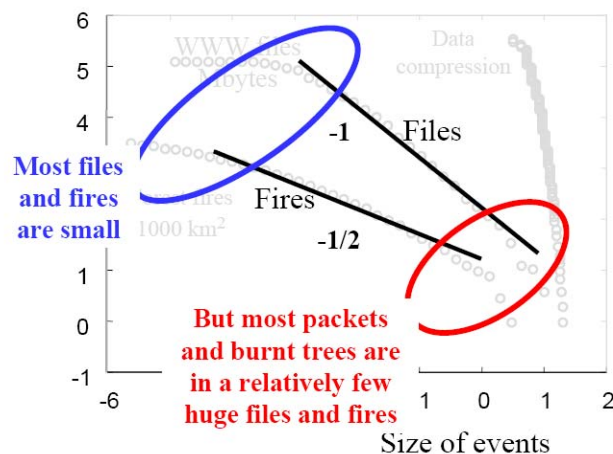


FIGURE 12

The big message (as shown in Figure 12) is that most files and fires are small and most earthquakes are small. You don't feel most earthquakes, but most packets and burnt trees are associated with those few largest events in these statistical distributions. If you sit around and wonder if there are earthquakes and fires happening, the answer is "yes" they are happening, but the ones we really notice are these large ones. In the end this high variability is the important message to take away from power laws. They are much more important than the power laws themselves and it's not necessarily a bad thing.

The fast protocol that John Doyle mentioned yesterday is really taking advantage of this in the context of protocol development for the Internet. These statistical distributions for fires and Web traffic have led us into looking at the Web problem and Internet traffic problem talked about yesterday. I went with John and some colleagues of mine in the geography department at the University of California, Santa Barbara, and looked at issues associated with real forest fires. We had this nice statistical fit—how does that compare to real fires? It's a very simplified mechanism to imagine trees burning with just assignments to the perimeter. We worked with some colleagues who are fire experts looking at statistics and fire simulators in real forests. For a real fire, hazard factors aren't just associated with proximity of trees and fire breaks. Things like the terrain matter, how dry the trees are, and whether or not they are dead or how old they are. Weather is a big factor, especially in California where you have the Santa Ana winds, which are very high wind conditions that create very dry and hot days. Those are dominant factors in California.

We took a model that had been developed for individual fires, HFire.net, and generalized it to account for long extended records of fires. It was based on satellite data for the topography, fuel measurements on the landscape, and models for how the fuel evolves once it is burned, or if it hasn't burned, and historical data from weather catalogues. We generalized this model so it could run for long periods of time, and then collected statistics. Figure 13 shows some work that we have done, which I'm not going to explain in huge detail, but we found similar kinds of fire footprints to the types that are seen in actuality.

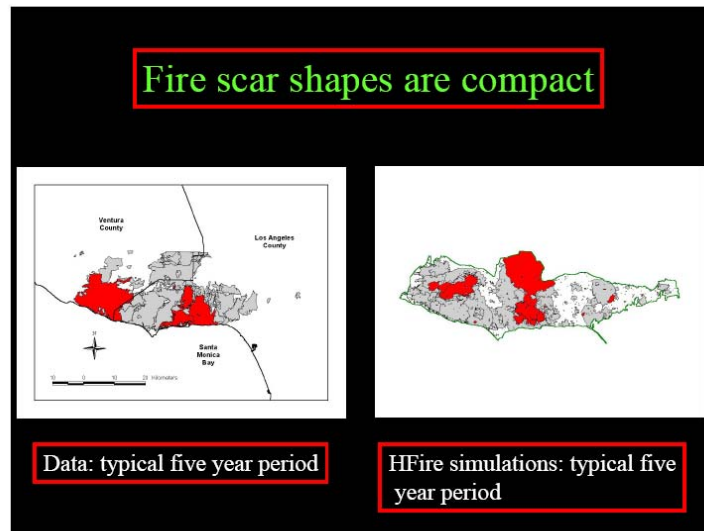


FIGURE 13

The statistics also provide an excellent statistical agreement between this very simplified HOT model and about the best data set that exists for fires, as shown in Figure 14. This is a starting point for lots of other things that you could do to think about more realistic fire models. It connects with a simple detailed picture and may be an interesting place to generally look at how these models could be used in terms of evaluating hazards and policy and social and economic impact. This is one aspect of this work.

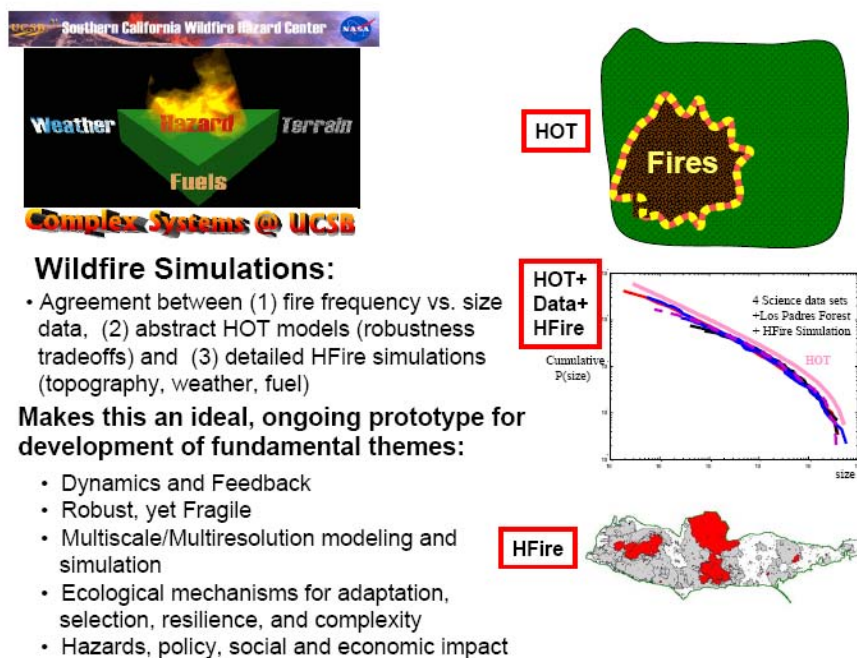


FIGURE 14

One of the questions that people often ask in terms of natural disasters like fires is are they getting worse? Clearly, in the case of fires, there has been 100 years or so of suppression policies mandated by the U.S. government, which has led to a build-up of fuels in our forests. In addition, we are urbanizing these fire-prone areas, urban-wildlife interfaces, which increase economic risks. Climate change may also play a role. Large wild fires are not a random and unexpected occurrence; they are to be expected and our policy has, if anything, made things worse.

Another interesting statistic is we spend about 100 to 1 in terms of reactive spending as opposed to proactive spending in terms of natural disasters. If we could find some ways to use information ahead of time, it's still a drop in the bucket. Another important point is that regular burn cycles for forests are an intrinsic part of the ecosystem dynamics. If you don't have regular burn cycles, when you do burn, biodiversity increases in the recovery period. Many plants in these areas are well adapted to fires and seeds will not germinate without them. You are going to have fires and if you don't there is a problem. That's part of the message, too.

I want to move beyond fire and talk about some of the other natural disasters. As shown in Figure 15, fires have power law statistics, as do hurricanes, floods, and earthquakes. Again, it

isn't the most important part that they have power law distributions; it's much more an issue that there is this high variability. Also, in looking at the global distribution of natural disasters, forest fires are really only a small piece of the pie. Floods dominate the losses. Earthquakes are also a relatively small piece of the pie. Windstorms are large, with Asia the dominate region in terms of losses.

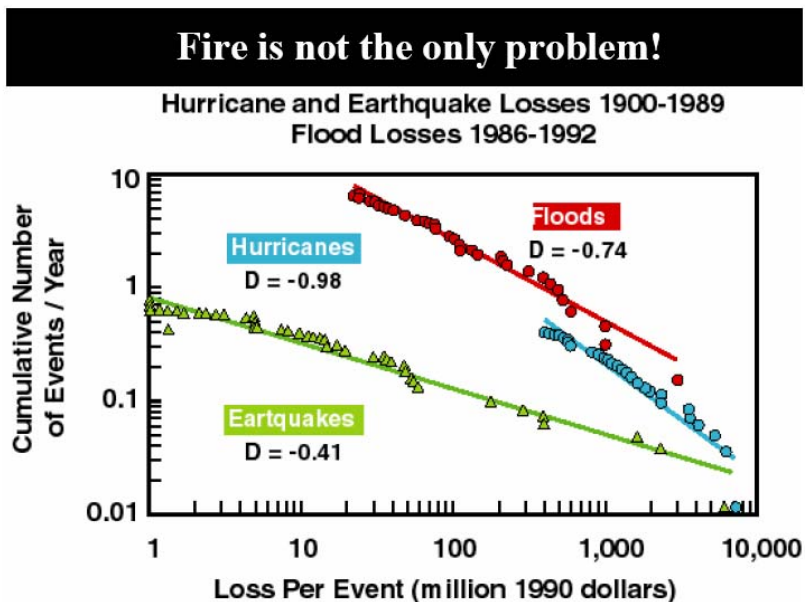


FIGURE 15

A large number of different approaches can be taken to break up this kind of data in terms of the damage associated with natural disasters. In dollars, geophysical phenomena are large, but when it comes to the number of people affected, these are fairly well localized in regions like California in the United States, and in terms of number killed.

Over the period 1970-2000, there was an increase an increase in the number of natural disasters recorded worldwide, with the increase in damage going up in a very steady way over that period. More people aren't dying overall but there has been an increase in loss. Again, a large part of the reason for these trends is that large populations are accumulating in coastal regions, which have high seismic activity, opportunities for flooding, hurricanes, and so on. Tectonic activity is in many cases also localized in coastal regions, and so damages from earthquakes have also gone up even though the number of earthquakes is not increasing.

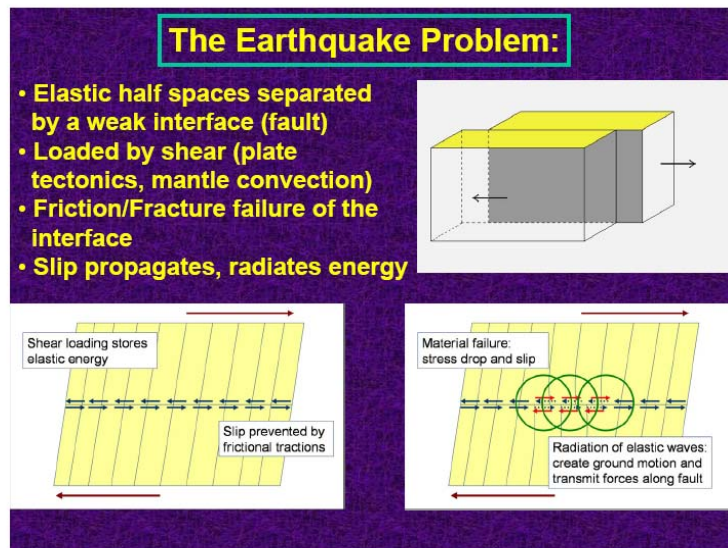


FIGURE 16

How do earthquakes happen? What's the physics? Figure 16 presents a cartoon-like picture. The idea is you've got plate tectonics, and you've got the crust of the earth that is like an egg shell on top of the mantle that is slowly convecting, which drives the relative motion of these plates. There are weak interfaces in between tectonic plates where stresses accumulate that create a slipping event when the material along the interface fails, and that sets off waves that radiate through the ground. Figure 16 shows a lateral fault, which is like the San Andreas Fault, but there are different kinds of faults. Again, this is a cartoon; the slip does not occur homogeneously. There are different things you can look at such as the dynamics, the complexity of the slip itself, and the fact that it's complicated. You can also look at the dynamic complexity of the radiation as shown via simulation in Figure 17. It doesn't have much interesting dynamics in the slip itself, but it is showing you what would happen in Los Angeles if you had a very simple slip pulse propagating down the San Andreas Fault, which is cartooned by that line.

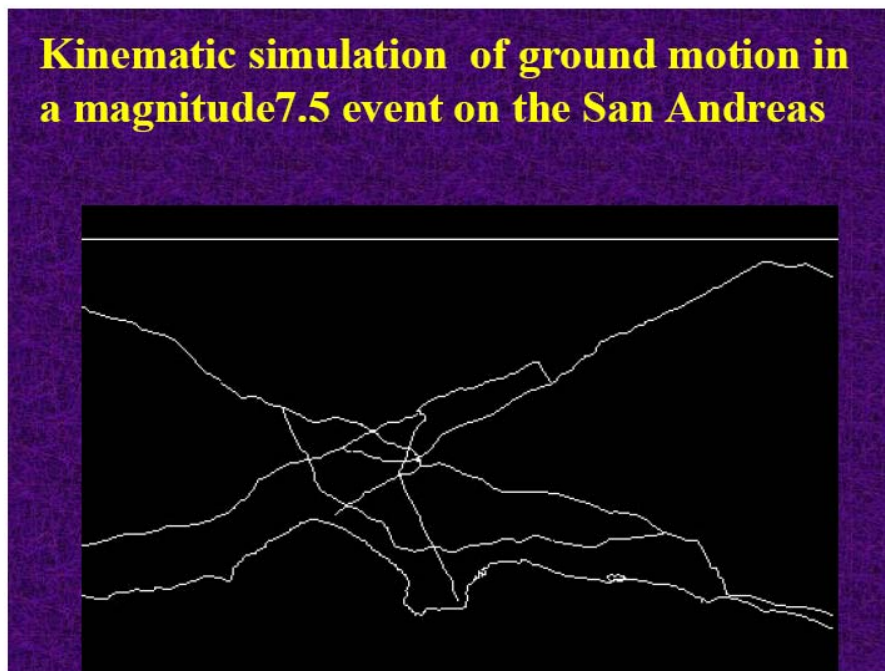


FIGURE 17

Figure 17 shows Los Angeles and its freeways. You can see that the ground motion is complicated, and the reason is because the hard rock basin underneath Los Angeles is complicated. What it looks like is known because of oil exploration. You can find that there are some places where you don't want to be, which has to do with such things as resonance effects. These kinds of models are the things that people use to try to set building codes in different parts of the city. I'll come back to earthquakes a little bit at the end, which is the area in which I have worked the most. But first I'll talk a bit about the Sumatran case and the tsunami. What caused the Sumatran earthquake? A simple answer is 200 million years of continental drift, as the Indian plate slides under the Asian plate. Figure 18 shows that collision, with the red area being the portion that slid and caused the Sumatran earthquake.



FIGURE 18

If you superimpose the region that slid on a map of California (as shown in Figure 19), it's an enormous magnitude 9 earthquake. This just shows how significant the Sumatran quake was.

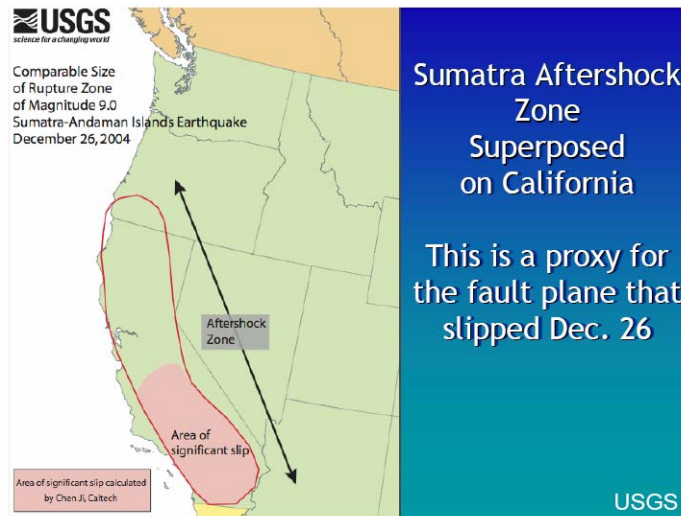


FIGURE 19

This particular earthquake is not the lateral kind, the kind of earthquake that makes tsunamis happen at subduction zones. Subduction zones are where you have two plates that, rather than slipping side-by-side, one is going under the other. They are out in the ocean where things are spreading up; the sea floor is spreading and there is a divergent zone. Material comes out and travels across the ocean very quietly. When it hits another plate it goes down, as shown

in Figure 20, and when it goes down something goes up. This happens underwater, so you have this water that doesn't want to sit like this and it has to cope with that. That is how you get a tsunami; waves are set off in both directions, as shown in Figure 21.

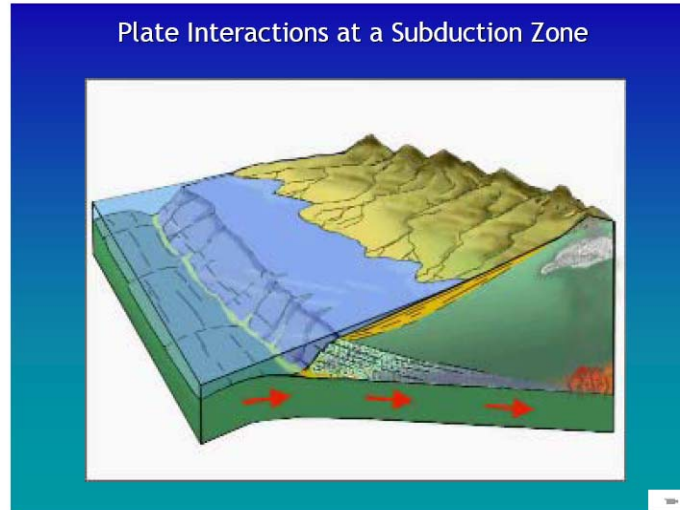


FIGURE 20

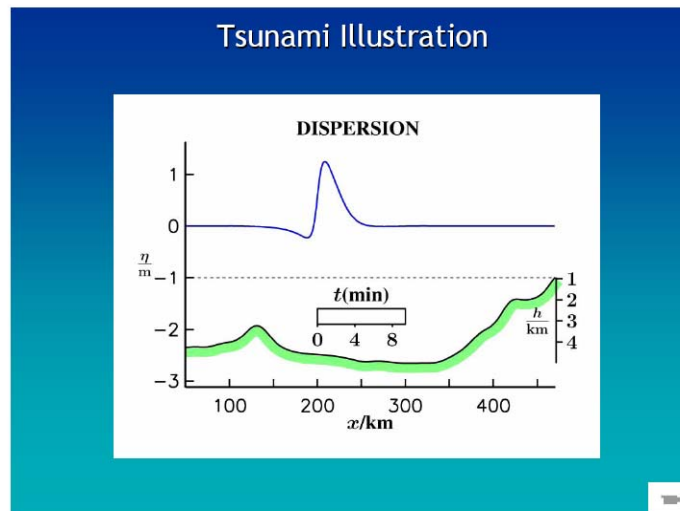


FIGURE 21

When the tsunami is a wave in the ocean, it is about one meter high—less than a meter high so it's nothing; it just goes along in the ocean. When it comes up against the coast, the slope amplifies the wave, as shown in Figure 22, and it can typically go to 10 meters, but it can also be

hundreds of meters. There are some known examples where it is higher. High does not necessarily mean high damage, but there is a wide range of heights of these waves. You know that they are coming, and you can estimate how long it's going to be. But tsunamis travel at hundreds of miles an hour across the ocean. The question is whether or not people know one is coming.

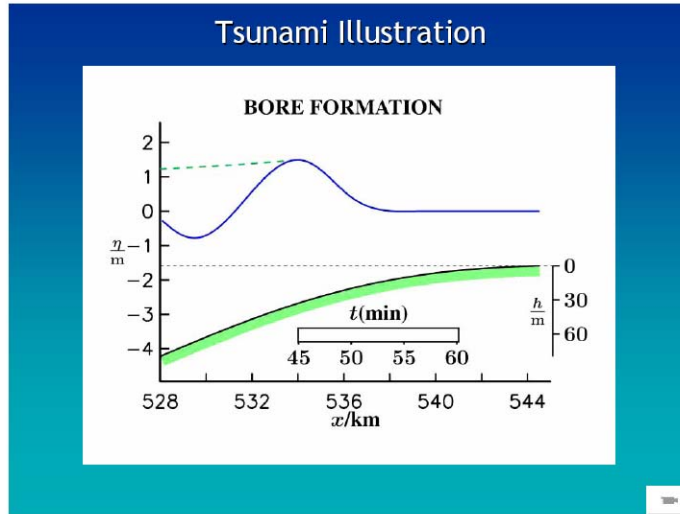


FIGURE 22

There are all kinds of simulations of the Sumatran tsunami. We have a simulation of the disturbance traveling across the ocean in these simulations, and you can estimate how long it will be until it hits various coastal places. Figure 23 shows the tsunami's progress in hours. It took about two hours to get to Thailand. It was Sumatra that didn't have very much warning. It's a lot more warning than you would ever have for an earthquake, but that might not be enough if you're out on the beach. So the question is whether or not we really have effective warning systems for some of these things.

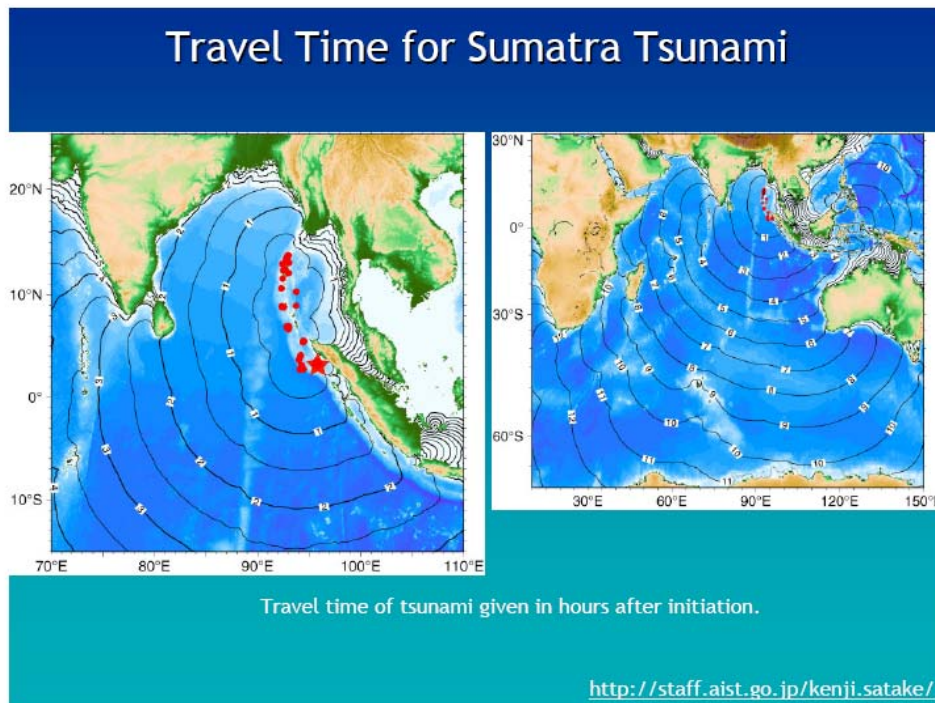


FIGURE 23

There are lots of tsunamis, and Figure 24 shows statistics for a bunch of recent ones. A magnitude 9 earthquake is a big earthquake, so there was a lot of damage. But not all tsunamis are damaging. The red ones in Figure 24 are the damaging ones and the white ones aren't.

Along the Pacific Coast in the United States we are always at risk for tsunamis. The biggest chance for tsunamis is most likely up in Alaska and in the Chile fault. We would have a lot of warning so we sort of sat around and said, well, look, what would you do? I guess what happens is that the show is on the TV, you probably get an e-mail hours in advance, but you might be out on the beach on vacation, and you might not know, so the policemen and fire departments go out and tell you and try to clear people off the beach. Some people are afraid to tell you because they fear people will go to the beach to see it. That's also taken into account.

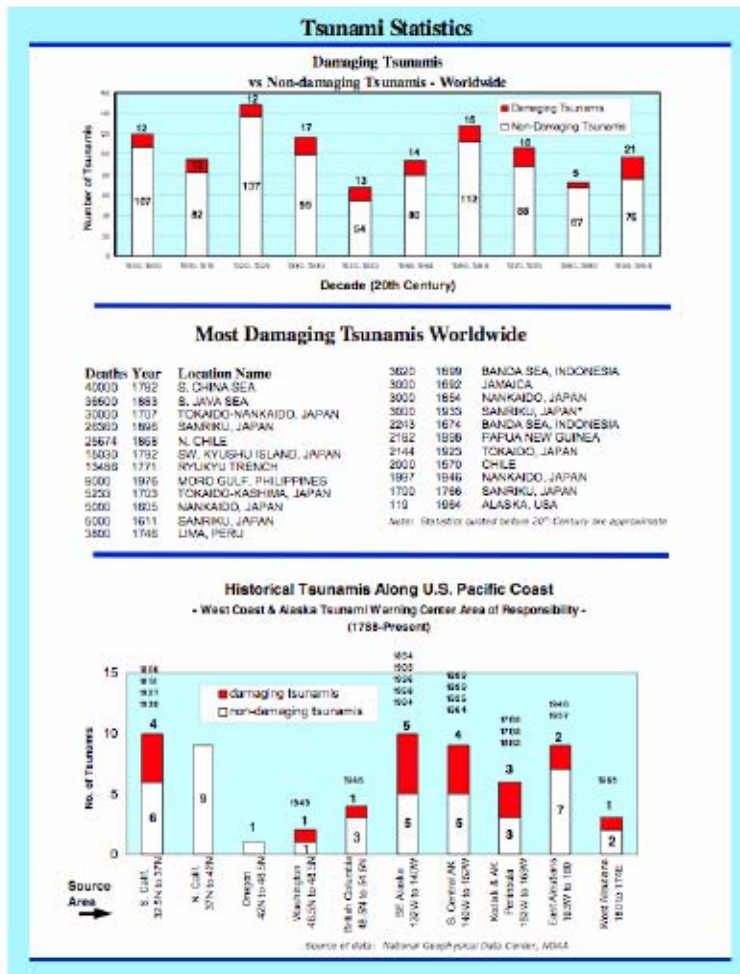


FIGURE 24

Hurricanes are another category of natural disasters that can be extremely costly. Scientists are not surprised that a large one like Katrina will happen, but it doesn't happen every day so we get used to thinking that it's not going to happen. There is a question of policy and where you allocate your resources. There is lots of research going on in predicting the intensity, how the intensity will change, and predicting the tracks of a tsunami or a tropical storm once it starts. There is also a lot of effort going on to predict what's vulnerable by modeling so there's a lot of interesting work in progress going on out there with the urban area topography. The thing that you can do here regarding something like earthquakes is couple with field research; go out and measure the hot spots in the ocean, and you can tell what is going to spin up the tsunami rate.

A lot of the vulnerabilities that we see are things that have to do with finite resource allocations and their impact over all kinds of scales on our sort of social network structure, as indicated by Figure 25. There is the geophysics and hydrodynamics, which has its impact on homes and families, infrastructure and energy. It puts stress on our hospitals. If they are already full or they are closing because of other stresses like insurance, you won't have Emergency Rooms for people to receive treatment. We won't have the military as much on the homeland to come and respond to our disasters if they are already allocated abroad. This impacts our transportation and communications systems. Fuel prices rise and it impacts airlines. Airlines are already going bankrupt; therefore, there is a huge stress on the system. It comes all the way up to global economic issues, politics and natural resources on large scales. I think the thing that is so striking is how a shock, like a hurricane, to a about robust-yet-fragile system can lead to cascading failures all the way up the chain.

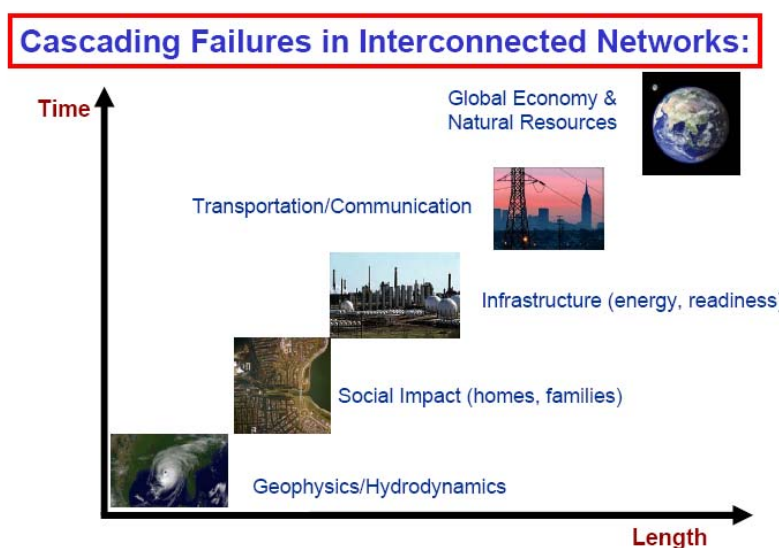


FIGURE 25

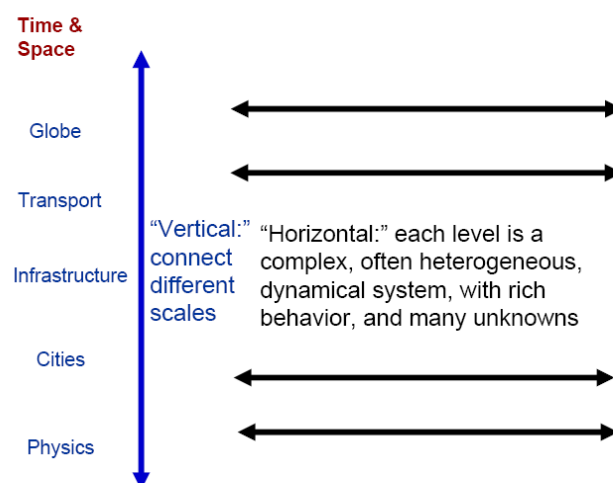


FIGURE 26

In Figure 26 I tried to cartoon the whole issue of scale, going up in time/space. Since most things happen on the diagonal, you might collapse that into a vertical time and space connecting these different scales, and horizontal issues associated with modeling the physics or geophysics on any one of these scales. I would say for natural disasters we focus enormously on the horizontal aspect and not very much on how we can transmit information from one scale to the other. I think it's a huge issue, and my plea is to the people here that are involved in sociology is how to think better about this problem.

I have been thinking about how to address the risk of earthquakes in that multi-scale way. One can model fine-scale geophysics, the impact of that on friction laws, the impact of friction on faults and networks, and all the way up to things that have to do with hazard evaluation policy and building codes. Part of the problem is that in this case, there is the issue of modeling on the horizontal scales of Figure 26 and then trying to connect the scales. When you get to the point of understanding how to set insurance rates and policies in terms of reaction to disasters, the vertical challenges dominate the issues.

So, dealing with uncertainty in seismic hazard analysis requires addressing the horizontal challenges—identifying the range of physical behaviors that are plausible—and also addressing the vertical challenges, such as uncertainty management and how to pass information between scales in a useful way. You might think of seismic hazard analysis as being represented by the elements in Figure 27, and to some extent they are there in the background. In the end, though, a lot of economics and policies are based on a single number, which in this particular case is a 62

percent chance of a magnitude greater than 6.7 happening in a 20-year period or something like that in the Bay area. There are all kinds of statistical problems associated with what this number is, and that's what my student Morgan Page and I have been looking at.

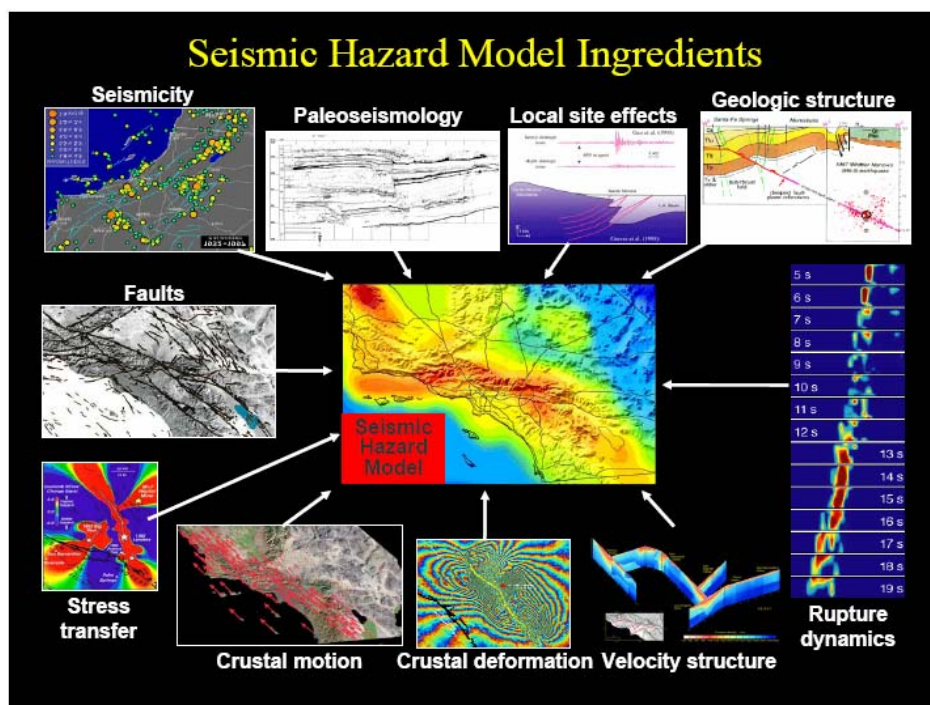


FIGURE 27

There needs to be more rigorous statistical methodology for combining data in these uncertain worlds, and also to incorporate physical constraints that come from modeling and simulation and ground motion estimates and so on. In this issue of dealing with uncertainty, vertical challenges really dominate our ability to estimate things like losses and risk to human life. If we could address these in a more systematic way maybe we would have a stronger impact on policy.

QUESTIONS AND ANSWERS

DR. GROSS: I'm Shula Gross from the City University of New York. My question is you showed a plot that suggested a power law for fire damages. It's funny, because statisticians and econometricians usually look at the tails, and we care more about the tails than about the center. You somehow did the reverse.

DR. CARLSON: I think there are cut offs partly because of physical sampling, and in some cases, if you look at data for a particular region, like Los Padres National Forest, there is a larger size that is a constraint in that particular forest based on terrain, and where we have got urban areas, or where you hit desert or where you hit rivers and so on. But the tails are really important. There is again, this very broad span. There is a cut off at the low end, too, of fires we just don't bother to measure, and so it's the fact that there is this broad span and natural cut offs at the two ends.

DR. BANKS: This goes a bit outside the purview of the conference, but I wonder if you have any comments on the following. When you have a country like the United States, which basically is self-insuring against natural disasters, one can usually look at the historical record, and one of your early slides did that, to sort of give you a forecast of what the total costs are going to be in any given year. That sets the level at which money must be collected in order to maintain a balance on that. But then one might very well use some of that money to invest in efforts to harden areas against disaster. I just don't know about the economic theory that drives self-insured agencies. Do you know if anybody is looking at that type of thing?

DR. CARLSON: I don't know about that, but I think it falls within this category of 100-to-1 reactive spending, where we don't invest very much in research, and we don't invest as much as we should in building stronger barriers against these kinds of events. I think that's huge. We know in many cases that we are operating at or near capacity. So, with things like the power grid, we know that we are operating at or near capacity. If we put more resources in, we would be okay, but instead we have power failures. It's going to get worse instead of better because of increased population, increased demand.

DR. SZEWCZYK: So, why do you live in California?

DR. CARLSON: Yes, I think that's a really good question. I grew up in Indiana, and I was afraid of tornados in Indiana. So, in Indiana you would watch the news, and there would be these tornados that come through, then I went to school on the East Coast, and I went to California. I think the first earthquake that occurred after I moved to California was the one in Canada that people felt in New York. California is a little bit crazy but it's beautiful though.

Stability and Degeneracy of Network Models

Mark S. Handcock, University of Washington

DR. HANDCOCK: There has been a lot of discussion related to the ideas I'll be talking about today. This is not a new topic at all but I think its worthwhile looking at how these particular issues relate to the particular problem of models and social networks. My main idea is to see how very old issues and very old concerns apply or play themselves out in the context of social network models.

This is joint work and I'm going to go through some things here, but if you really want to find it, please look at these working papers on the Web at the Center for Statistics and the Social Sciences (CSSS) at the University of Washington (www.css.washington.edu). Much of this work was with Martina Morris and with the University of Washington's Network Working Group, which includes Carter Butts, Jim Moody, and other folks that Martina mentioned yesterday. I particularly want to point out Dave Hunter of Pennsylvania State University, who is in the audience. Much of this work is done jointly with him.

I'll review some topics that have been gone through before. I won't belabor them now because they have basically already been covered. As I look around this room, the study of networks draws in a very diverse set of folks, and we have many varied objectives and a multitude of frameworks and languages used to express things. I don't think I'll spend much time on this. After one and a half days I think this is fairly clear.

One thing that is important to recognize, at least for the topics that I'll be looking at, is there are many deep and relevant theories here, and a deep statistical literature grouped into several communities. For instance, the social networks community has done a massive amount of work as you saw last night in the after-dinner talk. Two key references would be Frank (1972) and Wasserman and Faust (1994). The statistical networks community has also worked on this topic for a long period of time, and some key references are Snijders (1997), Frank and Strauss (1986), and Hoff, Raftery, and Handcock (2002). In particular, it's worthwhile to look at the work of David Strauss, which I think threads through much of what's going on and is very important from a technical contribution. Other contributions which I think haven't been discussed much at this workshop are from the spatial statistics community. There is a lot of very important work there and it's very closely allied. If you are looking to work in this area I think it's very important to read this stuff very closely. Good pointers into that literature include Besag (1974) and Cressie (1993). Another important literature, which I'll mention later in this talk, is

exponential family theory. It's been worked out a long time ago and the best place to start is a seminal book by Barndorff-Nielsen (1978). The last relevant literature is that coming from the graphical modeling community. They work on related ideas that are actually closer than it would seem by just judging the commonality in terms used. A good pointer is Lauritzen and Spiegelhalter (1988).

Networks are very complex things, and how we choose to model them will reflect what we have tried to represent well and what we are essentially not trying to represent well. That choice is going to be driven much by the objectives we have in mind. This is an obvious point, a point that drives through much of scientific modeling, but I have found that it's very important for the networks area because of the complexity of the underlying models and the complexity of the phenomena we are interested in.

What we are probably interested in first would be the nature of the relationships themselves, questions such as how the behavior of individuals depends on their location in a social network, or how the qualities of the individuals influence the social structure. Then we might be interested in how network structure influences processes that develop over a network. Dynamic or otherwise, the classic examples would include the spread of infection, the diffusion of innovations, or the spread of computer viruses, all of which are affected by the network structure.

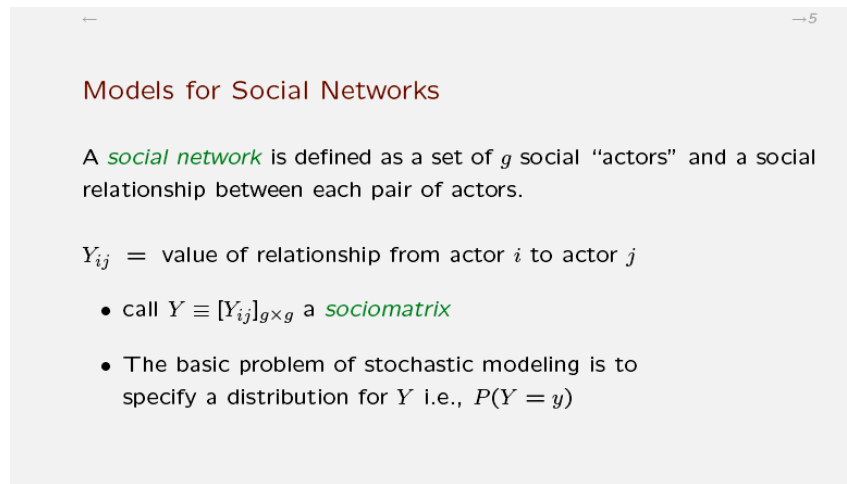
Lastly—and I think this is of primary importance and why many people actually study networks, even though they will look at it only after understanding the relationships and the network structure—is that we are interested in the effect of interventions. If we change the network structure and/or the processes that develop over that network, how will the network structure play itself out, and how will the process itself be actually changed? If you make changes to the network almost certainly the consequences of those changes will not be obvious.

Another important point is that our objectives define our perspectives. There is a difference between a so-called network-specific viewpoint and a population process. By network-specific I mean we look at a given network. It might be sampled or have missing data in it, but our scientific objective is that particular network. That is in contrast with a population viewpoint, where we view the data we have, whether it's complete or sampled, as a realization from some underlying social phenomena typically represented through a stochastic process, and we wish to understand the properties of that stochastic process. In that case, the observed network is conceptualized as a realization of a stochastic process.

I think a lot of the different approaches in the literature look very different because they have these different perspectives and objectives in that mind. For example, repeating a point

made last night by Steve Borgatti, one can compare a social-relations perspective with a nodal attribute compositional perspective. The former brings out interdependence/endogeneity and structure and positional characteristics, while the latter focuses on atomistic essentialism and reductionism. I won't say more about these things because I think they have been adequately covered.

Figure 1 defines a very basic model for a social network.



← —5

Models for Social Networks

A *social network* is defined as a set of g social “actors” and a social relationship between each pair of actors.

Y_{ij} = value of relationship from actor i to actor j

- call $Y \equiv [Y_{ij}]_{g \times g}$ a *sociomatrix*
- The basic problem of stochastic modeling is to specify a distribution for Y i.e., $P(Y = y)$

FIGURE 1

Hence, we can think of Y , the g -by- g matrix of relationships among the actors, as the sociomatrix. In a very simple sense, we want to write down a stochastic model for the joint distribution of Y , which is this large, multivariate distribution, often discrete, but it could be continuous also. All we want is a relatively simple “model” for the dependent structure of this multivariate random variable; that’s how a statistician would view this problem, and that view dominates the development.

← 6

Random Graph Distributions

Let \mathcal{Y} be the sample space of Y e.g. $\{0, 1\}^N$
 Any model for the multivariate distribution of Y
 can be *parameterized* in the form:

$$P(Y = y) = \frac{\exp\{\eta^T Z(y)\}}{c(\eta, \mathcal{Y})} \quad y \in \mathcal{Y}$$

Deag (1974), Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^q$ q -vector of parameters
- $Z(y)$ q -vector of *network statistics*.
- For a “saturated” model $q = 2^{|\mathcal{Y}|} - 1$
- $c(\eta)$ distribution normalizing constant

$$c(\eta, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta^T Z(y)\}$$

← Mark S. Handcock Stability and Degeneracy of Network Models →

FIGURE 2

Martina Morris showed Figure 2 yesterday but I’ll show it again just to bring it forward. Let \mathcal{Y} be the sample space of all possible graphs, for example, just presence or absence of ties, and we write down an exponential family model of the form shown. We also have the space of graphs that are possible. The denominator is a constant which is defined as the sum of the numerators.

One thing that I think is worthwhile to look at is a way of getting a general interpretation of the model through local specification. This is shown in Figure 3. We define Y^c as the graph excluding the ij component of the network. These are all the network elements, excluding the ij one. This is a standard formula, but I think it’s quite instructive. What it says is the probability of a tie between i and j given the rest of the network, divided by the probability of a non-tie between i and j with the rest of the network held fixed. The basic idea is that we consider the graph, and we hold the graph fixed and think about taking a particular tie: what happens when we change it from a 0 to 1? Looking at the odds of the tie conditional on the rest gives us a direct interpretation of these forms. It can also be interpreted in terms of a relative risk form. The basic idea gives us an interpretation of either, at least one or many others in terms of the local specifications of the actual form. How a particular dyad probably would tie in between a particular pair of actors is influenced by the surrounding ties. By specifying the local properties, as long as they are done in this way, you get the global joint distribution, going from the local specification to give you a global one. Again, this is relatively straightforward from an algebraic

standpoint. It's just helping us with the interpretation.

→ 1

Equilibrium: Local specification

Let \mathbf{I} be the set of indices of the unique elements of Y ,

$$Y_{ij}^c = \{Y_{kl} : kl \in \mathbf{I}/\{ij\}\} \quad y_{ij}^c = \{y_{kl} : kl \in \mathbf{I}/\{ij\}\}$$

$$y_{ij}^+ = \{y_{ij}^c \cup \{y_{ij} = 1\}\} \quad y_{ij}^- = \{y_{ij}^c \cup \{y_{ij} = 0\}\}$$

The full conditional distributions of Y_{ij} are

$$\frac{P_\eta(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P_\eta(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \exp[\eta^T \delta(y_{ij}^c)] \quad y \in \mathcal{Y}$$

where

$$\delta(y_{ij}^c) = Z(y_{ij}^+) - Z(y_{ij}^-)$$

- $\delta(y_{ij}^c)$ is the change in the graph statistics when y_{ij} changes from 0 to 1.

← Mark S. Handcock Stability and Degeneracy of Network Models →

FIGURE 3

I'll make a brief comment on models for the actor degree distribution, which is shown in Figure 4. We can write a particular model of this form where the statistics are just the proportion of actors with exactly k relationships, i.e., the proportion of the nodes with k relationship or degree k . Basically, we write just a linear predictor on these forms, this is the so-called degree-only model. In essence what it is saying is if we can condition on the degrees, and all of the graphs with those degree structures are equally likely, in that sense it's random mixing given a given degree sequence. This has a very long history in social networks community, and there are the two references shown in Figure 4 plus many others that are good places to start. This is receiving an enormous amount of work. The reason I'm pointing this out is because there has been a lot of work on these models, and they are being considered both from a technical and from a theoretical perspective. I often think that science is not helped a lot when that prior work is forgotten. The other thing you can do here is further parameterize the degree distribution, which essentially places nonlinear parametric constraints on the α parameters here. As most statisticians know, this moves it technically from just a straight linear exponential family to a curve exponential family. Dave Hunter and I did some work on models of that kind.

← -2

Models based on the degree distribution only

$$P(Y = y) = \frac{\exp\{\sum_{k=1}^{g-1} \alpha_k d_k(y)\}}{c(\alpha)} \quad y \in \mathcal{Y}$$

where

- $d_k(y)$ = the proportion of actors with exactly k relationships
- α $g - 1$ -vector of degree parameters
 - Long-history in social network community
 - Wasserman (1977), Snijders (1991)
 - Recent rediscovery and focus
 - Barabási & Albert (1999)
 - Newman, Strogatz and Watts (2001)
 - Further direct parametrization of the degree distribution
 - Forms with power-law behavior: Albert and Barabási (1999)

← Mark S. Handcock Stability and Degeneracy of Network Models →

FIGURE 4

We could then add in—Martina did this yesterday, and I’ll show this again briefly—co-variates that occur in this linear fashion, attributes of the nodes, attributes of the dyads. We can add some additional clustering component to a degree form in this way. This is shown in Figure 5.

← -3

Example: A simple model with arbitrary degrees and clustering

$$P(Y = y) = \frac{\exp\{y^T Z \beta + \sum_{k=1}^{g-1} \alpha_k d_k(y) + \rho C(y)\}}{c(\alpha, \rho)} \quad y \in \mathcal{Y}$$

- y is the N -vector of the unique elements of Y
- $d_k(y)$ = the proportion of actors with exactly k relationships
- $C(y)$ clustering coefficient of the network y
- α $g - 1$ -vector of degree distribution parameters
- ρ clustering parameter

FIGURE 5

I'll give you just a couple of illustrations of how this model might apply in particular instances. Say you wanted to represent a village-level structure with 50 actors, a clustering coefficient of 15 percent, and the degree distribution was Yule with a scaling exponent of 3. The Yule is the classic preferential attachment power law model, and we can just see how that looks. Figure 6 shows two different networks generated from that model.

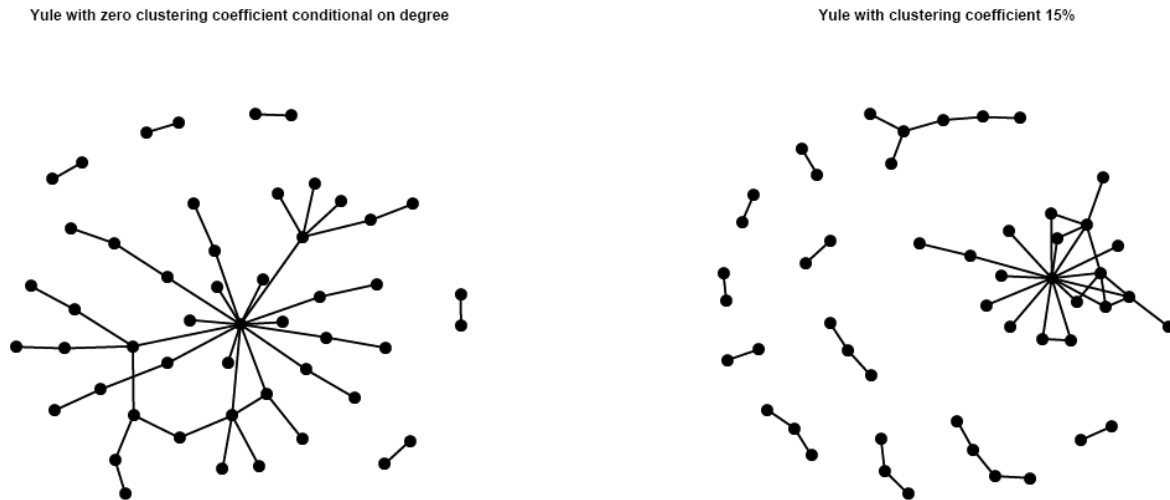


FIGURE 6

The network on the left side of Figure 6 is one with zero clustering, and on the right we see what happens when the mean clustering coefficient is pushed up to 15 percent with the same degree distribution. The basic notion is that these models give you a way of incorporating known clustering coefficients in the model, while holding a degree distribution fixed. Just to reiterate points made earlier that the degree distribution is clearly not all of what is going on.

Yule with zero clustering coefficient conditional on degree



Yule with clustering coefficient 15%



FIGURE 7

Figure 7 shows the same thing with 1,000 nodes, so it's a pretty big village. You can see to get the clustering going on with a given degree of distribution it's forcing a lot of geodesics in this form. Of course, with 1,000 nodes it's pretty hard to see.

Figure 8 is a bipartite graph; the same model form works for bipartite graphs. The network in the upper left is a heterosexual Yule with no correlation and the one in the upper right is a heterosexual Yule with a strong correlation triangle percent of 60 percent versus 3, which is the default one for random mixing given degree. There is a modest one here as well as one with negative correlation. I don't think I'm going to say too much more about these.

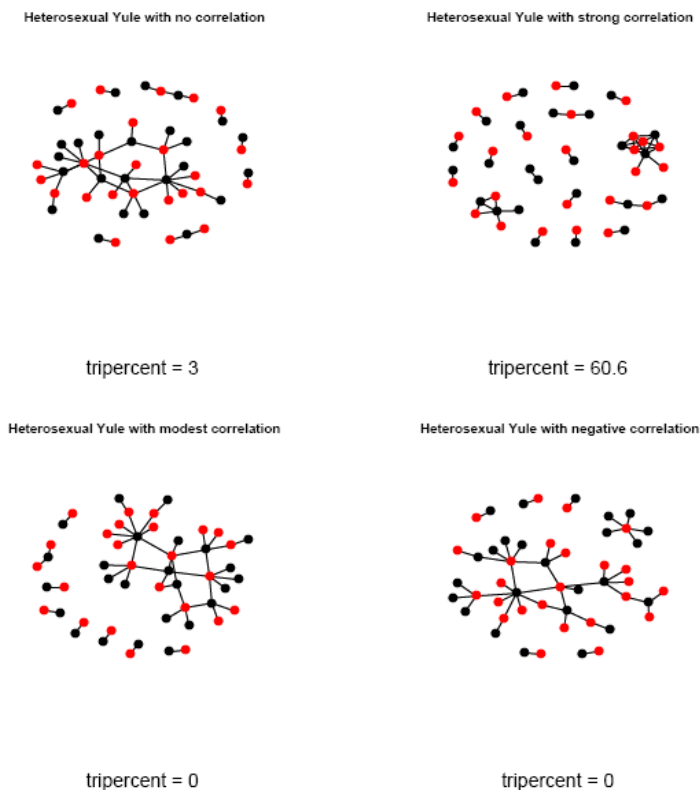


FIGURE 8

The main topic I would like to talk about today is to essentially address a canonical problem: how can we tell if a model class is useful? That is, if you write down a particular form of the kind in Figure 2, we can introduce statistics based on principles of underlying local specifications or local configurations, or we can just write down statistics that we believe would a priori, from a scientific perspective, explain a lot of the variation in the actual graph. But the natural question is, because these statistics will tend to be highly correlated with each other, highly dependent, it's not really clear for any given model exactly what the node qualities of that model would actually be. As Martina Morris showed yesterday, the natural idea of starting from something very simple can sometimes lead to models that aren't very good.

It is true the properties we saw were the properties of the model we wrote down. It's essentially saying that any implication that simple models have simple properties is not true. This has been known in statistical mechanics for a very long period of time. It's always a little bit of a surprise to see them occur in applied models or very empirically-based models. So, the basic question here is, is a model class itself able to represent a range of realistic networks? It's not the only question, but one question that could be asked here, and this is what I refer to as the issue of model degeneracy (Handcock, 2003). The idea is that some model classes will only represent a

small range of graphs as the parameters are varied. In circumstances where we like a fairly general model to cover a large range of graph and graph types that might not be a desirable property of a model.

The second issue is what are the properties of different methods of estimation, such as Maximum likelihood estimation, pseudo-likelihood, or some Bayesian framework? I'll make some comments on the computational issues where certain estimators do or don't exist in that many forms (e.g., see Snijders, 2002, and Handcock, 2002).

The last issue in assessing if a model class is useful is whether we can assess the goodness of fit of models. For example, we have a graph and we have a model, how well does the model actually fit the actual graph? I'll say some notes here about measuring this. I don't think I'll say a lot about this, but I think the application Martina went through yesterday was very interesting in terms of a stylized view of how that would be done. Some background on this topic may be found in Besag (2000) and Hunter, Goodreau, and Handcock (2005).

Figure 9 illustrates some points on model degeneracy. It is a property of a random graph model; it's nothing to do with data per se, but the model itself. We call it near degenerate if the model places all its probability mass, i.e., the likelihood of certain graphs on a small number of actual graphs. An example of this would be the empty graph, as shown in Figure 10. If we know that a model produces the empty graph with probability close to 1, that's probably not good, or the full graph or some mixture of them. In some sense, subsets of the set of possible graphs which are regarded for the particular scientific application is interesting. If we are interested in heterosexual graphs, and the model chosen places all of its mass on heterosexual graphs which is a large subset, that's a good thing. This idea is clearly application-specific.

→6

Model Degeneracy

idea: A random graph model is *near degenerate* if the model places almost all its probability mass on a small number of graph configurations in \mathcal{Y} .

e.g. empty graph, full graph, an individual graph, no 2-stars, mono-degree graphs

- Example: The *2-star* model

$$P(Y = y) = \frac{\exp\{\eta_1 E(y) + \eta_2 S(y)\}}{c(\eta_1, \eta_2)} \quad y \in \mathcal{Y}$$

is near-degenerate for most values of $\eta_2 > 0$

$$E(y) = \sum_{i < j} y_{ij} \quad S(y) = \sum_{i < j < k} y_{ij} y_{ik}$$

2-star
2-star configuration for undirected graphs

← Mark S. Handcock Stability and Degeneracy of Network Models →

FIGURE 9

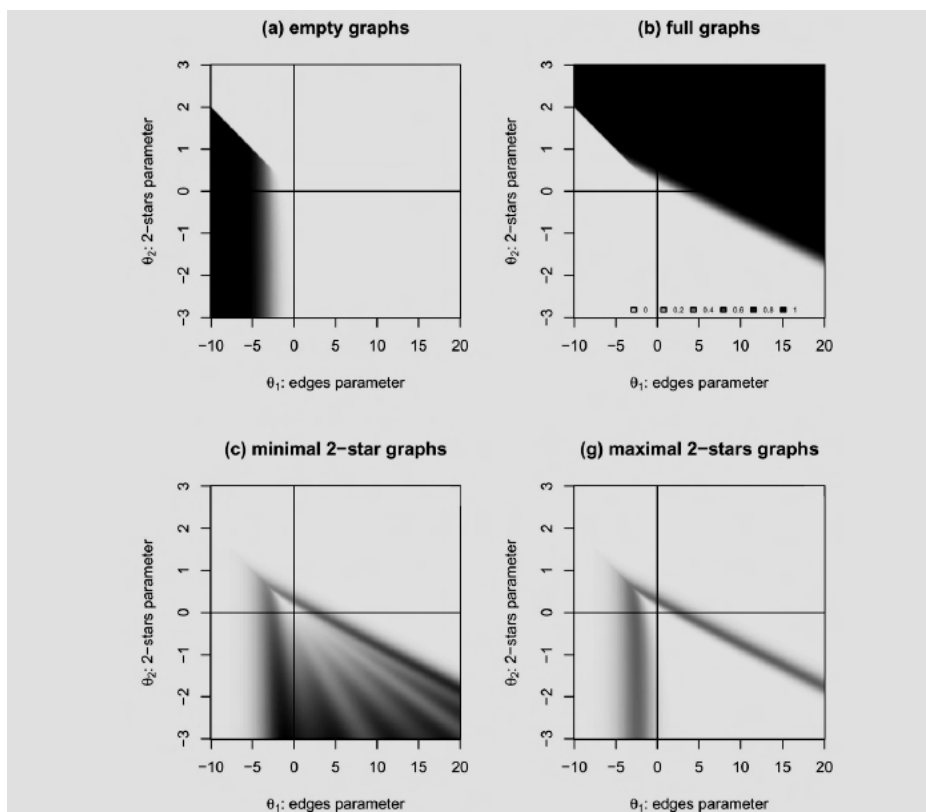


FIGURE 10

Let me illustrate this further with Figure 11, which is in some ways a toy model, but also in some ways an interesting model because of its connection to other fields. The horizontal axis relates to the number of edges in the graph, so it's a measure of graph density. The other axis is the so-called two stars, which I interpret as either the path between j and k going through a third node l , or another way to interpret it is a clustering of edges. How often do these two edges end up having a node in common?

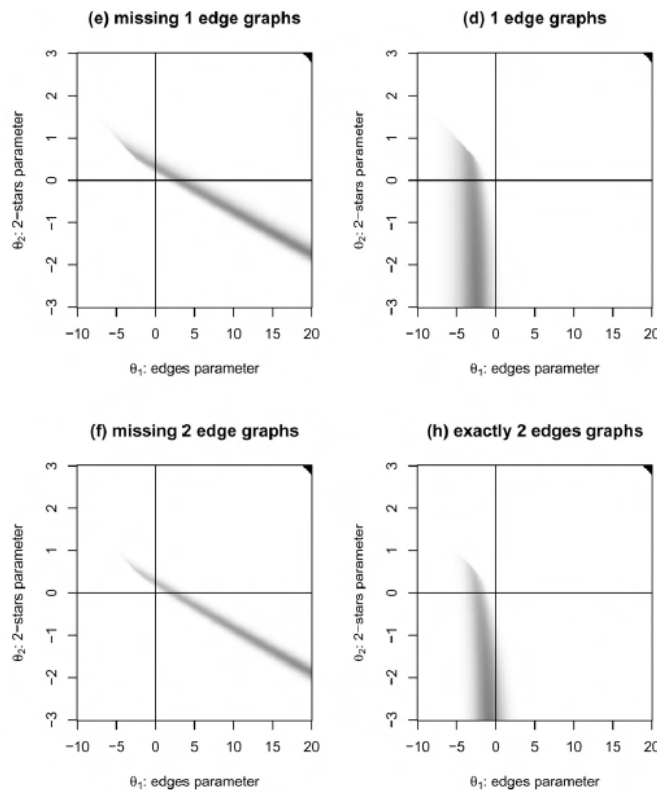


FIGURE 11

The reason I like the second interpretation is this model has analogues in other areas. One analog would be the Ising model for lattice-based processes. It also has an analog as a Strauss model for spatial point processes in the very least. Basically, what it has is an overall density parameter, and a single parameter dependence measure. The clustering is of a certain kind, of course, because it's simple. If this parameter of 2 is positive here, it's a sense of placing more probability on graphs with the positive cluster in here. If it's negative it actually places on "regular" graphs: graphs that essentially have edges that do not end up at the same node. What we actually see, although this looks relatively simple and relatively easy to work with, is that it is

near degenerate for most positive values of n_2 . While this is particularly galling to social networkers, as was pointed out yesterday for many social processes you have a positive clustering of ties, and as a result would expect most social process to have a positive parameter. This would seem like a good model to actually start, but in actual fact it is not a very good model at all. I'll show why this is true. I'll only focus on the plot shown in Figure 12.

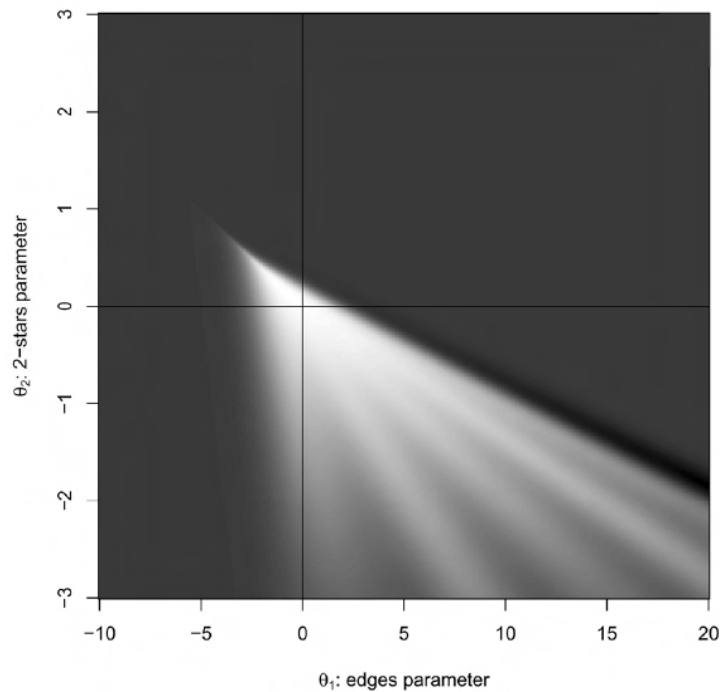


Figure 3: Cumulative Degeneracy Probabilities for graphs with 7 actors.

FIGURE 12

Figure 12 is just a plot of the 2-dimensional parameter space of this model. If the parameters are in the dark region, then all the models produce empty graphs. So, basically, it's saying the density is so low and the clustering is so low that you get empty graphs. You essentially have one area of the parameter space which is producing essentially all of the mass in a very small subset of models. It's only in this small area of the parameter space, around 0, which corresponds to just purely random graphs with equally likely ties, that you are about to get non-degenerate forms. Most of the time that θ_2 is positive here, this entire upper hemisphere or upper part of the graph you actually have mostly degenerate models. It's only the little small area here that you could actually get non-degenerate ones.

← → 1

Geometry of Exponential Random Graph Models

Consider the alternative parametrization of the models
 $\mu : \Lambda \rightarrow \text{int}(\mathcal{C})$ defined by

$$\mu(\eta) = E_{\eta} [Z(Y)] \equiv \sum_{y \in \mathcal{Y}} Z(y) \frac{\exp\{\eta^T Z(y)\}}{c(\eta)}$$

- The mapping is injective:

$$\mu(\eta_a) = \mu(\eta_b) \rightarrow P_{\eta_a}(Y = y) = P_{\eta_b}(Y = y) \quad \forall y.$$

- The mapping is strictly increasing in the sense that

$$(\eta_a - \eta_b)^T (\mu(\eta_a) - \mu(\eta_b)) \geq 0$$

with equality only if $P_{\eta_a}(Y = y) = P_{\eta_b}(Y = y) \quad \forall y.$

- Represents an alternative *parameterization* of the model

← Mark S. Handcock Stability and Degeneracy of Network Models →

FIGURE 13

Now I'll say a little bit about the geometry of exponential random graph models, as explained in Figure 13. For those of you who are very interested in this and are familiar with them, the papers on my Web site make this much more clear, so I'll go through quite a stylized form. The central idea is this. Rather than thinking about the natural parameterization of the exponential family, we can think about the classical mean value parameterization of that family defined by the following mapping. The parameter corresponding to the natural parameter of z is just the expectation over y of the graph statistics, explicitly of this form η of z . Note that this mapping is injective. It is also strictly increasing. If you look at this product here what you see is that if you have a positive increase in the mean value here you will always have this product positive here. So, it gives us some sense of the relationship between the natural and the mean value parameters. This is shown by Barndorff-Wilson and others, that it's just an alternative parameterization of the actual model. Why is that interesting? I think because this alternative parameterization has got a lot going for it, not necessarily to replace natural parameterization, but in a lot of scientific applications it's much more natural to look at.

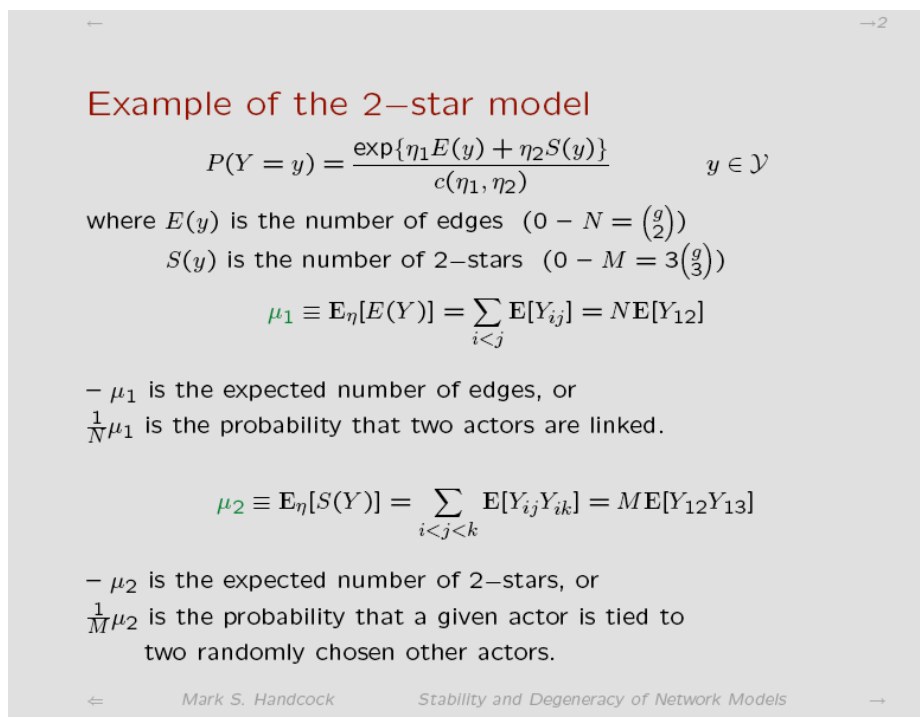


FIGURE 14

Let’s go back to the two star model as shown in Figure 14. Let E be the number of edges and S be the number of two stars. Then μ_1 , the value per parameter corresponding to number of edges is just the expected number of edges, or with the simple dividing by the number of dyads, it’s just the probability that the two actors are linked. As an interpretation I find this a very natural thing because if we just see this form here it gives it away. If the probably that the two actors are linked that is a very natural parameter of this model. Again, when you divide by the number of potential two stars here it is the probability that a given actor is tied to randomly chosen other actors. If we choose two randomly chosen other actors what is the probability they are tied to a third?

In Figure 15, we have a new parameterization in which the mean values are actually on scales that people usually think about in terms of social network forms, and I think they have a lot of value for that reason. The natural parameters in exponential families make the math really easy, but unfortunately their interpretation in terms of traditional log-logs can be very tortured at time because of the auto-logistic form. The auto dependence, meaning you are looking at some part conditional on holding some other parts of the graph fixed, which are closely dependent on them. And it is very, very tortured.

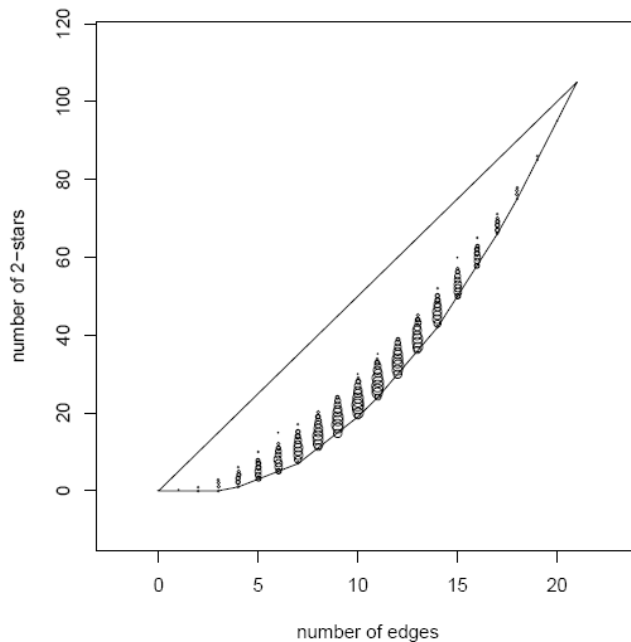


Figure 1: Enumeration of sufficient statistics for graphs with 7 nodes. The circles are centered on the possible values and the area of the circle is proportional to the number of graphs with that value of the sufficient statistic. There are a total of 2,097,152 graphs.

FIGURE 15

I'll briefly show some things here and then I'll move on. There are other parameterizations you can look at, such as that shown in Figure 16, which is just the mean value space. The natural parameter spaces across all of \mathbb{R}^2 . The mean value space is from the number of edges, so 0-21 in my simple example with 7 nodes, and from 0-105. It's a finite parameter space. The green area in Figure 16 is the interior of these points.

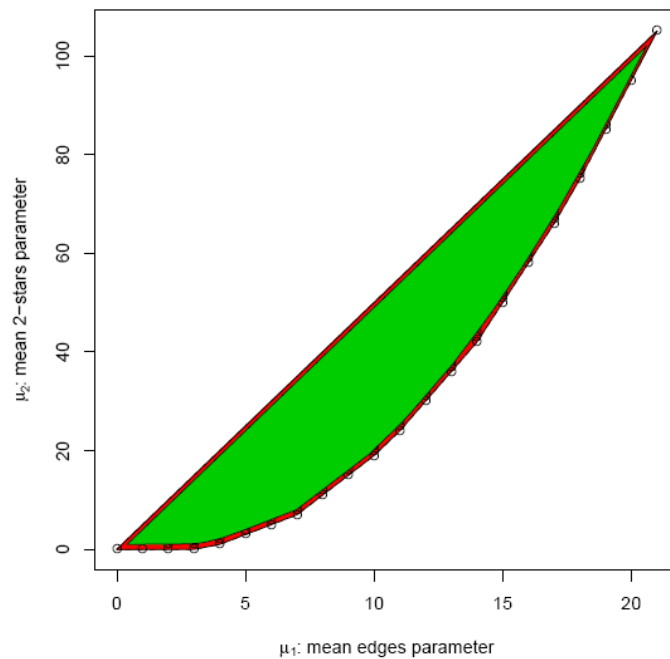


FIGURE 16

There are alternative mixed parameterizations which I find very useful that have a lot of very good statistical properties, and Figure 17 explains that. I want to say how this is relevant to the degeneracy issue, so we'll define that now explicitly: a model is near degenerate if $\mu(\eta)$ is close to the boundary of C . Let degree $Y = \{y \in Y : Z(y) \in \text{bd}C\}$ be the set of graph on the boundary of the convex hull. Based on the geometry of the mean value parametrization, the expected sufficient statistics are close to the boundary of the hull and the model will place much probability mass on graphs in degree Y . If Figure 18 shows our model space, then say you are close to the boundary of it, which I arbitrarily say is just this red rim around the corner. I'm essentially going to claim that models in this red rimmed boundary of the parameter space, the boundaries of the corresponding convex hull, that is going to be models which are degenerate. And models in the center are somewhat less degenerate, but it's a graded thing. Based on the geometry of the mean value parameterization, which is the expected statistics from the actual model, it is going to look very bad if you are right next to the boundary. This can be quantified in very many ways due to mathematical theorems, which I won't dwell on here.

→3

In some cases mixed parameterizations may be better

Let $(Z^{(1)}, Z^{(2)})$ be a partition of Z such that:

- $Z^{(1)}$ is interpretable as a mean value parametrization
- $Z^{(2)}$ is interpretable as the “natural” conditional log-odds

Consider similar partitions $(\eta^{(1)}, \eta^{(2)})$ of η and $(\mu^{(1)}(\eta), \mu^{(2)}(\eta))$ of $\mu(\eta)$.

Let $\Lambda^{(2)}$ be the set of values of $\eta^{(2)}$ for η varying in Λ and $C^{(1)}$ be the convex hull of $\{Z^{(1)}(y) : y \in \mathcal{Y}\}$.

The mapping $\zeta : \Lambda \rightarrow \text{int}(C^{(1)}) \times \Lambda^{(2)}$ defined by

$$\zeta(\eta) = (\mu^{(1)}(\eta), \eta^{(2)}) \tag{1}$$

is a *mixed* parametrization of the model (\mathcal{Y}, Z, η) .

The components $\mu^{(1)}$ and $\eta^{(2)}$ are variationally independent, that is, the range of $\zeta(\eta)$ is a product space.

←

Mark S. Handcock Stability and Degeneracy of Network Models

FIGURE 17

Figure 4: Regions of the parameter space of μ

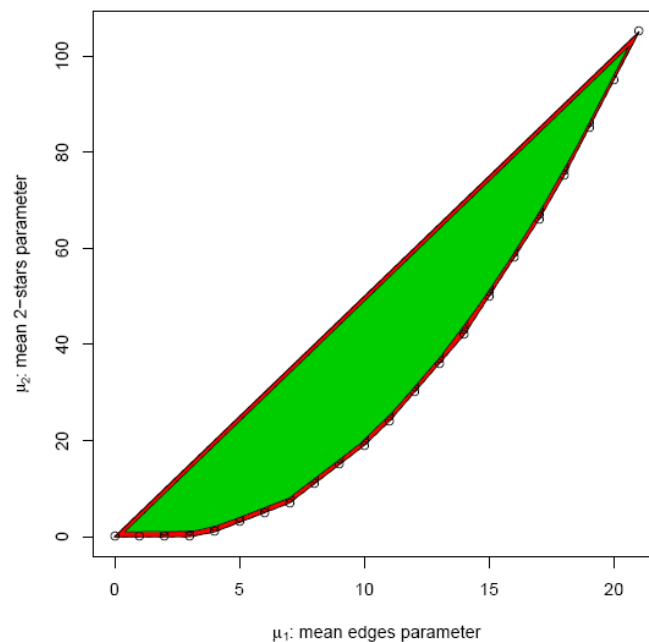


FIGURE 18

I will ask a natural question: what happens when you map this red region around the edge of the parameter space, what I'm calling the degenerate region, back to the natural parameter space, and what happens when you map the green? Figure 19 gives the idea here and I think this explains a lot why people have had a lot of trouble modeling in this class.

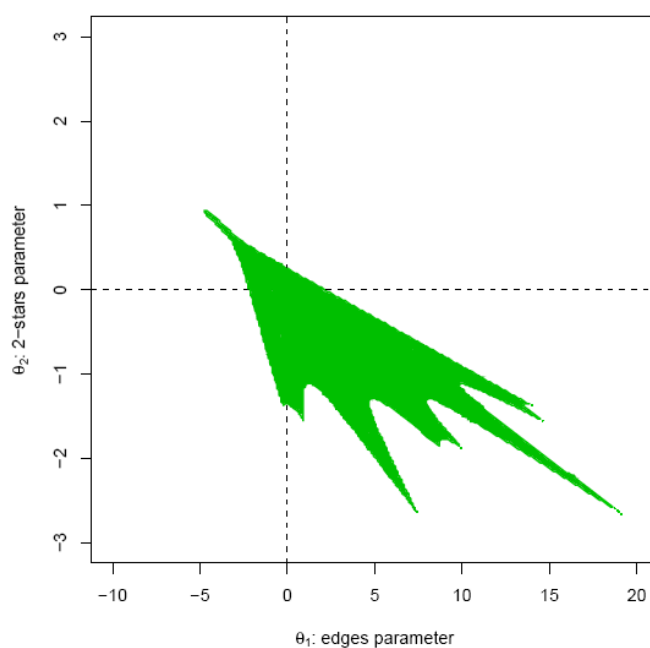


FIGURE 19

What I have done in Figure 19 is map the green region on the previous slide back to the natural parameter space here, which is all of \mathbb{R}^2 . What we see is that the green region is mapped to this very jagged form here, and that small boundary red region is mapped to every other place. In essence, what this is saying is that if your parameter is in the region outside the green region, it's essentially going to give you mean values on the edge of the parameter space, and hence very odd looking forms. The only reasonable values are the values in this small green area. What has gone on, in a nutshell is that people will be looking at changing the parameter values in the natural parameter space here, and we will be doing this fine. Moving it so slightly over and they will step off the edge into the other area here, which gives you very bizarre looking models. The geometry of this, which looks like some sort of "stealth bomber" gives us some sense that if you move around in a nice, smooth way in this space, you fall off an edge, you are in trouble.

To make matters worse, most of the interesting models are in the positive dependent area so they are out on this peninsula of models. In a nutshell, what people are doing is changing

parameter values in this peninsula; make a small change, turning off the edge into a very degenerate model form. There are a lot of problems in this example, but it follows through more generally due to the nonlinear nature of the mapping here, which leads to rapid changes in model behavior through small changes in the natural parameter space. This statement can be quantified in a number of ways as shown in Figure 20:

Result: Let e be a unit vector in \mathbb{R}^q and $\text{bd}(e) = \sup_{\mu \in \text{int}\mathcal{C}}(e^T \mu)$.

1. $\mu(\lambda e) \rightarrow \text{bd}(e)e$ as $\lambda \uparrow \infty$.
2. $P_{\lambda e, \mathcal{Y}}(Y \in \text{deg } \mathcal{Y}) \rightarrow 1$ as $\lambda \uparrow \infty$.
3. For every $d < \text{bd}(e)$, $P_{\lambda e, \mathcal{Y}}(e^T Z(Y) \leq d) \rightarrow 0$ as $\lambda \uparrow \infty$.
4. Let $\eta_0 \in \text{int}\mathcal{C}$.
 Then Kullback – Leibler divergence($\eta_0; \lambda e$) $\rightarrow \infty$ as $\lambda \uparrow \infty$.

FIGURE 20

I'll briefly speak about inference for social network models, although we can do inference based on the likelihood as before. We have a probability model. It's natural to think of likelihood biased inference as shown in Figure 21.

Inference for Social Networks

The log-likelihood is

$$\mathcal{L}(\eta; y) \equiv \log [P(Y = y; \eta)] = \eta^T Z(y) - \kappa(\eta) \quad y \in \mathcal{Y}$$

where $\kappa(\eta) = \log[c(\eta)]$.

Maximum likelihood is difficult as direct evaluation of $\mathcal{L}(\eta; y)$ requires calculating

$$c(\eta) = \sum_{y \in \mathcal{Y}} \exp\{\eta^T Z(y)\}$$

where $|\mathcal{Y}| \approx 2^{y^2}$

FIGURE 21

I will also say briefly about the existence and non-existence of maximum-likelihood estimators. The classical result, due to Barndorff-Nielsen (1978), is that we have necessary and sufficient conditions for the MLE to exist here. This is a classic result but it has enormous impact. If the observed graph statistics are in the interior of the convex hull of the discrete

support points for the sufficient statistics, the MLE exists, and it can be found. It is unique and can be found by directly solving equations or by direct optimization.

Pseudolikelihood estimation for Social Networks

A log-pseudolikelihood is

$$\mathcal{P}\mathcal{L}(\eta; y) \equiv \sum_{i < j} \log [P_{\eta}(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)] \quad y \in \mathcal{Y}$$

The pseudolikelihood estimator of η is then

$$\tilde{\eta} = \operatorname{argmax}_{\eta \in \Lambda} \mathcal{P}\mathcal{L}(\eta; z)$$

Easily implemented via a (formal) logistic regression

Statistical properties of pseudolikelihood estimators for social networks are only partially understood.

FIGURE 22

Figures 22 and 23 tell us exactly what and how to do in terms of maximum likelihood. On the other hand if it's on exterior of the convex hull, the MLE doesn't exist. So, if we use a method of optimization to find it we're in deep trouble. This caused a lot of problems in the past when people didn't realize that was what was actually going on.

Existence and uniqueness of MPLE

1. A necessary and sufficient condition for the MPLE to exist (i.e., to be finite) is: $\forall \alpha \in \mathbb{R}^d, \exists i, j$ such that

$$(2y_{ij} - 1)\alpha^T \delta(y_{ij}^c) \leq 0$$
 This occurs with positive probability.
2. If the MPLE exists, it is unique.
 When it exists, it is the unique solution to:

$$\sum_{i,j} y_{ij} \pi_{i,j} = \sum_{i,j} y_{ij} \delta(y_{ij}^c)$$
 where $\operatorname{logit}(\pi_{i,j}) = \eta^T \delta(y_{ij}^c)$.
3. If the MPLE exists the MLE exists, but not vice versa.

FIGURE 23

I could belabor this and say a lot about corresponding results like pseudo-likelihood, but reading the paper is probably the easiest thing.

Likelihood inference based on MCMC

– *idea*: If a large number of simulations from a social network in the same ballpark as that of the observed network can be generated then these can be used to approximate the MLE to a desired accuracy. Estimate the “population” value:

$$c(\eta) = \sum_{y \in \mathcal{Y}} \exp\{\eta^T Z(y)\}$$

by the “sample” value:

$$\bar{c}(\eta) = \frac{|\mathcal{Y}|}{M} \sum_{M \text{ sampled graphs } y} \exp\{\eta^T Z(y)\}$$

– Markov Chain Monte Carlo (MCMC) algorithms are a natural way to simulate social networks

FIGURE 24

I’ll make a brief comment on MCMC. For those who haven’t seen a lot of Monte Carlo this is the idea. I find it’s a simple way of thinking about likelihood estimation. The idea here, shown in Figure 24, is if you want to estimate the partition function or normalizing constant which is essentially the population mean of this over this set of all possible graphs. Being statisticians, if we have a very large population to measure the mean of what we would do is draw a sample from that population. We would then calculate the sample mean and use that to replace the population mean, which is one of the simplest ideas in statistics. In essence what we do is draw a sample of the possible graphs, compute the corresponding sample mean, and use this to approximate the partition function. The question is how do we get a sample of graphs? MCMC graphs are a natural way. This has been developed and I won’t say too much more about it. There is a result which corresponds to any MCMC likelihood used in that way actually converges with sufficient iterations which I won’t belabor here.

I always find interesting the relationship between a near degenerate model and MCMC estimation. Figure 25 addresses this idea, which goes all the way back to very nice work by Charlie Geyer. Basically, the idea is—there is some mathematics that goes behind it—if your model is near degenerate then your MCMC won’t mix very well. This might be a surprise to most folks who have ever tried this, but just to give you some sense in practice of how this works I’ll show you again this two-star model in Figure 26.

Effect of Near-Degeneracy on MCMC Estimation

- Closely related to nice properties of simple MCMC schemes (Geyer 1999).
 - If a random graph model is simulated using a MCMC based on a near-degenerate ψ it will very likely fail.
- Full-conditional MCMC with dyad update:

$$M(\psi) = \max_{y \in \mathcal{Y}} |\psi^T \delta(y_{ij}^c)|$$

where $\delta(y_{ij}^c) = Z(y_{ij}^+) - Z(y_{ij}^-)$

- As $\mu(\psi) \rightarrow \text{bd}(\mathcal{C})$, $M(\psi) \rightarrow \infty$
- There exists $y \in \mathcal{Y}$ with

$$\text{logit} [P(Y_{ij} = 1 \mid Y_{ij}^c = y_{ij}^c)] = \pm M(\psi)$$

- If ψ is near-degenerate then $M(\psi)$ is large and the MCMC will mix very slowly.

FIGURE 25

Example of degeneracy of the 2–star model

$$P(Y = y) = \frac{\exp\{\eta_1 E(y) + \eta_2 S(y)\}}{c(\eta_1, \eta_2)} \quad y \in \mathcal{Y}$$

- $M(\eta) = \max\{|\eta_1|, \eta_1 + 2(g-2)\eta_2\}$ MCMC will usually mix poorly.
- If $\mu(\eta)$ close to (3, 0) (e.g., $\eta = (4.5, -18.4)$) then $M(\eta) = 4.5$
 So an MCMC will approach (3, 0) and stay there (98.9% and 1.1% at $(2, 0) \in \text{bd}(\mathcal{C})$).
- If $\mu(\eta)$ close to (9, 40) (e.g., $\eta = (-3.43, 0.683)$) then $M(\eta) = 3.43$. The model places 50% of its mass on graphs with 2 or fewer edges and 36% on graphs with at least 19 edges.
- The model is also *unstable* e.g., $\eta = (-3.43, 0.67)$
 $\mu(\eta) \approx (4.4, 17.1)$ and the model places almost all its mass on empty graphs.

FIGURE 26

For example, suppose for the two-star models you want to choose a mean value with 9 edges and about 40 two stars, and you run your MCMC sampler. Figure 27 shows what you get. Many people have probably run into these if they have ever tried it.

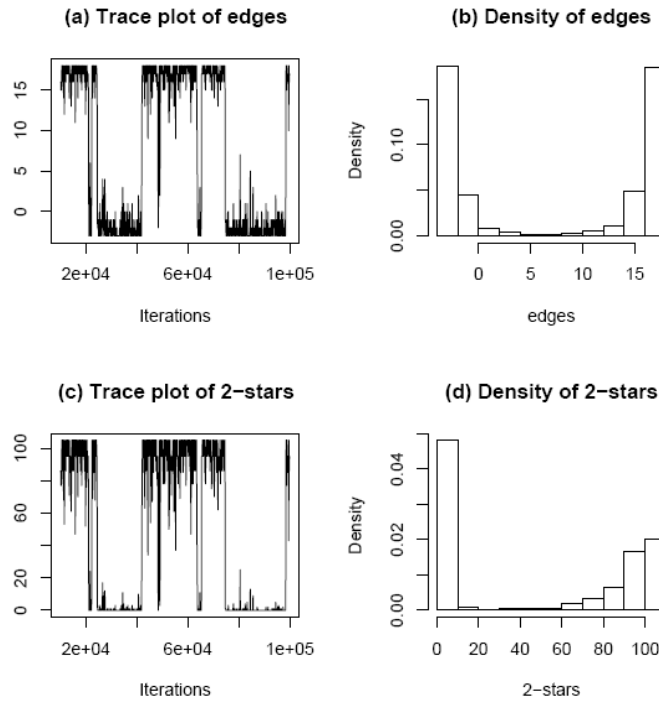


FIGURE 27

This is the trace part of the edges. I'm running the MCMC sampler for about 100,000 iterations of this simple 7 node process, and these are the trace-plotted number of edges as we draw from that mark off chain. It stays up here around 19, and dropping down to 15, 17, and suddenly jumps down to 3 or 2 or 0. It stays down there for maybe 20,000 and jumps back up and jumps back down. So, what we are actually seeing, if you look at the marginal distribution of these draws, is a profoundly polarized distribution with most of the draws from very low values and some of the draws from quite high. And of course, in such a way that the mean value is 9, which is exactly what we designed the model to actually do. This is another view of the example that Martina gave yesterday of the two star model. Note that this sampler is doing great. It is giving us samples back from the model we asked it for, but now that we looked at this we would probably say we don't want this model therefore bad mixing of an MCMC is highly related to these degeneracy properties.

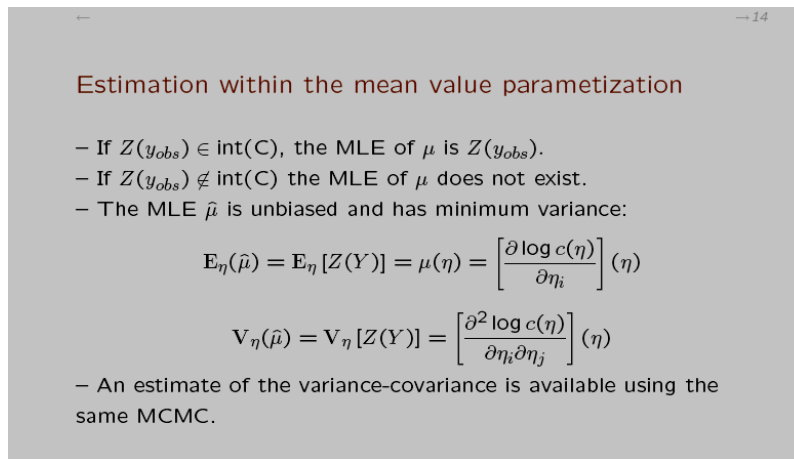


FIGURE 28

To come back very briefly, the estimation of the mean value parameterization is essentially just as hard as it is in natural parameterization, although the corresponding point estimate itself is trivial. It's just the observed statistics. Of course, to find the properties of that point estimator, and it's no good unless you know what its properties are, essentially you need the same MCMC methods as you need for the natural parameterization. It is much easier to do the natural parameterization from a computational perspective in terms of writing code because of slightly simpler forms. You don't need to solve some part of the inverse problem.

To finish, I'll say a little bit about network sampling because it has come up a bunch of times during this workshop. I'll give you some idea of the classical statistical treatment of it going back to Frank in 1972 and other work since. We think of our design mechanism as that part of the observation process that is under the control of the researcher. So, if we are thinking about doing network sampling, the design mechanism is how the researchers design the sampling process. Examples would be surveys using egocentric, snowball, or other link-tracing sampling. There is the out-of-design mechanism also, which is the unintentional non-observation of network information, meaning the mechanisms by which information is missed, such as the failure to report links, incomplete measurement of links, and attrition from longitudinal surveys. Note that for any sampling mechanism we need to deal with both these components of it.

It is sometimes convenient to cluster design mechanisms into conventional, adaptive, and convenience designs. So-called conventional designs do not use the information collected during a survey to direct the subsequent sampling of individuals.

For example, we'll sample everyone and do a network census, or we might do egocentric designs and randomly choose a number of individuals then look at the actual ties of only those individuals. That is, we don't use anything that we collect during a survey to look at subsequent

sampling. This might sound like an odd way to do it, but let me describe adaptive designs for contrast. In adaptive designs, we actually use information to direct the subsequent sampling; most network sampling is actually of this adaptive form. The idea is to collect ties. We then follow the links of the initial group and use the fact that we have observed the links of those individuals to essentially do contact tracing and move out in those forms. An example is classic snowball sampling link tracing around work designs. Many of the designs using computer science and physics fall under this form where you are using information taken during your particular survey to direct the subsequent sampling. There are also convenience designs where you might use something very intelligent but not very statistical. So, we are using convenient information; you just nab people who are close by and that are convenient to sample.

In Handcock (2003) I examined likelihood-based inference for adaptive designs—basically how to fit the models I described earlier, and Martina Morris described, when you have a conventional design and, probably more importantly, when you have adaptive data. You actually have massively sampled data due to link tracing. It turns out that you can fit these models in this way, and there is a computationally feasible MCMC algorithm to actually do it, which I think is actually pretty helpful in practice. This is being implemented in a statnet package for **R**.

As an example of the use of likelihood inference, I'll briefly mention the Colorado Springs "Project 90," which is familiar to many people in this room and which involved quite a broad sample. This study dealt with a population of prostitutes, pimps, and drug dealers in Colorado Springs; I will only focus on 1991. I'm looking at a heterosexual sexual network within the group of people who responded to a survey. Essentially what they did was a form of link tracing, which could be referred to a bi-link tracing, where you are recruited into the survey if two other respondents nominated you. It wasn't enough just to be nominated by one individual, you had to be nominated by two.

sexinternal

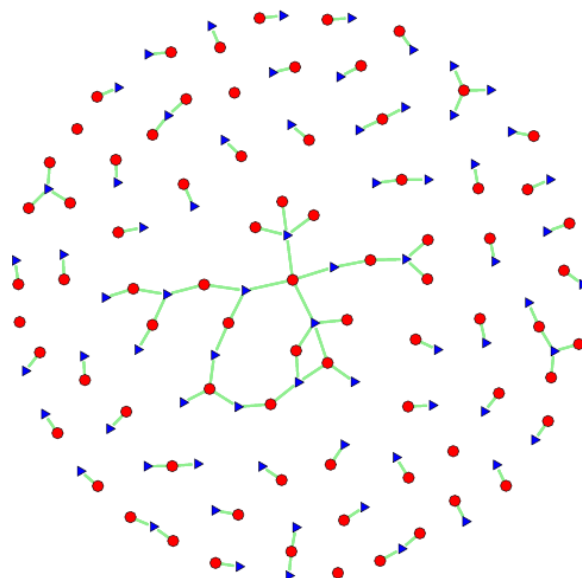


FIGURE 29

Figure 29 presents an example of that network. You see it has a relatively large component in the center. Many of the ties tend to be just monogamy here, just split up. You then have some small components around the side, and this core is around 300 folks. I think there are some isolates also. You get this sense that there is a large part that is connected but not massively, so individuals float around. Note this is a visualization, so there are some artifacts. For instance, it looks more central than it actually is. There is a lot of missing data here due to the sampling processes.

You are only seeing the actual part of it that is observed, and I won't go through the boring detail, but if you run the program here, Figure 30 shows the sort of result you will get. You put in all these covariates. We know a lot of the information about the individuals, their age, their race, whether they are a prostitute or a pimp. What we did was looked at different types of tie formation based on age, race, and occupation. We also looked at whether there was homophily based on race, a homophily based on age, and then these endogenous parameters that measures the dependency involved. I won't say too much about this particular model, but you can get the natural parameter estimates. I also measure the standard errors induced by the Markov of chain Monte Carlo sampling. In terms of goodness of fit, because this is an exponential model, we can compute the deviance here so we have the classic deviance statistics

for this process. Note that we do not have classical asymptotic approximations to their distributions!

Example statnet model fit

```
Monte Carlo MLE Results:
              estimate s.e.      p-value MCMC s.e.
edges          -2.0853 1.2294    0.08990 0.008610
nodefactor.age.22-28 -1.1113 0.3989    0.00535 0.003381
nodefactor.age.29-39 -0.6164 0.3379    0.06818 0.003289
nodefactor.age.>40  -0.4946 0.4057    0.22290 0.003803
nodefactor.black    0.5423 0.2599    0.03699 0.002555
nodefactor.prostitute 0.3230 0.4110    0.43190 0.003799
nodefactor.pimp     0.2919 0.3627    0.42100 0.003405
nodematch.black    1.1913 0.2773    < 1e-04 0.003406
nodematch.age.group 0.1304 0.2519    0.60470 0.003240
shared.partner.1   -1.6677 3.5160    0.63530 0.022006
shared.partner.2   -1.7923 7.0296    0.79880 0.044199
degree.exactly.1   2.0136 0.4126    < 1e-04 0.002698

Null Deviance: 7885.24 on 5688 degrees of freedom
Residual Deviance: 841.55 on 5676 degrees of freedom
Deviance: 7043.69 on 12 degrees of freedom

AIC: 871.55    BIC: 971.25
```

FIGURE 30

As we can see, attribute mixing does explain, although if you look at the p-values here, there is not a massive amount of difference based on attribute mixing. When we look at these dependence terms what we actually see is that there is a fairly large over supply of people with a degree of exactly 1. This parameter is quite large and positive, indicating there are a lot of people in monogamous relationships. There is also a fairly substantial homophily based on race involved in this process as well.

In my two remaining seconds I will say we can use those model parameters to generate processes that look a lot like Colorado Springs, and Figure 31 shows two of them. They don't have the larger component of the graph we saw in Figure 29, because this model doesn't naturally have that form, but if you go through other realizations you can get very similar looking forms.

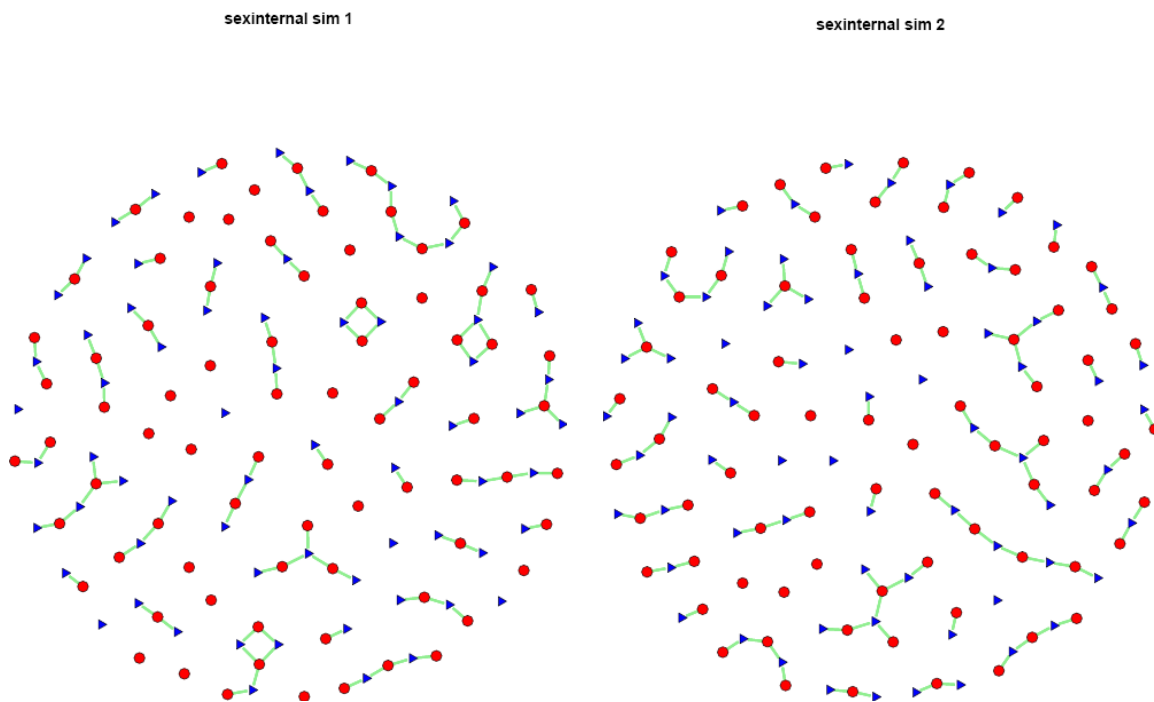


FIGURE 31

In conclusion, I'll reiterate that large and deep literatures exist that are often ignored; simple models are being used to capture clustering and other structural properties, and the inclusion of attributes (e.g., actor attributes, dyad attributes) is very important.

QUESTIONS AND ANSWERS

DR. KOLACZYK: Eric Kolaczyk, Boston University.

Mark, here is a quick question for you. My understanding is the nature of this degeneracy result is essentially relying on Barndorff-Nielsen's statement with the interior of what you were calling C , and whether you are in the boundary or not. That is always filtering in the sort of standard theory that you would see, but it always sort of floats by. Most of us don't end up really needing to worry about it. Or if you are going to end up working in discrete models and binomial type models or something, then you have to be aware of it a bit. But it almost never is as extreme as this. So, that's something I'm missing here. What is the intuition behind this? Is it the mapping involved? Why is it that it's so extreme here, whereas for the most part in most models that we would consider across a distribution of statistics, if you want to say roughly, people don't have to worry about that seemingly innocuous theoretical result?

DR. HANDCOCK: That's a good question. The underlying theory is the same, but here are actually two forms. One form is the computational degeneracy: that is, for a given observed graph, if you try to calculate the MLE, the MLE may or may not exist, and that is the standard Barndorff-Nielsen form. Although I agree that it is rarely seen for very complex models, and this is the first case I've seen beyond my undergraduate statistic classes.

The second is the same theory, but a completely separate idea is used for the model degeneracy, which has no data in sight. It's just that you've got a model. You are looking at it. You're saying what does this model do? You are treating it as a toy so you start playing with it and seeing how it behaves. The same underlying geometry is important, but it's not related to those results about existence of an MLE. The basic idea is that if we change the view from the natural parameter space, where interpretation of model properties is actually quite complex, to the mean value parameter space, which is expressed in terms of the statistics—which we chose to be the foundation of our model, and hence for which we should have good interpretation about—then it gives us a much better lens to see about how the model is working.

The last part of your question is that the nonlinear mapping is exactly the key here. And a stealth bomber is also an example plot that makes that clear. I should also point out, which I didn't earlier, Mark Newman has a nice paper where he looks at the mean-field approximation to exactly this two-ster model, and I think he has been able to produce a similar plot for the same model.

DR. HOFF: Peter Hoff, University of Washington.

Before I begin Mark, I just want to remind you that you were the one who taught me that I should be critical of these types of models. So, when we make univariate a parameter, and we have replications, and we make model-based inference, we have some way of evaluating whether or not a model is good or adequate. Or even if we are interested in the mean of a population, we have replicates. We can rely on a stochastic distribution, which relies only on the first and second moments of the actual population, and nothing else. So in a way we are not concerned with model misspecification, or we have ways of dealing with it, and in these situations, the way you formulate the model, you have one observation, so my question is your work, a lot of it shows that model selection is extremely critical. And the choice of these statistics you put in, that's the choice of the model, is extremely critical. So, I'm just kind of curious if you could comment on maybe the utility or how we can start thinking about getting replications to address this issue, and if there is any way to actually address this issue without replications, because you are putting up a table of P values and confidence intervals there. That is clearly going to rely on what other statistics you choose to put in the model.

DR. HANDCOCK: There are a number of different issues here. First, I think there are very valid points, very important points. Issue number one is single replication: well, that's clearly true. We've got a single graph, and what is essentially pseudo-replication, which we are using to compute the standard errors and other things like that, is in essence induced by the structure of the model. If I write down the given set of statistics that define that model, that implicitly gives us a set of configurations, which we can define the replication over. So, in essence, the standard errors are very dependent upon the specification of the actual model. Let me give a very simple example. If we just have a model with edges in it, and clearly all the edges are independent of each other we have essentially got replication of edges. If we put a model with two stars in it, then if we look at other forms, you look at all the possible configurations which are two stars, then the configurations which aren't of that form then give us a replication over then to look at the properties of the two stars involved. But note that that's very dependent upon the form so that's a partial answer to the question.

I think the other answers are important as well. How do we specify this model if it really matters how the form actually is? A natural answer to that, which is incomplete, is if you had replications of these graphs, the classic, purely independent forms would apply, and then we could work with them, and that would be one solution. The other answer I think is it's worth stepping back and asking ourselves whether to have pure independence is required. I'll just remind us that every spatial process has exactly the same issues. All the spatial statistics deal with spatial processes which are dependent across their full domain. I think a spatial lattice process, continuous field processes all have dependence, and you only have a single realization. So, this is more a property of dependent models rather than social network models on this particular model class.

Coming back to Peter's last point, model specification is extremely difficult here, because you are using the model specification to also find misspecification in that model. And I strongly emphasize the use of the goodness of fit methods that Martina described.

And why I think this is very helpful is that if you have a statistic which is not in a model, not in your specification, and you see how well the model actually can reconstruct that somewhat separate, although obviously typically dependent statistic, and that gives you a way of just seeing how it does represent properties you have not explicitly placed in your model. The other thing is just a standard statistical approach of doing exact testing, based on this model, is of a similar ilk, where you can do exact testing to look for model misspecification forms. I think Peter has raised a good point here. I have been very critical of these models in the past for the reasons I have said, and I'm actually using them. I'll leave it at that.

REFERENCES

- Barndorff-Nielsen, Ole. 1978. Comments on Paper by B. Efron and D.V. Hinkley. *Biometrika* 65(3):483.
- Besag, J.E. 1974. Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society, Series B* 36:192-236.
- Cressie, N. A two-dimensional random walk in the presence of a partially reflecting barrier. *Journal of Applied Probability* 11:199-205.
- Doreian, P., and Frans Stokman. 1997. *Evolution of Social Networks*. Amsterdam, The Netherlands: Overseas Publishers Association.
- Frank, Ove, and D. Strauss. 1986. Markov Graphs. *Journal of the American Statistical Association* 81(395):832-842.
- Holland, P.W., and S. Leinhardt. 1981. An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association* 76(373):33-50.
- Lauritzen, S.L., and D.J. Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B* 50:157-224.
- Leinhardt, Samuel. 1977. *Social Networks: A Developing Paradigm*. Burlington, Maine: Academic Press.
- Morris, Martina. 2004. *Network Epidemiology: A Handbook for Survey Design and Data Collection*. London: Oxford University Press.
- Newman, M.E.J. 2003. The structure and function of complex networks. *SIAM Review* 167-256.
- Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: U.K.: Cambridge University Press.

Visualization and Scalability

Characterizing Brain Networks with Granger Causality

Mingzhou Ding, University of Florida

DR. DING: Thank you for coming to this last session. It takes a lot of will power to stay on. I also belong to the neural club of this workshop, and I'm going to talk a little bit about brain networks and how to characterize that with Granger causality. The idea of Granger causality has been around for many years, and I will talk about our attempt at using this method to assign directions to interactions between neural ensembles. So far in this workshop we have heard a lot of graph theory, but the examples tend to be undirected graphs. So, hopefully, I'm going to present some directed graphs through my combination of this method with the brain.

We've heard a lot of excellent neural talks so far; therefore, the groundwork has already been laid for me. What I'm interested in is to put multiple electrodes in the brain and measure signals from various parts simultaneously. The particular measurements that we look at are local field potentials, as was the case for several previous talks at this meeting. This local field potential is also recorded with an electrode directly in the brain, but it's not looking at one neuron's activity, it is more like the activity of a population of neurons. People estimate that it reflects the summed dendritic activity of maybe 10,000 or so neurons. It's a population variable, and particularly suited for us to look at network formations during performance of some kind of experiment or task.

We've also heard a lot about rhythms in neural systems. If you look at the signal you record as a time series and look at the spectrum, very often you see very prominent peaks in various frequency bands; therefore in this particular study we are going to apply a spectral analysis framework to our experimental data. Our approaches are very simple; all of them are very standard statistical methodologies, so I have nothing new to offer in terms of statistical methodology.

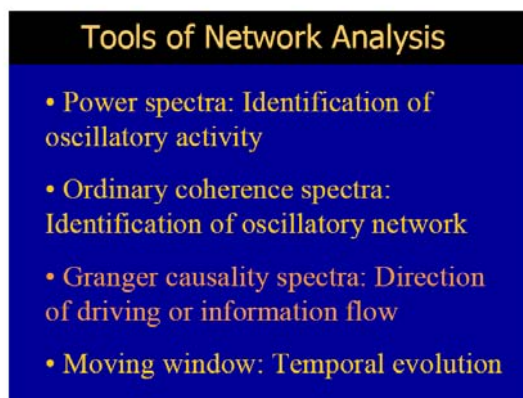


FIGURE 1

The tools of our network analysis are listed in Figure 1. The first technique we apply is simply the power spectrum. We collect data from each electrode and look at its power spectrum to identify whether there are any features like peaks in a band. Next, we try to find out which if any of those electrodes show similar features; for example, whether there are 10-hertz peaks at multiple recording sites. In the second step, we look at their statistical correlation by computing coherence to see whether those 10-hertz peaks measured in different parts of the brain actually belong to the same oscillatory network.

The network I'm talking about here is defined in a statistical sense. That is, data streams that are statistically co-varying during task performance are said to belong to a network. Sometimes this is called a functional network. This is in contrast to another way of talking about brain networks, which is anatomical networks, namely, how different parts of the brain are actually wired together, linked together through physical cables (axon projections). I also want to stress the indispensable relation between the two networks. From a graph theory perspective, our first step will allow us to identify similar features at different nodes, and the coherence will allow us to draw edges linking together nodes to indicate that activities at these different locations are co-varying at certain frequencies. The result is an undirected graph. Sometimes when you look at that graph you still don't really know the relative contributions of different nodes in this network, or in this case different brain areas.

The third step is more or less new in the neuroscience business, although it has been around for many years. Through Granger causality we are trying to introduce directions among different brain areas during a given task performance. Through those direction assessments you can sometimes get a better view of how different areas are organized together to produce meaningful behaviors that we experience every day.

The final tool we use is a moving window approach. This just allows us to move through our experimental recordings to look at various states in the brain because they change rapidly during cognitive performance. We want to look at the temporal evolutions of that.

A very quick introduction to Granger causality is given in Figures 2 and 3. This will be a familiar topic to the statisticians in the audience. Basically, you have two simultaneous time series recordings, X and Y, from which you build regressive predictors. You are predicting the current value of X based on a linear combination of a set of X measurements that has occurred in the past where the linear coefficients are selected in some optimal sense. When the actual value comes in you can look at the difference to see how good your predictor is compared to the real value. The difference is captured in an error term, and the magnitude of this error term is a signature of how good your predictor is. Obviously, the smaller it is the better your prediction.

Granger Causality I

Given: $x_1, x_2, \dots, x_n, \dots$
 $y_1, y_2, \dots, y_n, \dots$

Linear prediction:
 $x_n = a_1 x_{n-1} + \dots + a_m x_{n-m} + \varepsilon_n$
 $x_n = b_1 x_{n-1} + \dots + b_k x_{n-k}$
 $+ c_1 y_{n-1} + \dots + c_k y_{n-k} + \eta_n$

FIGURE 2

Granger Causality II

If

$$\frac{\text{Var}(\eta_n)}{\text{Var}(\varepsilon_n)} < 1$$

in some suitable statistical sense we say y -series has causal influence on x -series and represent it by the symbol $F_{Y \rightarrow X}$.

FIGURE 3

Then you try to see whether using the previous measurements of the X process, in combination with Y's previous measurements, produces a better prediction or not. In that sense you are looking at the error term one more time to see whether introducing the other measurement actually improved this error term. If that is true, then in some statistical sense we say Y has a causal influence on X. You can then reverse the role of the X and the Y to examine whether there is an influence in the other direction and assess the directional influence between these two time series.

The idea of Granger causality came about in the early 1950s in an article written by Wiener. He talked about the prediction of time series. That's where he first proposed this notion, but his article is very thick and full of measure theory; it's unclear how you could implement this in practice. In 1969 Clive Granger implemented this whole thing in a linear regression form as you see in Figure 1, and that led to a lot of applications in economics. Recently we and other people have begun to take this and try to apply it to our study of the brain. In 1982 Geweke found a spectral version of the previously introduced Granger causality, as indicated in Figure 4. Note that what we have seen earlier is a time-domain version. We also like to have spectral representations, because we want to study the rhythms in the brain. In this regard we found the spectral version by Geweke. He was able to have a very revolutionary way of looking at decomposing the interdependence between time series. The important thing he did was to prove a theorem that if you integrate this spectral version over frequency you get a time-domain version back. This relation alone shows that there is something right, so we will call this Granger causality spectrum.

Granger Causality Spectra

Geweke (1982) found a spectral representation of the time domain

Granger Causality:

$$F_{Y \rightarrow X} = \frac{1}{2\pi} \int I_{Y \rightarrow X}(f) df$$

$I_{Y \rightarrow X}(f)$ will be referred to as Granger Causality Spectra.

FIGURE 4

Geweke's basic idea is this. Given two signals, X and Y, you take X signal's power spectrum at a given frequency f and then ask whether it can be decomposed into two parts: an intrinsic part and a causal part. The first of these is due to some kind of intrinsic dynamics, while the other is due to the causal dynamics from the Y signal. After performing this decomposition, Geweke defined the causality as a logarithm of the ratio of total power to the intrinsic part of the power. Therefore, if the causal part is 0 (i.e., all of the power is in the intrinsic part), this ratio is 1 and the log of the ratio—the causality—is 0. Therefore, there is no causal influence. On the other hand, if the causal part is greater than 0, the ratio is greater than 1 and the log of that ratio will be greater than 0. You do a statistical test on whether this quantity is 0 or not.

We began to work with this set of measures about six years ago. We tested various cases trying to convince ourselves this thing is working, and it turns out we are quite convinced. We then began to apply it to experimental data. Today I'm just going to mention one experiment related to the beta oscillation in the sensorimotor cortex of behaving monkeys. Some of the results below have been published in *Proceedings of the National Academy of Sciences* in 2004.

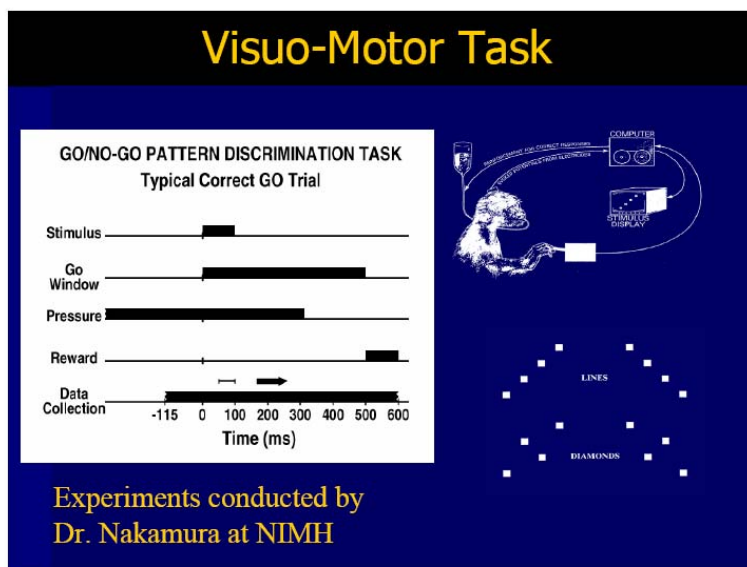


FIGURE 5

Figure 5 introduces you to the experiment itself. It's a very simple standard pattern discrimination go/no-go experiment. The monkey comes in, sits down, and when he feels ready to do the experiment he extends his hand and depresses a mechanical lever, depressing it and holding it steady. The depression of the mechanical lever triggers the electronics, and then the experiment commences. After a little while a stimulus will show up on the screen. He looks at that screen. The stimulus is a group of 4 dots in one of 4 configurations: right slanting line, left slanting line, right slanting diamond, or left slanting diamond, as shown in the bottom right of Figure 5. The monkey's job is to tell whether the dots are in a diamond configuration or in a line, without regard to orientation. If he is able to find out which pattern he is seeing he will make a response by indicating that he understood what is going on. The response is by releasing the lever if he sees one pattern, by holding onto the lever if he sees the other pattern. That's the way we know that he really understands the situation. In this case the response itself also has two types. One is a go response in which he lifts his hand from the lever, and the other one is a no-go response in which he does not move. This becomes very important for us later on when we try to understand our analysis.

The experiment was done with four monkeys, but today I'm going to report on the analysis of two. Figure 6 shows the location of the electrodes, with each dot representing one electrode insertion. You can see that the electrodes were distributed over many different parts of one hemisphere. That hemisphere is chosen based on the handedness of this monkey; if the

monkey is a right hander you record from the left brain because the hand is contralaterally controlled. If it's a left hander you do the right brain.

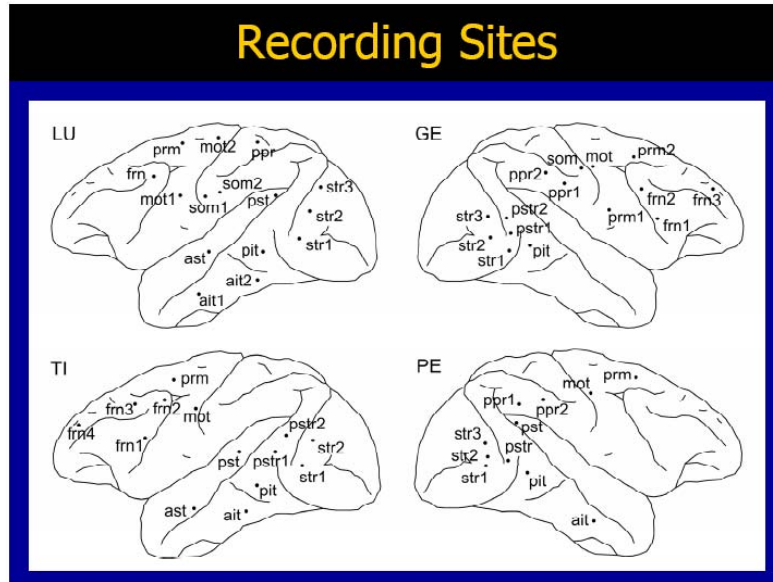


FIGURE 6

Bipolar recordings were used, meaning that two electrode tips are inserted into each location, one tip at the surface and the other at some depth. The difference between these two tips becomes the signal we amplify and then later analyze. It is important for us to take the difference so that the far field effect could be subtracted out. Likewise, the volume conduction and so on could also be subtracted out. Therefore, we are reasonably confident that the signals we are getting from these electrodes are indeed reflecting local activities. If you want to look at networks you don't want to have a huge common influence (e.g., far-field effect) on the analysis because that will make your analysis futile.

I want to show you some of the typical time series we get. Before that I want to comment on the mathematical framework we use to model these time series. When we study stochastic processes in the classroom we know that everything is defined first and foremost in terms of ensembles, but in the real world, for instance in economics, you get only one realization. Here you can have many realizations if you continue to do the same experiment over and over again. This is precisely the case in neurobiology. Figure 7 shows the overlaying of time series from many trials—that's over 400 trials superimposed on the same plot. In the traditional studies

people take all these trials and perform an ensemble average. This is the white curve you see; it's called the even-related potential, ERP. Traditional studies only study this average curve, and that's all. All these jiggly lines in Figure 7 are averaged away, branded as noise and then just forgotten. What we are going to do is bring them back and treat them as our signal.

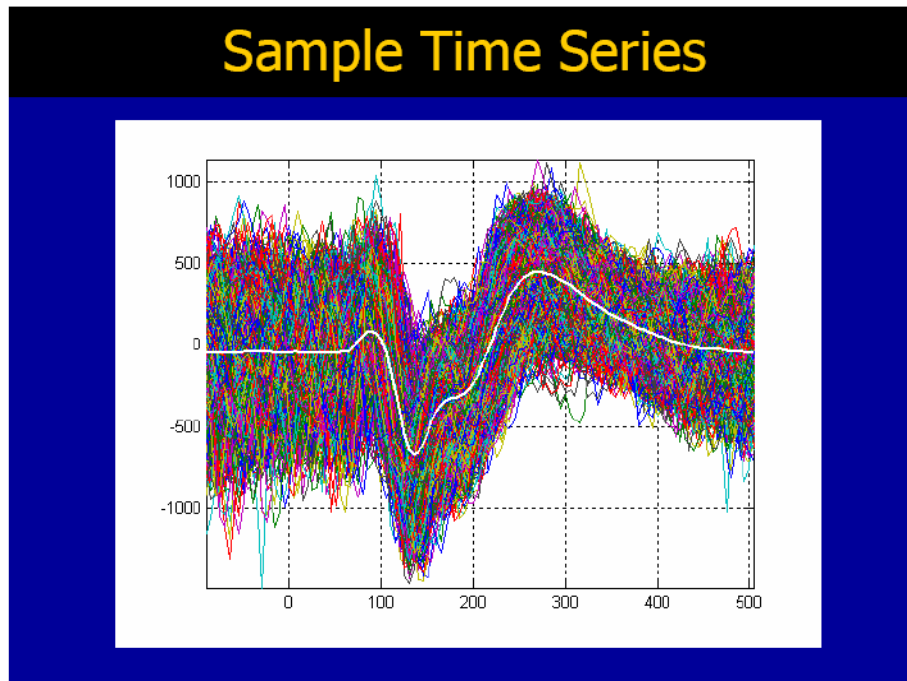


FIGURE 7

Time 0 in Figure 7 indicates the time when stimulus appears on the screen. The time to the left of that is when the monkey has already depressed the lever and is looking at the screen, getting ready to perform the experiment. Our first analysis was over this early period of the data, from -100ms to 20ms. During this period, the monkey is already prepared to do the job but the stimulus has not yet appeared, because the signal takes some time to transmit to the brain to trigger a response. This time period is sometimes referred to as the ongoing period, prestimulus, or fore period. We treat the data as coming from a stochastic process and each trial as one realization. Then we begin to fit the parametric models and do the signal analysis.

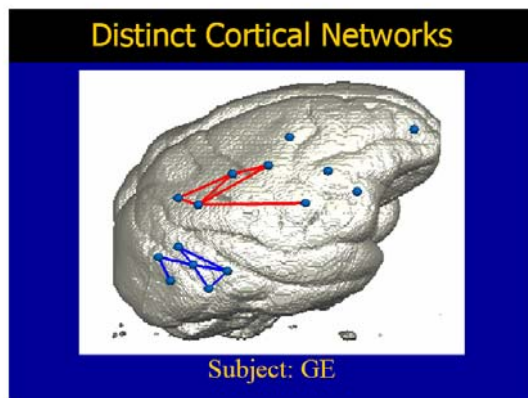


FIGURE 8

Among all the electrodes we simultaneously recorded, we first did a blind network analysis to see which time series correlates with which other and in what frequency range. Interestingly, two segregated networks popped out, as shown in red and blue in Figure 8. One is located in the sensorimotor area, and the other is in the visual system. Today I'm going to spend time analyzing the former, which is the most interesting one and the stronger of the two.

This network in the sensorimotor area is synchronized at about 20 hertz. That is, the nodes tend to communicate at about 20 hertz, and therefore it's a beta range oscillation. Among neuroscientists there is an approximate taxonomy to describe the different frequencies: a frequency in the range 1-3 hertz is called a delta wave, 3-7 is a theta wave, 7-14 is an alpha, 14-25 hertz is a beta wave, and higher frequencies are called gamma waves. We have heard other talks at this workshop dealing with gamma waves. But we are looking at a subclass of beta oscillations in the present study. Figure 9 shows the averaged power spectrum from all these locations; at each location we compute a power spectrum and then average them. The first plot in the top row shows the averaged power spectrum for the two different monkey subjects. Very clearly, they both show energy concentration in the 20 hertz range. The next plot in that row shows coherence results averaged together from the two different monkey subjects, and again we can see clear synchronized activities in the beta range, meaning that the signals are co-varying also in this 20 hertz range, that is, local oscillations also appear to be the oscillations that link all the sites together. They bind them together into one network.

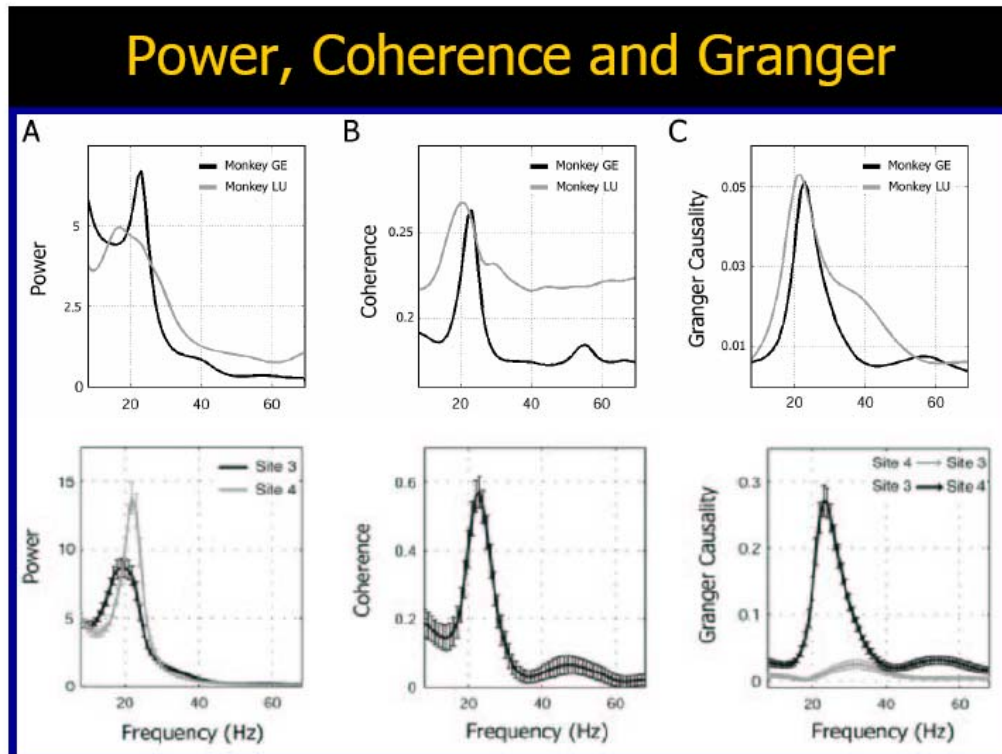


FIGURE 9

The top-right plot in Figure 9 shows our Granger causality result, which also shows quite clearly the concentration of Granger causal influences in the 20 hertz range. It appeared to be the same oscillation underlying the three phenomena. By the way, all the pair-wise Granger influences 1-2, 2-1, 1-3, 2-3, 3-1 and so on are averaged together for each subject. The bottom row shows just one pair from one subject. We are looking at the power, coherence and Granger. The interesting thing is that power spectra showed very clear beta oscillation. Coherence was very strong, 0.6, in the beta range, indicating that the two sites are strongly interacting but no directional information is available from coherence. But if you look at the Granger causality, it becomes immediately apparent that the interaction is unidirectional: one direction is very strong and the other direction is nearly flat.

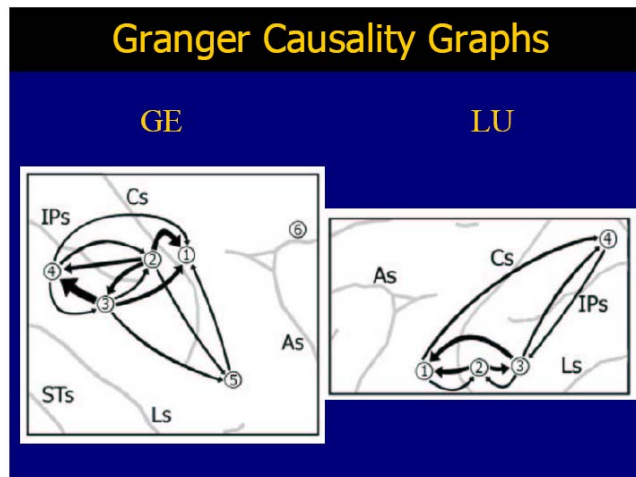


FIGURE 10

What we did next was to draw all the pair-wise Granger plots as shown in Figure 10 on a graph which we call the Granger causality graph which is a directed graph. A very interesting thing happens if you look at the primary somatosensory area. It tends to send out influences while receiving tiny ones. In contrast, the motor cortex receives all the input, and it is sending out little. In addition, I want to say that we did a significance test. A lot of this passed the significance test, but the strong ones are an order of magnitude bigger than the little ones.

To aid our understanding of the results, all the big and strong causal influences that are common to the two subjects are summarized in Figure 11. At the beginning when we began to do this study we had no idea what this network was actually doing. After we did all this directional analysis a very clear hypothesis jumped out at us: we believe this network is actually formed to support the monkey depressing the mechanical level and holding it steady, because holding something steady is in itself a function. The brain needs to be involved to support that function. The reasons for this hypothesis are several-fold. First of all, the primary somatosensory areas S1 are sitting center stage in this network, and are sending influences out to the other parts of the network. Think about this: holding something steady is like feedback control. You are stabilizing an equilibrium state, and when you are stabilizing an equilibrium state the key ingredient is the sensory feedback. That is, I need to know how I'm performing right now so that proper control signals can be exported to maintain that state.

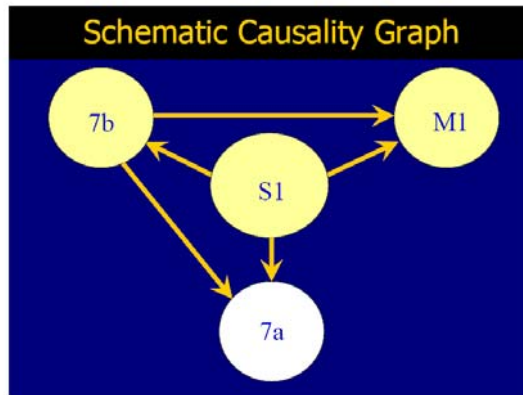


FIGURE 11

I want to quote a patient study that actually supports our work. In 1982 there was a group in London that studied a patient who suffered a neuropathy and lost all sensory feedback. But his motor system was totally intact. That is, the system that allows commands from the brain to move muscles was completely intact, but he could not feel anything. So if you ask this person to pick up a suitcase he would have no problem doing that. But if you ask him to hold the suitcase steady for an extended period of time, he cannot do it unless he looks at this hand. (If he looks at his hand, he is using visual information to supplement the lack of sensory feedback.) Therefore, this tells us that sensory feedback is critical in maintaining the steady state.

Area 7B is a very interesting area. We later learned that it is a very important area for non-visually guided motion. There is a network involving 7B and the cerebellum, which together form some kind of network that helps the brain maintain goals. When we are making a movement, the brain first has to have a goal about what to do. So we form a goal and tell the motor system to execute and fulfill that goal. In control terms, the model is here and an idea of what is to be achieved is here. This area receives sensory updates from S1, compares that to the internal model, and exports the error signals to the motor cortex for output adjustment.

The third line of evidence is coming from many other studies people have done in either humans or monkeys linking beta range neural oscillations and isometric contraction—maintaining a steady state pressure on some kind of a gauge. People who have studied the motor cortex have seen beta oscillation in this area very clearly. What we have done is extend the network into post-central areas. Through the directional analysis we are able to say that the post-central areas play an important role in organizing this network. Without the arrows we cannot really say who is playing the organizational role in this whole system, but adding the arrows we are able to

identify S1 as playing a more central role in organizing this network to support this behavior. At this point one can make a prediction. Namely, when the monkey makes a movement by releasing this pressure, this network should dissolve itself and go away. How do we test this hypothesis? Recall that in this experiment there are two response types, a go response—the monkey sees some stimuli he recognizes, and he lifts the lever—and there is a no-go response that consists of holding steady until the very end. Therefore, if this oscillation network is in support of holding something steady, then we should see this network continuing in the no-go conditions all the way to the end, while for the go conditions, because the monkey released halfway through it, the network should disappear halfway through the experiment.

This motivated the next study, which is the moving-window analysis, where you do a time-frequency analysis with an 80-millisecond window and compare go versus no-go conditions. Here the result is very clear. Figure 12 shows a time-frequency plot. This is time and this is the frequency. The magnitude is represented by color. As you can see, for the go trials, there is 20 hertz oscillation at the beginning. The oscillation then disappeared half way through the trial. This arrow is indicating the mean reaction time. That is the mean time at which the monkey released the hand, so prior to lever release, the oscillation just terminated. For the no-go trials this oscillation is maintained to the very end. This is evidence in support of our hypothesis that the network supports the pressure maintenance on the lever.

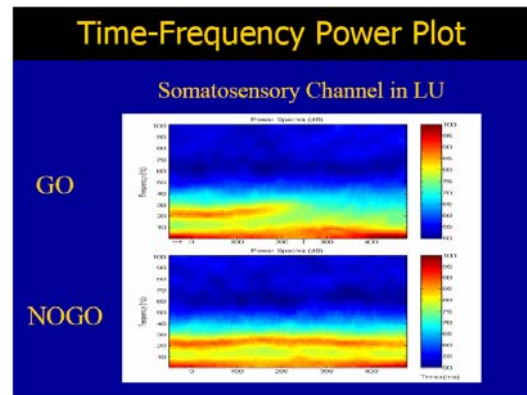


FIGURE 12

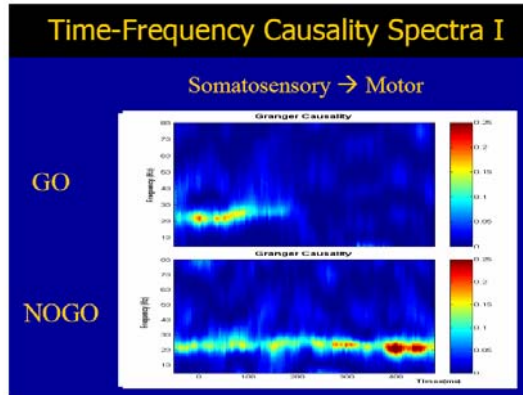


FIGURE 13

When we look at the time frequency plot of Granger causality it's clear that there is directional driving from S1 to M1 at the beginning (Figure 13). When the network disappeared the causal influences also stopped. In the no-go trials the causal influences are maintained all the way to the end. It even becomes stronger in the end. On the other hand, if you look at the causal influences in the opposite direction in this 20 hertz range in Figure 14 there is nothing going on. When we see some of this high frequency stuff we really don't know what this means. It's not very well organized but the 20 hertz definitely is quite unidirectional if we compare Figures 13 and 14. Our basic conclusion here is that the time course of these beta oscillations seems to be consistent with the lever pressure maintenance hypothesis.

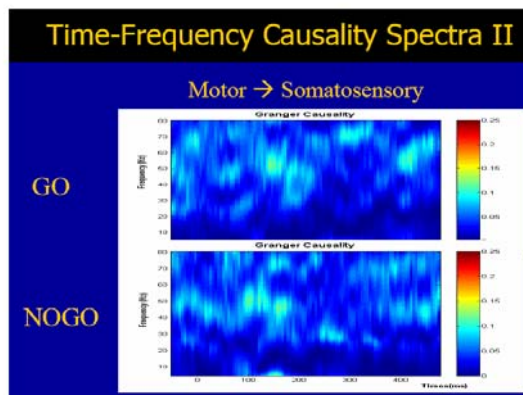


FIGURE 14

At this point I want to make a comment. That is, when people talk about networks in the brain, there are usually two ways to talk about it. One is functional network like what I'm discussing here. This is looking at the statistical relationships between different brain areas during a given task. On the other hand there are also anatomical networks in the brain. One can say that the anatomical network is there and not changing during this experiment. This experiment is only half a second long. The anatomical network didn't change a whole lot yet the dynamics that is traveling on the anatomical network disappears when the behavior it supports is no longer there.

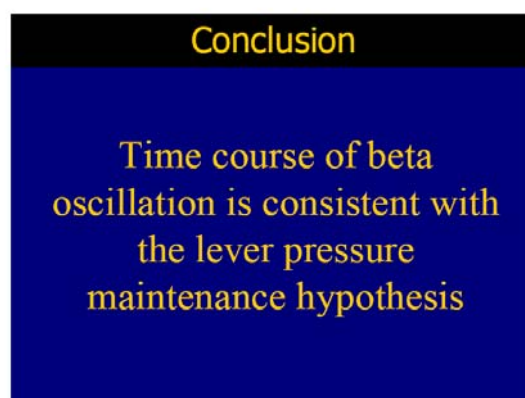


FIGURE 15

Therefore, the dynamics are completely flexible. It's like a highway system, the road is always there but the traffic pattern can be very different depending on the functions it serves. It's the same thing here. The network is in place but the dynamics traveling on that can be changing rapidly depending on the function it supports. This is an interesting thing.

Analysis Three

Conditional Granger Causality

FIGURE 16

The next thing is my last result concerning conditional Granger causality. Here I want to show how the anatomy can inform our functional network analysis. Specifically, this is to show that neuroanatomy can motivate us to look at a problem that we would otherwise not have thought of doing.

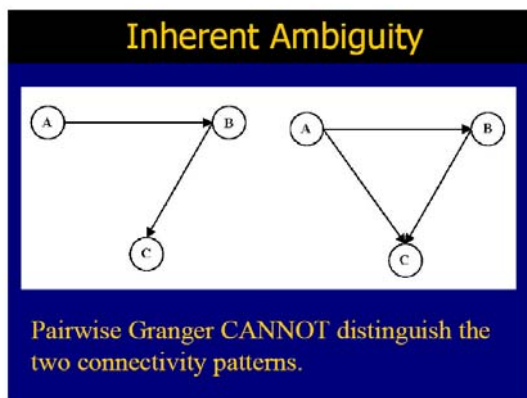


FIGURE 17

Mathematically, the problem is the following. Consider three time series. Suppose they interact as shown in Figure 17. A pair-wise analysis, as we have been doing so far, cannot tell these two patterns apart, because in both cases A drives C from a pairwise perspective. This issue has relevance for the analysis result reported in Figure 11.

Consider a subset of nodes in Figure 11 which is re-plotted in Figure 18. The arrows come from our pair-wise analysis. The question is, is the link from S1 to 7a real or is it the result of pairwise analysis? We care about this question because anatomically there is a reason for us to lean toward the latter. Specifically, Figure 19 shows the sensorimotor network that is published by 1991 by Felleman and Van Essen on the macaque monkey's brain.

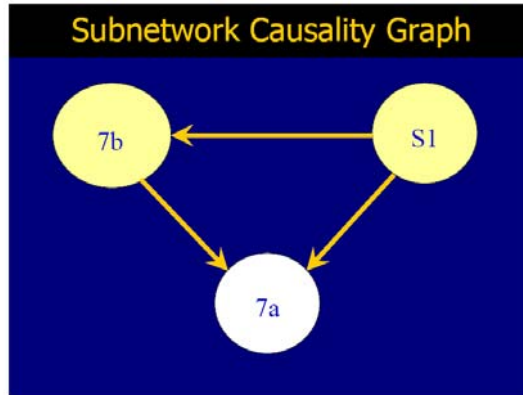


FIGURE 18

As you can see in this diagram 7A is not connected directly to S1. 7A is connected only to 7B in this network. This suggests the hypothesis that the causal influence from S1 to 7A is mediated by 7B.

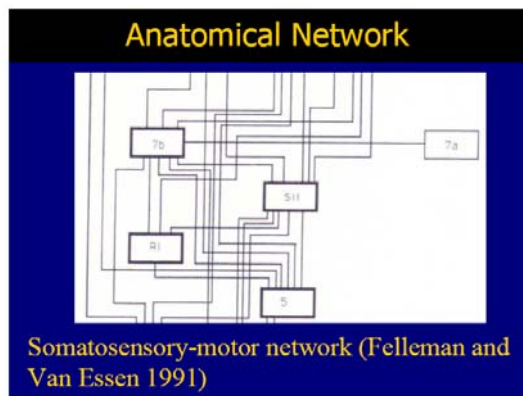


FIGURE 19

To test this hypothesis we need additional analysis ideas. This is accomplished by conditional Granger causality analysis. Like pairwise Granger it's also prediction-based. Refer to Figure 17. If, by incorporating A, we do not get any additional improvement in prediction, then we say we have the connectivity pattern like that on the left. On the other hand, if by including A we can still improve the predictor, we say that we have the pattern on the right. Therefore, these two cases will be distinguished based on these two different types of computation. Spectral versions of conditional Granger causality have been developed recently. We use the spectral version to test our hypothesis in Figure 20. The result is shown in Figure 21.

Hypothesis

$S1 \rightarrow 7a$ is mediated
by 7b

FIGURE 20

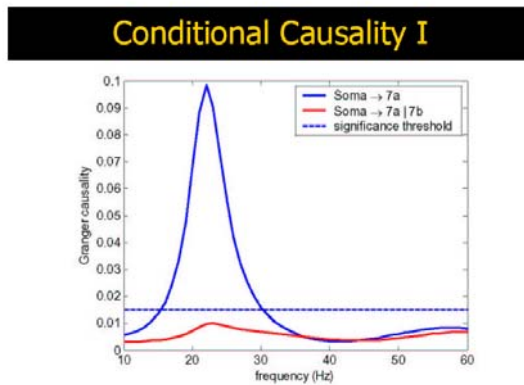


FIGURE 21

The blue curve shows the pairwise result from S1 to 7B which is very clearly above threshold. If you condition 7B out, then all of a sudden, the causal influence from S1 to 7B drops below threshold. In other words S1 to 7B appeared to be mediated by 7A.

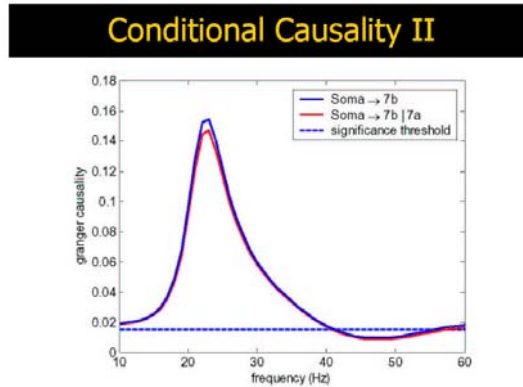


FIGURE 22

We then did a control study to rule out the possibility that the result in Figure 21 is a procedural artifact. We took the same three channels and conditioned 7A out to see the effect. As you can see in Figure 22 there is hardly any difference between the two curves. This means that what we see in Figure 21 is not simply a procedure related artifact. It, in fact, reflects anatomy-constrained neural communications between two areas.



FIGURE 23

Thank you very much.

[Applause.]

QUESTIONS AND ANSWERS

DR. WIGGINS: Some of what you are doing similar to things that have happened in genetic networks where people are trying to infer genetic networks from time series. So, just by points of comparison, one is I'm wondering if Granger tests, which I wasn't really familiar with, has been validated on synthetic data? Like people generate synthetic systems.

DR. DING: We have performed extensive test on synthetic data with excellent results. Regarding genetic networks, I have heard people talk about microarray data. We have never done any work in this area.

DR. WIGGINS: You might be interested to look at papers from the genetic network inference literature where they look at mutual information between different nodes, because in that case you can, with some confidence, eliminate spurious links, like in the triangle you had if you know for example that the mutual information between A and B is strictly less than either of the other two. Then you can argue that that's not a direct causation.

DR. DING: Certainly. It would be interesting to look at that. On the other hand, we are spectral people. We don't want just pure time domain analysis. There are many reasons with respect to the nervous system for the spectral bias.

DR. KLEINFELD: Just to make sure I understood the data was recorded with a 14 channel probe?

DR. DING: Each channel analyzed here represents a bipolar derivation. There are 14 or 15 of those.

REFERENCES

- Brovelli, A., M. Ding, A. Ledberg, Y. Chen, R. Nakamura, S.L. Bressler. 2003. "Beta Oscillations in a Large-Scale Sensorimotor Cortical Network: Directional Influences Revealed by Granger Causality." *Proc. Natl. Acad. Sci. USA* 101:9849-9854.
- Felleman, D.J., and D.C. Van Essen. 1991. "Distributed Hierarchical Processing in the Primate Cerebral Cortex." *Cerebral Cortex*, 1:1-47.
- Geweke, J. 1982. "Measurement of Linear-Dependence and Feedback between Multiple Time-Series." *J. Amer. Statist. Assoc.* 77:304-313.
- Geweke, J. 1984. "Measures of Conditional Linear-Dependence and Feedback Between Time-Series." *J. Amer. Statist. Assoc.* 79:907-915.

Visualization and Variation: Tracking Complex Networks Across Time and Space

Jon Kleinberg, Cornell University

DR. KLEINBERG: Thanks very much to the organizing committee for inviting me to come speak. This has been a very interesting workshop for me, especially seeing how the techniques across all these different areas make use of common principles. Obviously there is a lot we are going to have to think about afterwards.

I'm a computer scientist, so I'm coming at this from a computer science perspective. In particular, the kinds of networks that I'm interested in are the large information networks that are built up on top of the applications on the Internet, starting with things like the World Wide Web and spilling over into networks that affect the way we communicate through things like email and instant messaging, through the way we express interest in hobbies through online communities like Live Journal and so forth, to the way we do electronic commerce and so forth, as indicated in Figure 1.

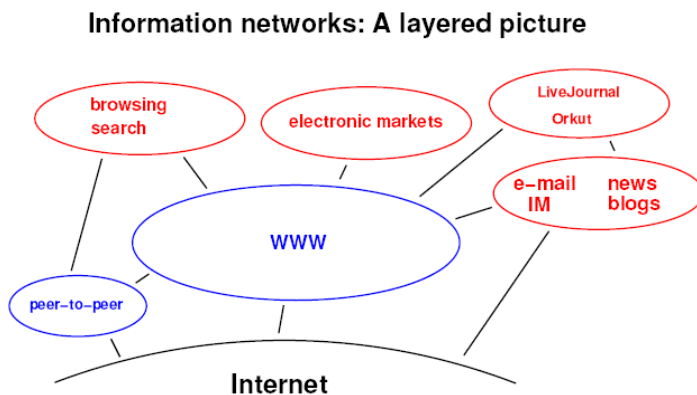


FIGURE 1

There is this whole, what some people call “speciation,” event on the Internet where we see all these new forms of media springing up very suddenly, looking at the emergence of not just the technological structure here, but also the social structure. This notion that there are interesting social structures going on in these online applications is something that has interested

me for quite a while, starting back with some work I did on the World Wide Web where the question was how to use the linked structure of the Web, and the fact that these are documents authored by people, to help improve a search. This is something that combines the social and technology aspect very clearly, although I haven't always been so good at conveying this.

I have been at workshops of this flavor for some time. A few years ago I gave a talk in a very early workshop that brought together some computer scientists and sociologists. We talked about links on the Web and kept casually referring to the Web as a social network without really trying to unpack what I was really saying. This was at the Santa Fe Institute, and during the break some of us were standing out on the patio looking out at this beautiful view of Santa Fe when a very eminent sociologist came up next to me smoking a cigar and asked, "Did you see that talk by the guy from Cornell? That was terrible. He thought he was talking about social networks, and all he was talking about was how Google worked." After explaining that I was that guy from Cornell, we actually had a reasonable conversation under the circumstances. That was what I view as one of the early steps in learning about some of the cultural gaps that exist between these different areas.

I think more than anything that what I have done, or social scientists have done, or what these workshops like this have accomplished, has simply been the unbelievable perfusion of these online media, which has just made it abundantly clear that all of these technological things we are seeing are in the end about content authored by people and people communicating with each other.

In reality it's not so much that they both exist but that they are really intertwined. You can't disentangle the technological networks from the social networks anymore. If I were to ask what are the networks and which email communication, instant messaging, or blogging or any of these other things take place, sure, your packets are traveling over the Internet, but if you are trying to configure a new instant messaging system you have to know a lot about the kinds of people that are going to be using it. You have to know the diurnal rhythms of them and whether it's going to be used by teenagers or by grown-ups and so forth. All this affects the way in which we configure the technology, so in a real sense these things are inextricable at this point.

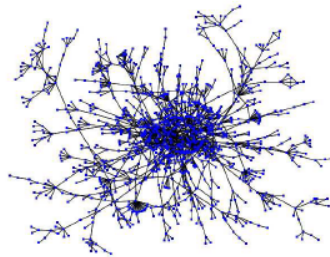
A particular thing I wanted to focus on here is in the context of these social and technological networks, the way in which they are embedded in space, embedded in time, and embedded in surrounding organizational structures. One thing that I wanted to get at is that we shouldn't just view the networks in isolation, but to think about them as being physically embedded as concretely as we view networks in neuron-anatomy as being physically embedded.

I'm going to be talking about network models, as have most of the other talks at this

workshop, reflecting our interest in the structural properties of networks, but also particularly for me, in the processes unfolding on networks, the algorithmic processes that unfold on networks, and trying to find models that relate these to each other. In the end, hopefully, even very simple toy models—and in this talk I'll try to go from a progression of toy models, up through more and more general models—they don't just help us form a vocabulary of patterns, although that is certainly a very useful thing, but also in the computer science context, it's often very surprising to me how quickly those turn into design principles. In the online world the ability to create new online media like peer-to-peer file sharing gives you a certain kind of flexibility such that things that had seemed like toy artificial models can very rapidly become principles by which you actually architect a system. That is one of the interesting things about this domain also.

I also wanted to mention one thing that has been lurking around the surface of some of the talks, and that is the social science component. It is the notion of working with model systems. I think, especially for the statistics aspect of this workshop, this is going to be something important to think about. So, if I could caricature the traditional research obstacle in studying complex networks, it would be that you'd like your networks to be large-scale, realistic, and completely mapped, but in actuality you can only achieve any two of those. For obvious reasons it is very hard to achieve all three at once.

Model Systems



Now: large on-line "model systems" can be tracked and studied.

- Web information and Blogspace (text, links, time evolution)
- Social-networking services: LiveJournal, Orkut.
- E-mail and instant messaging networks.
- Amazon, Internet Archive (reviews, discussions, downloads).
- arXiv (KDD Cup 2003 dataset [Gehrke-Ginsparg-Kleinberg]).

FIGURE 2

So, just as Mark Handcock stole a slide from Martina Morris, I'll steal a slide from Mark.

Figure 2 shows his Zachary Karate Club Network, and the only thing I want to point out here is that this is clearly a real network of real people, and it's completely mapped. We know a lot about it, but the nature of how data like that are collected inherently limits the size. In his after-dinner talk last night, Steve Borgatti mentioned that a trend over the last few years has been to see networks go from sizes in the hundreds to sizes in the ten thousands or hundred thousands. In the process of doing that, we have to ask what are we losing? We are obviously gaining a lot. Many of these systems are what I would call online model systems. When I say model systems, I don't mean that there is anything fake or synthetic about them. All of the people involved in the activities listed on Figure 2 are real people engaged in social interaction. It's simply that we don't know as much about them as we might have known about these smaller more self-contained systems such as the Karate Club Network. It's hard to draw their boundaries; I'm thinking about things like the links among Web pages or Web blogs in the communities of bloggers and how they interact—social networking services like Live Journal or Orchid or any of the other things, which are actually trying to create a social network on their members.

Email and instant messaging within organizations define network interaction. For online e-commerce sites, there has been some amount of interesting research on simply the behavior of people at these sites, because of course there is still the interaction of product reviews, forum discussions, replies, posts, and so forth, and then the download behavior of these people. We can then pick—this is a line of work that Mark really stimulated thinking—something like arXiv, which is a model system of, in this case, high-energy physicists, all publishing papers, co-authoring, citing each other, and so forth. It's a very rich domain in which we can try studying things on a very large scale.

It was this, for example, that caused three of us at Cornell—Paul Ginsparg, the author of the arXiv, Johannes Gehrke, and myself—to choose this as a data set for the annual KDD Cup Competition, basically saying here is a very rich system in which you can do data analysis in all sorts of dimensions, looking at real social data. There are also things that we would like to know about the people in this network that we don't, but for looking at very-large-scale data, so far this is the best we can do.

Figure 3 gives an overview of my talk. The hope was to show how all these themes can be traced out through a single particular topic, small-world networks and decentralized search. I could have easily picked several other topics.

Overview

Study some of these themes in a specific context:
small-world networks and decentralized search.

- **Initial model [Watts-Strogatz 1998]**
- **Searchability: an algorithmic issue [Kleinberg 2000]**
- **Abstracting a general pattern
[Kleinberg 2002, Watts-Dodds-Newman 2002]**
- **Identifying the pattern in large-scale network data**
 - ▷ **Web hyperlinks [Menczer 2002]**
 - ▷ **E-mail communication in an organization [Adamic-Adar 2003]**
 - ▷ **LiveJournal friendships [Liben-Nowell et al. 2005]**
- **The models as design principles: decentralized p2p systems.**
- **Time evolution [Leskovec-Kleinberg-Faloutsos 2005]**

FIGURE 3

We'll start with an initial model, the small-world model of Watts and Strogatz, which many of you know, and then we'll move from that, which is really a structural model, looking at the structure of the network, to a model about processes on top of the network. In particular, the problem of searchability of networks is really more about the process that operates on the network. We'll start with a very simple model of that. I'll try to pull back to create more general models. In particular, by abstracting a general pattern that's going on here, we'll start looking at both data and design principles. Then we'll try to find that pattern in large-scale network data. We'll see some rather surprising ways in which it has shown up in hyperlinks on the Web, email communication, and friendships on these online communities, and also how it's been used as a design principle in things like peer-to-peer systems. And finally, some recent work I've been doing with Faloutsos and Yuri Leskovec looking at how well these networks evolve over time, and some things that actually completely challenge the conventional wisdom we had had on this. And necessarily, since my time is short, I will go through all of these rather briefly.

Small-World Networks

Stanley Milgram's small-world experiment (1967):

Choose a target in Boston, starters in Nebraska.

A letter begins at each starter, must be passed between personal acquaintances until target is reached.

Six steps on average → six degrees of separation.

Watts-Strogatz model (1998):

**Take structured grid (e.g. 2-dim),
add random links**

(e.g. 1 per node).

**Diameter drops quickly while
network remains clustered.**

(e.g. [Bollobás-Chung 1988]).

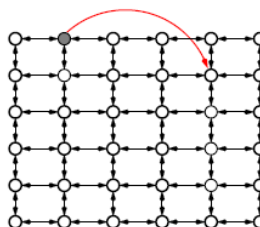


FIGURE 4

To introduce small-world networks, rather than reach back to the Watts-Strogatz model, I wanted to reach much further back to the original work in social psychology that stimulated a lot of this, Stanley Milgram's small world experiments. Both models are summarized in Figure 4. To briefly recap Milgram's experiment for those of you who haven't seen it before, the idea was to trace out short paths in a social network, and particularly the network of acquaintanceships in the United States. He picked a target person who his subject was supposed to reach, who lived in a suburb of Boston, and he picked a whole bunch of starting people, mainly drawing them from out in the Midwest. Each starter person would start with a letter that he had to forward to this target, and rather than mail it directly to them, they had to use what they were told about the target and mail it to one of their friends with the goal of reaching the target in as few steps as possible. Those were the rules of the experiment. You get a letter, you're trying to reach a stockbroker in Sharon, Massachusetts, and you think about the things that you might know about them and you know about your friends. You figure out who you should pass it to, and the famous outcome of this was that the median number of steps in the completed chains was six, a number that has entered into the phrase six degrees of separation.

There are a few things going on with this experiment, and I don't have time to unravel them all, but one is clearly the structural issue. Why is it that our social networks in which our friends all seem to know each other, which as a number of the talks have said before looks fairly clustered, why should it have these short chains? This highly influential *Nature* paper of Watts and Strogatz in 1998 can really be viewed as a proposal for a kind of network that we ought to be

studying. More than any particular result, it really said let's think about networks as superpositions; the notion of a random network embedded in some reference frame, in this case maybe a spatial reference frame. To take one of the flavors of their model, we can imagine that in a very simple version we have a grid with nodes, and these might represent the people in the social network, knowing something about their geographic coordinates. We then add random links, even as few as one per node; an example is shown in red in Figure 4. These random links are enough to make the diameter of the world very small, even while locally it looks very clustered. So, adding a very small number of random links, the network still looks relatively grid-like, even though we now know that its diameter has become very small. We can actually prove things like that using results that have been known for some time in the area of random graphs, in particular the results of Bollobás and Chung (1988), which studied almost exactly this network. So this was the idea, think of these networks as a superposition or, for our purposes, a random network embedded in some reference frame, be it spatial, organizational, or what have you.

When I first came across the Watts-Strogatz work—which wasn't hard, because they were also at Cornell at the time—the thing that really struck me about the original experiment was not just the structure. It was not just about the fact that there were these short chains, but it was about a process. As we teach our undergraduates in computer science classes, if you want to find the shortest path in a network you shouldn't have done this experiment. You should have given a letter to someone and said, send this to all your friends. Tell them to send it to all their friends, and eventually you will see what the shortest chain to the target is. You can think about the obvious reasons why Milgram wasn't able to do that experiment, therefore, he was forced to embark on this much more interesting one where you actually have to tunnel your way, you have make your best guess over the next hop. Essentially, what he was doing was decentralized routing.

Decentralized routing is a completely non-technological context on the social network of people. That's the first thing I would like to try modeling here. So, we would like to model a decentralized algorithm looking for a path through a network. Figure 5 describes a very basic model: imagine there is a current message holder who knows the message, the grid structure, their coordinates on the grid, and the coordinates of the target on the grid. They don't know all the random links, but they know their own random links. So, this simply says when you get the letter, you know that you are living wherever you live. You know the target is in Sharon, Massachusetts, and you know you have friends in New Jersey and Montana and Singapore and so forth, and you want to try to figure out how best to direct the letter. That's the idea. That's why

it is decentralized.

Decentralized Algorithms

Current message holder knows grid structure, destination, path so far.

Long-range contacts of node v only known if v has touched message.

**Gold standard: network has $n \times n$ nodes, but we want
a decentralized algorithm with exponentially better delivery time:
polynomial function of $\log n$, not n .**

Local search very successful at finding short paths.

Watts-Strogatz framework ideal as simple model for asking why:

- Need a network that is partially known and partially unknown.
- Known part has high diameter.
- Full network (known + unknown) has low diameter.
- Structure of known part provides cues for using unknown part.

FIGURE 5

The gold standard here, which is going to pop up through a lot of the results, is this difference between things that grow logarithmically in n versus things that grow linearly in n ; or more generally, things that grow like polynomials in $\log n$ versus polynomials in n . The idea is that the network has $n \times n$ nodes, but we want to be exponentially better. We want to be something like $\log n$ or $\log n$ squared in our delivery time. That would somehow tell us that the world is small. We can get there in a short number of steps.

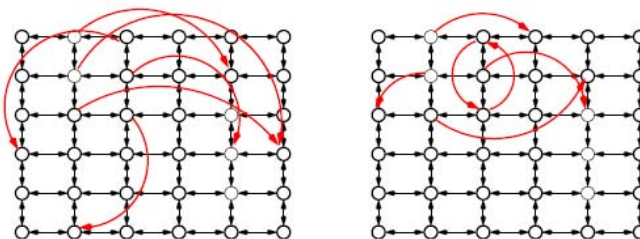
The funny thing here is that, although one could have asked this question immediately upon seeing the Milgram experiment, the Watts-Strogatz model, although not intended for this purpose, is really ideal as a kind of minimal model you would need if you wanted to ask this question. If you want to ask this question in an interesting way, what are the minimal ingredients you need? Well, you need a network that is partially known and partially unknown. A completely known network of course leads to the shortest-path algorithms, but that's a very different subject. Finding your way in a known network—that's basically the Mapquest problem. A completely unknown network is equally different. Then we have nothing to go on. In the known part you have high diameter, or else it would be too easy. For the full networks you have low diameters, so there is a solution to find. And we want to explore this relation between the

known and the unknown and the unknown part.

How does the random network relate to the way in which it's embedded in the spatial or organization structure or whatever? Again, if you are worried that this model is too simple, obviously, in a million ways it's too simple. Obviously, to the extent that the grid in Figure 4 is a proxy for anything, maybe it's a proxy for geographic information. We know people in the Milgram experiment also used occupational information and other things. I'll try folding some of these things in, but I want to start with an intentionally very simple model, as shown in Figure 6.

Generalizing the Network Model

Theorem: The search time of every decentralized algorithm in the Watts-Strogatz network is at least $\sim n^{2/3}$.



Generalizing the model: Add clustering exponent r .

- For each node v , add directed link to random long-range contact.
- Choose w as the contact with probability proportional to $d(v, w)^{-r}$ where $d(v, w)$ is the lattice distance from v to w .

FIGURE 6

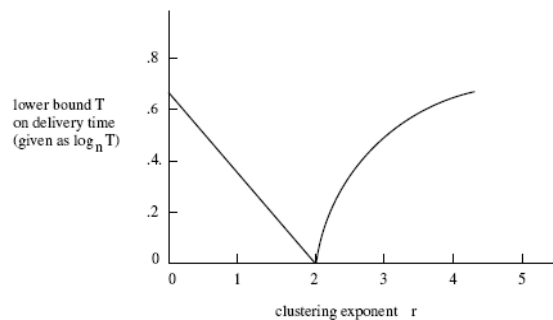
The Watts-Strogatz model is a bit too simple. So, the first theorem you can prove here is a negative result, an impossibility result which says that the network has $n \times n$ nodes. There really is a path of length of $\log n$. There was a solution to find, but probably with local information you can't find it. Any decentralized algorithm, however hard it computes, simply its lack of information is going to mean you need about at least n to the $2/3$ steps to find your way, and $2/3$ is actually the right answer here. You can also achieve this, but this is exponentially larger than they want. So, that's what happens when you have a simple toy model. You have this striking phenomenon, and try analyzing it, and sometimes there is not enough in there.

So, let's try generalizing the model just a little bit, just minimally to try to do something. We could add one more knob we could turn, and I'll call this a clustering exponent r . I'll still add

a random long-range link out of each node, as you see in Figure 6, but it's going to be tuned by this parameter r , which basically says so if I'm node v , I don't actually choose my long-range link uniformly at random, but rather, I take into account the spatial structure. I choose w as the other end, with probability something like lattice distance to the minus r . So, think about the effect r has: when $r = 0$, that's the uniform distribution and I'm back to the Watts-Strogatz model. As I crank r way up to some large number, I go from the network on the left in Figure 6 to the one on the right, where long links are very unlikely. Therefore, the network goes back to having a large diameter. My long-range links are not really helping me that much. I'm in this funny spectrum here where I know that when $r = 0$, the network is too confusing. There were the short paths, but I couldn't find them, and when r is very big, there are no short paths to find. The question is what happens in this middle region as we trace it out. At least we have a 1-dimensional family we can study.

The answer is that something kind of funny happens. There actually is an optimal operating point for this network as shown in Figure 7, and that's when $r = 2$. So when I choose links with probability according to an inverse square distribution, I can actually achieve this gold standard of polynomial \log squared n . As I move away from $r = 2$ in either direction, things get worse rapidly, at least in an asymptotic sense. Things become n to some power, where the power depends on how far away r is from 2. Two was somehow the right operating point for this network.

A Full Characterization



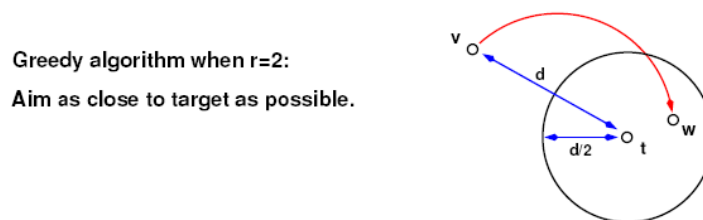
(a) $r = 2$: There is a decentralized algorithm with delivery time proportional to $\log^2 n$.

(b) $r < 2$: Decentralized algs need delivery time $\geq \epsilon_r n^{(2-r)/3}$

(c) $r > 2$: Decentralized algs need delivery time $\geq \epsilon_r n^{(r-2)/(r-1)}$

FIGURE 7

Let's think about that. Figure 8 gives some idea of how you prove something like that, and then we can go on to how we might extract out what we have actually learned from this beyond the fact that we are on a grid, beyond the fact that we have one long-range link, and so forth. But even in this very simple system there is obviously something to be proven. Referring to Figure 8, let me tell you how to prove the positive result with exponent 2, and I'll skip the negative results for the sake of time. The basic idea is that I say I'm trying to get to this target t and I'm at this node v . Rather than try to trace analyze the whole sequence of steps from here on, I want to ask a simpler question: What's the probability that in this very step I have my distance to the target? To have my distance to the target I have to go from d to jump into this ball of radius $d/2$. You can basically compute the probability mass in this ball and show that it's about $1/\log n$, independent of d . The d cancels out, which is the key here.



Basic outline of the analysis.

- Suppose message is at v , distance d from target.
- Show that with probability roughly $1/\log n$, message will enter ball of radius $d/2$ around t .
- Conclude: distance is halved roughly every $\log n$ steps.
- Distance can only be halved $\log n$ times \implies delivery time proportional to $\log^2 n$.

FIGURE 8

Basically, with $1/\log n$ probability you have your distance to the target right away. Therefore, every $\log n$ steps or so you expect to halve your distance. The distance can only be halved $\log n$ times, and so the delivery time is proportional to $\log n$ times $\log n$. That's the one-slide outline. My interest here is to try figure out whether we can conclude something more

generally about networks, since networks aren't $n \times n$ grids with random jumps and so forth. What is actually going on? Somehow, it was key that d was independent of your probability of halving. But somehow d just washed out completely. It had something to do with the fact that no matter what distance scale you were at, you could be sure of halving your distance, or expect to halve your distance. Think, somehow about this result qualitatively. Let's think about these exponentially layered distance scales around the nodes. All nodes that are within distance 1-2, 2-4, 4-8, 8-16 and all these powers of 2.

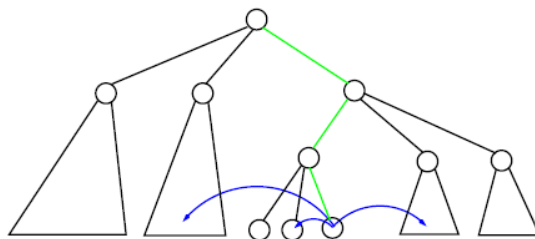
There are $\log n$ different distance scales in the network, and the great thing about $r = 2$ or, if I were in k dimensions, $r = k$ —somehow it's the dimension that matters here—is that nodes have about the same portion of links to each distance scale. You have about the same probability of a link to somebody very close as to one moderately close or one somewhat far away, very far away, and so forth. That was actually in one sentence what was going on in the proof. In the early version of this talk I had some hand-drawn picture trying to illustrate distance scales. And then I realized there are people who have done this much better than I have. For example, Sol Steinberg with his well-known 1974 *New Yorker* cover that was designed to show something else illustrates this perfectly. We devote the same amount of attention to 9th Avenue as we do to the rest of our neighborhood in Manhattan, as we do to all of New York City, to all the rest of the United States and of the world. These different layers of distances all occupying about the same amount of probability mass, and that's what is going on qualitatively.

Now, why is it useful to observe that? Well, that doesn't help us prove anything new, but it does help us go out and look for other models where we might see that happening. For example, we quickly conclude that the grid wasn't really very important to this behavior. The grid just gave us a system of balls, centered at points that had a certain way in which they grew in area. I could just as easily put the nodes on a hierarchy, and this is something that people ask about, because we also think about the world hierarchically. We even think about people hierarchically as we break them down, say, by occupation. It's fine to think about that kind of a model. So, now nodes would reside at the leaves of this hierarchy, as shown in Figure 9.

How is each node going to generate its random links? Here is a way of generating random links: first pick an ancestor at random in the tree, look at my path to the root and some distribution on the ancestors, and then pick a node uniformly at random. Suppose that's my generation process. Then the analog of searchability is explored in a 2002 paper of mine and also in a paper of Duncan Watts, Peter Dodds, and Mark Newman (2002), which also considered the case of multiple hierarchies. You get the best navigation in this case. Again, you have this uniformity over distance scales, which here would mean I should choose my ancestor uniformly

at random in this generation process. In a way, for hierarchies, the result is much more intuitive than this sort of mysterious inverse square law that you then have to explain. Here it makes sense that if I want to have good searchability to any target, I should be able to funnel down to any level of the tree as the search proceeds. If I'm trying to make my way over to some node over here, I first have to span this barrier, which requires a really long link. Once I'm into the subtree, I have to span the next level barrier, and I need to be able to do this at every level. This is actually the first indication that you could potentially look at real measurements doing that.

A Hierarchical Small-World Model



Instead of a lattice, suppose underlying model is a hierarchy.

- Nodes reside at leaves. Each node chooses ancestor at random, then chooses a leaf at random in that ancestor's subtree.
- Hierarchical analogue of small-world result:
[Watts-Dodds-Newman 2002, Kleinberg 2002]:
Best navigation when ancestor chosen uniformly at random.
- Web measurements, focused crawling algorithms [Menczer 2002].

FIGURE 9

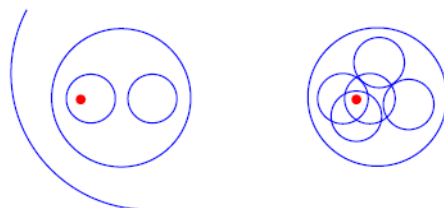
So, with the initial searchability result, you had the grid. You had these long-range links, and people said, okay, take searching the Web. You are browsing on the Web with local information. Where is the grid? Where are the random links? And the answer was, I don't know. Once you have a hierarchy, you can actually do things like take a hierarchical classification of Web pages. This is something that was done by Menczer at the University of Indiana. To look at how links are embedded in the structure, he took the open directory which classifies Web pages hierarchically (like on science/computer science/algorithms, biology/neuroscience, or art/music/Italian opera). Right now it's not a spatial structure, but a

taxonomical structure, and you could say if I actually wanted to do focused Web crawling, if I wanted to get a page on Italian opera, then as I look around at my links I should try to find ones that lead to pages on art. Once I'm within there, I assume there is going to be a slightly higher density of links into music, from there a higher density into opera and so forth. Not only did he partially validate this density of links decaying with height in the tree, but it also became the basis for some focused Web crawling experiments. Again, an example of the turnaround of seemingly simple models rapidly becoming design principles in this case for something like focused Web crawling.

There are a bunch of other proposed models here. For example Duncan, Peter and Mark had this model where you actually had several hierarchies. One hierarchy could be, as in the case of the Milgram experiment, geography. Another hierarchy could be occupation. You can choose among these different hierarchies, and there were some other proposed models. Maybe the right model for social networks is a hierarchy on top of a lattice.

At some point you think that, if these two results have very similar proofs, there must be some kind of more general statement that covers all of these. Figure 10 is actually a proposal for something that simultaneously includes searchability in grids, searchability in hierarchies, and a bunch of other things. In the end you could say nodes are embedded in some structure, some reference frame, unless it's imagined now as a collection of finite sets. Nodes form links because they belong to certain groups: I'm friends with you because we live on the same block, or because we work in the same field, or because we are fans of the same science fiction author. They all belong to some overlapping collection of groups.

Groups: A Common Generalization



Nodes form links because they belong to certain groups.

- In a lattice: groups are all the balls centered at points.
- In a hierarchy: groups are all the rooted subtrees.

Let node v link to node w with probability proportional to $\frac{1}{k^r}$,
where k is the size of the smallest group that contains both of them.

- Subject to some technical conditions, optimal searchability achieved when $r = 1$.
- Generalizes both lattice and hierarchy results.

FIGURE 10

If I were just in a lattice the groups could be balls, essentially of points. In a hierarchy, they would be the subtrees. In any set of groups, if you just draw some red dots as the nodes and some blue circles as the groups around them, I can generate a network by saying let v link to node w with probability proportional to $1/k^r$. Here r is my exponent again, and k is the size of the smallest group that contains both v and w . That's now going to be my very general proxy for distance here. It's somehow generalizing distance in a grid, and also least common ancestor height in a hierarchy. Simply how far apart you are is the smallest group you share. Sure, we both live in the United States, but the real point is that we are both fans of the same obscure author, and that's really the smallest group.

You can't take any set system and have this work. Subject to some technical conditions that include all these cases, often searchability is achieved when the exponent is 1, which is one very general way to see where this inverse square law is coming from also, because if I were on a grid and the groups were all the balls centered at points, then if you are at distance d from me, what's the smallest group containing both of those? How big is it? If you are distance d , then because area grows like the square, it's d^2 . So, 1 over group size to the first power is 1 over d^2 which gives us back the inverse square law. This is a way to see all those results under one common rubric, as in searchable networks can be achieved if you link with probably 1 over the size of the smallest group containing them. We now have a bunch of general kinds of results and now a few things happen in different directions. We have both a pattern that we can look for; we

also potentially have a proposal for ways to design networks.

Searching for a file with the original Napster was the equivalent of sending a letter not by four different acquaintances, but by using the U.S. postal system. It was a central index that knew everything, and you just asked it where the file is. But as noted in Figure 11, once Napster went away we were plunged from the world of the postal system or of directory assistance into the world of the Milgram experiment: if you want to find something, there is no central directory, you have to ask your network neighbors, who in turn ask their network neighbors, and maybe you find your MP3 file and maybe you don't. In fact, a lot of the early debates about things about Gnutella versus Freenet versus a lot of these early systems look very similar to different ways you might have run that experiment. For example, comparing Gnutella's flooding solution and Freenet's targeted solution, to take two of the very earliest ones of these decentralized things, is analogous to comparing a chain letter to the Milgram experiment. One is much more likely to get you a target, assuming it scales, but it scales very badly. Freenet was able to take advantage of some of the earliest of these results by configuring itself to look like a searchable small world. More recent work has actually tried to do this even more explicitly.

Peer-to-Peer Systems

File-sharing on the Internet (Napster, Gnutella, Freenet, Kazaa, ...)

- **After demise of Napster, centralized solutions no longer feasible:
File-sharing becomes a small-world search problem.**
- **Each node has some files and some neighbors it knows.
When file request arrives, node asks neighbors to help locate it.**
- **Gnutella: brute-force flooding of network.
Freenet: small-world-style directed search.
Recent approaches place nodes in "virtual" Cartesian space
and perform search relative to these coordinates.
Chord [Stoica et al. 2001], CAN [Ratnasamy et al. 2001],
Tapestry [Zhao et al. 2001], Pastry [Rowstron et al. 2001],
Viceroy [Malkhi et al. 2002], Symphony [Manku et al. 2003].**

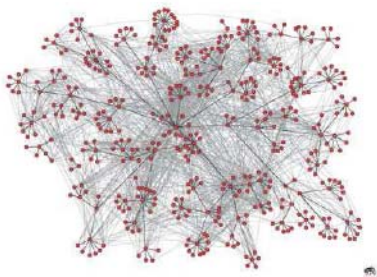
FIGURE 11

The online world gives us so much flexibility that there is no need to care that we don't live on an $n \times n$ grid with uniform spacing and random links. Once we build an overlay network on the Internet—as in CAN, for example—if you want the world to look like an $n \times n$ grid with

uniform spacing, you simply take all your nodes include in the protocol that each will be assigned two coordinates. That's the node's position in the grid, and it can then generate some random links.

For a lot of these peer-to-peer systems, it became remarkably easy to generate searchable networks explicitly by embedding them in some kind of Cartesian space, and some of these use these small world models quite explicitly. That's all I'm going to say about this as a design principle here.

Social Network Data



[Adamic-Adar 2003]: constructed social network on 436 HP Labs researchers, joining pairs who exchanged ≥ 6 e-mails (each way).

- Defined set of groups via sub-trees of organizational hierarchy.
- Let $g(x, y)$ denote size of smallest group containing x and y . Density of links scaled as $g^{-3/4}$. (Versus g^{-1} for optimal search.)

FIGURE 12

Let me go back to data. About a year and a half back there was this nice study by Lada Adamic and Eytan Adar where they looked at a particular surrogate social network, shown in Figure 12, on 436 researchers at Hewlett-Packard Labs in the Bay area. They simply asked Hewlett-Packard (HP) to contribute suitably anonymized logs of who communicate with whom by e-mail. They then built a network joining pairs of people who exchange at least 6 emails each way. That's our proxy for having a link on the network shown. Now, the question was what should we embed this into? There is a very natural embedding for the social structure of HP Labs, the organizational hierarchy. The lab director, area managers, and lower managers can be seen in the graph in Figure 12. Most organizations would look something like this. Because we have these more general ways of thinking about searchability, we can now ask how well this

maps onto some of these models of a searchable network.

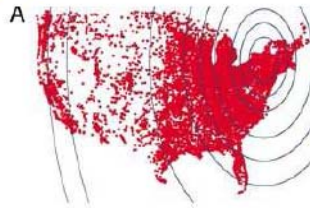
I can ask if $g(x,y)$ is the size of the smallest group containing nodes x and y , how does the density of links scale as I go up the hierarchy? As we know, optimal searchability should be one over group size. When I first saw they were doing this, I was a little concerned because it wasn't clear to me they would go down with group. That was the hope, but in fact while it doesn't give us g^{-1} , it gives us something that is in the ballpark: it gives us density of link scaling as $g^{-3/4}$. I don't want to conclude too much from this, because this is a single group of 436 people. They fit something as best they could here, and of course this network was not optimized for searchability. HP Labs didn't say let's set up everything so that we can do this Milgram experiment as efficiently as possible, although you can make qualitative arguments about why that might be a good thing for your organization. In fact, it says that the organization has something of that character, so it does give us a very general way of talking about how frequency of communication scales as you kind of move up layers, as you have to span larger and larger layers of an organization. Around this point in time there were these peer-to-peer networks. There was the open directory. There was actually something like a social network, and I began thinking it's nice that this pattern is appearing in real data. In the end the grid was just a metaphor. The grid with the random long-jumping links was just a way to get us thinking about this, but it was really about hierarchies and groups. There really was no grid, right? That, too, was not quite right.

There was a nice experiment done in the past year by David Liben-Nowell, my former student at Cornell, who is now at Carleton College, and a bunch of colleagues of mine: Robby Kumar, Jasmine Novak, and Andrew Tomkins, most of who are now at Yahoo. They looked for how links span actual geography. Remember that question, where was the grid, where are the long-range links? Well, for that you would need a large network in which people have somehow identified their friends and provided you geographic location information, which five years ago seemed way too much to hope for, but which we now actually have in things like LiveJournal.

So, you go into LiveJournal, which is an online community mainly populated by teenagers, and it's very easy to write a crawler that collects names, zip codes, and links. This is actually out there in public. So, as shown in Figure 13, you collect 500,000 members of LiveJournal with U.S. zip codes, giving you about 4 million links. The average number of friendships declared here is about 8. You may ask the question how it decays with distance. This gets you remarkably close; somehow you are still one step short of the goal, because the inverse square law was about the growth rate of balls, and the growth rate of balls was really on this assumption that the density is uniform. Within distance d , I have d^2 nodes no matter where I am. You look at the scatter plot of who is using LiveJournal and you see it's very non-uniform. The

first thing I thought when I saw the scatter plot was isn't that weird, no one out west is using LiveJournal. But actually this is the same thing you would see if I just plotted the population density of the United States. So, they are indistinguishable at the level of scatter plot.

Geographic Data: LiveJournal



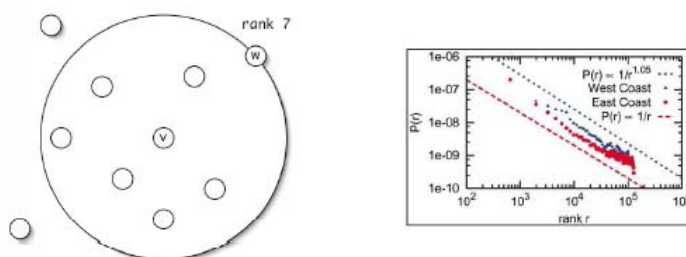
Liben-Nowell, Kumar, Novak, Raghavan, Tomkins (2005) studied LiveJournal friendship network.

- Source of large-scale network data with geographical embedding of nodes: 500K members with U.S. Zip codes, 4 million links.
- Analyzed how friendship probability decreases with distance.
- Difficulty: non-uniform population density makes simple lattice models hard to apply.

FIGURE 13

LiveJournal is sampling pretty broadly, but it's very non-uniform. The growth rate of the ball of somebody, whoever is using it out here in eastern Nevada, is going to grow very gradually compared with someone who is using it on a particular block in Manhattan. What do we do about that? One thing we do is use groups, because groups didn't require a grid. Liben-Nowell et al. simply said the groups are the first some number of people within some distance of me, all the balls centered on all the points. They took something that was slightly easier to analyze in their context. It's very similar; it's what they called rank-based friendship. Node v simply rank orders all their potential friends by everyone in the world by proximity. So the center of the circle in Figure 14 is node v , and their closest people are the small circles within that circle, the diameter of which is equal to the distance to person 7; that person is ranked 7th. Then what you do is link with probability somehow decaying in the rank: I'm more likely to link to the n^{th} most close person than the $(n+1)^{\text{st}}$ closest person. What they did was prove a different generalization. So, rather than generalize things into group structures, they generalized it into rank-based friendships and proved an analog of the inverse square results, as shown in Figure 14.

LiveJournal: Rank-Based Friendship



Rank-based friendship: rank of w with respect to v is number of closer friends that v has.

- Showed that efficient routing is possible for (nearly) arbitrary population density, if link probability proportional to $1/\text{rank}$.
- Generalization of lattice result (diff. from group structures).
- **Punchline: LiveJournal friendships approximate $1/\text{rank}$.**

FIGURE 14

Efficient routing is possible for a nearly arbitrary population density, again, mild assumptions here, if the link of probability is decay like $1/\text{rank}$. Why is this a generalization? Well, because if distance d would be ranked d^2 , there are d^2 people close to them. This gives me the inverse square law in that special case. Now they've got results saying what would be optimal for searchability? They have got 500,000 people with zip codes, and they look at some data. The punch line, which made me very happy, is that LiveJournal friendships actually approximate $1/\text{rank}$ quite closely. Time prevents me from getting into a lot of what they did here, but these lines are basically slope 1 and 1.05, and the chart in Figure 14 is one particular view of that data.

For Hewlett-Packard Labs you could have argued somebody sat down and drew up the order chart. Maybe they were thinking about this. This is weirder still; it somehow configured itself to be like this. I don't have a good explanation for that. I should also mention that independence of searchability is a little surprising because LiveJournal is a purely virtual online community, so if someone asked me what do you expect, I would have said a mixture model. I would have said you log on and you immediately link to all your high school friends who live in the same zip code as you. So, I get a big spike there, then I meet people utterly at random, and it's kind of like a uniform distribution on the rest. The point is that is not what happened. If you lived in Nevada, you were nonetheless more likely to link to somebody in Nebraska and

Wyoming than you were to someone in New York City, and that's bizarre. There is a lot more that should be looked at in this data, so one hesitates to draw sweeping conclusions yet, but it's very intriguing.

In the final few minutes I have remaining I want to talk about not just spatial embedding, but how things have evolved over time because here, too, there tend to be some surprises that actually make us go back and think about some things we need to be looking at. The question we wanted to ask was really a very simple one, stated in Figure 15, which is take one of these large model systems—we'll take the arXiv, because it really works very nicely for our purposes—and ask how do the degrees of separation change in the network as it grows over time? We could define degrees of separation to say the 90th percentile distance among all pair-wise distances, but we have tried other ways. Why in the world has no one studied this on some large-scale system? Well, it's computationally hard to answer this question. You want a network you can run up to any point in time and stop it. You would like to ask this question every month or every week, and every week that means you have to do an all-pairs shortest-path problem on a 100,000-node network. This is computationally somewhat demanding, so we were actually able to take advantage of some nice stuff that one of my co-authors wrote for approximate all-pairs shortest-paths in a large graph, some stuff Faloutsos worked on to track this effective diameter in several networks over time.

Evolution over Time: Network Distances

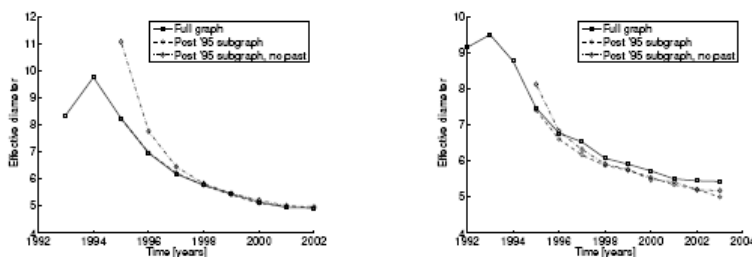
How do the “degrees of separation” change in a network as it grows over time?

- **Computationally intensive to answer this question.**
- **Leskovec, Kleinberg, Faloutsos (2005) tracked “effective diameter” in several networks over time**
(including arXiv co-authorship, arXiv citations, patent citations, Internet AS, ...)
- **Original plan: see if we could determine which slowly-growing function of n best fit the diameter.**

FIGURE 15

We looked at things beyond the arXiv of this model system, so on Figure 16 you will see plots. Time runs on the x axis, and effective diameter of the network runs on the y axis. The original plan was we’ll look at this, and we’ll try to see which slowly growing function of n best fit the diameter. Was it $\log^2 n$, was it $\log n$, or some other function? Given that was the original plan, we found these plots a bit surprising, because slowly growing functions as you know go up, and these were all going down. The point was, despite the fact that the network was growing—and some of these networks were growing at an exponential rate over time—the effective diameter, and this is basically a smooth version of the 90 percentile, which I can tell you more about offline, was going down as the network grew. The network was growing, but the degrees of separation were actually getting closer. I have a few lines here which are probably better explained in question and answer form, but basically, there is a big missing past issue here. We are coming and observing networks starting at a particular time.

Evolution over Time: Distances



- In fact: effective diameter is shrinking as these networks grow.
(Need to deal with “missing past.”)
- Related property: # edges grows superlinearly in # nodes
($e \sim n^{1.69}$).
- Progress on extending models to incorporate these behaviors.
 - ▷ Roughly: new person enters social network via point of contact;
they befriend each of contact’s friends with some probability,
and recursively. (An epidemic process ...)

FIGURE 16

With the arXiv, we technically have no missing past problem because we have the arXiv from its very first paper. But in a very real sense we have a missing past problem, because the archive got plopped down in the middle of a lively, thriving research community that has been exchanging pre-prints for a long time. We only started measuring it at time “zero.” One worries that what these results really show is the process of catching up with a stationary structure the network already had. I can’t say that that’s not what is going on, but one can look at the decrease in diameter in many different ways. For example, one can artificially cut off the past, any number of years you want. You can say supposed we only started measuring in 1995. That’s what the dotted line in the left-hand graph represents, and you get this initial massive swoop down as the network catches up with links that essentially were already there. That effect quickly goes away, and we still see a steady decline in diameter long after that effect appears to be negligible. One can also do things where you keep the network intact, but only start measuring—I would have to explain this more slowly—but it, too, catches up with this. So, what we seem to be seeing, long after a lot of the edge effects go away, we are still seeing this steady march downward. Maybe in retrospect I should have seen this coming, because back on Figure 5 I said our gold standard is $\log n$. We have n nodes and we want the diameter to grow like $\log n$. That’s sort of the graph

theorist's take on what the small world property means. It means that you grow exponentially slower than the size of the network. The common sense take on what the small world property means is that even as the world is growing, there are forces out there that are causing us to become more closely connected; in other words a decreasing diameter. Intuitively, maybe this has been consistent with the common sense intuition behind the small world property all along. It was simply our random-graph background that caused us to inject this log imagery. Having said that, $\log n$ and $\log^2 n$ and these slowly growing functions are a very useful modeling tool for the kinds of analyses I have been talking about up until now, which was not of a temporal trend nature but simply looking at static snapshots, and talking about asymptotics as a way of capturing these. But it says that there is a lot to do here. A related property, one reason that sort of makes it easier to believe why it might be going down is that the number of edges also decreases, so the average degree in the network is not constant.

If you compare now versus 10 years ago, people cite more papers, they author more papers, they co-author with more people. Everything in the arXiv network is densifying as we go forward, and clearly, any model is going to have decreasing diameter. There are models that we have been thinking about to do this. For example, one way, at least through simulation—this seems very hard to analyze—but through simulation if I have a model where people enter a social network via some point of contact, you are introduced into the network through a sort of epidemic process in which you befriend some of that person's friends, some of their friends, and so forth. Those are your links, and now we iterate, add a new person. Through simulation, one can set the parameters so that it densifies, the diameter goes down, but it seems very hard to prove anything about that kind of model. At least there are models out there that seem plausible, that might capture this kind of thing.

Reflections

- Network models \Rightarrow general patterns
 \Rightarrow design principles; discovery of patterns in real data
- Shrinking diameters: how much do we need to revise our thinking?
- Temporal patterns suggest more surprises ahead.
Large effort at Cornell to understand Web evolution over time.
(Macy, Strang, Arms, Huttenlocher, Kleinberg).
- Further applications of small-world networks:
 - ▷ When are short paths “meaningful?” [Faloutsos et al. 2004]
(E.g. if my next-door neighbor subletted an apartment from a person who took a class with the brother of a 9/11 hijacker.)
 - ▷ When are people willing to help forward messages on paths?
Incentive-based querying in p2p and social-networking systems [Kleinberg-Raghavan 2005].

FIGURE 17

Figure 17 presents some reflections on various things and suggests some other questions. As I said, the set of themes that I wanted to try tracing through this was somehow the way the network models, even very specific models, become more general patterns, they become design principles. In the end, with enough patience, you can actually find data sets that really start to bear some of these things out. The shrinking diameters thing is again very suggestive of more work that could be done here, and in general I think we are only starting to have the computational ability to look at very large networks and snapshot them densely over time. I think it suggests that there are more surprises coming up.

One network that we would like to try understanding over time is the Web itself. There is currently a large effort going on at Cornell involving some pure scientists, and also some people in the sociology department like Michael Macy and David Strang. In particular, we are trying to get a handle on Web evolution over time and some questions we can ask them. In small world networks, there are any number of other interesting questions that one can ask, which I haven't touched on here. One, for example, which came up in a paper a year ago by Cristos Faloutsos, Kevin McCurly, and Andrew Tomkins, was roughly the question of when are short paths meaningful. So intuitively there is this tempting connection between social networks and, for example, homeland security applications in which you discover statements like my next door neighbor once sublet an apartment from a person who took a class with the brother of a 9/11 hijacker. You read in the news that social network analysis is going to cause us to discover things

like this and that's going to be very useful. Are we really lacking any sense of how you score? Is this short path meaningful? Isn't the whole point of the small world phenomenon that I could say this about any instantiation? I would say that we are really lacking, and I think this is interesting direction for the statistics aspect of this workshop, some way of talking about when short paths in the network are meaningful, and when they simply suffer from what you could call the small world problem, that you see short paths everywhere.

Another area where this blends into economics and game theory—this is some work I have been doing with Prabhakar Raghavan, who is the head of research at Yahoo, where they are very interested in social media—is the notion that a lot of these onlines like LiveJournal and Orchid and Friendster and Linked In, and you can name any number of these, they attracted a lot of excitement a year ago based on the idea that they would become big marketplaces for information and services. You join one of these, and what's the value to you in joining something like Friendster or Link In or Orchid? Well, because maybe a year from now you are going to need to rent an apartment in Boulder, Colorado you ask your friends. You say you need to rent an apartment in Boulder, Colorado. Do you know anyone who has a room to rent? They ask their friends, and gradually an answer percolates back. Not only is it an answer, but it somehow has been vetted by your friends. That was some of the appeal of these networks, but once you start thinking about this as a marketplace for information and services, you have to ask about the incentive issues at work. If you are actually offering incentives to people, and they have to act essentially as middle men for these products and services, then when does this function efficiently? When do things break down simply because of the friction of everybody skimming off money that you offer them?

It turns out there are some interesting transitional behaviors here, which relate to some things about branching processes, and some sharp transitions between networks that function efficiently in this mode and networks that don't. There is something qualitatively important about an effective branching factor of 2, and something I can't really get into, but there is also a lot when you think about this interface of small world networks with the sort of economic game theory take on it.

In general, I think this is a window into a lot of interesting questions in these kinds of networks, a lot of which can be studied with some of the model system data that we have. We are trying to accumulate more data, and I think in general there is much to be done here. I would be happy to talk about it more offline, but for now I'll stop here.

QUESTIONS AND ANSWERS

DR. HANDCOCK: I have one question which has four related parts. This is coming back to the HP real world email data set. How did you fit the model? And then how did you assess the goodness of fit. Three, was there any missing data from the email responses, and how was that dealt with? And four, because it's an ad hoc model based on heuristics, how did it compare to other models that took into account the organizational structure?

DR. KLEINBERG: Those are all good questions, and actually I will start with an answer which applies to all four parts, which is I didn't do any of this, because this was actually the work of Adamic and Adar. So, I don't want to answer in detail on their behalf, because there was thinking behind that which I don't have access to all of. There was actually not that much missing data as far as I actually understood, because they essentially collected these from logs. They got people's permission to collect these from logs.

In terms of competing models, I would say there wasn't really a generative network model here, so, this was not a statement about a generative network model. It was a statement about simply trying to fit a decay of linkage frequency to group size. And so then it just comes down to a question of the goodness of fit for that particular data set, for which you can consult their paper. But there wasn't an attempt to then infer from that a general model.

DR. CROVELLA: I'm Mark Crovella from Boston University. I'm just sort of curious about that example. What do you think? Is it more likely that the e-mail messages are the result of the organizational structure, or the organizational structure is a result of the e-mail messages? If it's the latter, I would be interested in the social engineering aspects of that.

DR. KLEINBERG: Yes, I think that's a very interesting question. So, this may sound like a cop out, but I don't think it is that obvious. As we know from looking at organizations, they do change: groups get reassigned, groups split, groups merge, and that clearly is the result of increased bandwidth. So, here unfortunately, I'm not going to draw so much on social science studies since I haven't looked that much at the organizational theory of literature. But simply suggesting that obviously groups split and merge based on increased or decreased traffic between that, and that's a sense in which the e-mail traffic in the long-range sense is going to feed back into the organizational structure. But it's a good question. You could also argue that both of these are consequences of some hidden variable. So, in fact, people are working on similar or different things. That influences both the organization structure you see and who communicates with whom. In a sense, that's the single root cause of these, but I think it's true that as that changes over time, you are going to see changes in communication patterns in the organization.

REFERENCES

- Bollobas, B., and F.R.K. Chung. 1988. "The Diameter of a Cycle Plus a Random Matching." *SIAM J. of Discrete Math* 1.
- Faloutsos, C., K. McCurley, and A. Tomkins. 2004. "Fast Discovery of Connection Subgraphs." *Tenth ACM SIGKDD Conference*. Seattle, WA.
- Kleinberg, J. 2000. "Navigation in a Small World." *Nature* 406.
- Kleinberg, J. 2002. "Small-World Phenomena and the Dynamics of Information." *Advances in Neural Information Processing Systems (NIPS)* 14.
- Kleinberg, J., and P. Raghavan. 2005. "Query Incentive Networks." *Proc. 46th IEEE Symposium on Foundations of Computer Science*.
- Leskovec, J., J. Kleinberg, and C. Faloutsos. 2005. "Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations." *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*.
- Liben-Nowell, D., J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. 2005. "Geographic Routing in Social Networks." *Proc. Natl. Acad. Sci. USA* 102.
- Liben-Nowell, D., J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. 2005. "Geographic Routing in Social Networks." *Proc. Natl. Acad. Sci. USA* 102.
- Malkhi, D., M. Naor, and D. Ratajczak. 2002. "Viceroy: A Scalable and Synamic Emulation of the Butterfly." *Proceedings of 21st Annual Symposium on Principles of Distributed Computing*.
- Manku, G.S., M. Bawa, and P. Raghavan. 2003. "Symphony: Distributed Hashing in a Small World." *Proc. 4th USENIX Symposium on Internet Technologies and Systems*.
- Menczer, F. 2002. "Growing and Navigating the Small World Web by Local Content." *Proc. Natl. Acad. Sci. USA* 99(22):14014-14019.
- Milgram, S. 1967. "The Small World Problem." *Psychology Today* 1.
- Ratnasamy, S., P. Francis, M. Handley, R. Karp, and S. Shenker. 2001. "A Scalable Content-Addressable Network." *Proc. ACM SIGCOMM*.
- Rowstron, A., and P. Druschel. 2001. "Pastry: Scalable, Distributed Object Location and Routing for Large-Scale Peer-to-Peer Systems." *Proc. 18th IFIP/ACM International Conference on Distributed Systems Platforms (Middleware 2001)*.
- Stoica, I., R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. 2001. "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications." *Proc. ACM SIGCOMM*.
- Watts, D.J., and S.H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393.

Watts, D.J., P.S. Dodds, and M.E.J. Newman. 2002. "Identity and Search in Social Networks." *Science* 296:1302-1305.

Zhao, B.Y., J.D. Kubiawicz, and A.D. Joseph. 2001. "Tapestry: An Infrastructure for Fault-Tolerant Wide-Area Location and Routing." *UC Berkeley Computer Science Division, Report No. UCB/CSD 01/1141*.

Dependency Networks for Relational Data

David Jensen, University of Massachusetts

DR. JENSEN: What I'm going to talk about is some work that has been done in statistical modeling of relational data. I'll tell you what I mean by that in a moment. This is joint work with one of my students, Jennifer Neville, who is one of those fabulous students who the previous speaker referred to. She is also in the job market this year, so if you happen to be in computer science and looking for a good faculty candidate, she should definitely be on your list.

The main point of what I'm going to talk about is a new joint model for relational data: Relational Dependency Networks (RDNs). I'll also talk about it in the context of a particular application of this kind of statistical modeling—some fraud-detection work that we have been doing for the National Association of Securities Dealers. I will also try to serve as an exemplar of work in this area of computer science, particularly in what we generally think of as machine learning or knowledge discovery. And finally, I'll try to point to some interesting new research directions.

This talk is a bit of a Russian doll: I make multiple passes, going a little bit deeper each time. I'm a computer scientist, and we do iterative deepening search, so think of it that way. You may want to gauge your questions with that in mind.

National Association of Securities Dealers

- Largest private-sector securities regulator in the world, which monitors:
 - 5,200 securities firms
 - 100,000 branch offices
 - 600,000 securities brokers
- Responsibilities include: preventing and discovering serious misconduct among brokers (e.g., fraud)
- Last year NASD regulators:
 - Filed 1,400 enforcement actions
 - Barred or suspended 830 brokers
 - Collected \$104 million in fines

FIGURE 1

First, let me start out with this application. About two years ago we were approached by some people at the National Association of Securities Dealers (NASD). They are the world's

largest regulator of securities brokers; and they are essentially detailed by the Securities and Exchange Commission to regulate the market in much the same way as the American Bar Association and the AMA regulate their particular areas. Some background on the NASD is included in Figure 1. The particular interest that they had was to try to predict fraud, and to focus their examinations for fraud by accounting for the fact that fraud is a social phenomenon. They particularly came to us because we were involved in doing this kind of work, but we're taking into account the context, the relations between either people or Web pages or other things.

NASD has a large dataset, and Figure 2 is a rough schema of the data. Intriguingly, this is data that you can access online a single record at a time. We have the entire database, but you can access records on your particular stockbroker if you are interested in finding out if they have engaged in questionable conduct prior to you working with them. The website is <http://www.nasdbrokercheck.com>.

Central Registration Depository (CRD®)

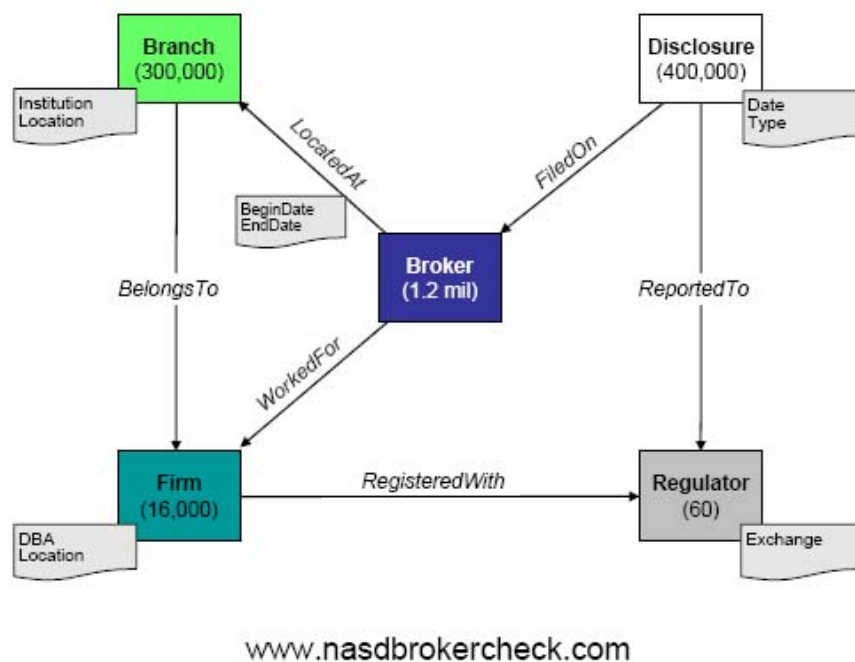


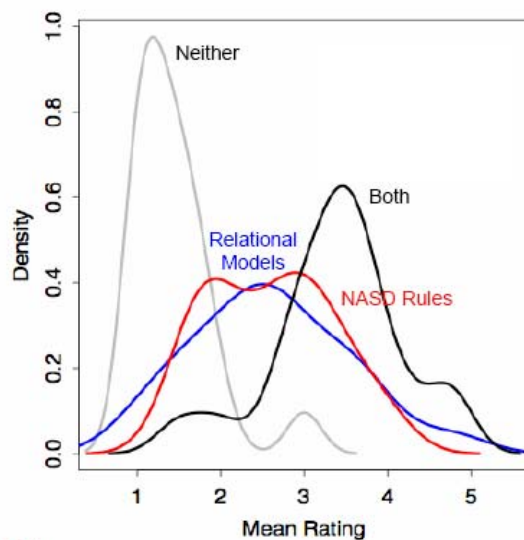
FIGURE 2

The NASD has data on roughly 1 million brokers, 300,000 branches, and 16,000 firms, so there is quite a large set of data and it is virtually complete. It is a wonderful network to look at because there are essentially no missing data in here. There is one aspect in which there is missing data, which is that brokers file disclosures indicating that they may have done something wrong. This can be a minor customer complaint; it can be a felony, and there is a large range of possible behaviors. This is a self-reporting system, so they file data on themselves, or their branch or the firm files data about them. Certainly in the case where they might have committed a felony and been fired, it would be the firm itself that would be filing the disclosure, but we know that those are fragmentary and are incomplete.

We built a model that I will tell you about in a little while. We took that model and predicted, prospectively, who was at high risk for fraud in the next 12 months. Despite people's general expectation, it's actually a very small percentage of all brokers that do anything fraudulent, but we knew that this would be a very weak predictor. We rank ordered all brokers by their probability to fraud and compared this to NASD's own rules. NASD has expert-derived rules which they use to select out a group of brokers for special attention by their examiners. We went down the list of brokers identified that way, which included about 125 brokers. We took the top 125 brokers on our list and compared the two sets. We took brokers who appeared on both lists, brokers on one list or the other list, and some brokers that were on neither list, and put them in front of human examiners, asking them to say how interested they were in those particular brokers, how much they would have wanted to know ahead of time about these brokers. We compared the scores that these four examiners gave to each of these cases. There were 80 cases, which each took about half an hour to examine, so for four examiners this was a person-week of effort per examiner.

Figure 3 shows what we ended up with. Brokers who were on neither list generally have a very low score near 1. (Five indicates the most interest on the part of an examiner, 1 the least interest.) If we used NASD's rules alone, or our rules alone derived automatically from the data, we would essentially have the same ability to predict brokers of interest to the examiners. If you combined the rule sets and said we're going to look at individuals who were on both lists, you would actually end up with a set of brokers of even more interest to the examiners.

Results: Evaluated by human examiners



(Neville et al. 2005)

KIDL

FIGURE 3

We also got a little bit of informal unsolicited feedback from one of the examiners. These are experts who have spent years examining individual brokers, and one of the examiners said, “Wow, if you got this one case right, I’ll be really impressed.” Because this examiner supervised barring this guy from the industry, he was so bad that he had used fraudulently obtained funds to attend a compliance conference that NASD runs on a regular basis. I wouldn’t be telling you about this if their rules hadn’t missed him, which was the case. In fact, our model ranked him in the top 100. So, he would have been one we would have recommended to the examiners to take a look at. That was a nice anecdotal result as well.

Let’s back up a minute: how did we do this? What are we in the business of doing? We were looking at relational data, data that tied brokers together based on the firms and branches that they worked for. We called that relational data. What do we really mean? As we have all been talking about here, a lot of traditional work in statistics assumes that you have a single table of data. You assume the instances are independent, and you are looking for relationships between the variables in a given instance. In contrast, we assume there are multiple tables, many different types of objects, and importantly, the statistical dependencies may run between instances in the

same table, or between tables. We might want to say that one broker influences the probability that another broker will engage in fraud, or something about that broker's firm may influence that probability. That's the sense in which the data we are working with are relational.

Our goals are to model attributes. Importantly, we are not trying to predict that a link will occur. We are trying to predict that some attribute or set of attributes exists for individual objects in the data, conditioned on the attributes of other objects and the relational structure among many object types. We are also after joint models so that we can do all the neat things you can do with a joint model. For instance, you can predict anomalies and other kinds of things. We are also under the strong impression that our inferences about one object should inform our inferences about another. We are doing what we refer to as collective inference. We want to do joint inference across all of the instances in our data. We can't just do that one at a time, we need to conduct inference on large and complex networks.

Also, because of the kinds of applications we are working with, NASD is a good example we want to follow to construct understandable models. The first thing that the NASD examiners asked of us when we first met them was, what does our model really say, how does it do what it does? We were able to put the models in front of them; they looked at them and said, yes, that makes sense. They were also able to say to us, "Oh, that thing over there . . .," which we said, "Yes, yes, ignore that, that's probably noise," they said, "No, that is actually something that is interesting to us, but there is a better way of expressing it." They were able to help us improve our models and our data sets. We are talking about large, complex networks over which we want to perform inference. So, what are the models that we are building? Figure 4 gives a high-level view of the model built from the data that I have just talked about. I'll explain it in detail in a minute. Some of the details aren't here because NASD, for obvious reasons, has said "Would you please not provide too many details of the exact model that you have built," although they are surprisingly willing to work with us on publication. In fact, three of the co-authors of a paper that we just presented at the Knowledge Discovery and Data Mining Conference were all from NASD, and that paper goes into some detail. What we have here are these plates, these boxes, indicating object types; circles indicating variables on those objects, variables, or attributes; and the edges indicating dependencies, statistical dependence that runs between variables.

RDN for broker fraud

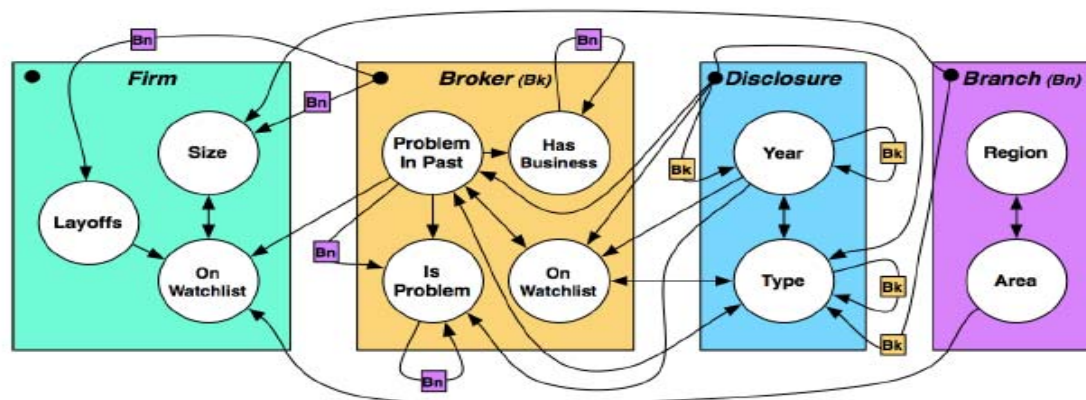


FIGURE 4

This is a type of graphical model that I will explain in more detail in a minute. It is a dependency network. So, if you want to know what an object depends on you can look at its parents. Look at things that have arrows going from them to the node that you are interested in predicting. For example, if you want to know whether the broker is a problem broker, you examine the sources of the arrows leading into that oval labeled “Is problem.” You’d want to look at whether they had problems in the past and have they had some sort of bad disclosures in the past? You would also want to look at that little loop under the oval in question, which asks whether there are known problem brokers in the same branch. There is an auto-correlation dependency, or homophily, among brokers who work in the same branch. You should look at whether other brokers in that branch have had problems in the past. That’s the dependency that goes out to the left side of the yellow box. Finally, there is one that goes up to the year of previous disclosures of that individual. You can look at these and get a sense of the overall types of statistical dependencies that exist in your data, and then the high-level summary of a model that actually has more detail sitting under each of the nodes in much the same way that Bayesian networks do. Also, there are these black dots in the upper left-hand corners of the four boxes, which indicate dependence on the mere count of those object types. For instance, the link between the dot in the blue rectangle and the “on watchlist” oval in the yellow box indicates that the number of disclosures influences whether a broker is on the NASD’s watch list. That black dot just indicates the number of, not some variable of that object type.

What are relational dependency networks? I'll tell you how they are constructed in a minute, but they are a very expressive model in that they particularly represent the cyclic dependencies, this auto-correlation or homophily. They are efficient to learn. They are essentially linear in the time needed to learn a conditional distribution, because they are actually composed of a large number of conditional models. They can perform this collective inference because the inferences about one object can inform the inferences about another. What you do when you are performing inferences is to do inference over the entire graph simultaneously. Finally, they are practical because of some nice qualities about them. You can actually use models that you know ahead of time and you don't have to learn all of the elements of those models. I should point out that the model you just looked at was learned without expert input other than the data set itself, so that model was learned automatically without any expert going in and turning off or adding dependencies.

For details you can see two papers that Jen Neville and I recently produced. Also, this builds on work by David Heckerman and co-authors at Microsoft Research published in 2000 on propositional dependency networks: networks that assume independence among instances. So the basic flavor of our work is in alignment with a set of other work that has been done in computer science and machine learning, specifically in what have come to be called probabilistic relational models, or just relational models. That includes work by Daphne Koller's group at Stanford on what they originally called PRMs, and now often called relational Bayesian networks, relational Markov networks, also out of Daphne's group, and Ben Taskar, one of her students, our own relational dependency networks, and most recently some work at the University of Washington by Pedro Domingos and his student Matt Richardson in Markov logic networks. Figure 5 lists some of this work.

Recent work in relational learning

- Joint models of attributes in semantic graphs
 - “PRMs” or relational Bayesian networks (RBNs) (Getoor, Friedman, Koller & Pfeffer 2001)
 - Relational Markov networks (RMNs) (Taskar et al. 2002)
 - Relational dependency networks (RDNs) (Neville & Jensen 2003, 2004)
 - Markov logic networks (Richardson & Domingos 2004)
- Statistical biases in relational learning
 - Autocorrelation & feature selection (Jensen & Neville 2002)
 - Aggregation & feature selection (Jensen, Neville, & Hay 2003)
- Collective inference
 - Hypertext classification (e.g., Chakrabarti, Dom & Indyk 1998)
 - General relational models (e.g., Neville & Jensen 2000; Taskar, Segal & Koller 2001; Jensen et al. 2004)

KIDL

FIGURE 5

As you can tell, there is an enormous alphabet soup here of different model types, and there are about 10 that I haven't mentioned. There has been an explosion of work in the last five years on graphical models for relational data—many different models being proposed. We are now in the process of trying to sort it all out, and figure out what the advantages and relative performance of these things are. We have done work over the last several years in statistical biases that can show up in these relational learning algorithms, and in this collective inference process. There's this idea that you've got a joint model over all of your data, and are making simultaneous inferences.

How do these RDNs work? Let's do one more Russian doll inside. Relational dependence networks are a pseudo-likelihood model. You have heard about some of these previously. A model is constructed by composing a set of conditional models that are each learned independently. This sounds like a really bad idea, because what you would like to do is learn your entire model, all of the elements of your model jointly, but if you have a sufficiently large data set, what you end up with is the data set performing a coordination function that allows you to have the individual conditional models be consistent enough with each other, that you end up forming something that looks a lot like a coherent joint distribution of the data. What you want to do, for instance, is learn these individual conditional models. You want to learn the probability that a broker is a problem broker. You look at a set of characteristics about that broker and the neighborhood surrounding them. You look at the past behavior of that broker, for

instance, which may be captured as an individual variable on that broker; relations to other problem brokers and attributes of those brokers; and other things like the location and status of the firms that the individuals have worked for. You need to learn each of those models in a way that makes it parsimonious and accurate.

The conditional models inside RDNs

- RDNs are pseudo-likelihood models learned by composing a set of conditional models $p(y|Y^-,X,S)$
- For example, $p(Is-Problem)$ given
 - Past broker behavior
 - Relations to other problem brokers
 - Location and status of current and past firms
 - ...
- To obtain advantages, the conditional models must be accurate and parsimonious

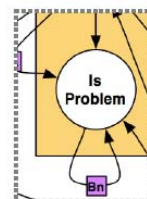


FIGURE 6

The types of conditional models that we learned are these relational probability trees. These are trees where you take, for instance, a broker and the surrounding neighborhood of objects, which we specify ahead of time, drop that into the tree, and it rattles down to one of several leaf nodes. The leaf node gives you a probability distribution over the variable that you are trying to predict. Again, this is learned automatically from the data, a separate algorithm. Part of the tree for our NASD work is shown in Figure 7, enlarged just to give you a better sense. Is the particular broker older than 27 years of age? The answer is no, so we go down the right-hand part of the tree. Then you ask something about the firm size of the broker and the current firm. You may then go down the right-hand branch and ask about the past co-workers. How many past co-workers do you have? If it's more than 35, you go down the left-hand branch of the tree, and you get down to this leaf node. You say, "Oh, well, there is sort of a roughly even probability of committing fraud or not committing fraud in the next 12 months."

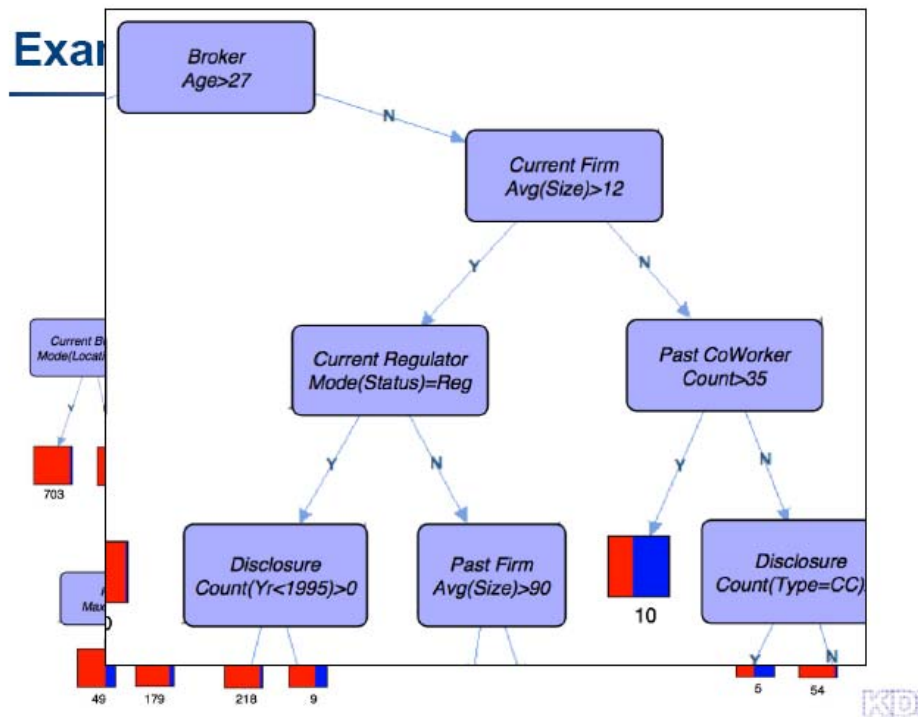


FIGURE 7

How are these relational probability trees learned? These are a tree-based classifier like those worked on for 30 years in the statistics and machine learning community. It's a very common way of constructing a conditional model. We take standard ways of building these trees, and say, "Well, the problem is we need to start with a graph. We need to start with a very large graph, consisting of lots of objects and lengths." What do we do? As shown in Figure 8, we take pieces of that graph—for instance, a broker and his or her surrounding neighborhood—and then say, okay, those are our instances. Now you have these little subgraphs that you have cut out of the graph, and you have these things that are sort of like instances, except they have widely varying structure. Some brokers may have many past co-workers, others may have very few. What you need to do is some aggregation over that to summarize the relational structure, which ends up giving you a somewhat different data set. There needs to be some adaptation of the learning to deal with problems of auto-correlation and some other things. That's addressed in detail in the paper. Finally, you can construct yourself one of these relational probability trees. Do this for every variable in your data. Compose these trees into a single model that ends up giving you the relational dependency network that I showed you earlier.

Relational Probability Trees (KDD'03)

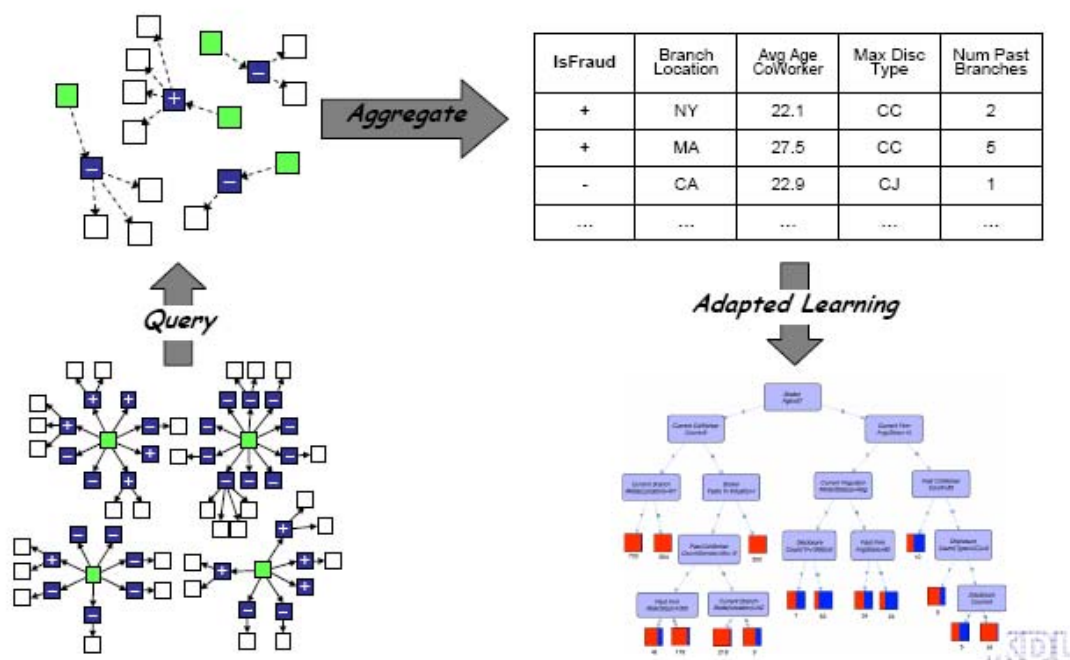


FIGURE 8

The key insight of this work and lots of work in relational data in the machine learning/CS realm has been this need to do aggregation, and to search over possible different aggregators. So, what do I want to know about the past co-workers of somebody? Do I want to know their average age; the number of disclosures that those past co-workers have had? What are the things that matter? You want to search over that space, and then choose those variables that help you most that allow you to form a highly accurate conditional distribution.

You can also have problems of what I would call pathological learning. An example is given in Figure 9. You can ignore the details of this particular network, but just to say we have a very nice relational dependency network here. If we do the wrong kind of learning of conditional models, we can end up with a relational dependency network that looks like this, that is essentially impossible to interpret, which makes inference much more complex. The reasons for that are there are at least two special problems you run into. One, it's been referred to several times as auto-correlation. You need to account for auto-correlation when you are doing statistical hypothesis tests and parameter estimation.

Pathological learning

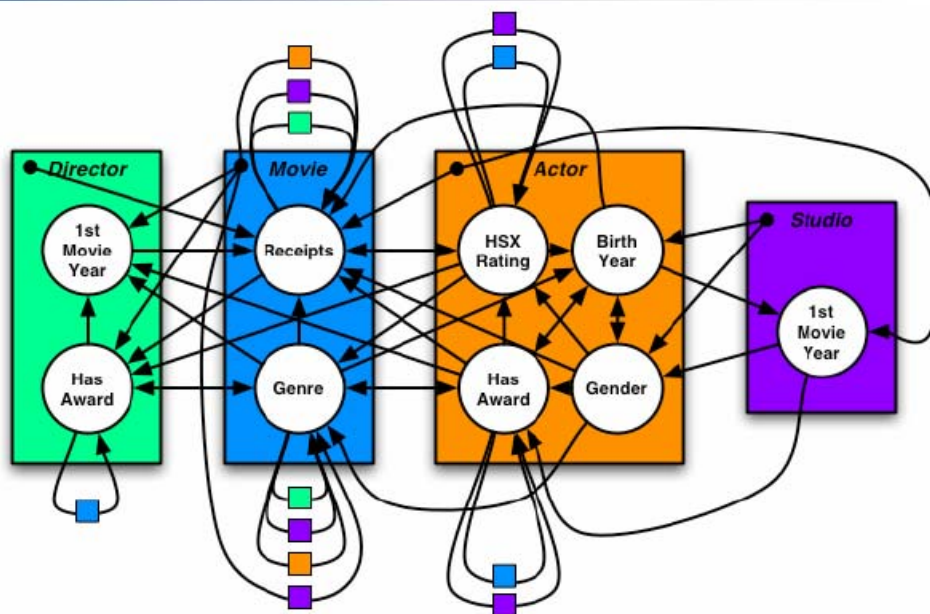


FIGURE 9

Figure 10 shows a lot of correlation. Let's say the green things there are firms and the blue things are brokers; you are looking at the characteristics of brokers and whether they will commit fraud. Firms tend to have either lots of these brokers or very few of them, so auto-correlation is a fact of life in many of these data sets.

Some causes for pathological learning

- **Autocorrelation**
 The value of a variable on one object depends on the values of the same variable on related objects
 $p(y) \neq p(y|Y_-)$
- **Structural dependence**
 The attributes and the structure of data are correlated
 $p(y) \neq p(y|S)$

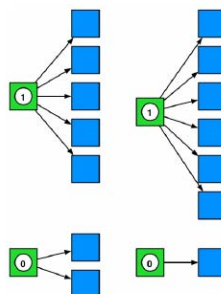


FIGURE 10

Another problem is what we call structural dependence, which is that you can often look at the mere structure, how many of something—for instance, how many brokers there are—and make some inference about the firm, where there is a correlation between the structure, the degree here in this case, and some attributes of one of the things in the graph. We have papers on both of these issues showing how they can affect both our algorithms and other algorithms in the field.

If you correct for these problems, auto-correlation and structural dependence, what you end up with are much smaller models. In each of the graphs shown in Figure 11, the far left-hand side is our approach to building a model. The far right-hand side is a standard approach in a community, C4.5. What we are showing is ending up with much more parsimonious models that are just as accurate. Among all of these models accuracy did not differ but size certainly did. That influences how many edges you end up having in that big network, which affects both its interpretability and also the efficiency of inference.

Corrections produce smaller models

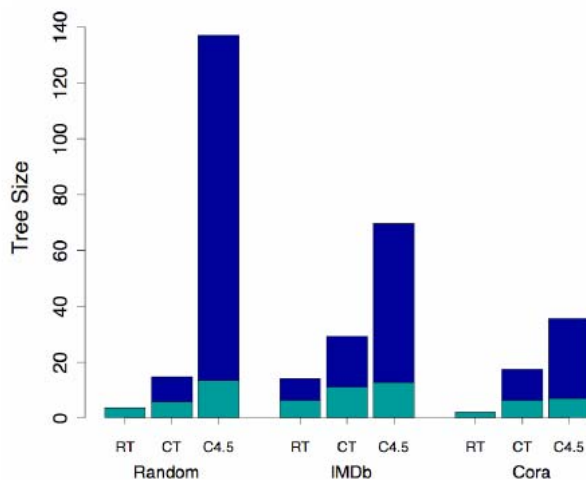


FIGURE 11

You now have this great model, and the question is, now what do we do? We've got this model but we need to actually apply it to a very large network. Figure 12 shows just notionally a citation network where you have papers and authors, and they are connected in the way that you would expect in terms of people authoring papers and papers citing each other. We have this network, and we also have attributes on these individual things: attributes on these papers and attributes on these authors. What our model says is if we want to make some inference about this

center node, center variable, what is the topic of that paper? We need to look at the topics of other papers published by that author, and we may need to look at characteristics of the authors as well. We need to go out into the network, grab the values of these variables and bring them back. We need to do that for every variable that we are trying to predict in the network, which means that this obviously can be a very intensive process. We also need to have the entire data set available to us to make these inferences simultaneously, because some of these nodes that are parents of the node we are trying to predict are, in turn, other things whose values we are inferring, other variables whose value we are inferring.

Rollout and inference

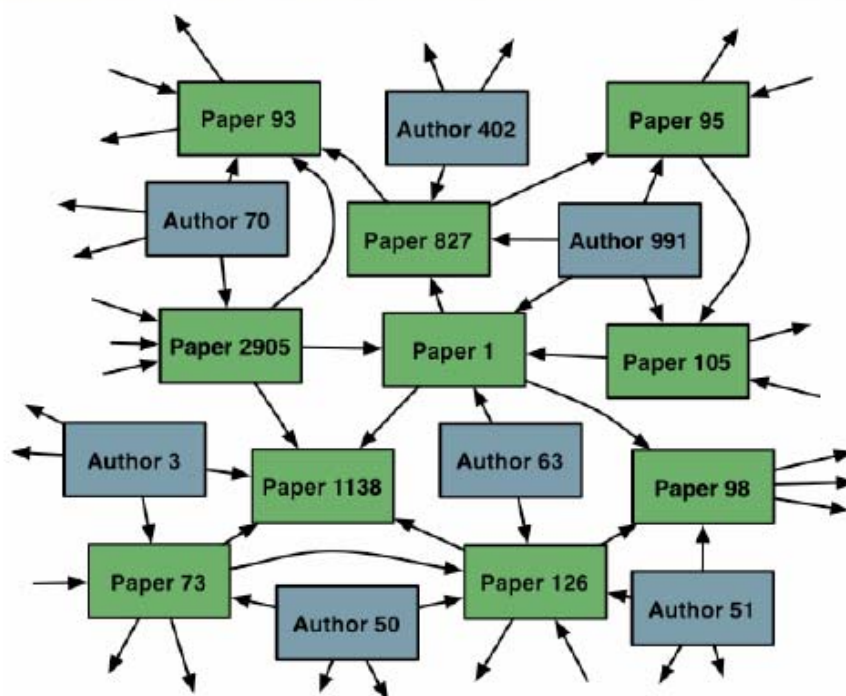


FIGURE 12

We used Gibbs sampling to do this inference process, and we have had very good experience with its convergence and other properties. What we can end up doing is a good job at predicting the value of unknown attributes wherever they reside in the network. Your data set can look essentially like Swiss cheese at some level on which you are trying to make inferences, and you can fill in those gaps over a wide variety of variables. Some of them are known and some of them are unknown.

We do this process we call collective inference, in contrast to conventional models (whether they are graphical models or something else) where you can just look in an individual instance and say we're going to make a prediction about this one, and now we are going to move onto the next one and make a prediction about this. Because of the interconnections among instances, you need to make all these inferences simultaneously. Doing this sort of inference has been shown to improve accuracy substantially in very specific settings. For instance, Chakrabarti et al. (1998) shows that collective inference can really improve your ability to determine the topics of Web pages, but now it's also been shown more generally. The influence of highly confident inferences can travel substantial distances in the graph. Collective inference exploits a clever factoring of the space of dependencies to reduce variance, thus improving performance over considering all relational attributes, as shown in (Jensen, Neville, and Gallagher, 2004).

Figure 13 shows another RDN that we constructed on a citation network in computer science, showing things that aren't terribly surprising, like authors who are well known tend to author with other people who are well known. That's the average rank of the author. Also, that the topics of papers tend to be related to the topics of other papers written by the same author. Again, this is not surprising, but it's nice that this fell out of the model without any of our input. This is the model that was learned directly from the data, so those sorts of dependencies give us some confidence that it is actually learning something useful, that it matches our intuitions.

Citation networks in Computer Science

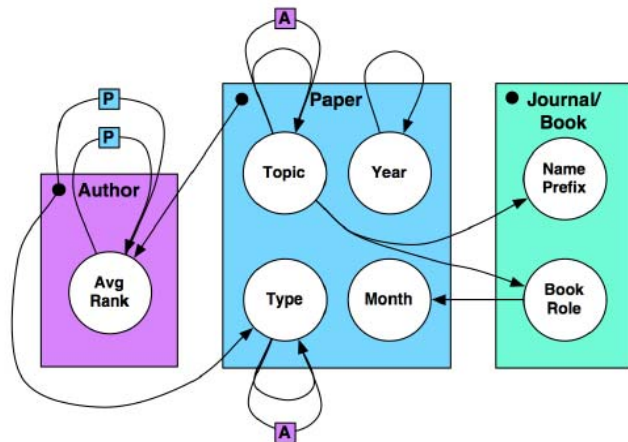


FIGURE 13

Another data set, shown schematically in Figure 14, is the Internet movie database. We

have movies, actors, directors, and studios linked up in the way that you would expect. Here we are learning large numbers of auto-correlation dependencies showing, for instance, that the genre of a movie is likely to be highly correlated with the genre of other movies that are made by the same director, made by the same studio, or starring the same actors. People tend to stay within genre. That's not a surprise, but again, nice to find that coming out of the model.

Internet Movie Database

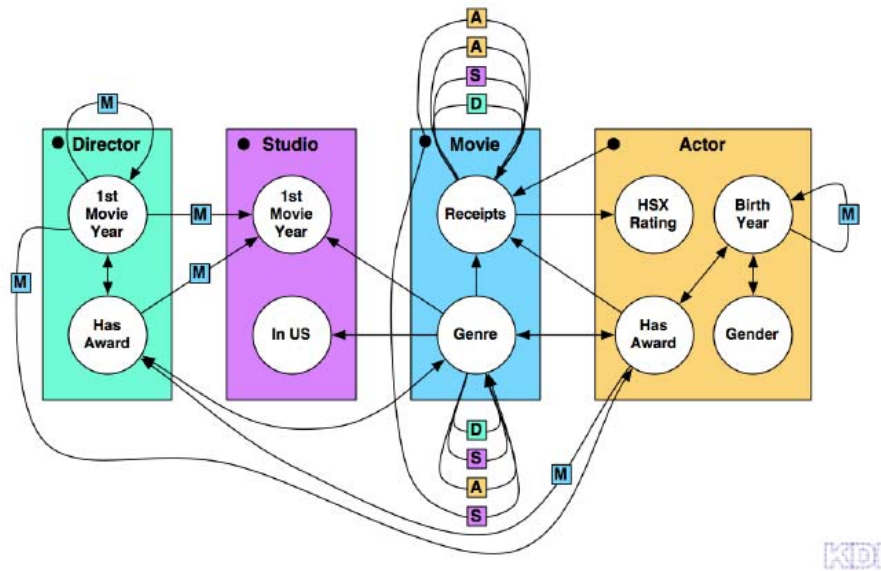


FIGURE 14

Figure 15 shows a basic gene regulatory network. This one is a bit more difficult for me to interpret, but it seems to make sense, and we have done experiments on simulated data to give us some confidence that the models are actually finding what we expect them to find.

Let me mention a couple of other pieces of work before I close. One is that in the course of doing this we needed a query language for graphs. SQL is what we started out using and it didn't work particularly well, so one thing we came up with was a visual query language for graphs. QGRAPH (Blau, Immerman, and Jensen, 2002), which was very useful. QGRAPH allows us to submit queries, written with a simple visual syntax, that return entire subgraphs with heterogeneous structure.

Gene regulatory networks

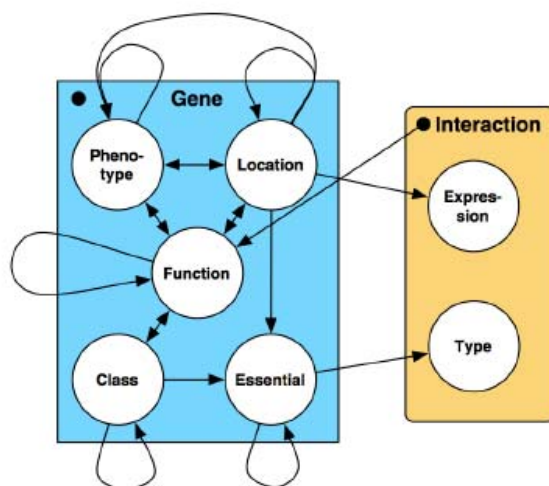


FIGURE 15

For instance, Figure 16 illustrates a query that we used in the NASD data. I'll blow up a portion of it. At the top this says, basically, get me all the brokers associated with all of their disclosures. The zero there says get me all the disclosures for this broker, and get me all of the past branches that they have worked with. The box in the center of Figure 16 indicates the equivalent of the sub-query where you say we want those structures that include the branch, the firm, the regulator of the firm, and all of the past co-workers of that firm. And by the way, we want all of the past branches. So, these get you structures that can be highly variable in size, but which have the same basic properties in terms of the kinds of objects that they contain. We use these queries both as an ad hoc tool for looking at big graphs, and for pulling out small pieces of them that are easy to actually display visually. We also use it as a way of conveying to the algorithm where it should look for possible statistical dependencies.

NASD Query

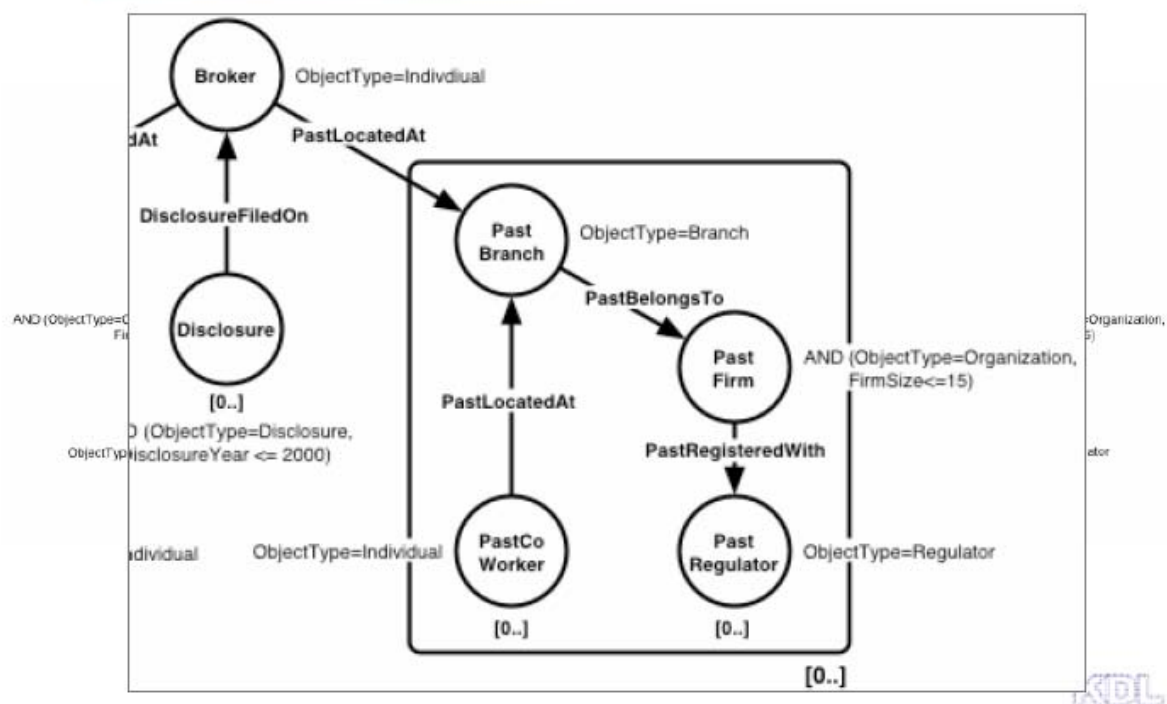


FIGURE 16

Also, just a quick ad for a paper that is on a very different topic. Jon Kleinberg set this up very nicely by discussing this problem of decentralized search in networks. We are able to construct paths through a network with an algorithm called “expected value navigation,” which we recently came up with and published at a conference in the summer of 2005. It does a remarkably good job of combining the information available for searching many of these networks, which is the degree of neighbors, as well as the attributes of neighbors, and using that to navigate in large networks.

We have open source software (PROXIMITY) available to build these relational dependency networks, and that implements the query language and other things. It is available from our Website. We released version 4.0 in April, and we have had more than 600 downloads, which is a nice thing. We are pretty sure that those aren’t just crawlers, but are people really downloading. I want to emphasize that this has an enormous potential for collaboration among all the kinds of people who are here, certainly people in computer science, but I think we need connections to people who work in theory, statistics, social networks, and physics.

Growing research network

- New work in CS areas of machine learning & KD
 - “Relational revolution” (Dietterich 2003)
 - Growing frequency of specialized workshops — AI&LA 1998, SRL 2000, MRDM 2001, SRL 2003, MRDM 2003, SRL 2004
 - Major topic area for technical conferences (ICML, KDD)
- Huge potential for interactions among researchers in ML/KDD, theory, physics, statistics, and social networks
 - Example: Domingos & Richardson 2001 (best paper KDD 2001); D. Kempe, J. Kleinberg, & E. Tardos (best paper KDD 2003)
 - Still a very long way to go...

FIGURE 17

Figure 17 summarizes my thoughts on our growing research network. There is the potential for some really great collaboration here, and there is all this disparate work that is going on that has crossovers. I have heard a lot of it today and yesterday, and really enjoyed it. It's been really good and I've certainly gotten a lot of input from this. One example I want to note is that of Pedro Domingos and Matt Richardson who put out a paper in 2001 that won the best paper award at the Knowledge Discovery and Data Mining Conference. Two years later Jon Kleinberg and co-authors picked up on that line of work, which had to do with viral marketing and some really interesting technical issues. That sort of work between two areas of computer science was nice to see. Now I hope we can branch out to other areas of science. Currently we model attributes, but we would like to move to being able to say we think links will be formed between these two objects, or we think there is a missing object or group in the data.

Figure 18 lists some of my ideas for potential future research. Sampling issues that have already been mentioned are active learning and knowing how to sample while you are building models. The problem of what I would call non-link relations are actually links that aren't instantiated, but rather are the kinds of things that we say, these two people are close in age, or they live next to each other. We don't want to have to instantiate all those links; we would like to be able to represent them in other ways. Spatial and temporal models are a set of other things that are more CS oriented.

Potential future research topics

- Expanding the scope of learned models
 - Objects
 - Links
 - Groups
- Accounting for realistic data sampling issues
 - Fragmentary sampling
 - Adversarial conduct
- Active learning
- Representing and using “non-link relations”
- Unifying with temporal & spatial models
- Unifying statistical modeling and more traditional system simulation
- Persistent knowledge bases
- Databases and query languages for “semantic graphs”

FIGURE 18

Finally, if you’d like further information, please contact me at jensen@cs.umass.edu or kdl.cs.umass.edu.

QUESTIONS AND ANSWERS

DR. BLEI: Hi, I’m Dave Blei from Carnegie Mellon. I didn’t quite see what the criterion was for determining dependence between two random variables. Maybe you could step us through from, say, actor to genre, how one decides to put an arc there.

DR. JENSEN: Structural learning in RDNs is merely learning conditional models, and looking at the variables that are in those conditional models. So, if I look at a model that is X conditioned on everything else, and the model says, Y_1 and Y_3 are in the model for X , I draw in that dependency.

DR. BLEI: How do we determine the independence of the variables? That’s really the question.

DR. JENSEN: Oh, how do you determine independence? Graph separation.

DR. BLEI: I have some data. It’s a big table, and I ask is row A dependent on row B ? What is the test for doing that?

DR. JENSEN: When you are building the relational probability tree, these are rows you

are talking about, not columns.

DR. BLEI: So, you are building a probability model of a set of random variables?

DR. JENSEN: Let's say variables here are columns.

DR. BLEI: When you have an arc between two random variables does that mean one is dependent on the other?

DR. JENSEN: Yes.

DR. BLEI: So, my question is just how to measure whether or not two variables are dependent on each other, since you are learning that arc.

DR. JENSEN: When we are learning the relational probability trees we are looking at individual variables as possible predictors of our dependent variable. We are scoring that with chi square in this case. These are discrete variables. If we have continuous variables we discretize them. We are using chi squares as our scoring function for individual conditional dependencies. Those are learned in the conditional models. We decide to select a feature as a split in a tree that may be the age of the broker. Let's use a simple one. The age of the broker is greater than 27; that's the root node of our tree. What we are predicting is the broker is a problem, so that will result in an arc in our RDN beginning at broker age, going to broker-is-a-problem.

DR. BANKS: I wasn't quite sure that I followed exactly what the SQL search for graphs would give you. Could you say a few more words about how you would instantiate an inquiry?

DR. JENSEN: Certainly. You are just asking what does the query language do for you.

DR. BANKS: What types of questions would you ask that would not be asked? I could ask who is in my social network that is not in Ed Wegman's social network, or something like that, but that is not what you are going after.

DR. JENSEN: Figure 19 illustrates a very simple query. Say we get red things that are associated with two or more structures of the following kind, a blue thing associated with three or more green things. Now, you could write that as an SQL query if you wanted to. What you would get back are rows, which are a red thing, a blue thing, and a green thing. What we want is to identify the subgraphs on the left and the right center as matches while skipping past subgraphs like the one in the lower right (which does not qualify because it has only two green things on the left end).

Example QGraph query

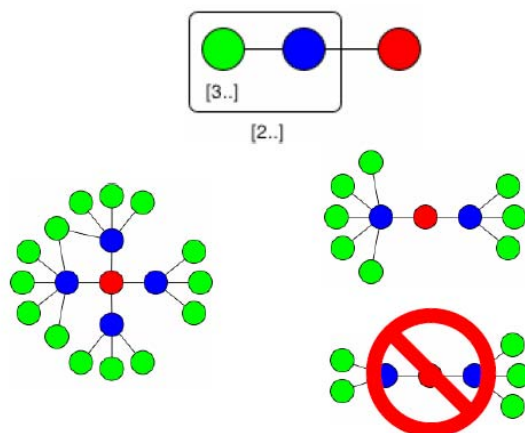


FIGURE 19

DR. GROSS: It's a hard thing to explain.

DR. JENSEN: It is a subtle different from SQL in the sense that you are getting back these subgraphs, and in these grouping constructs have some subtle differences, although what we are now finally doing is actually writing out the algebra, and showing where it differs ever so slightly from SQL.

What is nice about the language as a user is that it is visual. You draw these things that are much more pleasant than trying to write the convoluted SQL that we had to write to get these whole structures out. Also, it is easier to convey to people when you sit down and draw it out, because that looks vaguely like the results you are trying to get. It's certainly a lot better than trying to explain SQL to someone who doesn't know the language.

DR. GROSS: [Remarks off mike.]

DR. JENSEN: We have an implementation which I would definitely call a prototype, which is in PROXIMITY 4.0. It was also in the previous version of the system. We regularly query graphs that are on the order of 1 million to 2 million nodes. Very soon we are going to have the algebra complete, which allows you to easily implement it in whatever database system you choose. We have implemented it on top of both SQL in a previous implementation, and now on top of a vertical database called Monet. You wouldn't learn much from our implementation unless you were using this particularly bizarre database that we use for efficiency reasons. However, the algebra should provide you a really good starting point for an implementation that should be fairly straightforward, and many of the query optimizations that work for SQL will

work for this as well. The algebra is only slightly different.

REFERENCES

- Blau, H., N. Immerman, and D. Jensen. 2002. A Visual Language for Querying and Updating Graphs. University of Massachusetts Amherst Computer Science Technical Report 2002-037.
- Chakrabarti, S., B. Dom, and P. Indyk. 1998. Enhanced Hypertext Classification Using Hyper-Links. Proc. ACM SIGMOD Conference. Pp. 307-318.
- Dietterich, T. 2003. "Relational revolution." New work in CS areas of machine learning and knowledge discovery.
- Domingos, P., and M. Richardson. 2001. Mining the Network Value of Customers. Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining. San Francisco, CA: ACM Press. Pp. 57-66.
- Getoor, L., N. Friedman, D. Koller, and A. Pfeffer. 2001. Learning Probabilistic Relational Models. Relational Data Mining (S. Dzeroski and N. Lavrac, eds.). Berlin, Germany: Springer-Verlag.
- Jensen, D., and J. Neville. 2002. Linkage and Autocorrelation Cause Feature Selection Bias in Relational Learning. Proceedings of the 19th International Conference on Machine Learning.
- Jensen, D., J. Neville, and M. Hay. 2003. Avoiding Bias When Aggregating Relational Data with Degree Disparity. Proceedings of the 20th International Conference on Machine Learning.
- Jensen, D., J. Neville, and B. Gallagher. 2004. Why Collective Inference Improves Relational Classification. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Kempe, D., J. Kleinberg, and E. Tardos. 2003. Maximizing the Spread of Influence Through a Social Network. Proceedings of the 9th Association for Computing Machinery Special Interest Group and Forum for Advancement and Adoption of the Science of Knowledge Discovery and Data Mining International Conference.
- Neville, J., and D. Jensen. 2000. Iterative Classification in Relational Data. L. Getoor and D. Jensen (eds.). Papers of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data. Menlo Park, California: AAAI Press.
- Neville, J., and D. Jensen. 2003. Collective Classification with Relational Dependency Networks. Proceedings of the 2nd Multi- Relational Data Mining Workshop, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Neville, J., and D. Jensen. 2004. Dependency Networks for Relational Data. Proceedings of the 4th IEEE International Conference on Data Mining.

Richardson, M., and P. Domingos. 2004. Markov Logic: A Unifying Framework for Statistical Relational Learning. Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields. Banff, Canada: IMLS. Pp. 49-54.

Simsek, Ö., and D. Jensen. 2005. Decentralized Search in Networks Using Homophily and Degree Disparity. Proceedings of the 19th International Joint Conference on Artificial Intelligence.

Taskar, B., E. Segal, and D. Koller. 2001. Probabilistic Clustering in Relational Data. Seventeenth International Joint Conference on Artificial Intelligence (IJCAI01). Seattle, Washington.

Taskar, B., P. Abbeel, and D. Koller. 2002. Discriminative Probabilistic Models for Relational Data. Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02). Edmonton, Canada.

Appendixes

Appendix A

Workshop Agenda and List of Attendees

AGENDA

Workshop on Statistics on Networks

September 26, 2005
The National Academies
Washington, D.C.

9:00 a.m. – 10:00 a.m.

Keynote Address, Day 1

Network Complexity and Robustness

John Doyle, California Institute of Technology

10:15 a.m. – 12:15 p.m.

Network Models

Neurons, Networks, and Noise: An Introduction

Nancy Kopell, Boston University

Mixing Patterns and Community Structure in Networks

Mark Newman, University of Michigan and Santa Fe Institute

Dimension Selection for Latent Space Models of Social Networks

Peter Hoff, University of Washington

12:15 p.m. – 1:15 p.m.

Lunch

1:15 p.m. – 3:15 p.m.

Dynamic Networks

Embedded Networked Sensing (Redux?)

Deborah Estrin, University of California, Los Angeles

The Functional Organization of Mammalian Cells

Ravi Iyengar, Mount Sinai School of Medicine

Dynamic Network Analysis in Counterterrorism Research

Kathleen Carley, Carnegie Mellon University

3:30 p.m. – 5:30 p.m.

Data and Measurement

Current Developments in a Cortically Controlled Brain-Machine Interface
Nicho Hatsopoulos, University of Chicago

Some Implications of Path-Based Sampling on the Internet
Eric Kolaczyk, Boston University

Network Data and Models
Martina Morris, University of Washington

5:30 p.m. – 6:30 p.m.

Reception and Poster Session

6:30 p.m. – 8:30 p.m.

Dinner and Second Presentation

The State of the Art in Social Network Analysis
After-Dinner Speaker: Steve Borgatti, Boston College

September 27, 2005

9:00 a.m. – 10:00 a.m.

Keynote

Variability, Homeostasis, and Compensation in Rhythmic Motor Networks
Eve Marder, Brandeis University

10:15 a.m. – 12:15 p.m.

Robustness and Fragility

Dynamics and Resilience of Blood Flow in Cortical Microvessels
David Kleinfeld, University of California, San Diego

Robustness and Fragility
Jean Carlson, University of California, Santa Barbara

Stability and Degeneracy of Network Models
Mark Handcock, University of Washington

12:15 p.m. – 1:15 p.m.

Lunch

1:15 p.m. – 3:15 p.m.

Visualization and Scalability

Granger Causality: Basic Theory and Applications to Neuroscience
Mingzhou Ding, University of Florida

Tracking Complex Networks Across Time and Space
Jon Kleinberg, Cornell University

Dependency Networks for Relational Data
David Jensen, University of Massachusetts

3:15 p.m. – 4:00 p.m.

**General Discussion and Wrap-Up,
Adjournment**

LIST OF ATTENDEES

Dimitris Achlioptas, Microsoft Research
Deepak Agarwal, AT&T Research
Mirit I. Aladjem, National Cancer Institute
Reka Albert, Pennsylvania State University
David Alderson, California Institute of Technology
David L. Banks, Duke University
David M. Blei, Carnegie Mellon University
Stephen Borgatti, Boston University
Amy Braverman, Jet Propulsion Laboratory
Emery Brown, Massachusetts General Hospital and Harvard University
John Byers, Boston University
Kathleen Carley, Carnegie Mellon University
Jean Carlson, University of California, Santa Barbara
Alicia Carriquiry, Iowa State University
Aaron Clauset, University of New Mexico
Mark Coates, McGill University
Todd Combs, Air Force Office of Scientific Research
Jose Costa, University of Michigan
Mark Crovella, Boston University
Jim DeLeo, National Institutes of Health (NIH)
Mingzhou Ding, University of Florida
John Doyle, California Institute of Technology
Cheryl L. Eavey, National Science Foundation
Deborah Estrin, University of California, Los Angeles
Stephen E. Fienberg, Carnegie Mellon University
Linton C. Freeman, University of California, Irvine
Anna Gilbert, University of Michigan
Neal Glassman, The National Academies
Rebecca Goulsby, Office of Naval Research
Shula Gross, Baruch College
Mark Handcock, University of Washington
Nicho Hatsopoulos, University of Chicago
Alfred Hero, University of Michigan
Harry S. Hochheiser, NIH-National Institute of Aging
Peter Hoff, University of Washington
David R. Hunter, Pennsylvania State University
Ravi Iyengar, Mount Sinai School of Medicine
David Jensen, University of Massachusetts, Amherst
Jeffrey C. Johnson, East Carolina University
Karen Kafadar, University of Colorado, Denver
Alan Karr, National Institute of Statistical Sciences
Lisa A. Keister, Ohio State University
Jon Kleinberg, Cornell University
David Kleinfeld, University of California, San Diego
Wolfgang Kliemann, Iowa State University
Eric Kolaczyk, Boston University
Nancy Kopell, Boston University
Lun Li, California Institute of Technology

Eve Marder, Brandeis University
Wendy Martinez, Office of Naval Research
George Michailidis, University of Michigan
Milena Mihail, Georgia Institute of Technology
Jim Moody, Ohio State University
Martina Morris, University of Washington
Anna Nagurney, University of Massachusetts/Harvard University
Mark Newman, University of Michigan and Santa Fe Institute
Stephen C. North, AT&T Laboratories-Research
Robert D. Nowak, University of Wisconsin-Madison
Ramani Pilla, Case Western Reserve University
Steven Poulos, National Security Agency
J.T. Rigsby, Naval Surface Warfare Center
Yasmin Said, George Mason University
David Scott, Rice University
Neil Spring, University of Maryland
Monica Strauss, Harvard University
William Szewczyk, National Security Agency
Anna Tsao, Algo Tek
Eric Vance, Duke University
Mark Vangel, Harvard University
Vijay Vazirani, Georgia Institute of Technology
Christopher Volinsky, AT&T Laboratories-Research
Marta Vornbrock, The National Academies
Stanley Wasserman, University of Illinois, Urbana-Champaign
Edward Wegman, George Mason University
Scott Weidman, The National Academies
Barry Wellman, University of Toronto
Chris H. Wiggins, Columbia University
Rebecca Willett, Duke University
Walter Willinger, AT&T Laboratories-Research
Barbara Wright, The National Academies
Eric Xing, Carnegie Mellon University
Bill Yurcik, University of Illinois, Urbana-Champaign
Sandy L. Zabell, Northwestern University
Ellen Witte Zegura, Georgia Institute of Technology

Appendix B

Biographical Sketches of Workshop Speakers

Stephen P. Borgatti, Boston College, received his Ph.D. in mathematical social science from the University of California, Irvine, in 1989. His research interests are in shared cognition and social networks. He is the author of ANTHROPAC, a software package for cultural domain analysis, and UCINET, a software package for social network analysis. He is a past director of the National Science Foundation's Summer Institute for Research Methods, as well as a past president of the International Network for Social Network Analysis, the professional association for social network researchers. He currently serves as associate editor for a number of journals, including *Field Methods* and *Computational and Mathematical Organizational Theory*, and is senior editor for *Organization Science*. He is currently professor and chair of the Organization Studies Department at the Carroll School of Management at Boston College.

Kathleen M. Carley, Carnegie Mellon University, received her Ph.D. from Harvard University in mathematical sociology. She is currently a professor of computer science at Carnegie Mellon University. She also directs the center for Computational Analysis of Social and Organizational Systems (CASOS). CASOS is a university-wide center for understanding complex systems through the combined application of computer science, social science, and social networks. Her research combines cognitive science, social networks, and computer science to address complex social and organizational problems. Her specific research areas are computational social and organization theory; group, organizational, and social adaptation and evolution; social and dynamic network analysis; computational text analysis; and the impact of telecommunication technologies and policy on communication, information diffusion, and disease contagion and response within and among groups, particularly in disaster or crisis situations. Her models meld multiagent technology with network dynamics and empirical data.

Jean M. Carlson, University of California, Santa Barbara, received a B.S.E. in electrical engineering and computer science from Princeton University in 1984, an M.S.E. in applied and engineering physics from Cornell University in 1987, and a Ph.D. in theoretical condensed matter physics from Cornell in 1988. After postdoctoral work at the Institute for Theoretical Physics at the University of California, Santa Barbara (UCSB), she joined the faculty at UCSB in 1990, where she is currently a professor of physics. She is a recipient of fellowship awards from the Sloan Foundation, the David and Lucile Packard Foundation, and the McDonnell Foundation. Dr. Carlson's research interests include a combination of foundational work and a variety of practical applications of complex-systems theory, including earthquakes, wildfires, and optimization and design in networks.

Mingzhou Ding, University of Florida, received his B.S. in astrophysics from Peking University in 1982 and his Ph.D. in physics from the University of Maryland in 1990. He is currently a professor in the Department of Biomedical Engineering at the University of Florida. His main research interest includes cognitive neuroscience and related signal processing problems.

John Doyle, California Institute of Technology, is the John G. Braun Professor of Control and Dynamical Systems, Electrical Engineer, and BioEngineering at Caltech. He has B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology (1977) and a

Ph.D. in mathematics from the University of California, Berkeley (1984). His early work was in the mathematics of robust control, LQG robustness, (structured) singular value analysis, H-infinity, and there have been many recent extensions. He coauthored several books and software toolboxes currently used at over 1,000 sites worldwide, the main control analysis tool for high-performance commercial and military aerospace systems, as well as many other industrial systems. Early examples of industrial applications of his work include various airplanes—X-29, F-16XL, F-15 SMTP, B-1, B-2, 757; Shuttle Orbiter; electric power generation; distillation; catalytic reactors; backhoe slope-finishing; active suspension; and CD players. His current research interests are in theoretical foundations for complex networks in engineering and biology, as well as multiscale physics. His group led the development of the open source Systems Biology Markup Language (SBML) and the Systems Biology Workbench (SBW), which have become the central software infrastructures for systems biology (www.sbml.org), and also released the analysis toolbox SOSTOOLS (www.cds.caltech.edu/sostools). He was the theoretical lead on the team that developed the FAST protocol and shattered multiple world land speed records (netlab.caltech.edu). His prize papers include the Institute of Electrical and Electronics Engineers (IEEE) Baker Award (for the top research paper in all of the IEEE's approximately 90 journals, also ranked in the top 10 "most important" papers worldwide in pure and applied mathematics from 1981-1993), the IEEE Automatic Control Transactions Axelby Award (twice), and the American Automatic Control Council (AACC) Schuck Award. Individual awards that he has received include the IEEE Control Systems Field Award (2004) and the Centennial Outstanding Young Engineer (1984). He has held national and world records and championships in various sports.

Deborah Estrin, University of California at Los Angeles, holds a Ph.D. in Electrical Engineering and Computer Science (EECS) from MIT (1985) and a B.S. in EECS from the University of California at Berkeley (1980) and is a professor of computer science at the University of California at Los Angeles, where she holds the Jon Postel Chair in Computer Networks and is the founding director of the National Science Foundation Science and Technology Center for Embedded Networked Sensing (CENS). Dr. Estrin has been instrumental in defining the research agenda for wireless sensor networks. Her research focuses on technical challenges posed by autonomous, distributed, physically coupled systems. She is particularly interested in environmental monitoring applications and is on the National Ecological Observatory Network (NEON) design team. Earlier in her career she contributed to the design of Internet routing protocols. Dr. Estrin is a member of the NSF Computer and Information Sciences and Engineering (CISE) Advisory Committee and of the National Research Council's (NRC's) Computer Science and Technology Board.

Mark S. Handcock, University of Washington, is a professor of statistics and sociology, Department of Statistics, University of Washington, Seattle. His work focuses on the development of statistical models for the analysis of social network data, spatial processes, and demography. He received his B.Sc. from the University of Western Australia and his Ph.D. from the University of Chicago. Descriptions of his work are available at <http://www.stat.washington.edu/handcock>.

Nicho Hatsopoulos, University of Chicago, received his B.A. in physics in 1984 from Williams College. He received a master's (Sc.M.) in psychology in 1991 and a Ph.D. in cognitive science in 1992, both from Brown University. He was a postdoctoral fellow at the California Institute of Technology from 1992 to 1995 and then again at Brown University from 1995 to 1998. From 1998 to 2001, he was an assistant professor of research in the Department of Neuroscience at Brown University. From 2002 to the present, he has been an assistant professor in the

Department of Organismal Biology and Anatomy and on the Committees of Computational Neuroscience and Neurobiology at the University of Chicago.

Peter Hoff, University of Washington, is an assistant professor in the Departments of Statistics and Biostatistics, and a member of the Center for Statistics and the Social Sciences at the University of Washington in Seattle.

Ravi Iyengar, Mount Sinai School of Medicine, is the Rosenstiel Professor and Chair of the Department of Pharmacology and Biological Chemistry at Mount Sinai School of Medicine, New York, N.Y. Trained as a biochemist, Dr. Iyengar has used biochemical and molecular biological approaches to study cell signaling, with a focus on heterotrimeric G protein pathways. Over the past decade, the Iyengar laboratory has also used computational approaches to understand the regulatory capabilities of cellular signaling networks. The laboratory's most recent studies using graph theory approaches were published in *Science* in August 2005.

David Jensen, University of Massachusetts, Amherst, is an associate professor of computer science and director of the Knowledge Discover Laboratory at the University of Massachusetts. He received a doctor of science in engineering at Washington University in 1991. His research interests are knowledge discovery in relational data, social network analysis, and evaluation, social impacts, and government applications of knowledge discover systems.

Jon Kleinberg, Cornell University, received his Ph.D. in computer science from MIT in 1996; he subsequently spent a year as a visiting scientist at the IBM Almaden Research Center and is now a professor in the Department of Computer Science at Cornell University. His research interests are centered on issues at the interface of networks and information, with an emphasis on the social and information networks that underpin the Web and other online media. He is the recipient of an NSF Career Award, an Office of Naval Research Young Investigator Award, an Alfred P. Sloan Foundation Fellowship, a David and Lucile Packard Foundation Fellowship, teaching awards from the Cornell Engineering College and Computer Science Department, and the 2001 National Academy of Sciences Award for Initiatives in Research.

David Kleinfeld, University of California, San Diego, who now lives in La Jolla, is part of a generation of scientists who trained in physics in the 1980s and 1990s and now devote themselves to problems in the neurosciences. Dr. Kleinfeld focuses on feedback control in somatosensation, using the rat vibrissa sensorimotor system as a model, and on blood flow dynamics in vascular loops, using rodent neocortex as a model system. Aspects of the later work involve the use of nonlinear optics as a tool to measure and perturb flow. Dr. Kleinfeld takes particular pride in the advanced education of graduate and postdoctoral students through his involvement in summer programs on computational modeling, data analysis, and imaging held at Woods Hole and at Cold Spring Harbor.

Eric Kolaczyk, Boston University, is associate professor of statistics and director of the Program in Statistics in Boston University's Department of Mathematics and Statistics and a member of the Center for Information and Systems Engineering (CISE) at the university. His research focuses on the statistical modeling and analysis of various types of temporal, spatial, and network data, with a particular emphasis on the use of sparseness in inference. His work has resulted in new methods for signal and image denoising, tomographic image reconstruction, disease mapping, high-level image analysis in land cover classification, and analysis of computer network measurements. Professor Kolaczyk's publications have appeared in the literatures on statistical theory and methods, engineering, astronomy, geography, and computer science. His

work has been supported by various grants from the National Science Foundation and the Office of Naval Research.

Nancy Kopell, Boston University, has a Ph.D. in mathematics and has been working in neuroscience for about 20 years. Her mathematical focus is dynamical systems, especially geometrical theory of systems with multiple time scales. Scientifically, she has worked on a range of questions including pattern formation in chemical systems and central pattern generators for locomotion. She is currently focusing on how the nervous system makes use of its dynamics, especially its rhythmic dynamics, to help with sensory processing, cognition, and motor preparation.

Eve Marder, Brandeis University, is the Victor and Gwendolyn Beinfield Professor of Neuroscience in the Biology Department and Volen Center for Complex Systems at Brandeis University. She received her Ph.D. in 1974 from the University of California, San Diego, and subsequently conducted a 1-year postdoctoral at the University of Oregon and then a 3-year postdoctoral at the Ecole Normale Supérieure in Paris, France. She became an assistant professor in the Biology Department at Brandeis University in 1978 and was promoted to professor in 1990. During her time at Brandeis University, Professor Marder has been instrumental in the establishment of both undergraduate and graduate programs in neuroscience.

Professor Marder has served on the editorial board of the *Journal of Neurophysiology* since 1989. For almost 6 years she was a reviewing editor for the *Journal of Neuroscience*. Additionally, she now sits on the editorial boards of *Physiological Reviews*, *Journal of Neurobiology*, *Journal of Comparative Neurology*, *Current Biology*, *Current Opinion in Neurobiology*, *Journal of Experimental Biology*, and *Journal of Comparative Physiology*. She has served on numerous study sections and review panels for the National Institutes of Health, NSF, and other funding agencies. She also has served on the Council for the Society for Neuroscience, Council of the Biophysical Society, and several American Phytopathological Society (APS) committees.

Professor Marder is a fellow of the American Association for the Advancement of Science, a fellow of the American Academy of Arts and Sciences, a trustee of the Grass Foundation, and a member of the National Academy of Sciences. She was the Forbes Lecturer at the Marine Biological Laboratory (MBL) in 2000 and the Einer Hille Lecturer at the University of Washington in 2002.

She has studied the dynamics of small neuronal networks using the crustacean stomatogastric nervous system. Her work was instrumental in demonstrating that neuronal circuits are not “hard-wired” but can be reconfigured by neuromodulatory neurons and substances to produce a variety of outputs. Together with Larry Abbott, her laboratory pioneered the “dynamic clamp.” She was one of the first experimentalists to forge long-standing collaborations with theorists and has for almost 15 years of combined experimental work with insights from modeling and theoretical studies. Her work today focuses on understanding how stability in networks arises despite ongoing channel and receptor turnover and modulation, both in developing and adult animals.

Martina Morris, University of Washington, received a B.A. in sociology from Reed College in 1980, an M.A. in statistics from the University of Chicago in 1986, and a Ph.D. in sociology from the University of Chicago in 1989. She is the director of the Center for Studies in Demography and Ecology and holds the Blumstein-Jordan Chair in the Department of Sociology at the University of Washington. Her research is interdisciplinary, intersecting with demography, economics, epidemiology and public health, and statistics.

Mark Newman, University of Michigan, received his Ph.D. in theoretical physics from the University of Oxford in 1991 and worked at Cornell University and the Santa Fe Institute before moving to the University of Michigan in 2002. He is currently associate professor of physics and complex systems at the University of Michigan and a member of the external faculty of the Santa Fe Institute. He has research interest in network statistics and modeling, epidemiology, computer algorithms, and cartography.

