CHAPTER

Basics of Social Networks Analysis 12

INFORMATION IN THIS CHAPTER

- Social network analysis
- Gephi
- Components
- Analysis
- Application of SNA

INTRODUCTION

In one of the recent chapters we have discussed about the importance of data management and analysis. We also learned about certain tools which could be useful in the process. In this chapter we will deal with an associated topic, which is social network analysis (SNA). SNA is widely used in Information Science to learn various concepts. This is wide topic and has applications in many fields and in this chapter we will attempt to cover the important aspects of the topic and the tools required for it so that the readers can further utilize it depending according to personal needs.

As the name suggests, social network analysis is basically the analysis of social networks. The social network we are talking about is a structure which consists of different social elements and the relationship between them. It contains nodes which represent the entities and edges representing relationships. What this means is that, using SNA we can measure and map the relationships between various entities, these entities usually being people, computers, a collection of them, or other associated terms. SNA utilizes visual representations of the network for the purpose of better understanding it and implements mathematical theories to derive results. There are various tools that can be used to perform SNA and we will deal with them as and when required.

Let's deal with some basic concepts.

NODES

Nodes are used to represent entities. Entities are an essential part of social network as the whole analysis revolves around them. They are mostly depicted with a round shape.

Hacking Web Intelligence. http://dx.doi.org/10.1016/B978-0-12-801867-5.00012-4
Copyright © 2015 Elsevier Inc. All rights reserved.

EDGES

Edges are used to represent the relationships. Relationships are required to establish how one node connects to another. This relationship is very significant as it helps to perform various analyses such as how information will flow across the network etc. The number of edges connected to a node defines its degree. If a node has three links to other entities, it has degree 3.

NETWORK

The network is visually represented and contains nodes and edges. Different parameters of nodes and edges such as size, color etc., may vary depending upon the analysis that needs to be performed.

Networks can be directed or undirected, which means that the edges might be represented as simple lines or as directed arrows. This primarily depends upon the relationships between the edges. For example, in a network of mutual connection such as friends, can have an undirected network but for a network of relations such as who likes whom can have directed network.

Now we have a basic idea of SNA. Let's get familiar with one of the most utilized tools for it.

GEPHI

Gephi is a simple yet efficient tool used for the purpose of SNA. The tool can be downloaded from http://gephi.github.io/ and the installation process is pretty straightforward. Once installed, the tool is ready to be used. The interface is simple and is divided into different sections. There are three tabs present at the top left corner which allow working with the network in different manner. These three tabs are Overview, Data Laboratory, and Preview.

OVERVIEW

The Overview tab provides the basic information about the network and displays the network visualization. It is primarily divided into three sections which further have subsections. The left-hand side panel consists of sections which allow partitioning and ranking of nodes and edges, performing different layouts for the network based on different algorithms. The middle section consists of the space where the network is visualized and the tools to work with the visualization. The right-hand sections contain information about the network such as number of nodes and edges and operations such as calculating the degree, density, and other network statistics.

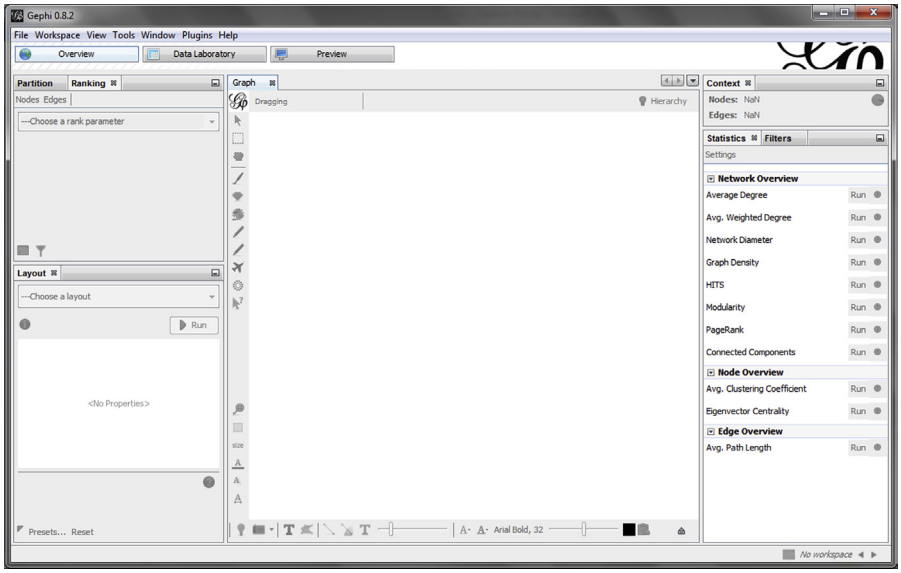

FIGURE 12.1

Gephi Overview.

DATA LABORATORY

Under the Data Laboratory tab we can play with the data in its raw form. In this tab, the entities and their relationships are displayed in the form of a spreadsheet. Here we can add new nodes and edges, search for existing ones, import and export data, and much more. We can also work on columns and delete them, copy them, duplicate them etc. The data present can also be sorted depending upon different parameters by simply clicking on the row names.

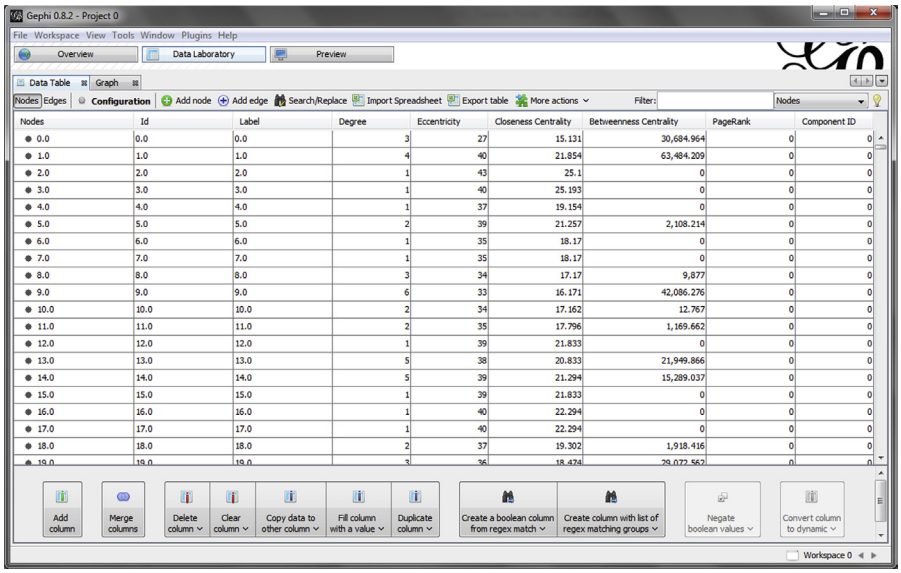

FIGURE 12.2

Gephi Data Laboratory.

PREVIEW

In Preview tab we can change various settings related to the properties of the network graph such as the thickness of the edges, color of the nodes, border width etc. This helps us to set different values for different parameters so that we can make recognizable distinction based on different properties of the graph. The settings can be made in the left-hand panel and the changes are reflected in the rest of the section available for preview.

There are many other tools available for SNA, some of which are SocNetV (http://socnetv.sourceforge.net/), NodeXL (http://nodexl.codeplex.com/), EgoNet (http://sourceforge.net/projects/egonet/) etc.

The term network is quite the same as we use in computers or any other aspect such as math or physics. The terminologies to get the proper definition might change in different areas of study but the bottom line is network which is the connection of different entities with relationships. As we discussed a bit about the network earlier, now it's time to dig a bit on the same. To make a simpler approach to network we will use term "NODE" for entities and term "EDGE" for the relationship.

To create a meaningful and easy to understand network or graphical representation of a network, we must need to focus on certain areas such as highlight the widely used and important nodes and edges, remove nodes with no data or edges, remove redundant data, group similar nodes based on geographical location, community, or anything that broadly relates those. These are the basic practices or points to be remembered while creating a meaningful and easy to understand network.

The components of a network such as the edge and the node have certain attributes based on that we can create a network. Those attributes play a vital role in understanding a network and its components better. Let's start with a node.

As discussed earlier, node has a property called degree. Degree can be used for calculating the likelihood of that node. It is nothing but the number of edges that are connected to the node. Though it also matters is that whether the edges are directed or undirected. Let's say the number of directed edges toward a node X is 5 and the directed edges away from X is 2. Then the degree of X is 7 because it's the combination of in-degree (5) + out-degree (2).

NODE ATTRIBUTES

Every node in a network can have a range of attributes that can be used to distinguish some properties of a node.

The attribute can be in binary form to explain in simple true/false, yes/no, online/offline, or married/unmarried. This is one of the easy representations of an node attribute where we have only choose one out of the two choices.

The attributes can be set categorical based on if options available are more than two such as if we want to set an attribute to a node called as relationship then we can use different category as an option to it, e.g., 1. Friend, 2. Family, 3. Colleague.

The attribute can also be set as continuous such as based on some of the information that cannot be same for every node. For example, date of birth, job position etc. We can also use the same as attributes of a node to distinguish as node quite easily.

EDGE ATTRIBUTES

DIRECTION

Based on direction, two major types of edges can be found.

1. Directed edges
2. Undirected edges

Directed edges

Directed edges are the edges with unidirectional relationship. The best example of a directed edge is X → Y. Here X is unidirectional related with Y. We can say that Y is a child of X or X loves Y or any such one-sided relationships.

Undirected edges

It can be used for establishing mutual relationships, such as X ← → Y or X — Y. The relationship can be anything like X and Y are friends or classmates or colleagues.

TYPE

It can be the type of relationship that put an edge in a group. Let's say that there are different nodes and edges but if some of the edges are similar by the type let say a group then we can distinguish them quite easily. Type can be anything such as starting from friends, close friends, colleague, relative etc. And it has a significant role in differentiating different edges.

WEIGHT

It can be the number of connection that two nodes can have. For example, if X←→Y shares more than one mutual/undirected or directed edge to each other then the weight of that edge is that number. For example, X relates with Y in five ways then the weight of that edge is 5. We can simply draw five edges between those two nodes or we can draw a deeper edge between them to make it easy to understand that these two nodes contain a higher weighted edge.

Weight can be also of two types:

1. Positive
2. Negative

Positive weight

It's based on the likelihood of a relationship. For easy understanding let's talk about a politician. There are many people who like a particular politician. So the relationship they establish with the same will be the positive weight.

Negative weight

Similarly as we came across, the negativity or the hate or the unlikelihood can be also a factor in a relationship. That can be measured by the negative weight.

RANKING

Based on the priorities of the relationship established between two nodes, edges can have different rankings. Such as X's favorite subject is Math, X's second favorite subject is Physics. So to differentiate between these priorities ranking comes in to existence for easy understanding in a network.

BETWEENNESS

There are certain scenarios where we can see that there are two different group of nodes connected to each other by an edge. So those kind of edges perceive a unique quality to combine two different groups or set of nodes and that can be called as betweenness. There are many other attributes that can be found situational. For the time being we can say that we have basic knowledge of network, its components, and its attributes so that if in future we get a chance to create a network or understand a given network then we can understand at least the basics of it properly.

The core basics of the network and about its components are covered above but still we haven't covered the main topic that is SNA.

As discussed earlier in the chapter, SNA is about mapping and measuring of relationships between different entities. These entities can be people, groups, organizations, systems, applications, and other connected entities. The nodes in the network are usually people but it can be anything based on what network we are looking at, while the links represent relationships or flows between the nodes. SNA provides both mathematical as well as graphical analysis of relationships using which an analyzer can deduce number of conclusions such as who is a hub in the network, how different entities connected to each other, and why they connected to each other with a proper logical and data-driven answer. The factors that come into act such as degree, betweenness, and others are already covered.

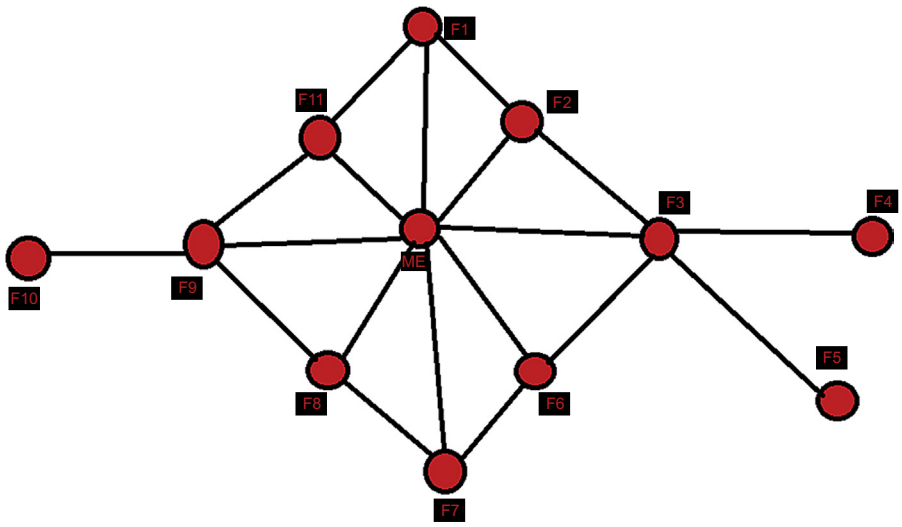

FIGURE 12.3

A small sample network to understand different components.

This is a sample network of friends where the edges define the level of communication shared between each one of them. It is a simple network to understand different aspects of SNA about how to find different important information by just looking at a network.

The first thing we can find quite easily is that the active node in the network. An active node or a hub is that node that has highest degree of edges. In this case as it is quite clear that node "me" has the highest degree of edges, and the degree is 8, we can conclude that node "me" is the most active node and connects almost all the other nodes. In any personal network generally it is said that the greater the number of friends the better the network, but it is not always true. Another thing we can conclude that node "me" connects to only those who have a common friend at least. These are some of the assumptions that can be derived from the network.

There are two nodes that play a vital role in the network. Node "F3" and "F9." These two nodes connect some nodes to the "me" node cluster. These "F3" and "F9" are the single points of contact in terms of connecting "F4"and "F5," and "F10" respectively. These two nodes decide what information to send in to the cluster and what information to send out according to the location that makes these two nodes extremely powerful in this network. We came across a term betweenness. Now it's the time to use it. The node with high betweenness has the greater influence on the data flow through that. So in our case as the node "F3" has the highest betweenness, and that is 2, it will influence the most and can be considered as the most powerful node in this network. And in other words it can be the one point of failure for rest of the nodes that is not directly connected in the cluster. In this way we can conclude that location plays a vital role in a network. It is the location that can make a node important as we have just seen in this case. These nodes can also be called as boundary spanners. As these nodes having ideas and information from both part of the network such as the cluster and the extended part, these nodes can be innovators that can think of new ideas and services by combining ideas from both part of the network.

If we look at the network centralities, then we can understand the network structure quite easily. It allows us to understand individual locations and its importance. If a network is very much centralized and has a single point of failure then the network can be easily fragmented by deactivating that node. So it's always better not to have a centralized network. In our case, it is a less centralized network. Though we have a hub and two nodes with betweenness, it's good to have a network like this because the deactivation of the hub will not affect the network directly because there are still paths to pass on information from one node to another. Though the failure of node "F3" and "F9" will create a sub network, the major part of this network or the "me" node cluster will still be unaffected.

Network reach is also a topic to discuss here. Network reach is nothing but using the shortest path that is generally one hop or two hop difference whether a node is able to communicate with any other node or not. This can be easily understood in live example of one of the popular social networking sites that is LinkedIn. It uses the same concept to look for the network reach. What it does, if we want to connect with a particular person then it will show the number of first, second,

and third degree connections so that we can use any of them to get introduced. As in LinkedIn it is always better to use the first degree connection to get introduced because it is the shortest path to reach to a connection, so it makes all our direct connections important in LinkedIn. Similarly in case of this network, it's always good to have a hub as a neighbor node. Because it's the neighbor node that plays a vital role in communication. If your neighbor node is a hub and connected to everyone then in a way we are also connected to everyone by him and it expands our network reach. Here in this network apart from node "F4," "F5," and "F10" rest of the nodes that come under "me" node cluster has very good network reach because of "me" node.

To get information from different sources differently, we have to be in a position where we have alternative shortest paths for a single node as this will allow us to get same information with different perspective in a network. This basically depends on the network integration but in our case it is very poor. Whatever information will flow, it must flow through the hub "me" node. So it will be very difficult to get different perspective on the information. If the nodes F1, F2, F3, F6, F7, F8, F9, F11 would have connected to each other like mesh topology then most of the nodes would have different alternative shortest paths and that opens an option to get same information flows from different paths adding different flavors to it.

Most of the time we don't value the extended network such as in our case node F4, F5, and F10. These nodes are not the part of the cluster and the same makes them very important. They are the one who will get very low information from the network cluster but for the network point of view they are the point of fresh ideas. They can provide outside valuable information into the cluster. For this network they might be foreign but for some network they must have been local and that information can be passed on to this "me" cluster using them. These kinds of nodes also called as peripheral nodes.

A network is never a result or report. It's more like a mirror which reflects a number of things about a network not necessarily good always. So try to understand the key components, key players of a network to understand the behavior of a network.

According to the position of a node, a node works in a specific way or vice versa. We can say that according to the role of a node we can find its position in a specific place in a network. So either way a position based on role or role based on location is quite important to understand a network. Now let's look into some of the roles and then try to figure it out how many roles were there in the previous network.

Roles based on location of the node.

Star/Hub

Star is an entity that is highly central. Previously we have used the term hub for the node that contains major number of connections, this is the same. So we can use term start or hub and from the previous example, we can see that we have a star node and that is "me."

Gatekeeper/Boundary spanners

An entity who mediates or we can say controls the flow between one portion of the network with another, earlier we named it different, we used boundary spanner for it. These are just different keywords with same definitions or are the same. In our previous example, "F9" and "F3" are the gatekeepers.

Bridge

It is the only edge which links/belongs to two or more groups. In the previous example, there are three bridges, (1) F9 → F10, (2) F3 → F4, (3) F3 → F5.

Liaison

An entity which has links to two or more groups that would otherwise not be linked, but is not a member of either group. In our previous example, it does not have any such node that was in a position of Liaison.

Isolate

As the name suggests, isolate is an entity which has no links to other entities; generally a linkless or edgeless node. In our previous example, we do not have any isolate nodes.

So there are certain roles that are not present in our example, so here is a new network that contains all the roles and highlighted properly for easy understanding.

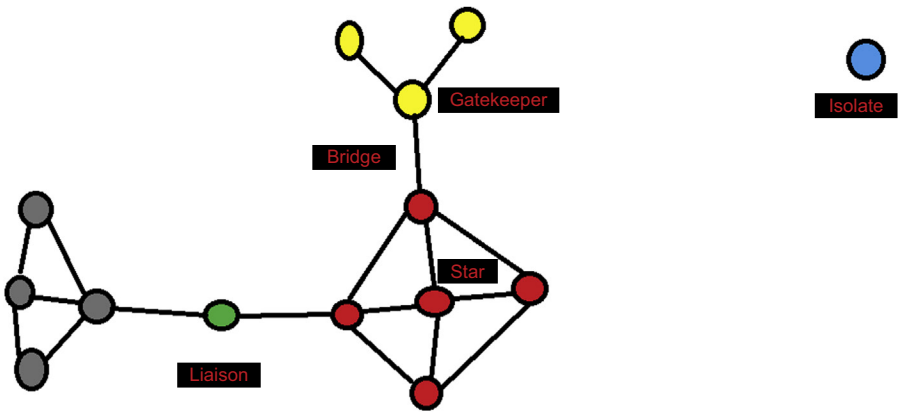

FIGURE 12.4

Network highlighting different roles.

SNA can be helpful in many ways to understand the information flow, we can use the same in variety of situations such as predict exit poll results from the verdict of online users or identify how and to what extent an information will flow in a network of friends, understand an organizational culture, or even find the loopholes in a process.

For a simpler example, to use it in generic scenario we can create a network of Twitter users of a community or organization and see who is following whom and

who is being followed. This would help us to understand who the key players are in that structure and create the most influence. Similarly we can also understand, who is more follower type and who are leaders. In a network of professionals in an organization, it can be used to identify the people who create a hierarchy and how path would be better if one professional needs to connect to another one who is not a direct connection.

Similarly it can be used to analyze a network of connected people to identify how a communicable disease would spread in the network and which links need to be broken before the whole network gets infected. Another example could be in a network of market leaders of an industry to identify who is the hub in that network and needs to be targeted to be influenced for a decision to be taken.

Most of the attributes and functions that we have discussed in the chapter can be automatically calculated using Gephi, it also has many algorithms which can be utilized to perform the layout, identify key elements, implement filters, and perform various other operations. It can also be extended utilizing various plugins option which is present under the tools button.

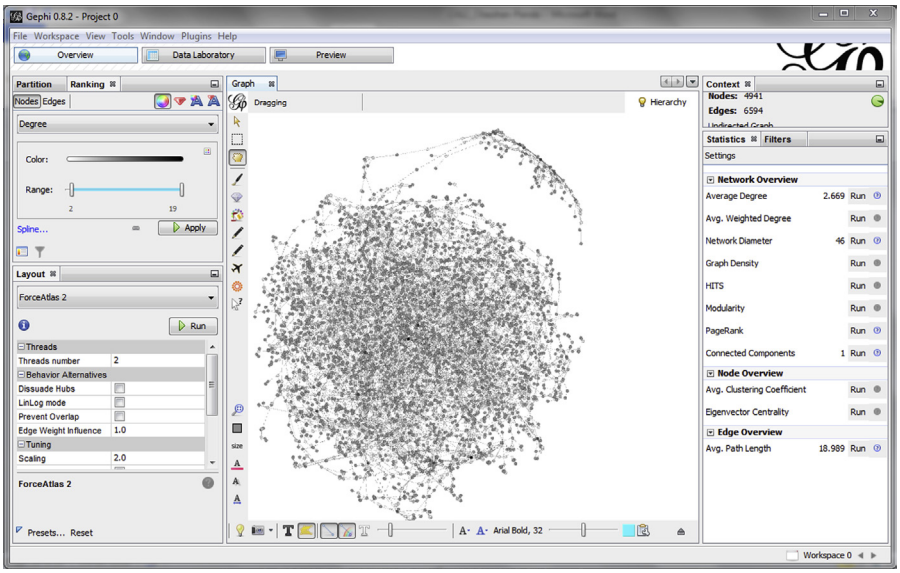

FIGURE 12.5

Sample network in Gephi with different values calculated.

SNA is used by various social network platforms and organizations which deal with connection between people; similarly it has application in many domains which depend on information science to flourish their market.

We learned something new in this chapter and can use the same in future for easy understanding of any complex system by creating a simpler network for that. Here

we have covered the very basics of the concept of SNA. This topic has very wide applications in different fields. Our aim here has been to introduce the topic so that readers can get familiar with it and understand its importance and hence explore it further for practical usage.

Moving on in the next chapter we will be learning about Python basics. Though programming basics will be covered, it is a good idea to brush up basic concepts before moving onto it.