The Dissertation Committee for George Fouad Nabih Shoukry
certifies that this is the approved version of the following dissertation:

Essays on Mechanism Design, Safety, and Crime

Committee:

Jason Abrevaya, Co-Supervisor

Maxwell Stinchcombe, Co-Supervisor

Marika Cabral

Laurent Mathevet

Thomas Wiseman

Essays on Mechanism Design, Safety, and Crime

by

George Fouad Nabih Shoukry, B.S.M.E.; M.S.M.E., M.S.Econ.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2014

To my parents, Fouad and Sonia, and to my brother Michael.

Acknowledgments

I wish to thank my advisors, Jason Abrevaya and Maxwell Stinchcombe, for their time, guidance, patience, and encouragement. My thanks also go to my committee members for their comments, time, and help: Marika Cabral; Laurent Mathevet, whose strong encouragement was fantastic; and Thomas Wiseman. I am indebted to all the professors I learned from and had conversations with, especially David Kendrick for his generous advice and support, as well the many seminar participants who gave me wonderful comments when presenting research in its infancy.

I thank the Economics department for making this research possible, and for the financial support I received over the years. Vivian Goldman-Leffler has been a wonderful knowledgeable source of information and encouragement since before I came to Austin.

I would also like to thank FreightWatch International, and especially Ron Greene and Danny Ramon, for providing the data used in part of this dissertation and for sharing with me their expertise and extensive knowledge about cargo security.

I would like to give a special thanks to all the friends I met in Austin, the Coptic Students of Texas group, and the many loving members of the Holy Cross Coptic Orthodox Church. Their support impacted me greatly and permeated every aspect of my life, including my research and this dissertation.

Finally, I wish to thank my parents and brother for their unconditional, continuous love, and for the many boxes of treats I received while away from home.

Essays on Mechanism Design, Safety, and Crime

George Fouad Nabih Shoukry, Ph.D.
The University of Texas at Austin, 2014

Co-Supervisors: Jason Abrevaya
Maxwell Stinchcombe

This dissertation uses theoretical and empirical tools to answer applied questions of design with an emphasis on issues relating to safety and crime. The first essay incorporates safety in implementation theory and studies when and how safe mechanisms can be designed to obtain socially desirable outcomes. I provide general conditions under which a social choice rule can be implemented using safe mechanisms. The second essay is an empirical study of how criminals respond to changing profitability of crime, a question that informs the policy debate on the most effective crime fighting methods. I find that the price elasticity of theft is about 1 in the short term and increases to about 1.2 over a seven-month horizon, suggesting that policies that directly affect crime profitability, such as policies that shut down black markets or those that reduce demand for illegal goods, can be relatively effective. The third essay shows that any standard implementation problem can be formulated as a question about the existence of a graph that solves a graph coloring problem, establishing a connection between implementation theory and graph theory. More generally, an implementation problem can be viewed as a constraint satisfaction problem, and I propose an algorithm to design simple mechanisms to solve arbitrary implementation problems.

Table of Contents

Chapter 1

Safety in Mechanism Design and Implementation Theory

Chapter Summary

A mechanism is unsafe if a small number of agents deviating unexpectedly can make the mechanism deliver an outcome regarded as bad by a large number of other agents. Under Nash behavior, the direct approach of designing mechanisms with only safe Nash equilibria is impossible in many environments. A weaker approach (double implementation in Nash and safe equilibria) making just some Nash equilibria safe is enough to guarantee safety of players in equilibrium with mild behavioral assumptions. In general, to be safe, a mechanism must restrict the set of outcomes that agents can achieve by deviating, introducing a tradeoff between safety and individual autonomy. Also, requiring a mechanism to have safe equilibria implement F may require that other equilibria deliver outcomes that are socially undesirable.

1.1 Introduction

We introduce a notion of safety as robustness of agents' welfare to unexpected deviations by others. Such robustness is crucial in the current global environment, where malicious intentions, seemingly irrational behavior, mistakes, accidents, or miscommunication can all lead to disasters. Without safety constraints, a mechanism leaves participants exposed to arbitrarily harmful outcomes resulting from unexpected deviations by others – arguably unacceptable, especially as mechanism design extends its reach to more applied

settings. Despite the considerable importance of safety, it has received very little attention in the mechanism design and implementation theory literature.

The canonical mechanism for Nash implementation Maskin (1999) lacks any safety properties. In an example in section 1.2.2, we will see that, in general, at *any* Nash equilibrium (NE) of Maskin's mechanism, a single player can impose the most undesirable outcome on all others by deviating. The lack of safety in Maskin's mechanism applies to many mechanisms in the literature.

Our precise understanding of safety, and exactly what aspects of robustness make a mechanism "safe", have important implications for the analysis and interpretation of this paper's results. We highlight two important aspects.

First, we aim to protect players from deviators, *regardless* of the incentives to deviate. A mechanism can sometimes count on players being self interested for safety of others. For example, many city roads do not have safeguards against drivers driving on the wrong side of the road because doing so would harm those drivers. In such settings our definition of safety may be too strong. However, there are many situations where a designer may wish to assure safety regardless of the incentives associated with deviating: to protect against actions of players intending to harm others without regard to their own utility, irrational behavior, mistakes, accidents, miscommunication; or in situations where deviations can be very costly as in high speed freeways or flying a passenger jet where one person can hurt many more. In any of those cases we insist on more regulations because incentives against deviating cannot be counted on to provide safety, making it necessary to adopt an idea of safety that focuses instead on the potential impact of deviations. This is precisely what our notion of safety allows us to do.

Second, we consider safety of players from *unexpected* deviations. We let players be entirely unprepared for deviations from equilibrium by others; players may wish to behave

differently if they suspected the existence of deviators, as is plausible in many realistic environments. This is in contrast to Eliaz (2002), where all players know that deviations from equilibrium may occur, and all have accurate and identical beliefs about the maximum number of players who may deviate. Players may be prepared for deviations in some problems, and it is interesting to consider robust mechanisms that work well in such settings, as in Eliaz (2002); however, our approach applies to situations where players are unprepared for deviations. For example, a kidney exchange designer may wish to make agents safe from unexpected behavior by others (e.g. intentionally transmitting deadly viruses) without relying on agents' expectations of such behavior. We also note that, in reality, mechanisms are often stopped and operations are halted if it is suspected that malicious agents who may harm others are present. Additionally, assuming best-responding players is cognitively demanding in the presence of deviators who may deviate in arbitrary ways. Our approach places the bulk of the safety burden on the designer, who aims to protect unprepared players from unexpected, harmful actions by others.

Combining these aspects results in a robust, intuitive notion of safety that can be applied to examine the safety properties in virtually any setting. We define a $(\bar{d}, \bar{v})$-safe message profile to be one at which a deviation by any $\bar{d}$ players (the *deviators*) can result in at most $\bar{v}$ victims among the non-deviators. We apply our safety concept to the standard implementation problem by defining a *victim* to be a player who attains an outcome within that player's least favorite K outcomes, where K is set by the designer.

One implication of the deviations being unexpected is that we do not count on the players to protect themselves in any way – they do not expect deviations from equilibrium. So, we must assume that they play a Nash equilibrium (or any other equilibrium notion of choice) without any regard for safety. In this paper we focus on safety in conjunction with the Nash equilibrium as a solution concept.

We show that the most direct approach of requiring all Nash equilibria to be safe is

impossible in many environments. We define strong $(\bar{d}, \bar{v})$-safe implementation to be Nash implementation with the added restriction that all Nash equilibria at any preference profile are $(\bar{d}, \bar{v})$-safe. Strong $(\bar{d}, \bar{v})$-safe implementation is an attractive form of safe implementation: every desirable social outcome is obtainable via a Nash equilibrium of the mechanism, and every Nash equilibrium is $(\bar{d}, \bar{v})$-safe and leads to a desirable social outcome. Behavioral assumptions on the players are weak because simply playing a Nash equilibrium assures safety. We show that, in the most basic environment, strong safe implementation imposes unnatural restrictions on preferences that are unlikely to hold in many cases.

Is all hope lost? No. The assumption that all Nash equilibria are safe is overly restrictive and it is not needed. If only *some* Nash equilibria are safe, and if the designer can incentivize players away from the unsafe Nash equilibria, then safety is guaranteed. This is the motivation behind our use of double implementation in safe and Nash equilibria, which we build up to by first considering a weak notion of safe implementation.

We consider next a weaker form of implementation, called $(\bar{d}, \bar{v})$-safe implementation, requiring the following two properties at any preference profile: all desirable outcomes can be obtained via Nash equilibria that are safe, and all Nash equilibria that are safe lead to desirable outcomes. We call a Nash equilibrium that is also safe a $(\bar{d}, \bar{v})$-safe equilibrium. $(\bar{d}, \bar{v})$-Safe implementation is equivalent to (full) implementation in $(\bar{d}, \bar{v})$-safe equilibria; it has the drawback that unsafe Nash equilibria may exist and may lead to undesirable outcomes. However, safe implementation forms the basis for implementing social choice rules in a safe way, and it is of theoretical importance for several reasons. First, it can inform us about the existence of social choice rules that are Nash, but not safe, implementable, and safe, but not Nash, implementable. This is important for understanding the trade-offs of safety. Second, by satisfying the minimum requirements that are needed for implementing social choice rules in a safe way, it can be used as a step towards more desirable forms of implementation. We give conditions that are necessary for safe implementation, and

sufficient with mild preference restrictions.

Finally, we consider double implementation in both $(\bar{d}, \bar{v})$-safe and Nash equilibria. Double implementation in this case is equivalent to safe implementation with the added requirement that all Nash equilibria (including unsafe ones) lead to desirable outcomes. Thus, it lies between strong safe implementation and safe implementation in restrictiveness: it allows for some unsafe Nash equilibria, but it constrains all of them to lead to desirable outcomes. We allow for simple transfers when considering double implementation. We give conditions that are both necessary and sufficient for double implementation when $n > 2$, $\bar{d} < n/2$, and the designer does not have access to precise cardinal preference information. Besides being of interest in its own right, double implementation guarantees that players will always be safe in equilibrium with very mild behavioral assumptions that are unrelated to safety.

We discuss the interplay between safety and freedom. We distinguish between negative freedom: freedom from harm, and positive freedom: freedom to obtain a different outcome. We measure positive freedom by how many outcomes a player can obtain via a potential deviation at equilibrium. We show that, in general, *every* player will have no more positive freedom in a safe mechanism than in one that disregards safety. We also show that positive freedom becomes more restricted as mechanisms become safer.

We show that, in general, there is a trade-off between safety and efficiency. There are situations where the designer must choose between Nash implementation and safe implementation, and where both are not possible together (double implementation fails). In those problems, if the designer chooses safe implementation, then it *must* come at the cost of Nash equilibria that lead to outcomes outside the social choice rule. To the extent that social choice rules attempt to select efficient outcomes, undesirable outcomes that are not in the social choice rule will be inefficient.

We prove a "no guarantees" result showing that, in a wide range of environments,

absolute safety ($\bar{d} = n-1$ and $\bar{v} = 0$) leads to non-existence of Nash equilibria, making Nash implementation with guarantees impossible. Throughout the paper, we use as an example the design of a hiring mechanism in which candidates have multidimensional quality and we contrast a safe mechanism with one that ignores safety.

The following subsection reviews the related literature, and the rest of the paper is organized as follows. Section 1.2 motivates the need for safety using an example and introduces our notion of safety. Section 1.3 presents our main results and section 1.4 discusses the relationship between safety, freedom, and efficiency, and presents the "no-guarantees" result. Section 1.5 expands on the details of our results and presents the mechanisms used in implementation. Section 1.6 illustrates the benefits of using safe mechanisms in a setting of hiring with multidimensional quality, and discusses safety in the context of Condorcet social choice functions. Section 1.7 discusses our safety concept and notions of implementation in more detail. Section 1.8 discusses further work and concludes.

1.1.1 Related Literature

Conceptually, there are two strands of robust implementation theory closely related to this paper. However, the robust implementation literature almost exclusively focuses on robustness of the outcome of the mechanisms, whereas we focus on robustness of players' welfare: (1) Eliaz (2002) defines robustness as the inability of a deviating minority to affect the outcome of the game used in implementation; in our approach, mechanisms do not have to be robust in this sense; rather, deviators can change the outcome of the mechanism, but we want to avoid them causing "disasters" for others. As previously mentioned, another essential feature distinguishing this paper from Eliaz (2002) is our focus on unexpected deviations. (2) Bergemann and Morris (2005) define robustness as insensitivity of the outcome of the mechanism to informational assumptions such as common knowledge, thus, they focus on a different concept of robustness and on the outcome of the mechanism.

This paper is the first in the implementation literature to consider robustness of players' welfare as opposed to robustness of outcomes. The distinction is important, because different robustness requirements lead to different sets of implementable social choice rules. In many applications (e.g. public good problems), the actual outcome of the mechanism is only of secondary importance to the designer and may only matter due to its indirect effect on the welfare of the players. It is then desirable to design mechanisms that have some robustness properties with respect to the welfare of the players directly.

In the game theory and computer science literature, some equilibrium concepts consider various forms of robustness to deviations as part of the solution concept. Halpern (2008) defines a k-resilient Nash equilibrium to be a Nash equilibrium where coalitions of up to k players do not gain by deviating. This is related to the literature on strong Nash equilibria and coalition proof Nash equilibria. Aumann (1960) defines a strong Nash equilibrium (NE) to be a message profile where there are no multilateral profitable deviations. In Aumann's definition, the deviations need not be resistant to further multilateral deviations. Bernheim et al. (1987) argue that Aumann's definition is too strong. They define a coalition-proof NE to be a message profile where there are no multilateral profitable deviations that are also resistant to further deviations by "subcoalitions". This is weaker than Aumann's definition because the set of deviations to be considered is smaller. These solution concepts focus only on incentives to deviate, whereas we focus on consequences of deviations from equilibrium.

Some equilibrium concepts that do discuss consequences of deviations from equilibrium do so from a *similarity* motivation: the idea that players should behave similarly in similar situations. Such refinements include trembling hand perfect equilibrium (Selten (1975)) and essential equilibrium (Wu and Jiang (1962)). The trembling hand perfect and essential equilibria require the existence of a NE close to the original one if the messages of the players and the payoffs, respectively, are perturbed slightly. The similarity motivation for refinements focuses on the effect of deviations from a NE on the behavior of the players, but

does not explicitly address the impact of deviations by some on utilities of others. Another notion of equilibrium in the literature that attempts to capture some of the effect of players on each other is risk dominance (Harsanyi and Selten (1988)). Risk dominance selects equilibria based on the willingness of rational players to deviate if they take the riskiness of different equilibria into account. It is based on the idea that players best-respond to beliefs about deviations by others, whereas we consider situations in which players are unprepared for deviations from equilibrium.

Game theoretic equilibrium concepts are ultimately concerned with prediction, but our objective is different: it is to consider when and how it is possible to design mechanisms with desirable robustness properties. The differing objectives imply that desirable aspects of robustness, and the motivations for studying them, may also differ, but this paper's focus on robustness to deviations from equilibrium is generally supported by similar concerns in game theory.

Robustness from arbitrary deviations by some players is an idea with a long history in computer science, where deviators are often referred to as "Byzantine" or "faulty." For example, Halpern (2008) discusses a robustness property called t-immunity, where a profile is t-immune if no player who does not deviate is worse off if up to t players deviate in an arbitrary way. This robustness property is very similar to the more general robustness concept we consider in this paper[1], and its use in computer science supports our view that such robustness is important to consider in applications.

[1]Note that t-immunity sets $\bar{v} = 0$ and uses a "relative" notion of victimization; see our discussion about relative notions of victimization in section 1.7.

1.2 Motivating Example and a Notion of Safety

This section motivates the need for safety in mechanisms and presents our safety concept.

1.2.1 Preliminaries

We first introduce some notation that will be used throughout the paper. Let $A = \{a_1, \ldots, a_x\}$ be the set of outcomes and $I = \{1, \ldots, n\}$ be the set of players. Each individual $i \in I$ has a preference relation R_i over the set of outcomes, A. The collection of preference relations for all individuals is called a preference profile and is denoted by $R \equiv (R_1, \ldots, R_n)$. Associated with each preference relation R_i is a strict preference relation P_i. The set of all permissible preference profiles is denoted by $\mathcal{R}$ and we let $\mathcal{R}_i$ be the set of all permissible preference relations for a specific player $i \in I$. A *mechanism* is a pair (M, g), where $M = \times_{i \in I} M_i$ is the product of message spaces for each player and $g : M \to A$ is an *outcome function*. A typical element of M_i is a message for player $i \in I$, denoted by m_i, and a collection of such messages is called a message profile and denoted by $m = (m_1, \ldots, m_n)$. For $B \subseteq I$, let $M_B := \times_{i \in B} M_i$ and let (m'_B, m_{-B}) denote a message profile in M with players in B reporting $m'_B \in M_B$ and those in $I \backslash B$ reporting $m_{-B} \in M_{I \backslash B}$. For $q \in \{1, \ldots, n\}$, let I_q be the set of all subsets of I of size q.

A mechanism (M, g), together with a preference profile R defines a normal form *game*. A *social choice rule* (SCR) or a *social choice correspondence* (SCC) is a function $F : \mathcal{R} \to 2^A \backslash \{\varnothing\}$. When F is single valued, i.e. $F : \mathcal{R} \to A \backslash \{\varnothing\}$, it is called a *social choice function* (SCF). A *solution concept* is a prediction specifying the message profiles, called *equilibria*, we expect players to play in a game. If S is a solution concept and $\Gamma = (M, g)$ is a mechanism, we denote the set of all equilibrium profiles under preference profile R by $S^\Gamma(R) \subseteq M$. The set of outcomes associated with $S^\Gamma(R)$ is denoted by $g(S^\Gamma(R))$. A SCR (or SCF) is implementable in a solution concept, S, if there exists a mechanism $\Gamma = (M, g)$ such

that $g(S^{\Gamma}(R)) = F(R)$ for all $R \in \mathcal{R}$. We use NE to denote the use of Nash equilibrium as a solution concept.

1.2.2 Motivating Example: Hiring with Multidimensional Quality

Consider a board of directors with $n \geq 4$ members, $I = \{1, \ldots, n\}$, where each member $j \in I$ has a different background, $b_j \in \{b^1, \ldots, b^n\}$. The board is deciding on hiring a CEO from among x candidates, $A = \{c_1, \ldots, c_x\}$. Candidates have skill levels s_{c_i} drawn from a continuous, atomless, distribution with support on $[0, m]$, and potentially different backgrounds, $b_{c_i} \in \{b^1, \ldots, b^n\}$, $i \in I$. We assume that there is at least one candidate with each of the backgrounds.

The members' preferences are as follows: each member $j \in \{1, \ldots, n\}$ prefers more skilled candidates, but gives a "bonus" premium p_j to candidates with a matching background. Hence, a member j sorts the candidates according to a score for each candidate c_i given by $s_{c_i} + p_j \mathbf{1}\{b_{c_i} = b_j\}$.

We assume that there are no ties in the skill levels of the candidates or the scores given to the candidates by any member so that all preferences of the members are strict.

The desirable social choice function (SCF), F, selects the candidate with the highest skill level. Hence, any mechanism that implements F must always select the highest skilled candidate.[2] Consider Maskin's mechanism (Maskin (1999)). The SCF satisfies no veto power and Maskin monotonicity. In this mechanism each member's message space is $A \times \mathcal{R} \times \mathbb{N}$,

[2]It must then be possible to deduce which candidate is the highest skilled from a given preference profile of the members. In this environment, a candidate c_i is more skilled than candidate c_j if and only if at least $n-1$ members rank candidate c_i higher than candidate c_j: if c_i is more skilled than candidate c_j then all members, except possibly for the member with the same background as c_j, will rank c_i higher than c_j; similarly, if at least $n-1$ members rank a candidate c_i higher than a candidate c_j, then it must be that c_i is more skilled than c_j. Hence, if the preference profile were known then a SCF can not only deduce the highest skilled candidate, but can also rank the candidates according to skill.

where $\mathcal{R}$ is the set of possible preference profiles. Let (a^i, R^i, z^i) denote a message reported by member $i \in I$. Given a message profile, $m = (m_1, \ldots, m_n)$, the mechanism selects the outcome $g(m)$ as follows:

Rule 1 If m is such that all members agree on a report, say (a, R, z), and if $a \in F(R)$ then $g(m) = a$.

Rule 2 Suppose all members agree on (a, R, z) with $a \in F(R)$, except for one member i who reports $(a^i, R^i, z^i) \neq (a, R, z)$. If a^i is weakly less preferred to a for member i under R_i then $g(m) = a^i$; otherwise, $g(m) = a$.

Rule 3 In all other cases, let $g(m) = a^{q^*}$ for $q^* = \max\{q \in I : z^q = \max_j z^j\}$.

This mechanism has virtually no safety properties: there are always $n-1$ members who can *each* get the mechanism to select the least skilled candidate by deviating alone from *any* Nash equilibrium. There are always $n-1$ members who can each deviate from any Nash equilibrium under Rules 2 or 3 and enforce, via Rule 3, any outcome. Suppose R is the true preference profile and (a, R', z) is a NE under Rule 1. By implementation a is the highest skilled candidate under R and R' (if it is not the highest skilled under R' then it would not be in $F(R')$ and (a, R', z) would not fall under Rule 1). Thus, the least skilled candidate is less preferred than a by at least $n-1$ members under R'. Rule 2 allows any member to deviate and get any outcome that is less preferred for that member under the preferences reported by others. Hence, at least $n-1$ members can *each* deviate to Rule 2 and force the mechanism to select the least skilled candidate. Therefore, it takes only one member reporting the least skilled candidate for the mechanism to pick the least skilled candidate instead of the highest skilled!

In practice such a failure can arise for many reasons: a member may intentionally try to harm others; a member may misjudge the least skilled candidate to be the highest skilled;

the least skilled candidate may be able to deceive one member into thinking they are highly skilled; or simply due to a mistake or a failure in communication in reporting the messages.

The lack of safety illustrated in this example is worrisome. In Maskin's mechanism and many others in the literature, it is generally true that a deviation by one agent, for any reason, can lead to unrestricted damage to others.

1.2.3 A Notion of Victimization

Safety can be viewed as limited risk, or limited exposure to harm. We must then define what it means for a player to be harmed. In much of the mechanism design and implementation literature, a player is mainly concerned about how the final outcome of the mechanism ranks in that player's preferences relative to other outcomes in some finite set.[3] Thus, we can say that a player is a *victim* if the final outcome of the mechanism ranks too low relative to other outcomes in that player's preferences. Formally, a player is a victim if the outcome of the mechanism is among the K least preferred outcomes for that player, where $K \in \{1, \ldots, x-1\}$ is a threshold chosen by the designer. This leads to the following definition.

Definition 1.2.1. Given a preference profile R and an outcome $a \in A$, player i is a victim if

$$|\{b \in A : aR_ib\}| \leq K \tag{1.1}$$

If (1.1) holds we also say that a is among the K least preferred outcomes for player i under R. Thus, given a preference profile and an outcome we can always determine whether a player is a victim or not.[4] Definition 1.2.1 has subtle implications for cases where players are indifferent between some outcomes, as the following example illustrates.

[3] We can easily accommodate infinite sets of outcomes as we explain in section 1.7.1.

[4] We note the inherent binary nature of the concept of a victim: a player is either a victim or not a victim. At first glance, it may seem like this limits the applicability of our safety definition in situations where a

Example 1. Suppose $A = \{a_1, a_2, a_3\}$, $K = 2$, and there are 3 players with the following preferences:

R_1	R_2	R_3
a_2	a_1, a_2	a_1, a_2, a_3
a_3	a_3	
a_1		

Then, the set of least preferred K outcomes for player 1 is $\{a_1, a_3\}$, for player 2 is $\{a_3\}$, and for player 3 is $\varnothing$. Hence, players who are indifferent between some outcomes will, in general, have smaller sets of outcomes that lead to their labeling as victims; players who are indifferent between all outcomes cannot be victimized.

Note: throughout the paper we suppress the dependence on the damage threshold, K, to simplify the notation. However, K is used to define a victim (definition 1.2.1), so any discussion of safety implicitly assumes a given $K \in \{1, \ldots, x-1\}$. Thus, changing the K affects the safety conditions. For example, the $\bar{v}$-safe outcome property, a necessary condition we formally define in section 1.3.2.1, simply states that any outcome in the social choice rule victimizes at most $\bar{v}$ players at the preference profile where it is chosen; if K is increased then a given outcome will victimize (weakly) more players, and, thus, a social choice rule that satisfied the $\bar{v}$-safe outcome property may fail to satisfy it after K is increased. To make the conditions explicitly depend on K, a "victim" can simply be replaced by "a player who obtains an outcome within the least K preferred for that player."

1.2.4 A General Notion of Safety

If safety is freedom from exposure to harm (i.e. freedom from risk), players must be limited in their ability to victimize others. It will thus be helpful to define a function that

designer cares about the level of damage players are exposed to. However, this can be accommodated by letting the definition of a victim depend on the level of damage and examining the safety properties at different damage thresholds. We allow for this possibility in our definition of a victim.

quantifies the ability of players to victimize others. We do this using the following definition of a *victims function*:

Definition 1.2.2. Given a game, (M, g, R), the victims function at message profile $m \in M$ is a function $V_m : \{1, \ldots, n-1\} \to \{0, \ldots, n-1\}$ defined by

$$V_m\left(\bar{d}\right) = \max_{B \in I_{\bar{d}}} \left[\max_{m'_B \in M_B} \left(\sum_{i \in I \setminus B} \mathbf{1}\left\{\text{i is a victim}\right\} \right) \right]$$

The victims function specifies the maximum number of victims that $\bar{d}$ players, which we refer to as *deviators*, can create by potentially deviating from the profile under consideration. Using our definition of a victim, we can write the victims function as,

$$V_m\left(\bar{d}\right) = \max_{B \in I_{\bar{d}}} \left[\max_{m'_B \in M_B} \left(\sum_{i \in I \setminus B} \mathbf{1}\left\{ \left| \{b \in A : g(m'_B, m_{-B}) R_i b\} \right| \leq K \right\} \right) \right]$$

We can now define safety in terms of the victims function.

Definition 1.2.3. Given a game, (M, g, R), a message profile, $m \in M$, is $(\bar{d}, \bar{v})$-safe if $V_m(\bar{d}) \leq \bar{v}$.

A profile is $(\bar{d}, \bar{v})$-safe if no group, $B \subseteq I$, of $\bar{d}$ players can obtain an outcome (by potentially deviating from m) that leads to more than $\bar{v}$ victims in $I \setminus B$ under R. Different combinations of $\bar{d}$ and $\bar{v}$ can lead to different safety levels that may be of interest depending on the context. Intuitively, a profile is safe if players are limited in their ability to victimize others, leading to a formal definition of safety as freedom from exposure to harm.

Note that if $\bar{v} \geq n - \bar{d}$ then the safety requirement imposes no restrictions. Throughout the rest of the paper, unless specified otherwise, $\bar{v}$ will always assumed to be strictly less than $n - \bar{d}$. Also, unless indicated otherwise, referring to a message profile simply as "safe"

implies that it is $(\bar{d}, \bar{v})$-safe where $\bar{d} \in \{1, \ldots, n-1\}$ and $\bar{v} \in \{0, \ldots, n-2\}$ are set by the designer.

It is easy to see why definition 1.2.3, using definition 1.2.1 of a victim, leads to desirable safety properties in an implementation framework. If a message profile is $(\bar{d}, \bar{v})$-safe, with victimization defined using definition 1.2.1, then no group of $\bar{d}$ players can deviate and cause more than $\bar{v}$ of the rest to get an outcome that ranks too low (among the least K favorite) in their preferences. Thus, groups of deviators cannot significantly hurt others, ruling out "disasters" and limiting how much players can reduce others' welfare. By choosing $\bar{d}$, $\bar{v}$, and K, the designer has great flexibility and control over the desired level of safety. Definition 1.2.3 is extremely general. Using variations on the definition of a victim, we can use definition 1.2.3 to examine the safety properties of virtually any game. Section 1.7.1 discusses this point, and our notions of safety and victimization, in more detail.

1.3 Main Results

1.3.1 Strong Safe Implementation

We first consider a strong form of safe implementation where we require the implementing mechanism to have the following properties: all Nash equilibria lead to desirable outcomes, all desirable outcomes can be obtained via Nash equilibria, and all Nash equilibria are safe. In this section we show that, in the most basic environment, strong safe implementation is restrictive, but in section 1.3.3 we relax some assumptions about the environment and show that safety in equilibrium can be achieved from a weaker notion of implementation under mild behavioral assumptions unrelated to safety.

Definition 1.3.1 (**Strong $(\bar{d}, \bar{v})$-Safe Implementation**). A social choice rule is *strongly $(\bar{d}, \bar{v})$-safe implementable* if a mechanism $\Gamma = (M, g)$ exists such that for any preference profile $R \in \mathcal{R}$, $g(NE^{\Gamma}(R)) = F(R)$ and all profiles in $NE^{\Gamma}(R)$ are $(\bar{d}, \bar{v})$-safe.

Strong safe implementation is simply Nash implementation with the added constraint that all Nash equilibria of the mechanism at any preference profile are safe.

Strong safe implementation is desirable because it implies that players will be safe as long as they play a Nash equilibrium. Unfortunately, the following Theorem shows that strong safe implementation implies either very strong restrictions on the implementable social choice rule or strong and unnatural restrictions on preferences that are unlikely to hold except in very specific environments. The proof for the theorem is in the appendix.

Theorem 1.3.1. Let F be a social choice rule that is strongly $(\bar{d}, \bar{v})$-safe implementable and let $\mathcal{R}$ be the set of admissible preference profiles. Suppose $a \in F(R)$ for some $R \in \mathcal{R}$. Then at least one of the following conditions is true:

1. $a \in F(R')$ for all $R' \in \mathcal{R}$.
2. If $a \notin F(R')$ for some $R' \in \mathcal{R}$, then there exists a pair $(i, b) \in I \times A$ such that:
   - aR_ib.
   - bP'_ia.
   - There is no preference profile where a weakly rises in every player's preferences relative to R and b victimizes more than $\bar{v}$ players in $I\backslash\{i\}$.

To see the intuition of Theorem 1.3.1, suppose $a \in F(R)$. Nash implementation implies the existence of a mechanism with a Nash equilibrium m with outcome a. Now, consider what some player can obtain by deviating from m; either: (1) no player can obtain any outcome other than a, or (2) some player can obtain some outcome other than a by deviating. In the first case, m must be a Nash equilibrium at every other preference profile, and implementation would necessitate that $a \in F(R')$ for all $R' \in \mathcal{R}$. In the second case (suppose player i can obtain outcome b by deviating from m), at every preference profile

where a weakly rises in the preferences of the players m must be a Nash equilibrium; requiring that all Nash equilibria be safe must then imply that no outcome obtainable by any player through a deviation from m (e.g. b) victimizes more than $\bar{v}$ of the others at the profile where a weakly rises in all players' rankings.

Condition (1) of Theorem 1.3.1 says that a, a desirable social outcome at some preference profile, must be a chosen at all other preference profiles; this is a very strong condition on the social choice rule. If condition (1) is true for all outcomes in the range of F then F is a trivial social choice rule: it is constant and does not depend on the preference profile.

Condition (2) of Theorem 1.3.1 is also restrictive, because it implies that relative preference movements of one outcome impose restrictions on *other* outcomes. It says that if a is not picked at some other preference profile, R', then there must be an outcome that is weakly less preferred by some player i at R, strictly more preferred by player i at R', and never victimizes too many players at *any* preference profile where a rises in everyone's rankings; this must hold for every profile that does not contain a as an outcome. This is a strict and unnatural requirement on the preferences. For any outcome a and player i, it is unlikely that a designer would a priori have information that some outcome (that is sometimes weakly less preferred to a by player i and sometimes strictly more preferred) is never too low in the rankings of other players whenever a rises in everyone's rankings. Corollary 1.3.2 establishes that no social choice rule is strong safe implementable if the set of preferences is sufficiently rich, and the following paragraph explains precisely which combination of preferences make strong safe implementation impossible.

Corollary 1.3.2. Suppose the set of preference profiles, $\mathcal{R}$, includes all possible strict preferences. Then no social choice rule is strongly safe implementable.

The assumption that $\mathcal{R}$ includes all possible strict preferences can be relaxed and the corollary would still hold. Intuitively, suppose $a \in F(R)$ for some $R \in \mathcal{R}$. If there is just *one* preference profile that violates condition (2) of Theorem 1.3.1 then condition (1) holds

and a must be chosen at all preference profiles. If condition (1) of Theorem 1.3.1 holds and a victimizes more than $\bar{v}$ players at *some* preference profile, strong safe implementation will fail because a is always chosen. Thus, the corollary holds as long as there are preferences where condition (2) of Theorem 1.3.1 is violated for some outcome, and as long as that outcome can potentially victimize many players.[5]

Note that corollary 1.3.2 holds for any arbitrary nontrivial combination of $\bar{d}$, $\bar{v}$, and K, so it is true even with the weakest safety requirements.

1.3.2 Safe Implementation

In this section we discuss a weaker notion of implementation, which we call $(\bar{d}, \bar{v})$-safe implementation. We provide necessary conditions of $(\bar{d}, \bar{v})$-safe implementation and show that they are also sufficient with mild preference restrictions when $\bar{d} < n - 1$.

For $(\bar{d}, \bar{v})$-safe implementation, we require the following two properties at any preference profile: all desirable outcomes can be obtained via Nash equilibria that are safe, and all Nash equilibria that are safe lead to desirable outcomes. This is equivalent to implementation in a solution concept that we refer to as a $(\bar{d}, \bar{v})$-safe equilibrium ($(\bar{d}, \bar{v})$-SE or simply SE):

Definition 1.3.2 (**$(\bar{d}, \bar{v})$-Safe Equilibrium**)**.** A $(\bar{d}, \bar{v})$-safe equilibrium is a message profile $m \in M$ that is both a Nash equilibrium and $(\bar{d}, \bar{v})$-safe.

Definition 1.3.3 (**$(\bar{d}, \bar{v})$-Safe Implementation**)**.** A social choice rule is $(\bar{d}, \bar{v})$-*safe implementable* if a mechanism $\Gamma = (M, g)$ exists such that for any preference profile $R \in \mathcal{R}$, $g(SE^{\Gamma}(R)) = F(R)$.

[5] A less general version of corollary 1.3.2 for social choice functions follows directly from the Muller and Satterthwaite (1977) Theorem, which states that, under some conditions, Maskin monotonic social choice functions are dictatorial. Theorem 1.3.1 allows for more generality in corollary 1.3.2 and makes clear how the assumption that $\mathcal{R}$ includes all possible strict preferences can be relaxed.

Safe implementation assures the designer that any outcome in the social choice rule can be obtained via a safe equilibrium of the mechanism. It is also logically consistent in the sense that if we expect players to play a Nash equilibrium that is also safe (i.e. a safe equilibrium), then all safe equilibria should lead to outcomes in the social choice rule at any preference profile. Thus, it achieves a minimal degree of safety and forms the basis for implementing social choice rules in a safe way.[6]

1.3.2.1 Necessary Conditions

This subsection shows that three conditions: $\bar{v}$-safe outcome, $(\bar{d}, \bar{v})$-safe monotonicity, and $(\bar{d}, \bar{v}, n)$-similarity are necessary for implementation in $(\bar{d}, \bar{v})$-safe equilibria. The first two are necessary conditions on social choice rules; the third is a necessary condition on the set of preference profiles when $\bar{d} \geq \frac{n}{2}$. We start first with the $\bar{v}$-safe outcome property, which intuitively follows from the $(\bar{d}, \bar{v})$-SE definition.

Definition 1.3.4 ($\bar{v}$**-Safe Outcome Property**)**.** If $a \in F(R)$ then at most $\bar{v}$ players have a as one of the least K favorite outcomes under R.

It is easy to see why the $\bar{v}$-safe outcome property is necessary for implementation: if implementation holds and it is not satisfied then there must be a $(\bar{d}, \bar{v})$-SE of the implementing mechanism at which there are more than $\bar{v}$ victims, directly contradicting the definition of $(\bar{d}, \bar{v})$-SE.

An implicit implication of the $\bar{v}$-safe outcome property is that the objectives of the designer (the social choice rule) must be consistent with choosing an outcome that is not too low in the rankings of many players. This is not restrictive in most of mechanism design,

[6] In this paper, we do not simply assume that players coordinate on a safe equilibrium. However, to the extent that a safe equilibrium may be an appropriate prediction in some environments (e.g. as in Halpern (2008)'s t-immunity), implementation in safe equilibria may be interesting in its own right.

which generally aims to combine individual preferences to reach a collective decision in some welfare maximizing way.

The $\bar{v}$-safe outcome property is similar to the "no worst alternative" property introduced by Cabrales and Serrano (2011). Cabrales and Serrano say that a SCC satisfies the no worst alternative property if, for all players, any outcome chosen in any preference profile is strictly preferred to some other outcome. If $\bar{v} = 0$, $K = 1$, and each player has a single worst alternative then the $\bar{v}$-safe outcome property is equivalent to the no worst alternative property.

The next property, $(\bar{d}, \bar{v})$-safe monotonicity, is a generalization of Maskin monotonicity (Maskin (1999)). $(\bar{d}, \bar{v})$-Safe monotonicity changes depending on $\bar{d}$ with different general conditions for the case of $\bar{d} = 1$ than other cases; the intuition is that, at a $(\bar{d}, \bar{v})$-SE, unilateral deviations must satisfy incentive (Nash) and safety constraints, whereas multilateral deviations need to satisfy safety constraints only.

Definition 1.3.5 ($(\bar{d}, \bar{v})$**-Safe Monotonicity Property**)**.** For any two profiles R, R' suppose $a \in F(R)$ and the following two conditions hold:

1. For all $i \in I$, if aR_iy and y victimizes at most $\bar{v}$ players in the set $I \backslash \{i\}$ under R, then aR'_iy and y victimizes at most $\bar{v}$ players in the set $I \backslash \{i\}$ under R'.

2. If $\bar{d} \geq 2$ then for all groups $B \subseteq I$ of size $n - \bar{d}$, if $x \in A$ victimizes at most $\bar{v}$ players in B under R, then x victimizes at most $\bar{v}$ players in B under R'.

then $a \in F(R')$

The first condition of the $(\bar{d}, \bar{v})$-safe monotonicity property states that for any player, i, if an element in A is weakly less preferred to a under R and causes at most $\bar{v}$ victims in the set of all players excluding i, then the same holds true for that element under R'. The

second condition states that for any group B of size $n-\bar{d}$, if an element victimizes at most $\bar{v}$ players in B under R then that element victimizes at most $\bar{v}$ players in B under R'. Section 1.5.1 discusses safe monotonicity in more details and clarifies its relationship with Maskin monotonicity.

The following Theorem shows that both the $\bar{v}$-safe outcome property and $(\bar{d},\bar{v})$-safe monotonicity are necessary for implementation in $(\bar{d},\bar{v})$-safe equilibria.

Theorem 1.3.3. Suppose F is implementable in $(\bar{d},\bar{v})$-safe equilibrium. Then F satisfies the $\bar{v}$-safe outcome property and the $(\bar{d},\bar{v})$-safe monotonicity property.

It is, again, easy to see why $(\bar{d},\bar{v})$-safe monotonicity is necessary. Suppose F is $(\bar{d},\bar{v})$-safe implementable and $a \in F(R)$ then there must be a $(\bar{d},\bar{v})$-SE, call it m, at R with a as the outcome. If both conditions of $(\bar{d},\bar{v})$-safe monotonicity are satisfied going from R to R' then m is a $(\bar{d},\bar{v})$-SE at R' as well. Hence, $a \in F(R')$. A formal proof is provided in the appendix.

Finally, we define a notion of similarity among the elements in a set of preference profiles that will be necessary for safe implementation only when $\bar{d} \geq \frac{n}{2}$.

Definition 1.3.6 ($(\bar{d},\bar{v},n)$**-Similarity**)**.** Given $\bar{d}$, $\bar{v}$, and n, let $q = \left\lfloor \frac{n}{n-\bar{d}} \right\rfloor$ be the integer quotient from dividing n by $n-\bar{d}$ (and ignoring the remainder). A set of preference profiles $\mathcal{R}$ is $(\bar{d},\bar{v},n)$-similar if whenever $\{(g^1,R^1),\ldots,(g^q,R^q)\}$, where $g^i \subseteq I$ and $R^i \in \mathcal{R}$, are any q pairs such that $|g^i| \geq n-\bar{d}$ and $g^i \cap g^j = \varnothing$ for any $i,j \in \{1,\ldots,q\}$, then there is at least one outcome in A that causes $\bar{v}$ or less players in g^i to be victims under R^i regardless of $i \in \{1,\ldots,q\}$.

If a set of preference profiles, $\mathcal{R}$, is $(\bar{d},\bar{v},n)$-similar then the preference profiles in $\mathcal{R}$ are not too different from each other. Specifically, if the players are partitioned in a way to maximize the number of groups of size $n-\bar{d}$ and each of those groups reported a different

preference profile from a $(\bar{d}, \bar{v}, n)$-similar set, then we can find an element that victimizes at most $\bar{v}$ players in each group under the profile reported by that group. The $(\bar{d}, \bar{v}, n)$-similarity condition is a somewhat restrictive constraint on preferences, but the following Theorem establishes its necessity when $\bar{d} \geq \frac{n}{2}$; it is not necessary for safe implementation when $\bar{d} < \frac{n}{2}$. The proof for the Theorem is in the appendix.

Theorem 1.3.4. Suppose F is implementable in $(\bar{d}, \bar{v})$-safe equilibrium with n players on a set of preference profiles $\mathcal{R}$ and suppose that $\bar{d} \geq \frac{n}{2}$. Then $\mathcal{R}$ is $(\bar{d}, \bar{v}, n)$-similar.

$(\bar{d}, \bar{v}, n)$-Similarity is necessary when $\bar{d} \geq \frac{n}{2}$ because for $(\bar{d}, \bar{v})$-safe implementation, any group of size $n - \bar{d}$ or larger must be able to protect its members, except at most $\bar{v}$ of them, from being victims. If $\bar{d} \geq \frac{n}{2}$ then the deviators can act as if they were a group of $n - \bar{d}$ or more "rational" players. The designer in this case would have no way of knowing which group of players to protect, so the designer must make sure that all groups of size $n - \bar{d}$ or greater are protected (in the $(\bar{d}, \bar{v})$ sense). If the set of preference profiles is $(\bar{d}, \bar{v}, n)$-similar then the designer can assure the safety of any group of size at least $n - \bar{d}$.

1.3.2.2 Sufficient Conditions

In this subsection we provide preference restrictions that, along with the necessary conditions presented previously, are sufficient for implementation in $(\bar{d}, \bar{v})$-safe equilibria when $\bar{d} < n - 1$.

Definition 1.3.7 (**$(\bar{d}, \bar{v})$-Exposure**)**.** A preference profile $R \in \mathcal{R}$ satisfies $(\bar{d}, \bar{v})$-exposure if for any subset of players $B \subseteq I$ of size $n - \bar{d}$, the set of outcomes in A that cause more than $\bar{v}$ players in B to be victims is nonempty.

$(\bar{d}, \bar{v})$-Exposure rules out profiles where some groups of $\bar{d}$ players cannot victimize more than $\bar{v}$ of the other players even if the $\bar{d}$ deviators were allowed to pick any outcome in A. These are profiles where the question of safety is vacuous for some groups: profiles

where some groups of $n - \bar{d}$ or more players are "invincible" in the sense that no outcome in A can victimize more than $\bar{v}$ in those groups (e.g. profiles where all players are indifferent between all outcomes). In economic environments, where any group of $\bar{d}$ deviators can choose to deprive all the rest from all economic goods and split the goods among themselves, the $(\bar{d}, \bar{v})$-exposure condition is not restrictive.

The following Theorem establishes that the $\bar{v}$-safe outcome and the $(\bar{d}, \bar{v})$-safe monotonicity properties are sufficient for $(\bar{d}, \bar{v})$-safe implementation under the appropriate preference restrictions that depend on $\bar{d}$.

Theorem 1.3.5. Suppose F satisfies the $\bar{v}$-safe outcome property and the $(\bar{d}, \bar{v})$-safe monotonicity property, admissible preference profiles satisfy $(\bar{d}, \bar{v})$-exposure, and $n > 2$. If $\bar{d} < \frac{n}{2}$, or if $\frac{n}{2} \leq \bar{d} < n - 1$ and the set of preference profiles is $(\bar{d}, \bar{v}, n)$-similar, then F is $(\bar{d}, \bar{v})$-safe implementable.

The proof for Theorem 1.3.5 is constructive and is provided in detail in the appendix. Section 1.5.2 describes the mechanism used for implementation and outlines the steps of the proof.

1.3.3 Double Implementation

By itself, safe implementation does not guarantee that all Nash equilibria will lead to desirable social outcomes, because it places no restrictions on unsafe Nash equilibria. In this section we explore double implementation in both safe and Nash equilibria. We will show that double implementation guarantees safety in equilibrium with mild behavioral assumptions.

Definition 1.3.8 (**Double Implementation in $(\bar{d}, \bar{v})$-Safe and Nash Equilibria**). A social choice rule is *double implementable in $(\bar{d}, \bar{v})$-safe and Nash equilibria* (or simply double implementable) if a mechanism $\Gamma = (M, g)$ exists such that for any preference profile $R \in \mathcal{R}$, $g(SE^{\Gamma}(R)) = F(R)$ and $g(NE^{\Gamma}(R)) = F(R)$.

Double implementation requires both Nash and safe implementation: it requires that all Nash equilibria lead to outcomes in the social choice rule, and all outcomes in the social choice rule be obtainable via a safe equilibrium. Because safe equilibria, when they exist, are always a subset of the set of Nash equilibria, Nash implementation already implies that all safe equilibria lead to outcomes in the social choice rule. Imposing safe implementation additionally requires that any outcome in the social choice rule be obtainable via a safe equilibrium of the implementing mechanism.

First, we introduce a necessary monotonicity condition on social choice rules that must hold for double implementation when transfers are not allowed.

Definition 1.3.9 ($\bar{v}$**-Safe-Nash Monotonicity**)**.** For any two profiles R and R', suppose $a \in F(R)$ and the following condition holds: for all $i \in I$, if aR_ib and b victimizes at most $\bar{v}$ players in the set $I\backslash i$ under R then aR'_ib. Then $a \in F(R')$.

Theorem 1.3.6. Suppose F is double implementable in $(\bar{d}, \bar{v})$-safe and Nash equilibria. Then F is $\bar{v}$-safe-Nash monotonic.

To see why $\bar{v}$-safe-Nash monotonicity is necessary, suppose F is double implementable and $a \in F(R)$. There must be a $(\bar{d}, \bar{v})$-SE, call it m, at R with a as the outcome. If $\bar{v}$-safe-Nash monotonicity holds going from R to R' then m is a Nash equilibrium at R' as well. Nash implementation then implies that $a \in F(R')$.

$\bar{v}$-Safe-Nash monotonicity is more restrictive than Maskin and $(\bar{d}, \bar{v})$-safe monotonicity, and it implies both.[7] However, we show next that if side payments are allowed and

[7]To see this, we can write $\bar{v}$-safe-Nash monotonicity as follows: if $a \in F(R)$ but $a \notin F(R')$, then for some $i \in I$ and some $b \in A$, aR_ib and b victimizes at most $\bar{v}$ players in the set $I\backslash i$ under R, but bP'_ia. This directly implies the contrapositive of $(\bar{d}, \bar{v})$-safe monotonicity (see section 1.5.1), showing that $(\bar{d}, \bar{v})$-safe monotonicity holds. It also implies that if $a \in F(R)$ but $a \notin F(R')$ then at least one player had a preference reversal between a and a weakly lower ranked outcome at R going from R to R', showing that Maskin monotonicity holds.

the $\bar{v}$-safe outcome property holds, then $\bar{v}$-safe-Nash monotonicity is sufficient for double implementation in safe and Nash equilibria when $\bar{d} < n/2$.[8]

In the rest of this section we allow the designer to specify a vector of transfers $t(m) = (t_1, \ldots, t_n) \in \mathbb{R}^n$ for each $m \in M$ in mechanism (M, g). We use the convention that transfers are paid to the players, so if t_i is negative then player i pays $|t_i|$. We impose no assumptions except that there exists a large enough transfer $L \in \mathbb{R}$ so that any player would prefer to avoid paying L (having transfer $-L$ as opposed to 0), and would prefer a transfer of L instead of 0, regardless of the outcome or the preference profile.

For the following, we extend preferences to be defined over sets of outcomes and transfers $A \times \mathbb{R}$. If $(a, 0)R_i(b, 0)$ then player i weakly ranks outcome a above b at preference profile R; we often drop the transfers if they are 0 to simplify notation. We assume positive transfers are valued positively by players at all preference profiles.

Before we proceed, we discuss some slight modifications that are needed in environments with transfers. First, to avoid trivially eliminating victims by compensating them for bad outcomes, we constrain our mechanisms throughout to have no transfers in equilibrium. Second, our definition of a victim requires a slight modification, because the notion of the least K preferred outcomes may not be well defined when negative transfers are allowed. We define a victim to be a player who obtains an outcome within their least K preferred outcomes among the outcomes in A without consideration for transfers or who has a negative transfer (or both). This definition also accounts for players attempting to cause harm by imposing negative transfers on others.

The following Theorem establishes that the $\bar{v}$-safe outcome and the $\bar{v}$-safe-Nash monotonicity properties are sufficient for double implementation when mechanisms with transfers

[8] $\bar{v}$-safe-Nash monotonicity is, in general, not necessary for double implementation when mechanisms utilizing transfers are allowed, unless the designer does not have precise cardinal information, as we explain section 1.3.3.1.

are allowed. In section 1.3.3.1 we show that they are necessary *and* sufficient in the absence of precise cardinal information.

Theorem 1.3.7. Suppose $n > 2$, $\bar{d} < \frac{n}{2}$, and F satisfies $\bar{v}$-safe-Nash monotonicity and the $\bar{v}$-safe outcome property. Then F is double implementable in $(\bar{d}, \bar{v})$-safe equilibria and Nash equilibria using a mechanism with transfers.

The proof for Theorem 1.3.7 is constructive and is provided in the appendix. We use the same mechanism used for safe implementation (described in section 1.5.2), but we additionally impose a particular transfer structure, which we describe in section 1.5.3.

The following corollary shows that double implementation ensures safety in equilibrium with mild assumptions on behavior unrelated to safety.

Corollary 1.3.8. Suppose F is double implementable in Nash and safe equilibria using the mechanism described in the proof of Theorem 1.3.7. Then the following is true:

1. Any unsafe Nash equilibrium is untruthful (all players report a false preference profile).
2. For any unsafe Nash equilibrium there exists a truthful, safe equilibrium that yields the same outcome as the unsafe Nash equilibrium.

Thus, in the mechanisms we use, players gain nothing by playing an unsafe Nash equilibrium over a safe one, and they all must lie to play an unsafe Nash equilibrium. An important implication is that mild assumptions on collective honesty of players, or assumptions on truthful equilibria being focal when untruthful equilibria are payoff-equivalent, are enough to ensure that players are always safe in equilibrium.

1.3.3.1 Necessary and Sufficient Conditions

We have shown that $\bar{v}$-safe-Nash monotonicity is necessary for double implementation in the absence of transfers, but is it necessary when transfers are allowed? In this section we

show that if the designer does not have access to precise cardinal preference information then $\bar{v}$-safe-Nash monotonicity remains necessary, even if mechanisms with transfers are allowed. This will lead to necessary and sufficient conditions for double implementation.

Definition 1.3.10. Suppose $(a,0)R_i(b,0)$ and $(a,0)R_i'(b,0)$. A designer has *precise cardinal information* about preferences for player i and outcomes $\{a,b\}$ at preference profiles $\{R,R'\}$ if transfers $t_1,t_2 \in \mathbb{R}$ can be determined so that $(a,t_1)R_i(b,t_2)$ and $(b,t_2)P_i'(a,t_1)$, or if it can be determined that such transfers do not exist.

Access to precise cardinal information implies a designer can determine if it is possible to set transfers that maintain a player's ranking of two outcomes at one preference profile, but reverse that ranking at another preference profile. Intuitively, this implies the designer knows how much more a player prefers one outcome to another at each of the two preference profiles.

Theorem 1.3.9. Suppose F does not satisfy $\bar{v}$-safe-Nash monotonicity, but F is double implementable in $(\bar{d},\bar{v})$-safe and Nash equilibria. Then the designer has precise cardinal information about preferences for some $(i,a,b,R,R') \in I \times A \times A \times \mathcal{R} \times \mathcal{R}$, where $a \neq b$ and $R \neq R'$.

The intuition for Theorem 1.3.9 is as follows. If double implementation is possible and $\bar{v}$-safe-Nash monotonicity fails, then, in the absence of transfers, there is an outcome in the social choice rule at some preference profile where every safe equilibrium yielding that outcome of the mechanism is also a Nash equilibrium at another preference profile where that outcome is not in the social choice rule. To make double implementation possible, the designer must then use transfers and give at least one player a profitable deviation from the undesirable Nash equilibria at one preference profile while maintaining their safe equilibrium properties at another preference profile. This implies knowledge of cardinal information about preferences as defined in definition 1.3.10.

Theorem 1.3.9 implies that $\bar{v}$-safe-Nash monotonicity is necessary for double implementation if the designer does *not* have precise cardinal information about preferences – as is true in most common environments. Therefore, when precise cardinal preference information is not available, mechanisms with transfers are allowed, $n > 2$, and $\bar{d} < n/2$, $\bar{v}$-safe-Nash monotonicity and the $\bar{v}$-safe outcome property are both necessary *and* sufficient for double implementation.[9]

1.4 Safety Trade-offs and Other Results

1.4.1 Freedom: the Price of Safety

Freedom can be classified into *negative freedom*: freedom from harm, and *positive freedom*: freedom of players to obtain different outcomes in equilibrium. In Lemma 1.4.1 (section 1.4.3) we will see that safety requires the players to have certain negative freedoms; this section shows that safety restricts players' positive freedom. In general, *every* player will have no more positive freedom in a safe mechanism than in a mechanism that disregards safety. In addition, positive freedom becomes more restricted as mechanisms become safer.

Let the positive freedom of a player be parametrized by the number of options available to that player in equilibrium (i.e. the outcome selected at equilibrium and all outcomes the player can obtain by deviating). A player with more options in equilibrium has more positive freedom. Let $O_i(m)$ be the set of options available to some player i at message profile m and let $L_i(a, R) \equiv \{b \in A : aR_ib\}$ be the lower contour set of player i at outcome a when the preference profile is R (i.e. it is the set of all outcomes that are not strictly preferred by player i to a under R_i). Suppose the preference profile is R and a is chosen by the social choice rule at R. At any Nash equilibrium m that implements a of any mechanism, it must be true that

[9]The necessity of the $\bar{v}$-safe outcome property follows directly from the definitions, regardless of whether transfers are allowed or not, given our constraint that no transfers exist in equilibrium and our notion of victimization.

$O_i(m) \subseteq L_i(a, R_i)$; that is, the set of options available to any player does not contain any outcome more preferred than the one selected in equilibrium. Any safe equilibrium, m', is a Nash equilibrium and has the property that no player can victimize more than $\bar{v}$ players of the rest by deviating. Thus, letting $G_i(R)$ be the set of outcomes that do not victimize more than $\bar{v}$ players in the set $I \backslash i$ under R, it must be true that $O_i(m') \subseteq L_i(a, R_i) \cap G_i(R)$. This shows that if mechanisms are designed to maximize the positive freedom of players subject to the desired constraints regarding incentives and/or safety then $|O_i(m')| \leq |O_i(m)|$ and a player has no more positive freedom at a safe equilibrium than at a Nash equilibrium.

Now suppose we generalize the definition of positive freedom to groups of players. Let $O_B(m)$ be the set of options available to a group of players B through a multilateral deviation from m (including the outcome when m is played). Nash implementation imposes no restrictions on $O_B(m)$, but at a safe equilibrium players in B must not be able to victimize more than $\bar{v}$ of the rest if $|B| \leq \bar{d}$. Hence, as $\bar{d}$ increases, mechanisms that $(\bar{d}, \bar{v})$-safe implement social choice rules will be more restrictive in terms of the sets of options available to groups of players. Furthermore, as $\bar{v}$ decreases and mechanisms become more safe, guaranteeing the victimization of less players, the set of outcomes that must be ruled out through deviations from equilibrium increases. This restricts the positive freedoms of all the players.

Using the notation in this section, the Appendix illustrates the connection between $(\bar{d}, \bar{v})$-safe monotonicity and the set of options available to a player i at equilibrium m, $O_i(m)$.

1.4.2 Safety vs. Efficiency

In some situations, the designer will face a choice between safe implementation and Nash implementation, and both will not be possible simultaneously. The advantage of safe implementation is that it guarantees that some Nash equilibria will be safe, and, more importantly, that every desirable outcome can be obtained by a safe Nash equilibrium. The

drawback is that it imposes no restrictions on unsafe Nash equilibria, meaning that they may lead to undesirable outcomes outside of the social choice rule. Social choice rules are often chosen in a way to maximize overall welfare, so, to the extent that this is true, outcomes outside the social choice rule will be inefficient.

To illustrate this in our hiring example, in section 1.6 we show that double implementation in safe and Nash equilibria is indeed possible if the designer has some information about candidates' relative skill. If the designer does not have this information, double implementation is impossible and the designer must make a choice between safe and Nash implementation. If the designer chooses safe implementation, then there *must* exist Nash equilibria that select a less skilled candidate than the highest skilled. This trade-off between safety and efficiency extends more generally in situations where the designer must choose between safe and Nash implementation.

1.4.3 No Guarantees

An interesting case of safe implementation not covered in Theorems 1.3.5 or 1.3.7 is that of $\bar{d} = n - 1$, and $\bar{v} = 0$. If a social choice function is implementable in a $(n-1, 0)$-SE then the designer can *guarantee* each agent an outcome other than the agent's least favorite K ones. Many institutions and laws attempt to guarantee players a minimum level of payoff if they follow particular messages, regardless of what other players do. This section illustrates the difficulty with $(n-1, 0)$-safe implementation in a wide range of environments when transfers cannot be used.

The following lemma provides a necessary condition on mechanisms that implement social choice functions in a $(n-1, 0)$-SE.

Lemma 1.4.1. Suppose the set of preference profiles contains all strict preferences over A. Any mechanism, (M, g), that implements a social choice function in a $(n-1, 0)$-safe equilibrium must be such that $\forall i \in I$ and for all K-size subsets $D \subseteq A$, $\exists m_{Di} \in M_i$ such that

$g(m_{Di}, m_{-i}) \notin D, \forall m_{-i} \in M_{-i}$.[10]

The lemma states that any mechanism used in $(n-1,0)$-safe implementation must give every player possible messages that allow him to rule out any K outcomes in A from being implemented regardless of the messages of the other players. The proof of the lemma is in the appendix, but the intuition is simple. If a mechanism does not allow some player to rule out some K elements in A, then there will be a profile of preferences where those outcomes are the least preferred for that player but the other $n-1$ players have messages that force the mechanism to select one of those outcomes, violating the conditions of $(n-1,0)$-SE.

The following theorem uses the previous lemma to establish an impossibility result for implementation of social choice functions in $(n-1,0)$-SE when preferences include all strict preference profiles.

Theorem 1.4.2. Suppose $n \geq 2$ and the set of preference profiles, $\mathcal{R}$, contains all strict preferences over A. If $x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$ then no social choice function is implementable in a $(n-1,0)$-safe equilibrium.

Note that the number of outcomes can be more than nK so the theorem does not rely on the fact that ruling out the least favorite K outcomes for each player could exhaust all the possibilities. Remarkably, this is true *regardless* of $K \in \{1, \ldots, x-1\}$; the theorem does not count on K being large enough. Thus, the planner cannot make guarantees even in cases where $K = 1$, the weakest guarantee possible.

What is most interesting about Theorem 1.4.2 is that it holds not precisely because a designer cannot make safety guarantees, but because of an interplay between guarantees and freedom that causes non-existence of Nash equilibria. Indeed, a designer can design a mechanism that keeps players safe even if $x > nK$: simply allow them messages that rule

[10] This lemma can be easily extended to groups of players in general $(\bar{d}, \bar{v})$-safe implementation.

out their least favorite K outcomes and pick an outcome that is not ruled out. However, the theorem shows that if $x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$ then such a mechanism will, in general, have no Nash equilibria because the power given to any player (to rule out any K outcomes) is too great that players may be able to use it to deviate and obtain better outcomes. The following example makes the intuition clear.

Example 2. Suppose a committee, $I = \{\text{Member 1, Member 2, Member 3, Member 4}\}$, of $n = 4$ members is deciding on hiring a candidate from among $x = 6$ candidates: $A = \{a_1, \ldots, a_6\}$. Let $K = 1$. In this case $x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$ and we know by Theorem 1.4.2 that no social choice function is implementable in a $(n-1, 0)$-SE. Even though there are more candidates than members, *no* mechanism can *guarantee* each member a candidate better than his least favorite. To see this clearly, let the preference profile, R, be as follows:

Member 1	Member 2	Member 3	Member 4
a_2	a_1	a_1	a_1
a_3	a_3	a_2	a_2
a_4	a_4	a_4	a_3
a_6	a_6	a_5	a_5
a_1	a_2	a_3	a_4
a_5	a_5	a_6	a_6

The choices under the solid line are those that victimize the members. Any potentially implementable social choice function in a $(n-1, 0)$-SE must never pick a_5 or a_6 under R. So, consider the case when $F(R) = a_1$. To see that no mechanism can have a $(n-1, 0)$-SE in which the outcome is a_1, let the message profile $m = (m_1, m_2, m_3, m_4)$ be a $(n-1, 0)$-SE candidate. m_2 must rule out a_5 otherwise we could find m'_{-2} such that a_5 is forced as the outcome by three members at (m'_1, m_2, m'_3, m'_4) and m would not be a $(n-1, 0)$-SE. Since m_2 ensures that a_5 will not be selected, Member 1 would rather deviate and choose a message that ensures that a_1 is not selected in order to get one of a_2, a_3, a_4, or a_6. We know that such a message exists by Lemma 1.4.1, so (m_1, m_2, m_3, m_4) is not an equilibrium because Member 1 has a profitable deviation.

The same intuition extends to cases where $F(R) \in \{a_2, a_3, a_4\}$. Using similar logic, Member $i \in \{2, 3, 4\}$ would have a profitable deviation from any $(n-1, 0)$-SE if $F(R) = a_i$, leading to a contradiction.

Despite the proof of Theorem 1.4.2 relying on richness of the preference space, the problem with non-existence of Nash equilibria does not disappear in environments with more restricted preferences. The essence of the problem is that each player will have "free" veto power over any $l \leq K$ outcomes if l of that player's least favorite outcomes are also among the K least favorite of some others. Thus, if a player does not like the outcome chosen, they will (under very weak preference conditions) be able to veto it. Even if $K = 1$, the weakest safety guarantee possible, weak preference restrictions will imply that in general many players can veto outcomes they don't like. This can easily become problematic when attempting to implement desirable social choice functions.

Note also that the assumption of all strict preferences being admissible can be relaxed. Lemma 1.4.1 and Theorem 1.4.2 only need that the planner does not ex-ante know which outcomes are among the least favorite $2K$ for each player. This can be seen from example 2; the first 4 rows of the preferences can be arbitrarily modified and the example would still hold.

1.5 Discussion of Results

1.5.1 Safe Monotonicity

To understand $(\bar{d}, \bar{v})$-safe monotonicity better, we can express it differently. For any two profiles R, R' suppose $a \in F(R)$ but $a \notin F(R')$. Then F is $(\bar{d}, \bar{v})$-safe monotonic if one of the following conditions hold:

1. For some $i \in I$ and $y \in A$, aR_iy and y victimizes at most $\bar{v}$ players in the set $I \backslash \{i\}$

under R, but either $yP_i'a$ or y victimizes more than $\bar{v}$ players in the set $I\backslash\{i\}$ under R'.

2. $\bar{d} \geq 2$ and there is some group $B \subseteq I$ of size $n - \bar{d}$ and some $x \in A$ such that x victimizes at most $\bar{v}$ players in B under R but victimizes more than $\bar{v}$ players in B under R'.

It is useful to compare $(\bar{d}, \bar{v})$-safe monotonicity with Maskin monotonicity. Maskin monotonicity says that if an outcome is chosen at a particular profile and this outcome is not chosen in another profile then some player must have had a preference reversal where this outcome fell relative to another between the two profiles.

$(\bar{d}, \bar{v})$-Safe monotonicity is equivalent to Maskin monotonicity when $\bar{d} = 1$ and $\bar{v} = n - \bar{d}$. In this case the second condition in $(\bar{d}, \bar{v})$-safe monotonicity never plays a role and it is impossible for a player to victimize more than $\bar{v}$ of the rest so the first condition reduces to simple preference reversals, which is the essence of Maskin monotonicity. For more general combinations of $\bar{d}$ and $\bar{v}$, Maskin monotonicity does not imply $(\bar{d}, \bar{v})$-safe monotonicity and $(\bar{d}, \bar{v})$-safe monotonicity does not imply Maskin monotonicity, but both have a nonempty intersection. The following examples illustrate these facts.

Example 3. Let $A = \{a, b, c\}$ and $K = \bar{d} = \bar{v} = 1$ and consider a social choice function given by $F(R) = a$ and $F(R') = c$, where R and R' are given by

R_1	R_2	R_3	R_1'	R_2'	R_3'
a	a	b	c	c	a, b, c
b	b	a	a	a	
c	c	c	b	b	

F satisfies Maskin monotonicity trivially because players 1 and 2 have preference reversals between the chosen outcome and a less preferred one across profiles in both directions.

Next, we check $(\bar{d}, \bar{v})$-safe monotonicity. Note that because $\bar{d} = 1$, the second condition of $(\bar{d}, \bar{v})$-safe monotonicity does not play a role. Also, since $K = 1$, a player is victimized by some outcome if that outcome is the least preferred for that player. Because $a \in F(R)$, condition (1) of $(\bar{d}, \bar{v})$-safe monotonicity holds (going from R to R') if for each player i, the set of outcomes weakly less preferred to a that victimize at most $\bar{v} = 1$ players in $I \backslash \{i\}$ weakly grows going from R to R'. We check this next:

- For player 1 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{2, 3\}$ is $\{a, b\}$. At R' that set is $\{a, b\}$.
- For player 2 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{1, 3\}$ is $\{a, b\}$. At R' that set is $\{a, b\}$.
- For player 3 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{1, 2\}$ is $\{a\}$. At R' that set is $\{a, c\}$.

Thus, condition (1) of $(\bar{d}, \bar{v})$-safe monotonicity holds going from profile R to R'. However, $a \notin F(R')$, so $(\bar{d}, \bar{v})$-safe monotonicity is not satisfied and F is not (1,1)-safe implementable with $K = 1$. Hence, Maskin monotonicity does not imply $(\bar{d}, \bar{v})$-safe monotonicity.

Note: if $F(R')$ is modified to be $\{a, c\}$, then $(\bar{d}, \bar{v})$-safe monotonicity is satisfied, showing that Maskin monotonicity and $(\bar{d}, \bar{v})$-safe monotonicity have a nonempty intersection.

Example 4. Let $A = \{a, b, c\}$ and $K = \bar{d} = \bar{v} = 1$ and consider a social choice function given by $F(R) = a$ and $F(R') = c$, where R and R' are given by

R_1	R_2	R_3	R'_1	R'_2	R'_3
a	a	b	a, c	a, c	a, c
b	b	a	b	b	b
c	c	c			

Maskin monotonicity is not satisfied: a weakly rises in all players' ranking going from R to R', but $a \notin F(R')$. However, $(\bar{d}, \bar{v})$-safe monotonicity is trivially satisfied: for player 1 at R, the set of outcomes weakly less preferred to a that victimize at most one of the players in the set $\{2, 3\}$ is $\{a, b\}$. At R' that set is $\{a, c\}$. Because $\{a, b\} \not\subseteq \{a, c\}$, condition (1) of $(\bar{d}, \bar{v})$-safe monotonicity fails going from R to R'. Going from R' to R, a is weakly less preferred to c and victimizes no players at R', but it is strictly more preferred to c by all players at R. Thus, condition (1) also fails going from R' to R and $(\bar{d}, \bar{v})$-safe monotonicity holds trivially. This shows that $(\bar{d}, \bar{v})$-safe monotonicity does not imply Maskin monotonicity.

1.5.2 Safe Implementation Mechanism

We describe here the mechanism used to prove Theorem 1.3.5 and outline the steps of the proof. The same mechanism is also used in the proof of Theorem 1.3.7, but we add additional transfers to each of the rules; we discuss those transfers in section 1.5.3. The mechanism depends on whether $\bar{d} < \frac{n}{2}$ or $\frac{n}{2} \leq \bar{d} < n - 1$. We describe first the mechanism for the case of $\bar{d} < \frac{n}{2}$. Each player reports an outcome in A, a preference profile from $\mathcal{R}$, and an integer in $\{1, \ldots, n\}$. Hence, the message space for each player is $A \times \mathcal{R} \times \{1, \ldots, n\}$, and a message for player i is denoted by (a^i, R^i, z^i). The outcome function, $g(\cdot)$, is determined as a function of the message profile, m, as follows:

Rule 1 If m is such that all players agree on a report, say (a, R, z), and if $a \in F(R)$ then $g(m) = a$.

Rule 2 Suppose all players agree on (a, R, z) with $a \in F(R)$, except for one player i who reports $(a^i, R^i, z^i) \neq (a, R, z)$. If a^i is weakly less preferred than a for player i under R and if a^i does not victimize more than $\bar{v}$ players in the set $I \backslash \{i\}$ under R then $g(m) = a^i$; otherwise, $g(m) = a$.

Rule 3 Suppose $\bar{d} \geq 2$ and all players agree on (a, R, z), except for a set of players B who submit different reports and $1 < |B| \leq \bar{d}$. Let B' be the set of players in B who report an outcome that does not victimize more than $\bar{v}$ players in $I \backslash B$. If B' is not empty, then sort the players in B' from lowest to highest according their index, rename player $i \in B'$ by the new sorted order $q_i \in \{1, \ldots, |B'|\}$, and let $g(m) = a^{q_i^*}$, where $q_i^* = 1 + ((\sum_{i \in B'} z^i) \mod |B'|)$ is the q_i^*'th player in B'; otherwise $g(m) = a$.

Rule 4 In all other cases, $g(m) = a^j$ for $j = 1 + ((\sum_{i \in I} z^i) \mod n)$.

Rule 1 allows players to get any outcome in F if they all agree on the report submitted. Rule 2 allows a player to deviate from a consensus and receive any outcome if that outcome is weakly less preferred for that player and if that outcome does not victimize more than $\bar{v}$ of the others under the profile reported by the other players. Rule 3 allows a set of players of size at most $\bar{d}$ to deviate and receive any outcome as long as that outcome does not victimize more than $\bar{v}$ of the other players under the profile reported by the others. Finally, Rule 4 is a modulo game allowing any player to get any outcome they desire.

The proof must show that any outcome in F can be obtained as the outcome from a $(\bar{d}, \bar{v})$-SE of the mechanism above and any $(\bar{d}, \bar{v})$-SE of the mechanism maps into F. The first part is simple; suppose R is the true preference profile and suppose $a \in F(R)$. The profile m given by $m_i = (a, R, 1)$ is a $(\bar{d}, \bar{v})$-SE of the mechanism under Rule 1 that leads to $g(m) = a$. To see that m is a $(\bar{d}, \bar{v})$-SE, note that no player has an incentive to deviate due to Rule 2, and no set of players of size at most $\bar{d}$ can victimize more than $\bar{v}$ of the rest due to Rules 2 and 3. Hence, any outcome in F can be obtained as the outcome from a $(\bar{d}, \bar{v})$-SE of the mechanism.

An observation crucial to completing the second part of the proof is that no $(\bar{d}, \bar{v})$-SE of the mechanism above can exist under Rules 2, 3, or 4. This is because under these rules there is always a set of players of size $\bar{d}$ that can, via Rule 4, obtain any outcome in A.

By assumption, the preference profile satisfies $(\bar{d}, \bar{v})$-exposure so there is some outcome in A that victimizes more than $\bar{v}$ players in each set of players of size $n - \bar{d}$. This implies that a set of $\bar{d}$ players can victimize more than $\bar{v}$ of the rest, violating the definition of a $(\bar{d}, \bar{v})$-SE. Hence, the only $(\bar{d}, \bar{v})$-safe equilibria of the mechanism are under Rule 1. Now, for any $(\bar{d}, \bar{v})$-SE $(m_i = (a', R', z'))$ under Rule 1, Rules 2 and 3 imply that the conditions of $(\bar{d}, \bar{v})$-safe monotonicity are satisfied for R, which shows that $a' \in F(R)$.

If $\frac{n}{2} \leq \bar{d} < n - 1$ the proof remains unchanged except that the set of possible profiles is assumed to be $(\bar{d}, \bar{v}, n)$-similar and rule 3 of the mechanism is modified to be:

Rule 3 A group of players is called a *consensus group* if all the players in that group agree on a report and no player outside the group agrees with the players in the group. Suppose rule 2 does not apply and suppose there are p *consensus groups*, each of size $n - \bar{d}$ or more players.

- Case 1. If there is only one consensus group, call it B_1, then let $g(m) = a^q$ for $q = \max\{j \in I \backslash B_1\}$ if no element in the set $\{a^j : j \in I \backslash B_1\}$ victimizes more than $\bar{v}$ players in set B_1 under the profile reported by players in B_1; otherwise set $g(m) = \bar{a}^1$, where $\bar{a}^1$ is the outcome reported by players in B_1.
- Case 2. If there are multiple consensus groups then pick an outcome that does not victimize more than $\bar{v}$ players in any of the consensus groups under the profiles reported by the players in the groups. Such an outcome exists because of the assumption that the set of preference profiles is $(\bar{d}, \bar{v}, n)$-similar.

The function of the first case in the modified rule 3 is to allow a group of $\bar{d}$ deviators from rule 1 to achieve any outcome they desire as long as that outcome does not victimize more than $\bar{v}$ players of the rest. This is needed to allow the $(\bar{d}, \bar{v})$-safe monotonicity condition to work. The second case is for profiles where it is impossible for the designer to

distinguish between the deviators and the non-deviators. We use the assumption that the set of preference profiles is $(\bar{d}, \bar{v}, n)$-similar to allow the designer to pick an outcome that does not victimize more than $\bar{v}$ players in any consensus group of size $n - \bar{d}$ or more.

1.5.3 Double Implementation Mechanism

The mechanism we use in the proof of Theorem 1.3.7 is the same as the one used for safe implementation, except that we also add the following transfer scheme modification:

Transfer Scheme 1

Rule 1 All transfers are zero under rule 1.

Rule 2 Suppose all players agree on (a, R, z) with $a \in F(R)$, except for one player i who reports $(a', R', z') \neq (a, R, z)$. If $R' \neq R$ and $z' \neq z$ then all transfers are zero; otherwise player i is fined L.

Rule 3 Under this rule $\bar{d} \geq 2$ and all players agree on (a, R, z), except a set of players B who submit different reports and $1 < |B| \leq \bar{d}$. Impose a fine L on all players in B unless the following is true: $|B| = 2$, call them players i and j, with player i reporting (a', R', z') and player j reporting (a'', R'', z''), with $R' \neq R$, $z' \neq z$, $R'' = R$, and $z'' = z'$, then impose a transfer L from player i *to* player j.

Rule 4 Each player reporting some $z \in \{2, \ldots, n\}$ obtains a transfer L from each player reporting $z - 1$, and any player reporting $z = 1$ obtains a transfer L from any player reporting $z = n$.

The proof of Theorem 1.3.7 (see the appendix) shows that no Nash equilibria exist under rules 2,3, or 4 of the mechanism described in section 1.5.2 when Transfer Scheme 1 is imposed. For example, consider rule 2. Player i can choose whether to be fined or not

without affecting the outcome. If a Nash equilibrium exists under rule 2 then no players will be fined, but in that case the transfer scheme allows a "whistleblower" to reveal the unsafe situation by deviating to rule 3 and obtaining a transfer of size L from player i. Thus, no Nash equilibria exist under rule 2. The proof also shows that no Nash equilibria exist under rules 3 or 4.

1.6 Example: Hiring with Multi-Dimensional Quality

For the motivating example in section 1.2.2, consider safety with $\bar{d} = \bar{v} = K = 1$. With the environment described for the problem, the least skilled candidate will be the least preferred for at least $n-1$ members. Hence, protecting members from deviations by any one ($\bar{d} = 1$) that cause them to get their least favorite outcome ($K = 1$) and allowing at most 1 member to be victimized ($\bar{v} = 1$) results in protection from the least skilled candidate being selected through a one-member deviation.

We first show that the assumptions of Theorem 1.3.5 are satisfied.

Suppose $F(R) = a$ so that a is the candidate with the highest skill. Because there are at least 4 candidates in 4 different backgrounds and members rank candidates with backgrounds other than their own based on skill only, no member will have a as their least favorite outcome. Hence, the $\bar{v}$-safe outcome property is satisfied.

To show that $(\bar{d}, \bar{v})$-safe monotonicity holds, assume $F(R) = a$, the first condition of $(\bar{d}, \bar{v})$-safe monotonicity holds[11] moving from R to R', but (to show a contradiction) suppose $F(R') = b \neq a$. Because $\bar{v} = 1$, the first condition of $(\bar{d}, \bar{v})$-safe monotonicity implies that any candidate that is not the least skilled at R and is weakly less preferred than a by some member must remain weakly less preferred than a by that member at R' and also not be the least skilled at R'. Two implications follow. First, a is not the least skilled at R'. Second, it

[11] Only the first condition applies because $\bar{d} = 1$.

must be that b was the least skilled at R, otherwise it could have only fallen relative to a in ranking at R' and a would be ranked higher than b by at least $n-1$ members at R' causing b to not be chosen at R'. Because b is chosen at R', it must be more preferred than a by at least $n-1$ members. At R', the least skilled candidate is a candidate other than b or a, call it c. As mentioned earlier, the least skilled candidate will be the least preferred by at least $n-1$ members. Because a was the highest skilled at R, there must exist a member to whom c was less preferred than a at R and not the least preferred (because b is the lowest skilled candidate at R), but c becomes the least preferred at R', victimizing at least $n-1$ members. This violates the monotonicity condition leading to a contradiction.

Finally, we show that any preference profile arising from the environment described will satisfy $(\bar{d}, \bar{v})$-exposure. For any subset of members B of size $n-1$, the lowest skilled candidate must be ranked lowest by at least $n-2$ members in B. Because $n \geq 4$, the lowest skilled candidate victimizes at least 2 members in any group of size $n-1$, satisfying $(\bar{d}, \bar{v})$-exposure.

Because all the conditions of Theorem 1.3.5 are satisfied, we can directly use the mechanism given in the proof to $(\bar{d}, \bar{v})$-safe implement F. $(\bar{d}, \bar{v})$-safe implementation of F does not rule out the existence of undesirable Nash equilibria, but it implies that the highest skilled candidate can always be selected as an outcome from a safe profile, where deviations due to bad intentions, mistakes, or errors of judgment cannot cause disasters like selecting the lowest skilled candidate.

In this example, $\bar{v}$-safe-Nash monotonicity fails to hold, so double implementation is impossible[12] (this is true even if transfers are allowed, because we do not have precise cardi-

[12]To see this, consider two profiles R and R' with the property that candidate a is highest skilled at R, candidate b is the least skilled at R, and at R' all candidate skills remain the same, except that candidate b's skill becomes higher than all other candidates. Thus, $F(R) = a$ and $F(R') = b$. However, for any player i, any candidate that is less preferred than a at R and victimizes at most 1 player in $I \backslash \{i\}$ is less preferred than a for player i at R' because candidate skills do not change except for candidate b that victimizes at

nal preference information). However, double implementation is possible with a preference restriction.

Suppose we restrict the preferences as follows. If $R, R' \in \mathcal{R}$ and a and b are the highest and lowest skilled candidates at R, respectively, then either: (1) b is not the highest skilled at R' or (2) a is not the second highest skilled at R'. One way to justify such restriction is that the designer has some knowledge about relative ranking of candidates' skill levels – enough to rule out two candidates being the highest two skilled if one being the highest and one being the lowest skilled is also a possibility. This assumption would be restrictive if it had to always hold for double implementation, but it does not and the reason it is necessary in this example is because the social choice function we consider is a Condorcet social choice function. In section 1.6.1 we explain the restrictiveness of double implementation with Condorcet social choice functions more generally; the preference restrictions here precisely rule out preference profiles that cause a failure of double implementation as described in Theorem 1.6.1.[13] If we allow such a restriction on preferences then $\bar{v}$-safe-Nash monotonicity is satisfied[14], and double implementation is possible using the mechanism in the proof of Theorem 1.3.7.

Double implementation in safe and Nash equilibria implies that safe implementation holds, and no unsafe Nash equilibria result in undesirable candidates. If, in addition, we are willing to assume that truthful equilibria, that are payoff-equivalent to untruthful ones,

---

least 3 players. Thus, $\bar{v}$-safe-Nash monotonicity fails, because $a \notin F(R')$.

[13] Note that the preference restrictions in Theorem 1.3.1 for strong safe implementation with Nash behavior and no transfers must hold regardless of the form of the social choice function. In this example, even with the preference restriction we assume, strong safe implementation with Nash behavior and no transfers is impossible.

[14] To see this, consider two profiles R and R' with $F(R) = a$. Suppose the least skilled candidate at R is b. For each member, the set of outcomes weakly less preferred to a at R that victimize at most $\bar{v} = 1$ of the other members is simply $A\backslash\{b\}$. Suppose condition (1) of the preference restriction holds. By assumption, b is not the highest skilled at R', and if a weakly rises in the ranking of each player among the set $A\backslash\{b\}$ going from R to R', then a remains the highest skilled at R' and $F(R') = a$. If condition (2) holds then there was a preference reversal for some player between a and some outcome in $A\backslash\{b\}$ going from R to R', so $\bar{v}$-safe-Nash monotonicity imposes no restrictions and holds trivially.

are focal then players will *always* be safe in equilibrium and the mechanism in the proof of Theorem 1.3.7 achieves the same objectives of strong safe implementation.

1.6.1 Double Implementation of Condorcet Social Choice Functions

Several papers discussed the possibility of Nash implementing social choice rules that select Condorcet winners. Maskin (1999) shows that Nash implementation is possible when Condorcet winners exist at every preference profile; Healy and Peress (2013) show that Maskin monotonicity fails if preference profiles with no Condorcet winner are allowed. In this section we show that social choice rules that select a strict Condorcet winner, whenever it exists, cannot, in general, be double implemented in Nash and safe equilibria without preference restrictions, even if all admissible preference profiles have a strict Condorcet winner.

Definition 1.6.1 (**Condorcet Social Choice Function**)**.** F is a Condorcet social choice function if $F(R) = a$ whenever $|\{i \in I : aP_ib\}| > |\{i \in I : bP_ia\}|$ for all $b \in A$.

We say that a is a Condorcet winner if $|\{i \in I : aP_ib\}| > |\{i \in I : bP_ia\}|$ for all $b \in A$.

Theorem 1.6.1. Suppose F is a Condorcet social choice function and there are two preference profiles profiles $R, R' \in \mathcal{R}$ and outcomes $a, b \in A$ such that:

1. a is a Condorcet winner at R.
2. b victimizes at least $\bar{v} + 2$ players at R.
3. b is a Condorcet winner at R'.
4. For all $i \in I$ and $c \in A$, if aR_ic and c victimizes at most $\bar{v}$ players in $I\backslash\{i\}$ at R then aR'_ic.

Then F is not double implementable in $(\bar{d}, \bar{v})$-safe and Nash equilibria.

The intuition of the proof is simple. If a is a Condorcet winner at R then safe implementation implies the existence of a safe equilibrium with outcome a. From this safe equilibrium, no player can obtain outcomes that victimize more than $\bar{v}$ of the others. If one of those outcomes, say b, becomes the Condorcet winner at R' and the ranking of a relative to other outcomes is maintained, then a safe equilibrium with outcome a at R must be a Nash equilibrium at R'. But a is not chosen at R', because b is a strict Condorcet winner at R', leading to a contradiction.

Thus, achieving safety with Condorcet social choice functions requires preference restrictions. In environments where such restrictions are plausible double implementation in $(\bar{d}, \bar{v})$-safe and Nash equilibria may be possible, in which case strong safe implementation is also possible with mild behavioral assumptions, as in our hiring example.

1.7 Discussion

Safety can be incorporated in mechanism design in various ways. This section discusses some important aspects of our safety concept and our notions of implementation.

1.7.1 Discussion of the Safety Concept

Our notion of safety is intuitive, it simply imposes restrictions on how much players can harm others. To make our definition more realistic and flexible, we allow for groups of players to attempt to cause harm and for some players to be victimized. In our notion of victimization, we also allow for different thresholds of safety by letting the designer choose the relative position in a player's ranking that defines a victim.

Our safety concept, as stated in definition 1.2.3, is very general. It applies to virtually any implementation or mechanism design problem. To illustrate this point in an auction setting, suppose we define a victim as *"a player with the highest private value who pays some*

amount of money but loses an auction due to irrational behavior by other players". Using this definition, it is easy to see that an auction with a reserve price is "safer" than one with an entry fee. Even though both auctions may be theoretically equivalent in some cases, an irrational player in an auction with an entry fee can submit high bids and win the auction causing the player who should have won to lose and pay the entry fee, whereas an irrational player in a reserve price auction cannot victimize others. For another example, a "victim" in a sponsored online ad auction could be a firm who is forced to pay above $\$K$ amount by fraud; with such a definition we can examine the safety properties of equilibria of different mechanisms, for instance, a cost-per-action mechanism may be safer than a cost-per-click mechanism where a competitor can attempt click-fraud.

Ordinality and finiteness

We stress that, although our definition of a victim relied on the finiteness of the set of outcomes, it extends naturally to infinite sets of outcomes: we can simply consider the K^{th} quantile instead of the least K preferred outcomes when defining a victim. Doing so would extend our results naturally to problems with an infinite set of outcomes.

Our definition of a victim uses ordinal preference information, making no use of cardinal information about the preferences of the players. This assumption may be violated in some settings of interest in mechanism design. However, many important subfields of mechanism design, such as voting and matching, rely on ordinal preferences (even though cardinal preferences may exist). To address safety in such settings, it is necessary to adopt a notion of safety that naturally accommodates ordinal preferences on finite sets of outcomes.

Our notion of safety can also be used as a basis for safety criteria in more general settings. For example, if utility functions are introduced, definition 1.2.3 can remain unchanged with the main change being in definition 1.2.1 of a victim; a victim can be defined

to be a player who obtains an outcome leading to a utility level below a certain threshold. The utility threshold will, in general, be different for every player and be dependent on the precise utility function.

Finally, we note that it is common for policy makers to assess a mechanism's success (or failure) using ordinal welfare criteria very similar to the one used in this paper. One common criterion is the rank distribution, or how many players get their top choice, how many get their second choice, ..., and how many get their last choice. Some examples are the San Francisco Unified School District, Teach for America, and Harvard Business School (Featherstone (2011)). Our safety criterion, along with our definition of a victim, is equivalent to using the rank distribution as a welfare measure–precisely what often occurs in many practical problems.

Thus, our safety criterion and notion of victimization address a large and important body of problems that are most naturally studied using ordinal welfare criteria. Our safety notion also opens new avenues for further research by providing a basis for developments of safety criteria in other settings.

Relative vs. Absolute Notions of Victimization

With our notion of safety, victims may be present at a "safe" message profile. An alternate definition can specify $\bar{v}$ to be the maximum number of victims from a deviation *relative* to the number of victims in equilibrium. Such a definition can easily be accommodated by simply changing the notion of victimization, but we do not use it precisely because we do not want many victims to be present at a "safe" message profile. To illustrate the drawbacks of a relative notion of victimization, consider two message profiles: m has no victims, and a deviation from m can lead to at most two victims, but m' has numerous victims, and a deviation from m' can lead to at most one more victim; a relative notion of safety may well

consider m' to be "safer" than m, but our notion of safety would classify m' to be no safer than m. Also, a relative notion of safety can allow a designer to make mechanisms seem more "safe" by simply letting the number of victims be high at all profiles, thus guaranteeing that profiles are not too bad, relatively.

Another possibility is to define a victim to be a player who does worse off (more generally: by a certain amount) from the equilibrium payoff, as in the t-immunity solution concept (Halpern (2008)). Again, a change in our notion of victimization can accommodate this definition, but this definition has some drawbacks. First, it does not allow for a clear statement of safety in terms of payoffs unless we know the payoffs of players at equilibrium, and mechanisms generally have multiple equilibria with different payoffs. Thus, a "safe" mechanism with this notion of safety may very well classify a payoff vector for agents to be *less* desirable than another vector of *strictly lower* payoffs, simply because the former may be *relatively* further from one equilibrium than the latter is from another equilibrium. Second, this notion of victimization may also artificially lead low-payoff equilibria to seem more safe. To illustrate some of the difficulties with this definition, consider a matching mechanism with two equilibria: one where all players get their top preferred outcome, but where a deviation can cause at most one non-deviator to get their second highest preferred outcome, and another equilibrium where all players get their least preferred outcome. A notion of victimization that focuses on payoffs of players *relative* to their equilibrium payoffs may consider the first equilibrium unsafe, but the second safe. In contrast, our notion of victimization would classify the second equilibrium to be no safer than the first.

1.7.2 Safe Implementation Discussion

Is safe implementation easier than Nash implementation?

There is a common notion in the literature (e.g. Jackson (2001)) that a stronger (i.e.

more refined) solution concept allows more social choice rules to be implemented, and thus makes the implementation problem "easier". This notion may be true in some cases, but not in general. Particularly, it is not true in the case of implementation in $(\bar{d}, \bar{v})$-safe equilibrium, a more refined solution concept than Nash equilibrium.

This can easily be seen using the simple social choice function in Example 3 in section 1.5.1. As the example shows, the social choice function satisfies Maskin monotonicity. It also satisfies no veto power and $n \geq 3$, so it is Nash implementable by Theorem 3 in Maskin (1999). However, the example shows that it does not satisfy $(\bar{d}, \bar{v})$-safe monotonicity, a necessary condition for $(\bar{d}, \bar{v})$-safe implementation. Thus, a solution concept that is a Nash refinement does not necessarily allow more social choice rules to be implemented.

Existence of $(\bar{d}, \bar{v})$-Safe Implementable and Double Implementable Social Choice Rules

Theorem 1.4.2 is reminiscent of the Gibbard-Satterthwaite Theorem (Gibbard (1973) and Satterthwaite (1975)), which says that, under some conditions, a social choice function is strategy-proof if and only if it is dictatorial. Dictatorial social choice functions cannot be safe with a rich preference space (one player can always get any outcome they prefer), and strategy-proofness implies players have no incentive to deviate from reporting their true preferences. Theorem 1.4.2 says something similar: if a social choice function *is* safe (not dictatorial), then players will, in some preference profiles, have an incentive to deviate at any safe message profile of any implementing mechanism.

Though seemingly related, the Gibbard-Satterthwaite Theorem differs from Theorem 1.4.2 in important ways; the most important of these is perhaps the space of mechanisms each of the theorems operates on.[15] Theorem 1.4.2 says that no mechanism can implement

[15]The results also concern different solution concepts: Gibbard-Satterthwaite Theorem is based on dominant messages, but the solution concept in Theorem 1.4.2 is fundamentally based on Nash equilibrium.

a social choice function with safety guarantees in the space of *all* mechanisms (a very large space!). On the other hand, the Gibbard-Satterthwaite Theorem is only concerned with one particular mechanism: one where players report their preferences and the outcome is the social choice function at the reported preferences.

If dictatorial social choice functions are not safe with rich preference spaces, should we worry that the Gibbard-Satterthwaite Theorem implies non-existence of $(\bar{d}, \bar{v})$-safe implementable social choice rules? No; again, the Gibbard-Satterthwaite Theorem considers implementation using only one mechanism (truth-telling) and there may be other mechanisms that make implementation possible.

The necessary and sufficient conditions in Theorem 1.3.5 are not too restrictive. Simple social choice rules exist that are $(\bar{d}, \bar{v})$-safe implementable, such as the one in example 4: the example shows that $(\bar{d}, \bar{v})$-safe monotonicity holds trivially; the $\bar{v}$-safe outcome property holds (a victimizes no players at R and c victimizes no players at R'); and $(\bar{d}, \bar{v})$-exposure is satisfied (c victimizes all members in any group of two players at R and b victimizes all members in any group of two players at R'). Because $n = 3$ and $\bar{d} < \frac{n}{2}$, Theorem 1.3.5 implies that the social choice function in this example is $(\bar{d}, \bar{v})$-safe implementable, showing the existence of simple $(\bar{d}, \bar{v})$-safe implementable social choice rules. Also, the SCF in the hiring example of section 1.2.2 may or may not be dictatorial depending on the bonus premiums members assign candidates in their fields, but section 1.6 shows that this SCF is $(\bar{d}, \bar{v})$-safe implementable regardless of the bonus premiums of the members. For double implementation in safe and Nash equilibria, section 1.6 shows that the SCF in our hiring example is double implementable with the appropriate preference restrictions, showing that natural social choice rules exist that are double implementable in safe and Nash equilibria.

1.8 Further Research and Conclusions

1.8.1 Further Research

Our victimization concept (definition 1.2.1) allowed us to apply our safety notion to the standard implementation problem, but, as mentioned before, our safety concept (definition 1.2.3) is much more general and can apply to virtually any mechanism design problem. A systematic study of safety (using definition 1.2.3) in various applications of mechanism design, perhaps with flexible context-dependent definitions of victimization, would be very useful in understanding safety more broadly. It would be interesting to examine and potentially improve the safety properties of common mechanisms in various applications, such as auctions, matching, voting, and market design.

In reality, society is often concerned with multiple safety objectives simultaneously, such as protection from irrational and dangerous individuals and larger, more organized groups deciding to harm others (e.g. with discriminatory policies). Our notion of safety imposes just one restriction on the victims function: $V_m(\bar{d}) \leq \bar{v}$, where m is the equilibrium profile.[16] This can be extended in various ways: more restrictions at multiple points of the victims function can reflect more complicated safety requirements; the victims function at non-equilibrium profiles may also be restricted to improve robustness; and finally, multiple definitions of victimization (e.g. multiple K's or fundamentally different definitions of victimizations) can be used, leading to multiple victims functions, each with its own set of $\bar{d}$'s and $\bar{v}$'s.

Furthermore, in reality we may not always be able to eliminate disasters, but we can potentially make them unlikely. Adopting a probabilistic approach to safety would be a fruitful area for future research. Our victims function is concerned only with the worst case scenario at a given $\bar{d}$, but a probabilistic approach can examine the whole distribution of the

[16] If $\bar{v} < n - \bar{d}$ then $V_m(\bar{d}) \leq \bar{v}$ also implies that $V(\beta) \leq \bar{v}$ for any $\beta < \bar{d}$.

number of victims.

Several papers in the implementation theory literature explored solution concepts other than Nash equilibrium and different forms of implementation. For example, Abreu and Matsushima (1992) show that virtual implementation in iteratively undominated messages is easier to achieve, and that it can be done with simpler mechanisms. In this paper we studied safety in conjunction with the Nash equilibrium as a solution concept, and we used standard notions of implementation. The solution concept is not merely a decision under the designer's control, but should instead be an accurate prediction of players' behavior. Predictions may depend on the set of players, and on the context of the implementation problem, so studying the constraints implied by safety in conjunction with an arbitrary solution concept would be an interesting question for future research. Also, considering other notions of implementation (e.g. virtual implementation) and the extent to which using simple mechanisms limits safety are interesting possibilities for future research.

1.8.2 Conclusions

For system designers and policy makers safety is of utmost importance. Despite this fact, many mechanisms in the implementation theory and mechanism design literature have *no* safety properties: for most mechanisms in the literature, an unexpected deviation by just one player can cause widespread disasters. Introducing safety requirements can change the set of implementable social choice rules, so it is important to address safety rigorously and examine the constraints implied by safety.

This paper bridges the gap between safety and the mechanism design and implementation theory literature. We introduced safety as robustness of the welfare levels of the players to actions of others, and we characterized social choice rules that can be implemented in a safe way. We introduced three notions of safe implementation: (1) strong $(\bar{d}, \bar{v})$-safe implementation is Nash implementation with mechanisms where all Nash equilibria are $(\bar{d}, \bar{v})$-safe;

(2) $(\bar{d}, \bar{v})$-safe implementation, which is a weaker notion of implementation that allows for some undesirable Nash equilibria; and (3) Double implementation in $(\bar{d}, \bar{v})$-safe and Nash equilibria, which maintains safe implementation and ensures that all Nash equilibria lead to desirable outcomes. We provided necessary conditions for $(\bar{d}, \bar{v})$-safe implementation, and we showed that they are sufficient under some conditions on preferences. We also provided conditions that are both necessary and sufficient for double implementation when transfers are allowed.

Although strong safe implementation, the most direct approach to designing safe mechanisms, is impossible in many environments, double implementation ensures safety with mild behavioral assumptions unrelated to safety. We proved a "no guarantees" result showing the impossibility of implementation in $(n-1, 0)$-safe equilibria in a wide range of environments. Finally, we discussed the connection between safety, freedom, and efficiency showing that, in general, safety restricts players' positive freedom and reduces efficiency.

Chapter 2

Criminals' Response to Changing Crime Profitability

Chapter Summary

How do criminals respond to changes in profitability of crime? This is an important question, relevant for understanding the effectiveness of various crime fighting policies, such as reducing demand for illegal goods, disrupting black markets, and any policy that makes crime more costly by eliminating cheaper avenues for criminals. The crime literature has not sufficiently addressed this question, partly because it is difficult to find a reliable measure of crime's profitability. Using confidential data on cargo theft, we match historical prices of various goods with their thefts to find the price elasticity of theft. Conditioning on good, state, and time fixed effects, we estimate the price elasticity of theft to be about 1.002 in the short run and 1.192 over a seven-month horizon. We show how the main empirical results can arise in a simple model of crime with strategic interactions.

2.1 Introduction

Crime is an important costly social phenomenon. The U.S. spends more than $260 billion on the justice system annually, $80 billion of which are spent on corrections (Kyckelhahn and Martin (2013)). Other, less quantifiable costs of crime include damages to victims and their families; costs to society through spending on security; economic activity disruptions; and a large incarcerated population (1 in every 104 Americans (Glaze (2011))) that cannot work or support their, often at-risk, families.

With a desire to reduce crime, the crime literature has focused on how different factors affect crime levels. Becker (1968) identified three factors influencing a rational agent's decision to commit crime: (1) expected gain from committing crime (what we call crime profitability), (2) expected punishment, and (3) the agent's outside option. Each of the three factors suggests a different set of policies that differ widely in their mode of operation, costs, and potential to reduce crime. Understanding which of the policies is the most effective given their costs hinges on an important question: how do criminals respond to incentives related to each of the three factors? Two of the three main factors involved in the crime decision, punishment and outside options, have been extensively studied in relation to crime. The literature on how criminals respond to changes in the gain from committing crime, perhaps the factor driving crime levels most directly, is surprisingly small, possibly due to data limitations.

In this paper we study how criminals respond to changes in profitability of crime. We use confidential data on cargo theft from FreightWatch International, an international logistics security firm. The data is unique in that it records the precise nature of the cargo stolen, allowing us to examine the relationship between prices of specific goods (e.g. apples, copper, etc...) and their thefts. By matching prices of goods to their thefts, we estimate the price elasticity of theft to be about 1.002 in the short run and 1.192 over a seven-month horizon. We examine substitution of criminals between goods, and we find no strong evidence of a substitution effect. We analyze the heterogeneity in the price elasticity of thefts across goods and across locations. We characterize the most responsive (to prices) locations for theft of some major goods in the data, and we present a simple model explaining the heterogeneity in responsiveness by spatial heterogeneity in connections to black markets. Finally, we show how the main empirical results can be explained in a simple model of crime with strategic interactions between owners and (potential) thieves.

This paper is organized as follows. Section 2.2 reviews the related literature. Section

2.3 discusses the data and presents descriptive statistics. Section 2.4 presents the main results. Section 2.5 discusses heterogeneity of the results by good and by location. Section 2.6 presents a simple theoretical model of crime with strategic interactions that can explain the main results. Section 2.7 concludes and discusses further work.

2.2 Related Literature

As we mentioned previously, Becker (1968) identified three main factors involved in the decision to commit crime: expected gain from committing crime, expected punishment, outside options. We discuss the literature related to the three main factors next.

Punishment and Policing

Since the work of Becker (1968), a large literature tried to test and quantify the main relationships in the economic model of crime with particular emphasis on how policing and punishment affect crime levels. Though the literature has a long history, it has had various persistent difficulties. Until recently, good data on crime was lacking, so much of the literature used aggregate crime data. General problems include endogeneity of policing/deterrence policies, differentiating between deterrence and incapacitation, and simultaneity. Cameron (1988) reviews the early literature. Levitt (1998) shows that the majority of the observed negative relationship between arrests and property crime is due to a deterrence effect. Using monthly observations on crime in New York City for nearly 30 years, Corman and Mocan (2000) find strong evidence supporting deterrence.

More recent literature uses exogenous variation or quasi-experiments to estimate the effect of punishment and policing on crime. Di Tella and Schargrodsky (2004) find a large, but very localized, deterrent effect of observable police on crime using an exogenous shock to allocation of police after a terrorist attack. Bar-Ilan and Sacerdote (2004) use a series of

experiments changing traffic fines in Israel and San Francisco and estimate a large negative effect of fine increases on red light running. Lee and McCrary (2009) exploit the discontinuity in U.S. sanctions at age 18 and find a small decline in the log-odds of offending at age 18. Drago et al. (2009) use a unique exogenous change in expected punishment at the individual level due to a clemency Bill in Italy and estimate that increasing the expected sentence by 50 percent reduces recidivism rates by about 35 percent in 7 months.

McCormick and Tollison (1984) follow a unique approach; they utilize a quasi experiment in the Atlantic Coast Conference (ACC) basketball tournament and find a significant negative effect of the number of referees on the number of fouls called, but Hutchinson and Yates (2007) find that the results of McCormick and Tollison (1984) are driven entirely by data error. Following McCormick and Tollison (1984), Levitt (2002) and Heckelman and Yates (2003) estimate the effect of referees on infractions in the National Hockey League (NHL) exploiting a randomization of the number of referees in the 1998-2000 seasons. Levitt (2002) finds inconclusive evidence from the experiment in the 1998-1999 season, and Heckelman and Yates (2003) find that increasing the number of referees did not significantly reduce infractions in the 1999-2000 season.

Outside Options

Raphael and Winter-Ebmer (2001) and Donohue and Levitt (2001) find that unemployment has a significant effect on property crime. For the declining crime trend in the 90's, Levitt (2004) argues it was not because of the strong economy, demographics, police strategies, gun laws, concealed weapon laws, or the death penalties; instead the decline was affected mainly by the number of police, prison population, decline of the crack epidemic, and legalization of abortion. Freeman (1999) shows that unemployment is not an overwhelming determinant of crime, even though it is related, and that more inequality is associated with

more crime. Zhang (1997) finds that welfare payments reduce crime and Foley (2011) finds that the timing of welfare payments affects criminal activity.

Rewards from Crime

The literature on how crime levels respond to changing rewards from crime is small. To our knowledge, only two papers examine the effect of prices on thefts. Reilly and Witt (2008) show that declining real prices of audio-visual goods contributed to the decline in burglaries in the UK. Identification in that case is complicated by the fact that prices of audio-visual goods have been declining steadily over time with no interesting variation. Sidebottom et al. (2011) show that prices of copper are correlated with copper cable theft in the UK. Both papers focus on the UK, the results do not necessarily generalize to the US because the punishment systems are different so criminals may respond differently to price incentives.

2.3 Data and Descriptive Statistics

Transit of goods effectively through the supply chain is crucial for the functioning of the economy as a whole. Cargo theft is a major threat for supply chains. Cargo trucks often carry goods worth hundreds of thousands or millions of dollars. Criminals often operate in specialized theft gangs to steal cargo trucks and sell the stolen goods.

We use confidential data from FreightWatch International, an international logistics security provider. FreightWatch offers security services for cargo in transportation via various methods, such as real-time covert electronic tracking. Most of the thefts in the data are not from FreightWatch clients, but rather from external sources like security councils and law enforcement agencies. Security councils are organizations comprised of industry professionals, law enforcement and government agencies, cargo insurers, carriers, and risk

management professionals. They formed because of the high value nature of cargo thefts and the effectiveness of information sharing in mitigating the losses. When cargo is stolen, members of a security coalition can report it, leading to quick dissemination of information via "theft alerts" to all members of the coalition, and increasing the likelihood of a safe recovery of the cargo. Of the thefts in the data with a known external source of the information, 79.6% are from security councils, 17.6% are from law enforcement agencies, and 2.8% are from news sources.[1]

The data contains information on 5865 cargo theft incidents in the U.S. and Canada spanning more than 7 years from January 2006 to June 2013. Information about each theft can be classified into three main categories: date, location, and theft information; we explain each in more detail next.

Date Information

Entries record the date on which the theft happened and a small percentage contain time of day information. About 11% of the data contain a date range instead of one date; the date range is typically a few days. For analysis of monthly counts, entries with ranges spanning more than one month will be dropped. Figure 2.1 shows the total monthly counts of cargo theft from January 2006 to December 2012 and a cubic function of time as a trend. FreightWatch started recording data in January of 2006, which may partially explain the initially low theft volume.

[1] Source information was only recorded for most thefts starting in January 2013.

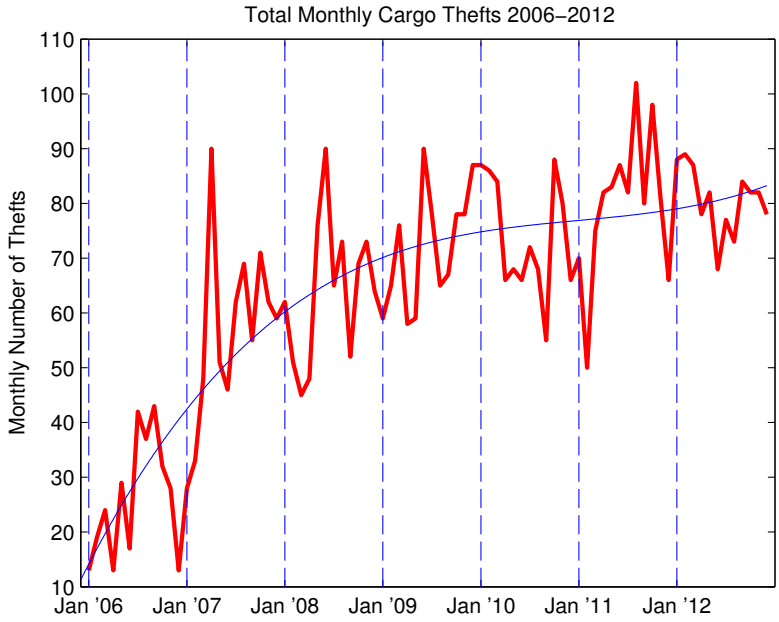

Figure 2.1: Total Monthly Cargo Theft Counts

Location Information

Entries contain city and state where the theft occurred. About 28% of the thefts also record an address. Figure 2.2 shows the cities where cargo thefts occurred.

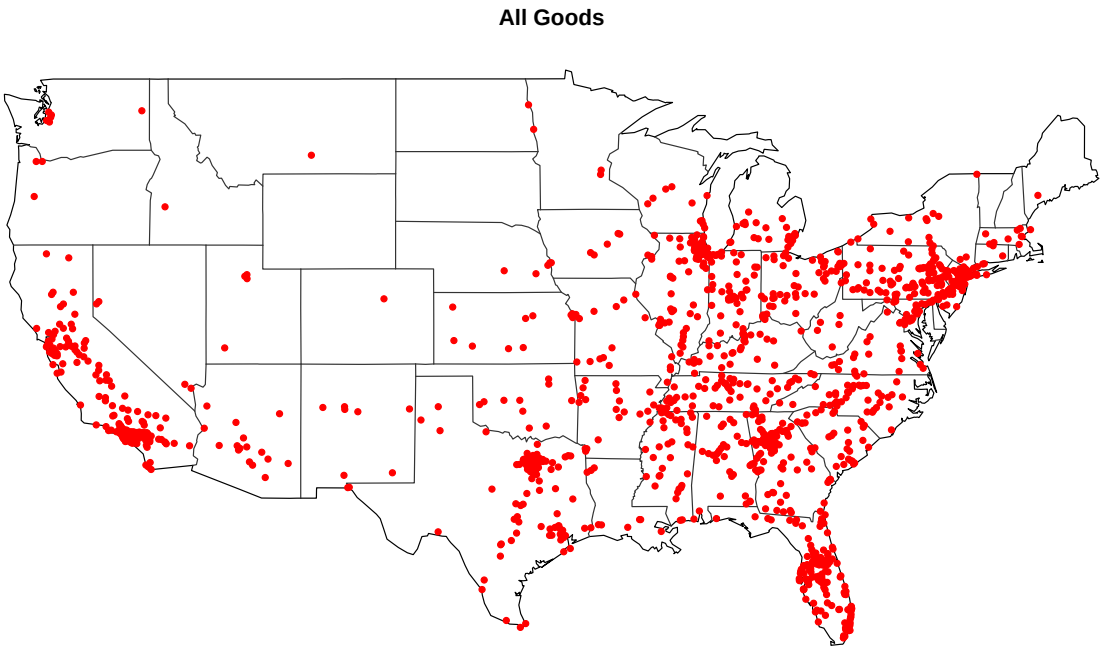

Figure 2.2: Cargo Theft Cities

The following table lists the top 10 states where cargo thefts were recorded.

State	Thefts	% of Total
CA	1495	25.5%
FL	807	13.8%
TX	687	11.7%
NJ	441	7.52%
GA	412	7.02%
IL	269	4.59%
TN	214	3.65%
PA	158	2.69%
IN	103	1.76%
NC	90	1.53%
Top 10	4676	79.7%

Table 2.1: Cargo Thefts by State

Additionally, about 50% of the theft records contain a description of the location of the theft, indicating whether the theft occurred at secured parking, unsecured parking, a truck stop, a warehouse, etc...

Theft Information

The type of cargo that was stolen is recorded for each theft. Cargo type is recorded in two fields: the first records the type of cargo in one of 14 major types of goods (e.g. electronics, food/drinks, etc...), and the second field records a more specific description of the cargo. To make this information available for analysis, I assigned codes for each of the goods occurring frequently and hand-coded each theft. More than one hundred different goods were identified and coded in the data. Appendix B.1 contains a list of all the goods that were coded.

About 58% of the records contain the dollar value of the stolen goods. This value ranges from \$160 to \$76 million; the mean value is \$354,650 and the median is \$100,000. The total for the records containing the value is more than \$1.2 billion. Figure 2.3 shows the empirical cdf of the value of the stolen goods. We note the apparent lack of truncation in the value associated with a reported theft.

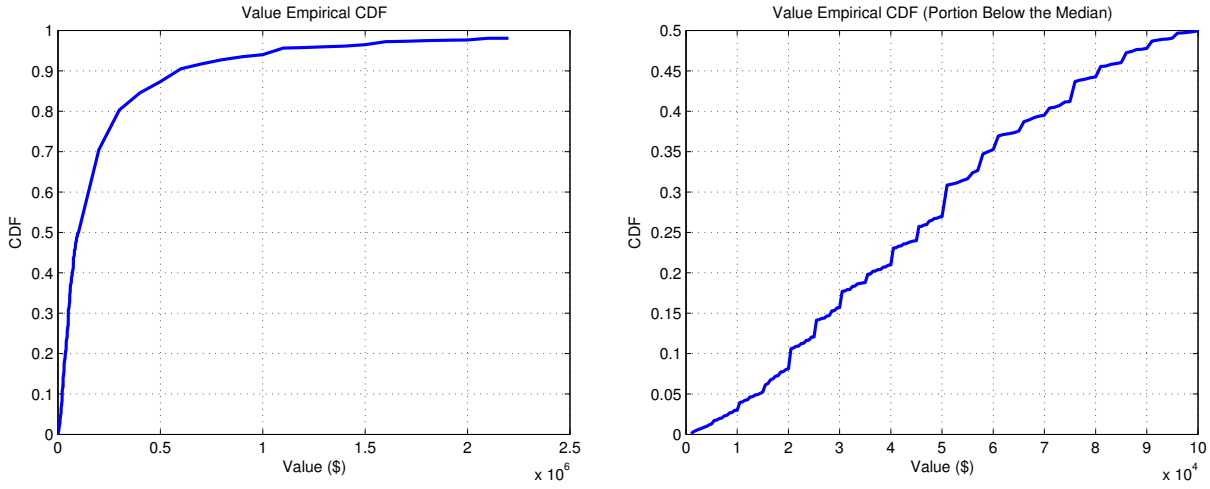

Figure 2.3: Stolen Value Empirical CDF

Finally, the data also contains information about how the theft occurred (e.g. theft of trailer, facility burglary, deceptive pickup, hijacking, etc...).

Other Data Sources, Price-Good Matching, and Deletions

Price data was obtained from the Bureau of Labor Statistics (BLS) Consumer Price Indexes (CPI) and Producer Price Indexes (PPI); except for section 2.4.1 all prices were deflated by the average CPI for all items (series ID CUUR0000SA0). Data on monthly highway miles traveled in each state was collected from various issues of the Federal Highway Administration (FHWA) monthly Traffic Volume Trends reports, Table 5. Data on rankings of U.S. ports by container traffic for 2009 was obtained from the U.S. Census Bureau, Statistical Abstract of the United States: 2012.

To check the robustness of our results, we needed data on shipments for various goods. Because domestic shipments are mostly composed of domestic production and imports, in some cases we use data on production and/or imports to approximate domestic shipments.

We collected monthly data on U.S. shipments of steel from various press releases of the American Iron and Steel Institute (AISI), and imports of steel were obtained from the U.S. Census Bureau International Trade Data; domestic and import shipments were summed to get total steel shipments. The U.S. Census Bureau's Manufacturer's Shipments, Inventories, and Orders (M3) Survey was used to obtain value of shipments for aluminum and tobacco (obtained from the Federal Reserve Bank of St. Louis (FRED), series AANMVS and A12BVS); the value series were then divided by the price series from the BLS to obtain shipments. For beer, we used monthly data on production and imports from the Beer Institute (Brewer's Almanac and the Beer Institute Industry Update), which were added together to get shipments. United States Department of Agriculture (USDA) data shows that U.S. imports of meat and poultry are a small percentage (less than 10, and much less for pork and poultry) of domestic consumption, so shipment data for beef, pork, and poultry was approximated by monthly U.S. production data from the USDA Economic Research Service. Shipments of apples, bananas, oranges, tomatoes, and potatoes were collected from various issues of the USDA Agricultural Marketing Service's Fresh Fruit and Vegetable Shipments reports, which include domestic shipments and imports; shipments of bananas were approximated by imports, as imports account for more than 99% of the U.S. banana consumption.

Price data from the BLS was obtained for 41 goods that were present in the data. Prices were then matched with the appropriate goods as described in appendix B.2. Appendices B.1 and B.2 contain lists of the good codes, BLS price series IDs, and the matchings between the price series and good codes.

We aggregated thefts by good at a monthly frequency; thefts with a date range spanning two or more months were not counted. For the empirical analysis, we used thefts occurring only up to December 2012, because information about thefts from some external sources can take several months to be recorded. Finally, two thefts were not assigned codes because the type of cargo stolen was missing. After the deletions, there were 2,435 unique

incidents in which at least one good stolen was among the 41 goods we obtained price data for from the BLS. In addition, we eliminated thefts with unknown locations and thefts that happened in Canada for regressions utilizing state-level data, resulting in 2,319 thefts that were used in analyses utilizing location information.

2.4 Main Results

2.4.1 Nonparametric Evidence

This section shows nonparametric graphical evidence that prices are correlated with thefts. We also use a simple correlation placebo test to show that the positive correlation between thefts and prices is significantly higher than what would be expected if goods and prices were randomly matched.

The following plots show smoothed thefts and prices over time for six goods: copper, steel, all metals, food and drinks, fuel, and appliances. The dash-dot (blue) lines are prices measured along the left axis and the solid (red) lines are thefts measured along the right axis. Time is on the horizontal axis with the vertical dashed (blue) lines separating the years. Prices are nominal CPI prices from the BLS measured relative to 1982-84 dollars with 1982-84 dollars = 100.

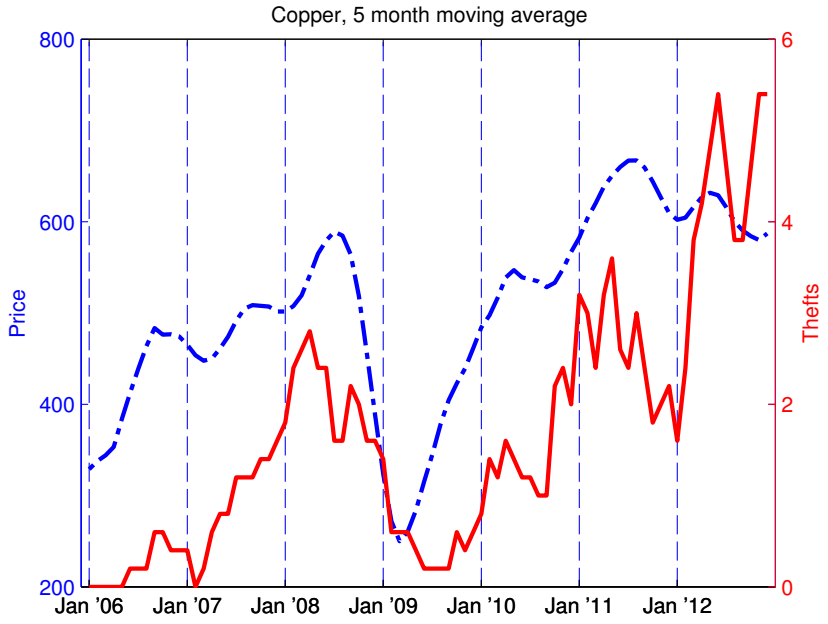

Figure 2.4: Copper Thefts

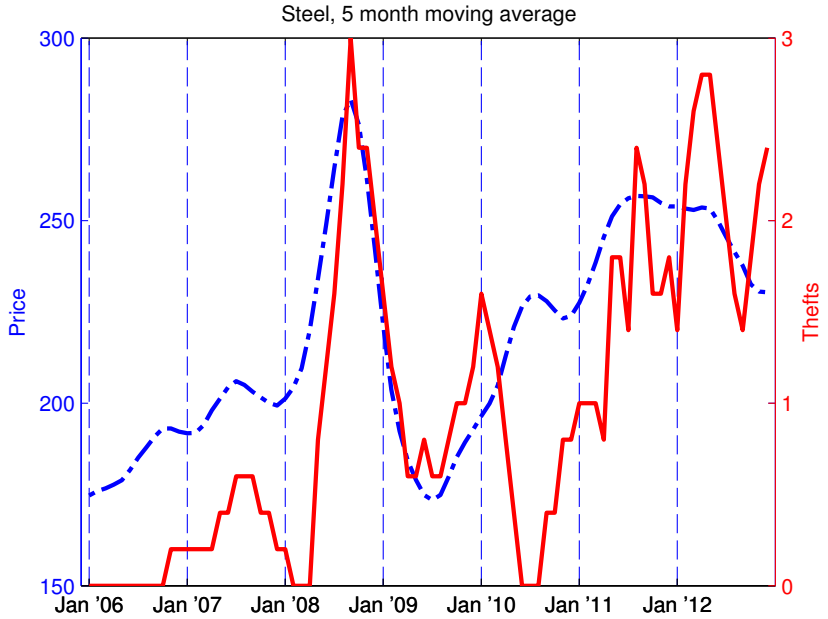

Figure 2.5: Steel Thefts

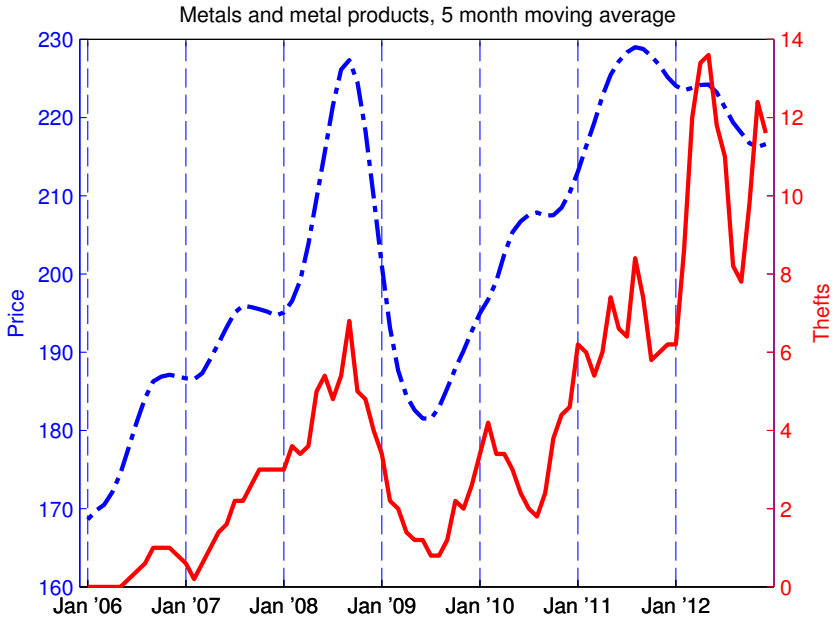

Figure 2.6: Metal Thefts

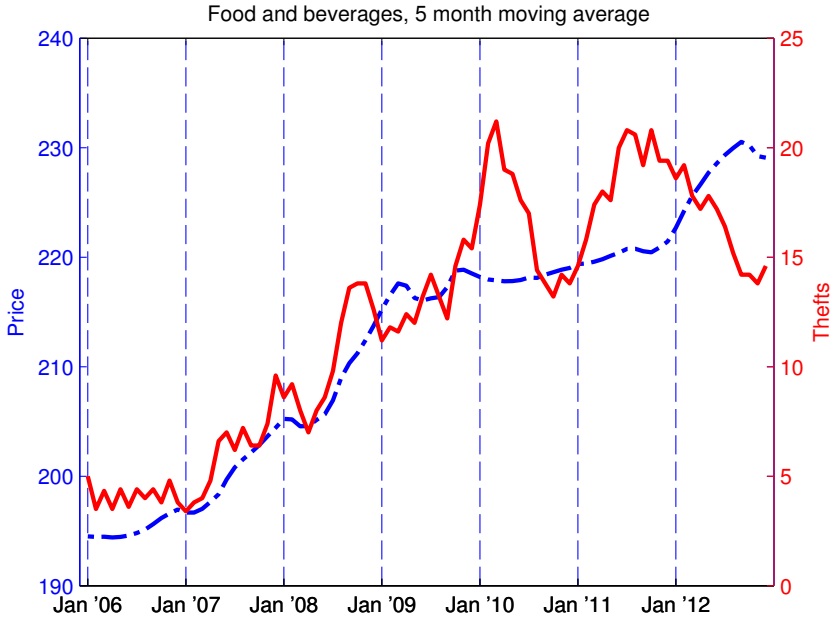

Figure 2.7: Food and Beverages Thefts

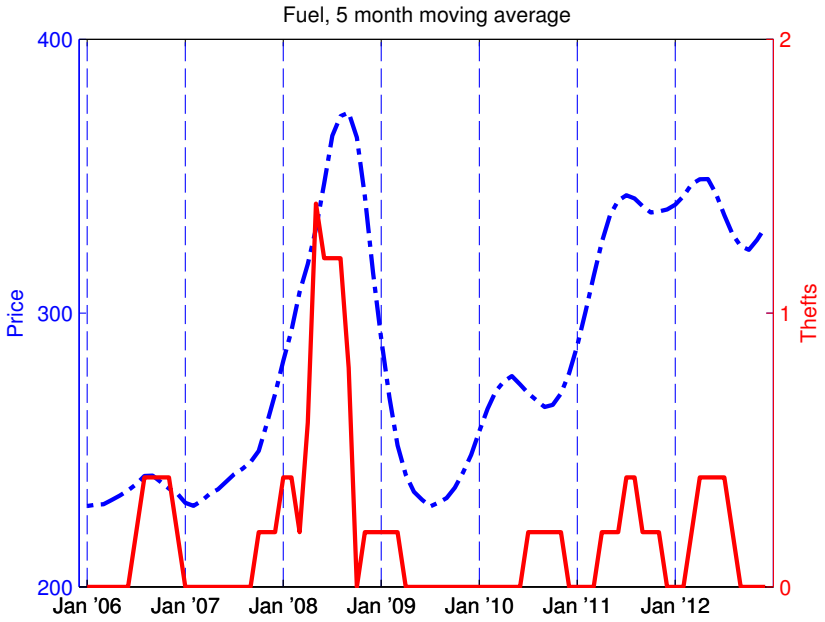

Figure 2.8: Fuel Thefts

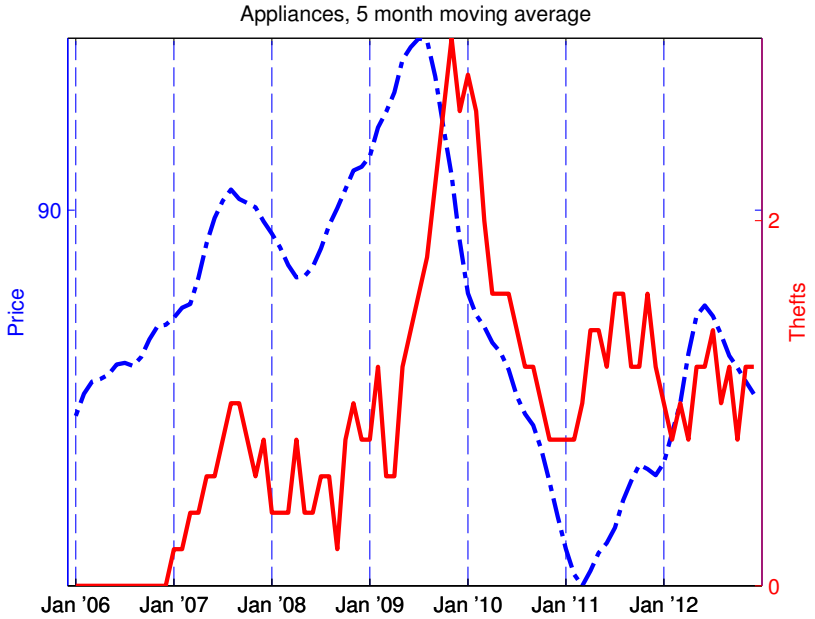

Figure 2.9: Appliances Thefts

The positive correlation between prices and thefts, possibly with some lag, is clear from the figures. The average correlation between prices and thefts for the 41 goods used in the empirical analysis is 0.15363. We ran a correlation placebo test by randomly matching prices to goods within the 41 original pairs and calculating the average correlation across the 41 shuffled pairs. Figure 2.4.1 shows the histogram of the average correlation for one million such shuffles. The p-value of the actual average correlation between goods and prices is about 0.008. Thus, the observed positive correlation between prices and goods is not simply due to other factors such as aggregate time trends.

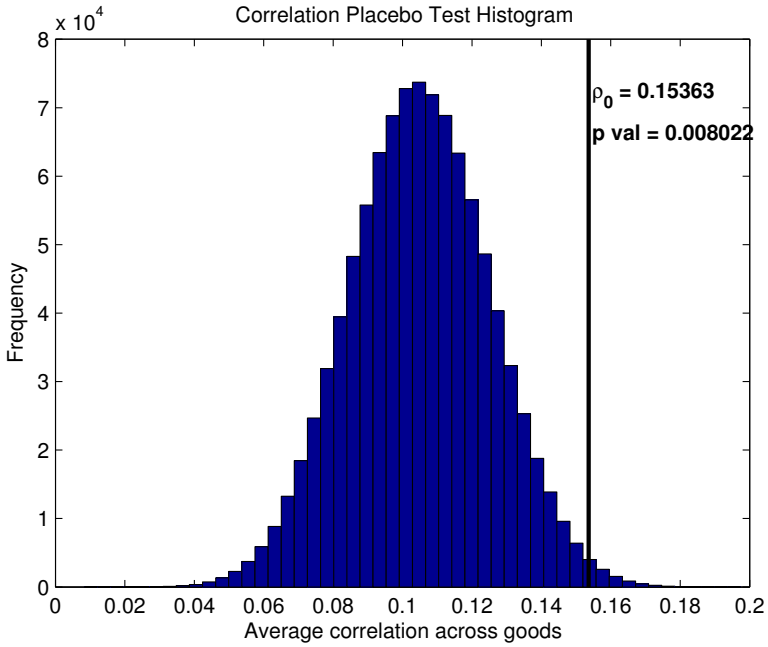

Figure 2.10: Correlation Placebo Test

2.4.2 Panel Data Estimates

By aggregating thefts to a monthly level and matching the appropriate goods to their monthly prices for 41 goods (see appendix B.2) we construct a panel of thefts and prices

over time. We first use OLS to obtain preliminary estimates of the price elasticity of theft. Some goods did not have any thefts in some months, making obtaining an elasticity estimate by a log-log regression infeasible. Therefore, we report instead the semielasticity (change in thefts from a 1% increase in prices) as a percentage of the average monthly number of thefts. This quantity can actually be interpreted as an elasticity and can be directly obtained from the regression in (2.1).[2]

We run the following fixed effect regression for $K \in \{0, \ldots, 6\}$:

$$y_{it} = \alpha_0 + \sum_{j=0}^{j=K} \gamma_j \bar{y} \log P_{i,t-j} + c_i + d_t + \epsilon_{it} \tag{2.1}$$

Where y_{it} is the number of thefts of good i in month t, $\bar{y}$ is the average number of thefts across goods and months, P_{it} is the price of good i in month t, c_i is a good fixed effect, d_t is a month fixed effect, and ϵ_{it} is the error term. As mentioned previously, the γ_j coefficients can be interpreted as elasticities. Table 2.2 shows the joint price coefficient (the sum of the price coefficients: $\sum_{j=0}^{j=K} \gamma_j$) from the regression in (2.1) for $K \in \{0, 6\}$. The joint price coefficient for $0 \leq K \leq 6$ is increasing in magnitude and significance.

[2] To see this, consider the model: $E(y|x) = \beta E(y) \log x$, $y \geq 0$, $x > 0$, $E(y) > 0$. $\partial E(y|x)/\partial x = \beta E(y)/x$. So $\beta = (\partial E(y|x)/E(y))(x/\partial x)$, which holds for every x and particularly at x^* such that $E(y|x^*) = E(y)$. To see that such x^* exists, define $f(x) := E(y|x)$ and assume that x has compact support $[a, b]$ and that f is continuous and monotonic, as is the case in most regression models. Under the assumptions, $f(a) \leq E(f(x)) \leq f(b)$ and the Intermediate Value Theorem implies the existence of x^* with $f(x^*) = E(f(x))$. The Law of Iterated Expectations then implies $E(y|x^*) = E(E(y|x)) = E(y)$. At $x = x^*$, $\beta = \partial \log E(y|x)/\partial \log x$, the elasticity of $E(y|x)$ with respect to x.

Under the assumption of constant elasticity, knowing the elasticity at $x = x^*$, or β, is sufficient. Intuitively, β is the increase in $E(y|x)$ resulting from a 1% increase in x, as a percentage of $E(y)$; it is $\delta/\bar{y}$ where δ is the coefficient from the regression based on the model $E(y|x) = \delta \log x$ and $\delta/100$ is the semielasticity.

VARIABLES	$K = 0$	$K = 6$
Joint Price	0.424	1.201**
	(0.529)	(0.571)
Observations	3,444	3,198
R-squared	0.097	0.081
Number of goods	41	41

Robust standard errors clustered
at the good level in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.2: OLS; Good and Time Level

The "Joint Price" row of Table 2.2 shows the cumulative effect of the price and its lags on thefts. The price elasticity of theft ranges from 0.424 in the short run to 1.201 over a 7-month horizon.

With positive count data that take on small values, OLS regression models have several shortcomings. Count data necessarily violates the normality assumption on the errors in OLS and covariates can usually take on plausible values that result in negative predictions. For count data that take on large values the discreteness of the data can often be approximated by a continuous model, but this is not so for count data with small values and a significant number of zeros, as in our case. Therefore, we use the Poisson fixed effects (FEP) model pioneered by Hausman et al. (1984). As Wooldridge (2002) notes, the FEP estimator has very strong robustness properties; the only assumption required for consistency is that the conditional mean, $E[y_{it}|\mathbf{x}_{it}, c_i]$ is correctly specified (where $\mathbf{x}_{it}$ are regressors), so no assumptions are made on the distribution of y_{it} given $(\mathbf{x}_{it}, c_i)$, and it allows for arbitrary dependence between y_{it} and y_{ir}, $t \neq r$. This model is of the form:

$$E[y_{it}|\mathbf{x}_{it}, c_i, d_t] = \exp\left(\alpha_0 + \sum_{j=0}^{j=K} \gamma_j \log P_{i,t-j} + c_i + d_t\right) \tag{2.2}$$

VARIABLES	$K = 0$	$K = 6$
Joint Price	1.024***	1.118***
	(0.211)	(0.231)
Observations	3,444	3,198

Robust standard errors clustered at the good level in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.3: FEP; Good and Time Level

The price elasticity of theft in row "Joint Price" of Table 2.3 is much more stable, and significant, than in the OLS regression.

2.4.2.1 State Level Estimates

Because the location of every theft is recorded, we can aggregate the thefts for each state and obtain panel data estimates at the good×state×time level. This allows us to additionally control for state fixed effects.

One difficulty with obtaining estimates at the state level is the heterogeneity in size, location, and highway miles of different states. State fixed effects are helpful, but to mitigate such problems further we normalize thefts using monthly data on number of highway miles traveled in each state from the Federal Highway Administration.

We run the following FEP regression:

$$E[y_{ist}|\mathbf{x}_{ist}, c_i, f_s, d_t] = \exp\left(\alpha_0 + \sum_{j=0}^{j=K} \gamma_j \log P_{i,t-j} + c_i + f_s + d_t\right) \tag{2.3}$$

Where y_{ist} is $10^5 \times$thefts of good i in state s in month t divided by miles traveled in state s in month t in millions. Multiplying by 10^5 leads to a mean value across all states, goods, and time periods with the same order of magnitude as in the previous regressions, but it does not affect the elasticity estimates. The results from regression (2.3) are in Table 2.4.

VARIABLES	$K = 0$	$K = 6$
Joint Price	1.002***	1.192***
	(0.207)	(0.266)
Observations	175,644	163,098

Robust standard errors clustered at the good level in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.4: FEP; Good, State, and Time Level

We regard the elasticity estimates in Table 2.4 as the most accurate because specification (2.3) utilizes state level data and uses a model that naturally accommodates count data. The price elasticity of theft ranges from about 1.002 in the short term to 1.192 over a seven-month horizon.

2.4.3 A Robustness Check for Price Endogeneity

Prices in all the above regressions may be endogenous. To see this, consider the following simple supply and demand diagram:

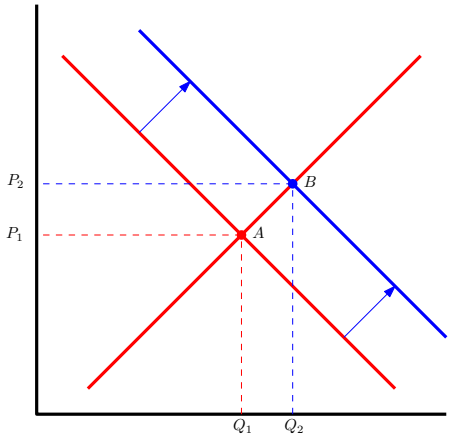

Figure 2.11: Demand Shift

Starting from point A, a demand shift may cause a movement to point B raising both prices and quantity supplied. If thefts are random (not targeted) and the price of some good changes because of a demand shift as in Figure 2.11 then thefts of that good may increase, not because thieves are stealing that good more, but simply because there are more shipments of that good on the road. Therefore, normalizing thefts by shipments may result in very different time series plots, and accounting for shipments may reduce, or eliminate, the effect of prices on thefts.

We obtained shipment data from various sources, sometimes approximate (see section 2.3), for twelve goods: aluminum, steel, beer, tobacco, beef, pork, poultry, apples, bananas, oranges, tomatoes, and potatoes. Because shipment data is in various different units (e.g. pounds for food and gallons for beer), we normalize the shipments for each good to have mean 1 by dividing by the average monthly shipments for that good. We then account for shipments in two ways: in one we divide monthly thefts by normalized monthly shipments, which we refer to as the shipments normalization; and for the purpose of the regressions we also account for shipments by adding shipments as a control without changing the theft dependent variable. The panels in Figures 2.12 and 2.13 show thefts and their normalized

versions for some of the goods we could obtain shipment data on.

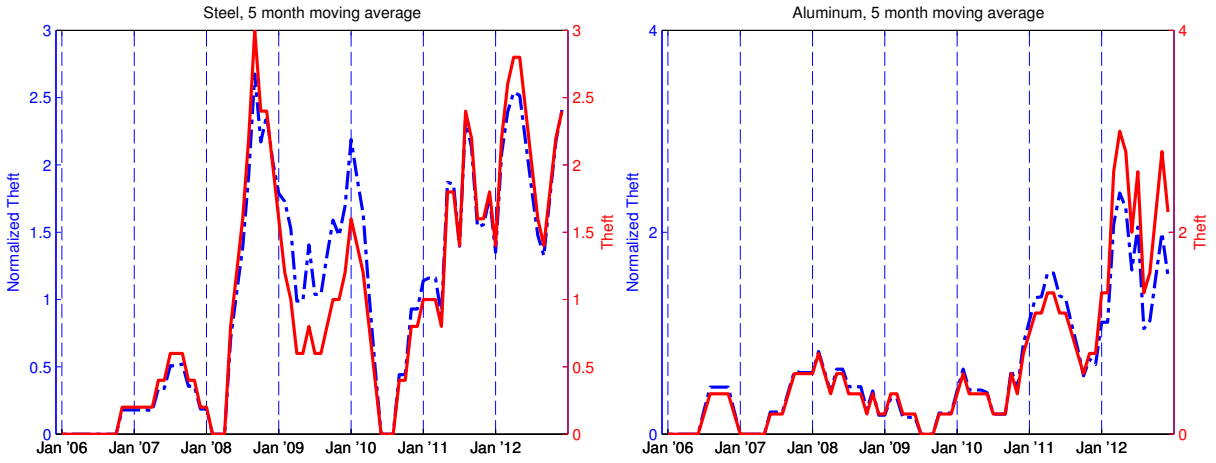

Figure 2.12: Theft and Normalized Theft for Steel and Aluminum

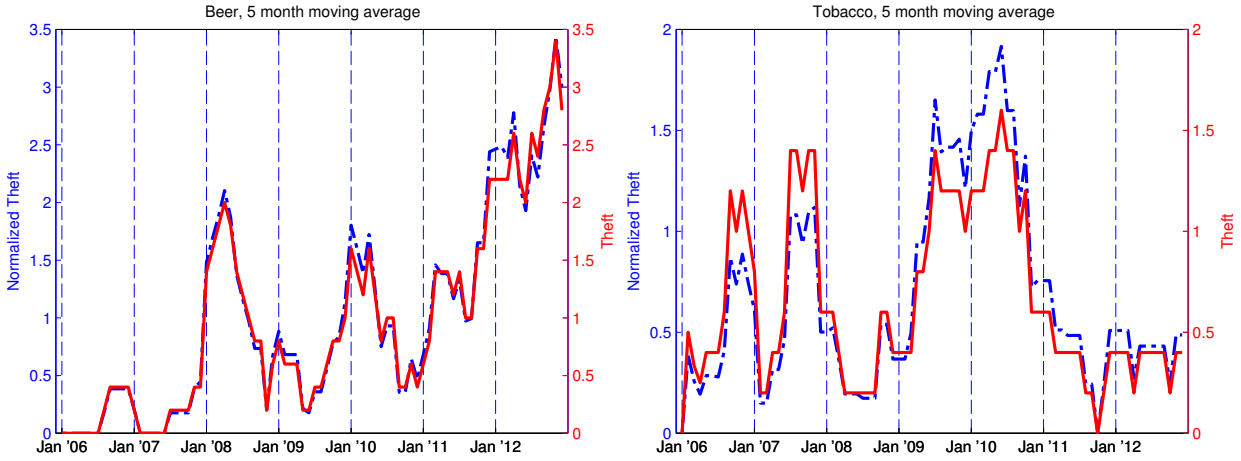

Figure 2.13: Theft and Normalized Theft for Beer and Tobacco

Figures 2.12 and 2.13 show that normalizing thefts by monthly shipments does not change the time pattern. This suggests that the nonparametric evidence for a correlation between prices and thefts presented in section 2.4.1 would continue to hold when thefts are normalized by shipments.

To examine the effect of accounting for shipments on the regression results, we limited the original 41-good panel data set to the twelve goods with available shipment data, and we ran specifications (2.1) and (2.2) for $K \in \{0, 6\}$ for those goods with and without accounting for shipments. The results are in Table 2.5.

		$K = 0$	$K = 6$
		Joint Price	
OLS	Without Shipments	0.612	0.786
		(1.018)	(1.453)
	Shipments Normalization	1.112*	1.391
		(0.553)	(0.952)
	Shipments Control	1.001	1.247
		(0.897)	(1.218)
FEP	Without Shipments	0.0200	-0.0952
		(0.858)	(1.012)
	Shipments Normalization	0.543	0.490
		(0.631)	(0.863)
	Shipments Control	0.582	0.557
		(0.688)	(0.961)
	Observations	1,008	936
	Number of goods	12	12

Robust standard errors clustered
at the good level in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.5: Accounting for Shipments

Comparing the joint price coefficients in Table 2.5 before and after accounting for shipments, for any of the models, we can see that the price effect does not disappear or even diminish; it actually becomes stronger! This suggests price movements *along* the demand curve, illustrated in Figure 2.14 by a movement from point A to point B, as opposed to

prices being mostly driven by demand shifts.

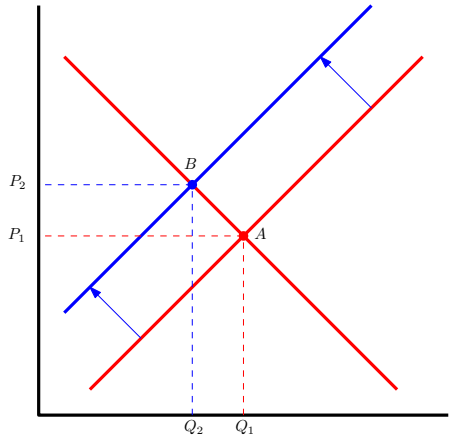

Figure 2.14: Movement Along the Demand Curve

2.5 Good-Specific and Spatial Heterogeneity

2.5.1 Heterogeneity by Good

There is likely to be a large amount of heterogeneity in how prices affect thefts across goods. The available consumer and commodity price data for some goods like metals may approximate black market prices much better than other goods like electronics. Electronics prices can vary greatly within a single product group (e.g. cellphones), and while average prices may be declining over time, thieves may be stealing the latest high value technology. Additionally, some goods may be exported for sale in a foreign country after being stolen making international prices and demand conditions more important than domestic prices.

To estimate the good-specific effect of prices on thefts, we use a Poisson quasi-maximum likelihood approach and analyze each good separately. The Poisson quasi-maximum likelihood estimator (QMLE) is a popular method for analyzing count data.[3] It has strong

[3] Grogger (1990) is one example where the Poisson QMLE was applied to time series of crime data.

robustness properties (Wooldridge (1997)): it is consistent under the assumption that $E(y|\mathbf{x})$ is correctly specified to be *any* function of $\mathbf{x}$ and the parameters, so the distribution of y_i given $\mathbf{x}_i$ does *not* have to be Poisson; consistency does not require any assumptions (subject to regularity conditions) on the variance of y_i given $\mathbf{x}_i$, in particular, the Poisson assumption that the variance is equal to the mean can be relaxed; and consistency holds under arbitrary serial correlation in the observations (subject to regularity conditions). The consistency of the Poisson QMLE with arbitrary serial correlation in the observations is desirable in our application because we apply the Poisson QMLE to time series data. We use the fully robust asymptotic variance matrix (Wooldridge (2002)), which leads to standard errors that are valid under any conditional variance assumption.

To facilitate comparisons across goods we chose to run the same regression for each of the goods. For each of the goods we use the Poisson QMLE based on the following model:

$$E[y_{it}|\mathbf{x}_{it}] = \exp\left(\alpha_0 + \gamma_i \log P_{i,t} + \beta_i \mathbf{x}_{it}\right) \tag{2.4}$$

Where $\mathbf{x}_{it}$ includes terms for a cubic time trend. This specification was chosen by checking the Akaike Information Criterion (AIC) for models with various combinations of lagged thefts, lagged prices, and time polynomials of various orders. The AIC calculations were based on the sum of the AIC from the regressions across all goods.

Grogger (1990) applied the Poisson QMLE to time series homicide data to study the deterrent effect of capital punishment.

Pet Food	-15.36	(10.66)	Beer	-10.10	(7.812)
Apples	19.47*	(11.81)	Spirits	-0.596	(8.310)
Baby Food	-3.271	(14.38)	Wine	2.335	(10.86)
Bananas	2.806	(6.235)	Iron and steel	3.574***	(0.965)
Beef and veal	1.521	(11.29)	Copper base scrap	2.680***	(0.943)
Bread	86.80*	(48.41)	Aluminum base scrap	1.559	(0.978)
Butter	-4.583	(6.532)	Men's apparel	19.72	(14.19)
Candy and chewing gum	-3.372	(8.393)	Women's apparel	-5.617	(4.767)
Breakfast cereal	-4.789	(14.93)	Infants' and toddlers' apparel	-12.73	(23.23)
Poultry	-1.901	(11.96)	Footwear	-0.793	(5.185)
Coffee	-1.356	(4.637)	Tobacco and smoking products	8.754***	(2.841)
Dairy	-8.599	(6.949)	Personal care products	-1.541	(4.335)
Fish and seafood	-0.655	(7.038)	Fuel oil and other fuels	6.747***	(2.292)
Juices and nonalcoholic drinks	-10.12	(9.264)	Tires	-2.868	(4.416)
Oranges, including tangerines	-2.226	(3.302)	Furniture and bedding	10.57*	(6.168)
Pork	-6.735	(9.518)	Appliances	6.371	(5.089)
Potatoes	-9.417	(8.968)	Tools, hardware, and outdoor	-1.743	(7.953)
Tomatoes	0.352	(4.531)	Televisions	-7.072**	(2.765)
Carbonated drinks	-1.294	(12.84)	Audio equipment	3.590	(9.655)
Sugar and artificial sweeteners	3.924	(17.99)	Computers and accessories	-2.018	(3.564)
			Toys	-9.332	(15.21)

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.6: Good Specific Regressions

Table 2.6 reports the log-price coefficient from the regression for each of the goods. The large heterogeneity across goods is immediately clear. We mentioned several real-life factors that may cause heterogeneity in the price-elasticity of theft across goods, but there are also more technical ones.

Because the variation in the regressions of this section is only at the time level, sufficient movement in prices over time is crucial in obtaining reliable estimates. Goods with little time variation in the price are likely to produce more volatile, and less reliable, estimates. Figure 2.15 shows a scatter plot of the coefficient estimates vs. the index of dispersion (variance divided by the mean) of the price series.

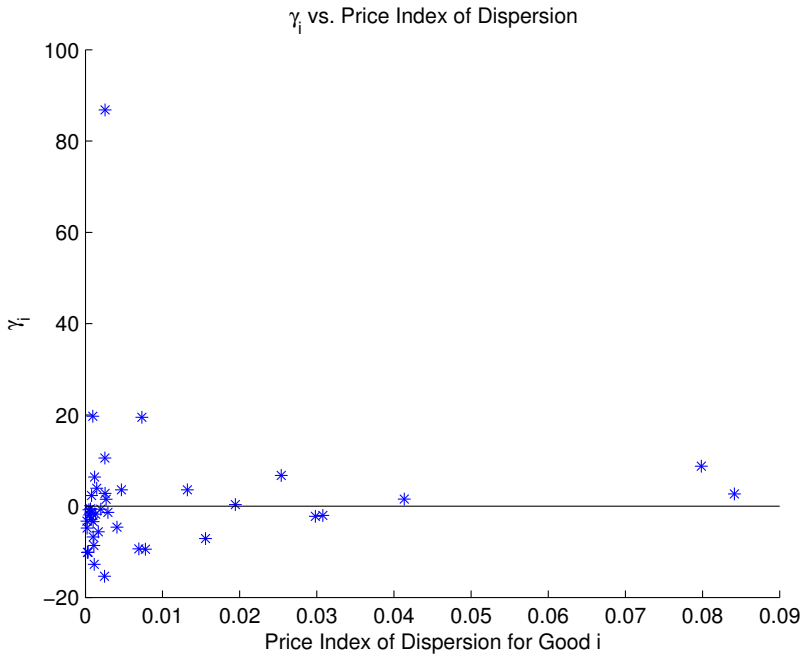

Figure 2.15: γ_i vs. Price Index of Dispersion

Also, there is likely to be more volatility in the estimates for goods with few thefts over the 84 month period, because the precise timing of the thefts for those goods can greatly affect the estimates. This can be seen in Figure 2.16, which shows a scatter plot of the coefficient estimates for each of the 41 goods vs. the total number of thefts over the time horizon.

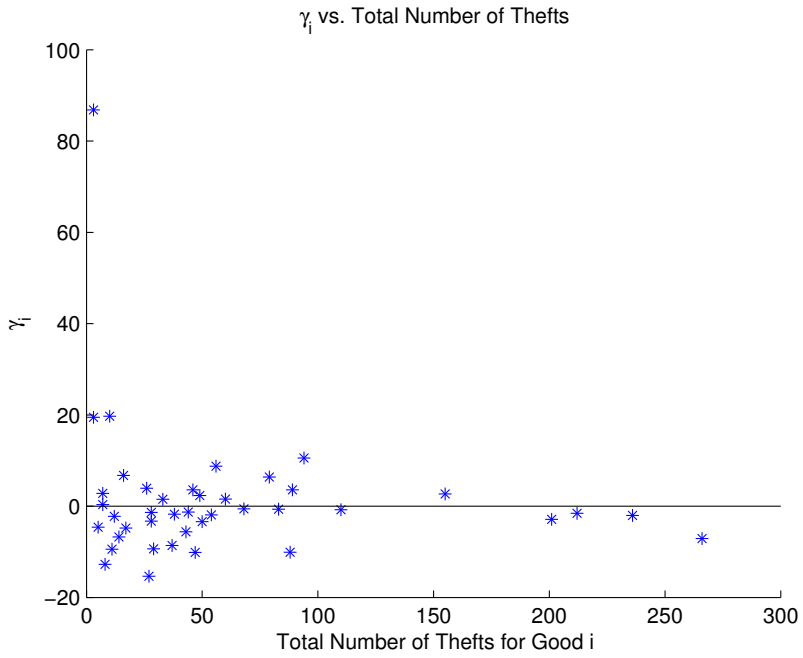

Figure 2.16: γ_i vs. Total Number of Thefts

Thus, goods with a higher price index of dispersion and a higher total number of thefts are likely to produce more accurate estimates. Figure 2.17 shows the kernel density estimate of the price elasticity of theft for the 5 goods with the highest number of total thefts among the 10 goods with the highest price index of dispersion. We call those goods the "informative subset" because the variation in their prices (measured by the index of dispersion) is more informative about the price elasticity of theft, and because the relatively high total number of thefts makes the estimates less volatile.[4]

[4]The informative subset goods are iron and steel, copper, aluminum, computers, and televisions.

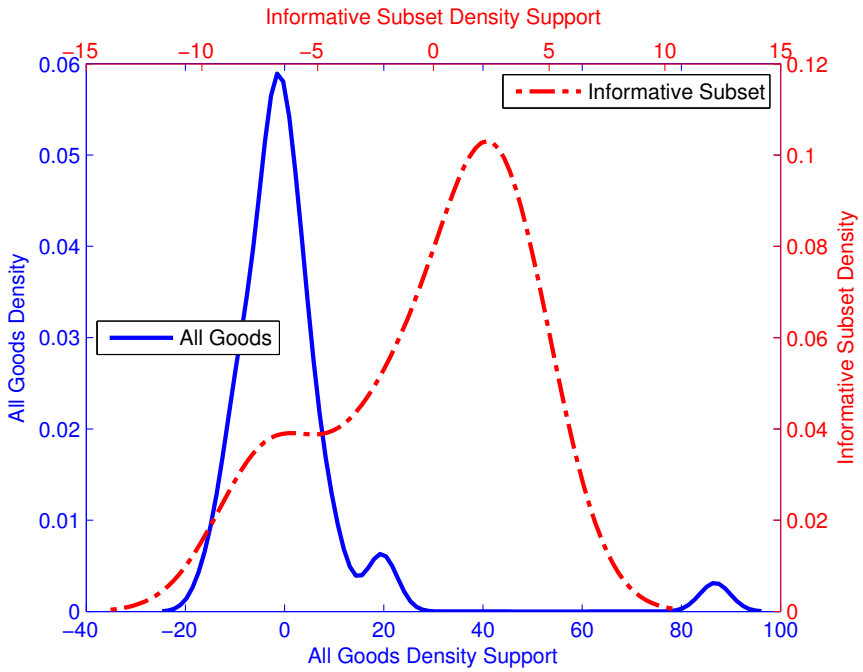

Figure 2.17: Kernel Density of γ_i Estimates

2.5.2 Theft Substitution Across Goods

Knowing how much criminals substitute across goods depending on relative prices is important from a policy perspective. The elasticities found previously may overestimate the efficacy of a policy that reduces the effective black market price. A 1% decrease in the price of a good may cause a reduction in thefts of that particular good, but if criminals substitute across goods based on their relative price then the decrease in overall crime may be less. Thus, it is interesting to know how thefts of good i are affected by the price of good j. In this section we focus on substitution between metals.

Metals make a suitable good group to analyze substitution for several reasons. Unlike goods like electronics, metals cannot be easily tracked and they are more homogeneous, so they can be sold domestically or exported. Metals are easy to sell as scrap and can be sold

virtually anywhere with potentially less market frictions than other goods. Prices for metals are fairly spatially homogeneous and are very easy to find (e.g. prices are continually updated at metal scrap yards based on market conditions). Finally, metal prices are more volatile than other goods and metal cargo thefts are common in our data, which make analysis of metal thefts more informative.

We use the QMLE model in section 2.5.1 (equation 2.4). Table 2.7 lists the own and cross price elasticities for the three metals analyzed in the data. The entry in row i and column j is the coefficient on the price of the column j good from a regression of thefts of good i on the price of good i, the price of good j (if $i \neq j$), and other controls as in equation 2.4.

	Steel	Copper	Aluminum
Steel	3.574***	-0.579	-1.003
	(0.965)	(0.709)	(1.106)
Copper	0.667	2.680***	2.226
	(1.221)	(0.943)	(1.508)
Aluminum	1.155	1.288	1.559
	(2.580)	(2.994)	(0.978)

Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.7: Substitution Across Metals

The elements on the diagonal in Table 2.7 have the expected sign. The results suggest that there is no strong evidence for a substitution effect between metals. It is also likely that this extends to other goods, because cross-price elasticity estimates for metals are likely to be higher and more significant than other goods for the same reasons that make metals a suitable group to study substitution.

2.5.3 Heterogeneity by Location

Location plays an important role in theft decisions; supply of goods, black market demand, and transportation costs (possibly to another country) can vary greatly by location. In this section we explore the relationship between location, prices, and thefts.

Figures 2.18 through 2.23 show differences across locations in the quantity and types of goods stolen. We place a circle on each city where a theft occurred and scale the circles according to the number of thefts in a particular city. Except for Figure 2.18, all circles are scaled linearly and are comparable across goods.

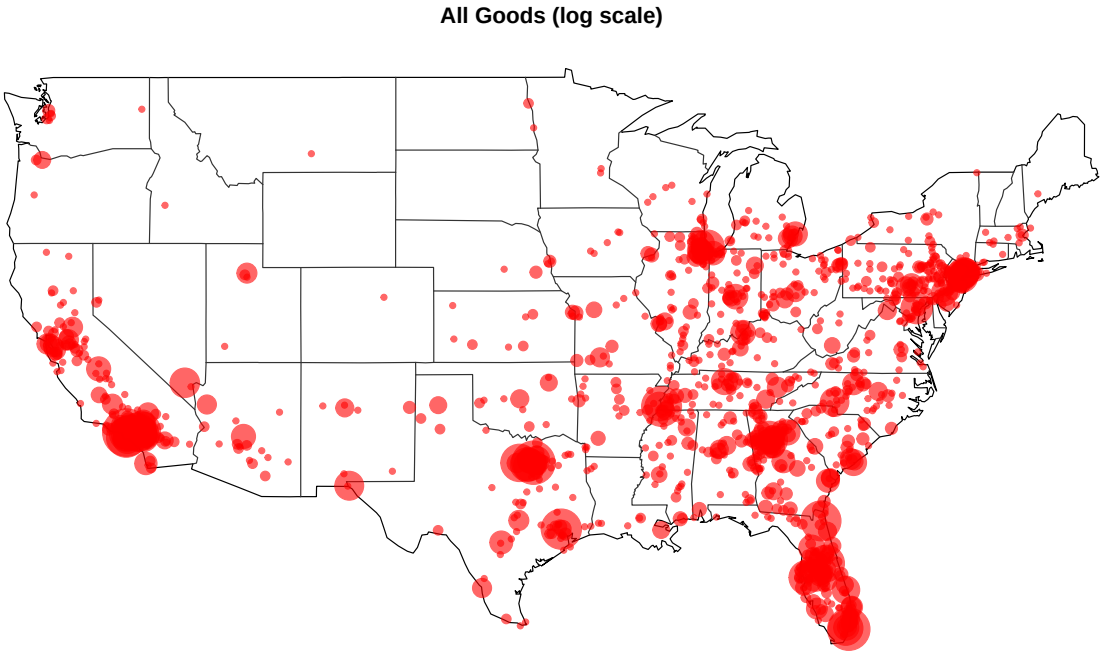

Figure 2.18: Location of Thefts of All Goods

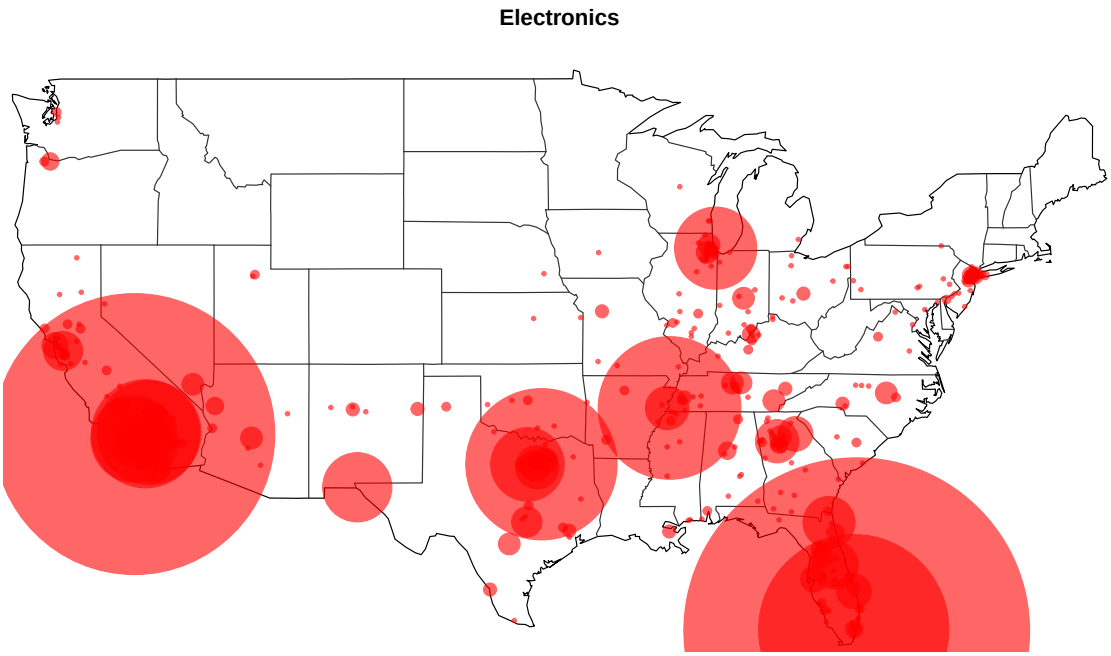

Figure 2.19: Electronics Theft Locations

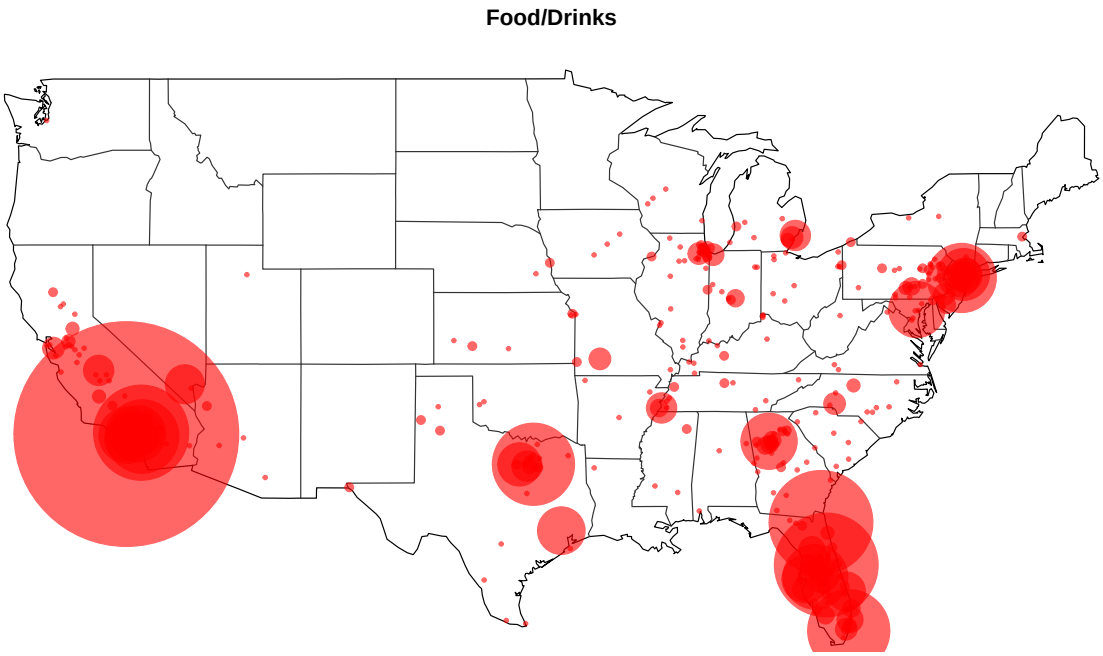

Figure 2.20: Food and Drinks Theft Locations

Figure 2.21: Clothing and Shoes Theft Locations

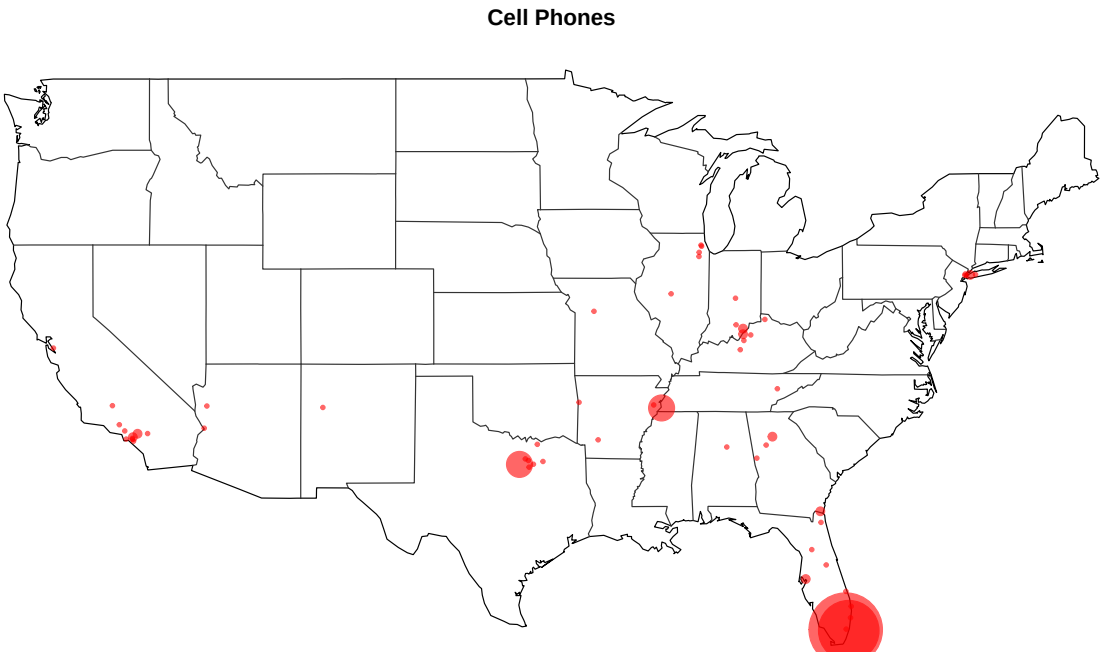

Figure 2.22: Cellphone Theft Locations

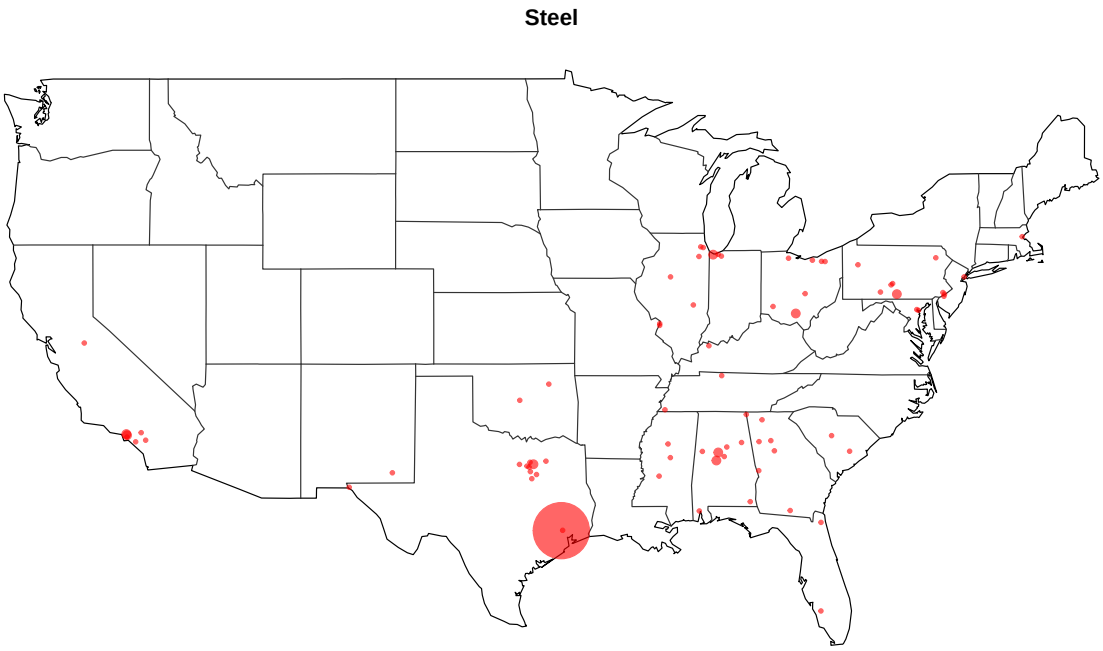

Figure 2.23: Steel Theft Locations

It is apparent in Figures 2.18 through 2.23 that some spatial patterns in thefts exist. Those patterns may be due to many factors, such as heterogeneity in market demand conditions, transportation costs to black markets, transportation patterns of different goods, crime levels, etc... It is difficult to isolate specific causes for the spatial patterns in *levels* of theft with the data available. However, by examining the *changes* in thefts of a particular good when the price of that good increases, we may be able to identify likely locations where specialized thieves (or gangs) operate.

First, we note that some cargo thefts are targeted and some are random. When targeting a specific good, cargo thieves have been known to "case" a location, gather intelligence and information on shipping routes, and follow trucks for hundreds of miles if needed, until the driver stops, and steal the cargo; random thefts are also common, making "bait trailers"

that are left unattended a common method by which police catch cargo thieves.[5]

We assume that if no thieves exist in location s with connections to brokers/fences specializing in good i then thefts of good i in location s are largely not targeted and are instead a random function of other factors of location s such as overall levels of crime, frequency of shipments carrying good i in location s, etc... On the other hand, if a gang in location s has connections to easily sell a shipment of good i on the black market then they are more likely to target that good when its price increases. This leads to the following model:

Model:

$$T_{is} = I(c_{is}, p_i) + R(s, \theta_{is}).$$

Where T_{is} is the number of thefts of good i in location s; c_{is} is the number of thieves with connections to dispose of good i in location s; p_i is the price of good i; I is the amount of targeted (intentional) thefts; and R is the amount of random or non-targeted thefts of good i in location s, which is a function of a random term, θ_{is}. We assume the following:

$$\frac{\partial R}{\partial p_i} = 0, \quad \frac{\partial^2 I}{\partial c_{is} \partial p_i} > 0$$

The second assumption implies that a location with more connections to dispose of a particular good will see a higher rise in thefts of that good when its price increases than another location with less connections. The assumption that $\partial R / \partial p_i = 0$ implies:

$$\frac{dT_{is}}{dp_i} = \frac{\partial I(c_{is}, p_i)}{\partial p_i}$$

[5]Interviews with cargo theft ring insiders (Keteyian (2010)) reveal that "brokers" who sell the stolen goods to buyers are a crucial part of a cargo theft ring. Burges (2012) notes that brokers often place "steal order" with thieves to obtain specific goods. Brokers and fences are of special importance in cargo theft gangs because of the large quantities of product involved in a single cargo theft. Non-targeted thefts are also common; discussing the variety of goods stolen, a cargo thief notes in an interview that "a truck full of yogurt is a truck, and somebody is going to pay for it."

Given this model, we can identify the relative value of c_{is} for a particular good in different states from comparing dT_{is}/dp_i for that good across states. Intuitively, if the price of a good increases and thefts of that good in state s_1 do not change, but thefts in state s_2 respond much more readily to prices, then state s_2 is more likely than state s_1 to be the location of gangs with the connections and means to dispose of that good, even though state s_1 may have higher *levels* of theft of the good (which may occur because of random state-specific factors).

Rather than obtaining good-specific estimates of the elasticity for each state, which would lead to volatile estimates due to infrequent thefts, we instead get an average for each state by estimating a state-specific version of regression (2.2). We only estimate the regression for states with over 50 thefts. Table 2.8 reports the ranking of the most responsive five states.

Rank	**State**	**Price Elasticity of Theft**	
1	Florida	1.977***	(0.516)
2	Texas	1.696***	(0.362)
3	Illinois	1.346***	(0.474)
4	California	0.724***	(0.121)
5	Tennessee	0.335	(0.601)

Robust standard errors clustered
at the good level in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.8: Most Responsive States

2.5.4 Heterogeneity by Proximity to Ports

Ports are a major transportation point for cargo, and behavior of criminals operating at or close to ports may differ from those operating elsewhere. In this section we separately consider thefts occurring close to a major U.S. port. Figure 2.24 shows the location and

relative size of the largest twenty U.S. ports ranked by cargo traffic.[6]

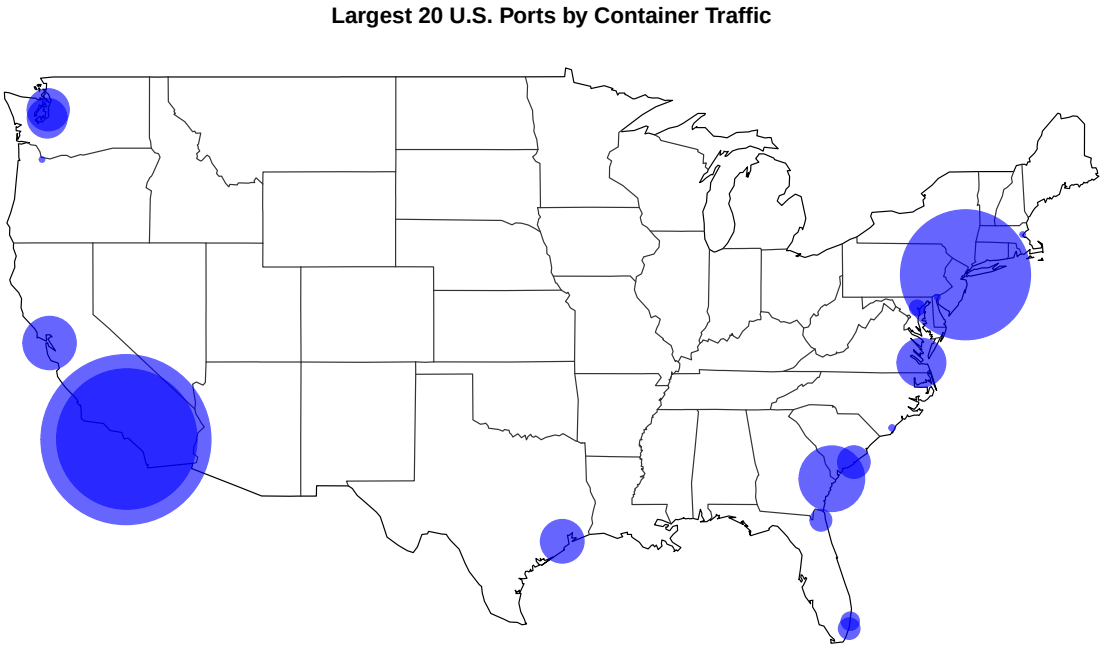

Figure 2.24: Largest Twenty U.S. Ports by Container Traffic

The center of each circle in Figure 2.24 represents a port and the circle's size represents the amount of trade at that port by cargo traffic. Appendix B.4 lists more detailed information on each port.

We construct a separate panel at the good×county×time level for thefts happening within one of the counties overlapping or adjacent to one of the top 10 ports. Table 2.9 reports the results from the FEP specification analogous to equation 2.4 (the dependent

[6]The largest twenty ports include two in Alaska and Hawaii that are not shown on the map. Two ports in Puerto Rico and Louisiana were excluded: data on highway miles traveled in Puerto Rico is missing so it was excluded to make the elasticity estimates from this section consistently comparable to those in previous sections; port of New Orleans was excluded to avoid bias due to the damage from hurricane Katrina.

variable is number of thefts and we condition on good, port region, and time fixed effects) for thefts occurring close to one of the largest twenty U.S. ports.

VARIABLES	$K = 0$	$K = 6$
Joint Price	1.285***	1.314***
	(0.247)	(0.224)
Observations	175,644	163,098

Robust standard errors clustered at the good level in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.9: Elasticities for Thefts Around Major Ports

2.6 Theoretical Analysis

The model proposed by Becker (1968) has intuitive implications, but it does not take strategic interactions into account. Strategic interactions are prevalent in cargo security and play a central role in crime: owners often secure their belongings depending on their value, and private and public security often depends on the crime rates. The question we address in this section is: can the empirical results be explained in a model of crime with strategic interactions?

In this section we show that accounting for strategic interactions can easily reverse the results of the basic model, a point made previously by Tsebelis (1990), and propose extensions that explain the empirical results. We first discuss the basic model and the comparative statics it implies. We then describe a simple model with strategic interactions, and finally we show how observed empirical patterns can arise out of a model with strategic interactions. The strategic interactions model we use is a version of the "inspection game", which has been applied to crime previously (e.g. in Tsebelis (1990)).

2.6.1 The Basic Model

An agent has outside option $r \sim F$, where F is a strictly increasing cdf with pdf f. The agent faces a decision to steal a good of price p or get the outside option r. The probability of getting caught (security) is s, and the individual loses the stolen good and pays a fine y if caught.

The agent will choose to steal the good if the outside option is less than the expected value of stealing. That is,

$$r \leq (1-s)p - sy$$

Thus, the expected ex-ante probability of a theft is

$$T = F\left((1-s)p - sy\right) \tag{2.5}$$

This simple model has intuitive comparative statics: increasing punishment and the outside option reduces the theft probability, and increasing the price of the good increases the theft probability. We can derive the following comparative static from equation (2.5):

$$\frac{dT}{dp} = (1-s)f((1-s)p - sy) > 0 \tag{2.6}$$

Thus, an increase in price increases the theft probability because it increases the gains from stealing.

In reality, security often depends on the value of the item in consideration and fines depend on the severity of the crime. Taking those two factors into account, equation (2.6) becomes,

$$\frac{dT}{dp} = f((1-s)p - sy)\left(1 - s - p\frac{ds}{dp} - y\frac{ds}{dp} - s\frac{dy}{dp}\right) \tag{2.7}$$

and it is not clear what the sign of dT/dp is. If prices increase but security and penalties do not increase fast enough to offset the increase in gain from stealing, then thefts can increase; on the other hand, if penalties and security increase too fast when prices are higher then

an increase in prices may actually reduce thefts. This shows that incorporating strategic interactions is crucial for a complete understanding of crime.

2.6.2 Strategic Interactions

Many models can incorporate strategic interactions. Here we propose a simple one and we discuss extensions that explain the observed empirical facts.

Suppose the owner of a good with price p can choose to either secure the good at cost c or not secure it. If the good is secured then any attempt at stealing it fails and if it is not secured then any attempt at stealing it succeeds. An agent decides whether to steal the good or not as before. The agent and the owner play the following complete information game:

	Secure	Not Secure
Steal	$-y, -c$	$p, -p$
Not Steal	$r, -c$	$r, 0$

Where we assume that $-y < r < p$ and $0 < c < p$. Let α be the probability that the agent plays "Steal" and β be the probability that the owner chooses "Secure". Note that the probability of capture, what we called "s" earlier, is now β, determined endogenously as the outcome of an equilibrium. The above game has a unique Nash equilibrium (α^*, β^*) given by,

$$(\alpha^*, \beta^*) = \left(\frac{c}{p}, \frac{p-r}{p+y}\right) \tag{2.8}$$

In contrast to the basic model with no strategic interactions, the equilibrium in (2.8) predicts that an increase in price does not increase, but rather *decreases*, crime (α^*) and leads instead to an increase in security (β^*). This is consistent with findings that inexpensive items, such as candy (McNees et al. (1980) and Carter (1995)), are often shoplifted. Because the

fixed price of security (e.g a lock or a camera) becomes relatively cheaper as an item's price increases, retailers often secure more expensive items much more heavily than cheaper ones. It is thus plausible that a price increase may *reduce* thefts, as captured by this model.

Empirically, we observe that prices (at least in cargo theft crimes) are positively related to thefts. Can a model of crime with strategic interactions lead to predictions consistent with the data? We propose some extensions in which the simple model we described can lead to results consistent with the data.

Incomplete Information

Suppose the owner does not know the outside option of the agent, r, but instead knows that $r \sim F$, where F is a strictly increasing cdf with pdf f. In equilibrium, if the owner chooses to secure the good with probability β, the agent will choose to steal the good if the outside option is less than the expected value of stealing, or $r \leq (1-\beta)p - \beta y$. Thus, the probability the agent chooses "Steal" is $F((1-\beta)p - \beta y)$. The owner will choose an equilibrium strategy, β^*, to maximize his expected utility:

$$\beta^* = \max_{\beta \in [0,1]} \{-\beta c - p(1-\beta)F((1-\beta)p - \beta y)\} \tag{2.9}$$

Let $\overline{T}$ denote the ex-ante expected probability of theft. The equilibrium comparative statics, particularly $d\overline{T}/dp$, depend on the parameters of the problem. To see this, let F be a uniform distribution with support on $[-y, \overline{r}]$ for some $\overline{r} > p$. Equation (2.9) becomes:

$$\beta^* = \max_{\beta \in [0,1]} \left\{ -\beta c - p(1-\beta)\frac{(1-\beta)p - \beta y + y}{\overline{r} + y} \right\} \tag{2.10}$$

The solution to (2.9) in this case is:

$$\beta^* = \max\left\{\frac{2p^2 + 2py - c(\overline{r} + y)}{2p^2 + 2py}, 0\right\} \tag{2.11}$$

Equation (2.11) characterizes the equilibrium strategy of the owner, the probability of securing the good. The agent follows a bang-bang strategy of comparing his outside option with the expected value from stealing, given β^*, and chooses "Not Steal" if his outside option is above the expected value from stealing and "Steal" otherwise. Thus, the ex-ante expected probability of theft is

$$\overline{T} = \frac{(1 - \beta^*)p - \beta^* y + y}{\overline{r} + y} \tag{2.12}$$

$d\overline{T}/dp$ depends on the parameters of the problem and is not always negative as with the complete information case. To illustrate, let $c = 1$, $y = 2$, and $\overline{r} = 8$. Figure 2.25 shows the equilibrium strategy of the owner, β^*, and the expected ex-ante probability of theft, $\overline{T}$. As the price increases $d\overline{T}/dp$ changes from being positive to negative.

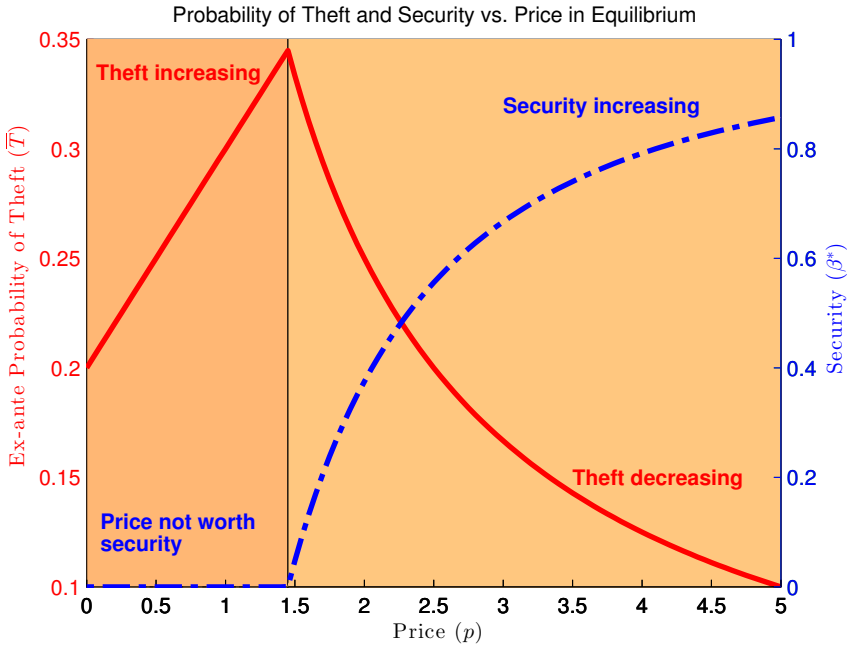

Figure 2.25: Incomplete Information Equilibrium

This illustrates the deeper theoretical fact that introducing incomplete information to a complete information game can *reverse* the equilibrium comparative statics.

Imperfect Security

Security decisions in the cargo transportation industry are much more fluid than the sharp binary decision we have in our simple model. Security can range anywhere from hiring a reputable shipping company and employees with a clean background check to watching the cargo real-time via covert electronic tracking. Similarly, cargo thieves are not always captured when goods are secured and can be extremely sophisticated – they can gather intelligence, conduct surveillance, hire insiders, and use GPS jammers – especially when the cargo can be worth millions of dollars.

We consider a setting in which security is imperfect, so that a theft can succeed with probability θ despite the good being secured.[7] This leads to the following model:

	Secure	Not Secure
Steal	$-y(1-\theta)+\theta p, -c-\theta p$	$p, -p$
Not Steal	$r, -c$	$r, 0$

The equilibrium structure in this game depends on the parameters. If $-y(1-\theta)+\theta p < r$ and $c+\theta p < p$ then there is a mixed Nash equilibrium given by:

$$(\alpha^*, \beta^*) = \left(\frac{c}{(1-\theta)p}, \frac{p-r}{(1-\theta)(p+y)} \right)$$

In this equilibrium the probability of a theft, α^*, is decreasing in p. However, if an increase in p is large enough then "Steal" becomes a strictly dominant strategy for the agent and $\alpha^* = 1$. Also, if $p < c/(1-\theta)$ then "Not Secure" is a dominant strategy and, again, $\alpha^* = 1$. Thus, the equilibrium behavior, and particularly the probability of stealing, is non-monotonic in p, as in the case with incomplete information.

2.7 Conclusions and Further Work

The results in this paper highlight the importance of potential rewards from crime in determining crime levels. Using confidential data on cargo theft, we estimate the price elasticity of theft to be about 1.002 in the short run and 1.192 over a seven-month horizon. By analyzing the heterogeneity in the price elasticity of theft across goods and across locations, we characterize the most responsive locations, and we present a simple model explaining the heterogeneity in responsiveness by spatial heterogeneity in connections to black markets.

[7] In reality θ is not exogenous, because it can be influenced by the owner securing the good more heavily. However, this can be countered by thieves using more sophisticated means to steal the good.

We find no evidence of a large and significant substitution effect across goods. We present a simple model of crime with strategic interactions to explain the main empirical results.

The elasticity we find for the response of crime levels with respect to the gain from committing crime is large relative to other elasticities in the literature (e.g. those related to punishment and policing). To the extent that our results generalize to other crimes, they suggest that crime fighting policies that act through the direct channel of reducing the potential gain from committing a successful crime (e.g. policies that shut down black markets or reduce demand for illegal goods) may be more effective than policies that attempt to reduce crime through punishment and policing. This is especially likely given that costs of incarceration and increased policing can be very substantial. Examining the relative effectiveness of the policies in a setting that allows for such an analysis could have important policy implications.

Chapter 3

A Graph Theoretic Characterization of Implementation Theory Problems

Chapter Summary

We show that any implementation problem can be formulated as a question about the existence of a graph that solves a (nonstandard) graph coloring problem. We apply this fact by proposing a simple backtracking algorithm with forward checking to find the simplest (in an arbitrary notion of simplicity) mechanisms, if they exist, to solve any implementation problem in an arbitrary solution concept.

3.1 Introduction

The simple, yet fundamental, insight of this paper is that any mechanism with a countable message space can be represented as a weighted, colored graph. This connection to graph theory opens the door for using graph theoretic methods, which have been extensively studied and developed, to answer questions in mechanism design and implementation theory. Algorithms for coloring graphs can be used to design mechanisms; complexity results from graph theory, which are relatively well established, can be used to understand the complexity properties of implementation and mechanism design problems; and graph theoretic methods can be used to prove and illuminate results in mechanism design and implementation theory.

In a seminal paper, Maskin (1999) presented general results for Nash implementation.

Several papers have studied if more social choice rules can be implemented with reasonable modifications to the solution concept (e.g. Palfrey and Srivastava (1991)). Other papers consider implementation under additional behavioral assumptions (e.g. Eliaz (2002) and Dutta and Sen (2012)). Different assumptions about the solution concept, and restrictions on the mechanisms used, can lead to vastly different results regarding which social choice rules can be implemented. Thus, establishing general results that are not specific to just one solution concept in a unifying framework is crucial in addressing design problems with nonstandard solution concepts and in making mechanism design apply more generally.

We show that any implementation problem, in an arbitrary solution concept, can be reformulated as a problem about the existence of a graph with a specific coloring, establishing a connection between implementation theory and graph theory. Taking an applied approach, we illustrate one of the applications of this connection by proposing an algorithm, based on the backtracking with forward checking algorithm for solving Constraint Satisfaction Problems, to design simple mechanisms to solve an implementation problem in an arbitrary solution concept. Specifically, we let δ be an arbitrary simplicity criterion based on a mechanism's strategy space, and for any implementation problem with an arbitrary solution concept and any simplicity criterion threshold, we propose an algorithm to: (1) determine whether a mechanism δ-simpler than the given threshold exists that solves the implementation problem, and (2) determine the δ-simplest such mechanism. Despite the generality of our exposition, the algorithms we propose can be easily applied to construct mechanisms in small problems. By allowing for an arbitrary solution concept, our results can be used to construct mechanisms (or rule out the existence of simple ones) for complicated solution concepts, which the literature may not address and for which no implementation results are known.

Our results also illuminate and generalize some results in implementation theory. For example, we present an intuitive "no-conflict" condition on the constraints implied by an

implementation problem that is necessary for implementation. The no-conflict condition is independent of the solution concept, but we show that it is equivalent to Maskin monotonicity when the solution concept is Nash equilibrium. In a general implementation problem, the no-conflict condition can be viewed as the equivalent monotonicity condition for that particular problem and it offers a first check for whether the problem can be implemented or not.

Our approach makes it clear that graph coloring problems, or constraint satisfaction problems more generally, underlie much of implementation theory. Vohra (2011) shows how many mechanism design problems can be formulated as linear programming problems. For the most part, Vohra (2011) focuses on environments with incomplete information and where the revelation principle applies. In contrast, we focus on implementation theory problems, where the revelation principle does not apply. This makes the problem considerably more difficult, because (assuming a mechanism exists that solves a particular problem) without the revelation principle the message space of the mechanism is unknown and the outcome at each message profile reported is also unknown. The problems we show are most related to implementation theory, graph coloring and constraint satisfaction problems, are both computationaly hard and cannot be solved by efficient linear programming methods in general.

This paper is organized as follows. Section 3.2 describes the basic implementation problem. Section 3.3 establishes the connection to graph theory and presents the main results. Section 3.4 discusses constraint satisfaction problems, of which graph coloring problems are a special case. Section 3.5 presents algorithms to solve arbitrary implementation problems using simple mechanisms. Section 3.6 discusses special cases (e.g. Nash implementation) and provides examples. Section 3.7 concludes and discusses further work.

3.2 The Standard Implementation Problem

Let A be a set of outcomes and $I = \{1, \ldots, n\}$ be the set of players. Each individual $i \in I$ has a preference relation R_i over the set of outcomes, A. The collection of preference

relations for all individuals is called a preference profile and is denoted by $R \equiv (R_1, \ldots, R_n)$. Associated with each preference relation R_i is a strict preference relation P_i. The set of all permissible preference profiles is denoted by $\mathcal{R}$ and we let $\mathcal{R}_i$ be the set of all permissible preference relations for a specific player $i \in I$. The preferences of each player $i \in I$ are representable by a utility function, $u_i : A \times \mathcal{R}_i \to \mathbb{R}$. A *social choice correspondence* (SCC) or a *social choice rule* (SCR) is a function $F : \mathcal{R} \to 2^A \backslash \{\varnothing\}$. When F is single valued, i.e. $F : \mathcal{R} \to A \backslash \{\varnothing\}$, it is called a *social choice function* (SCF). A mechanism is a pair (M, g), where $M = M_1 \times \cdots \times M_n$ is the product of message (or strategy) spaces for each player and $g : M \to A$ is an *outcome function*. A typical element of M_i is a message for player $i \in I$, denoted by m_i, and a collection of such messages is called a message profile and denoted by $m = (m_1, \ldots, m_n)$. A mechanism (M, g), together with a preference profile R defines a normal form *game*. A *solution concept* is a prediction specifying the message profiles, called *equilibria*, we expect players to play in a game. If S is a solution concept and $\Gamma = (M, g)$ is a mechanism, we denote the set of all equilibrium profiles under preference profile R by $S(\Gamma, R) \in M$. The set of outcomes associated with $S(\Gamma, R)$ is denoted by $g(S(\Gamma, R))$. A SCC (or SCF) is implementable in a solution concept, S, if there exists a mechanism $\Gamma = (M, g)$ such that $g(S(\Gamma, R)) = F(R)$ for all $R \in \mathcal{R}$.

We assume that renaming the strategies in a mechanism does not affect the set of equilibria: suppose $\Gamma = (M, g)$ and $\Gamma' = (M', g')$ are two mechanisms with the property that a bijection $h_i : M_i \to M_i'$ exists such that $g(m_1, \ldots, m_n) = g'(h_1(m_1), \ldots, h_n(m_n))$; then for each preference profile $R \in \mathcal{R}$, $(m_1, \ldots, m_n) \in M$ is an equilibrium in Γ if and only if $(h_1(m_1), \ldots, h_n(m_n)) \in M'$ is an equilibrium in Γ'.

We say that a collection $(I, A, \mathcal{R}, F, S)$ of a set of players, outcomes, preference profiles, a social choice rule, and a solution concept form an *implementation problem*. We say that a mechanism, $\Gamma = (M, g)$, solves the implementation problem if $g(S(\Gamma, R)) = F(R)$ for each $R \in \mathcal{R}$.

3.3 The Connection to Graph Theory

A *graph* is a pair of sets $G = (V, E)$ with $E \subseteq [V]^2$, where $[V]^2$ denotes all subsets of V of size 2; we refer to V as the set of vertices and E as the set of edges. We will also refer to A as the set of outcomes or colors interchangeably. A graph coloring is a function: $C : V \to A$. We will sometimes use $C(G)$ to denote the coloring associated with graph G, and $C(B)$ to denote the set of all colors of a subset of vertices $B \subseteq V$. For any colorings $C^1, \ldots, C^k$, we define their union to be a function $C^{\bigcup} : V \to 2^A \backslash \{\varnothing\}$ such that $C^{\bigcup}(v) = \bigcup_{i=1}^{k} C^i(v)$ for any vertex v, and their intersection to be function $C^{\bigcap} : V \to 2^A$ such that $C^{\bigcap}(v) = \bigcap_{i=1}^{k} C^i(v)$. We call any function with domain V and range 2^A a *multi-coloring*. Unions and intersections of multi-colorings are defined analogously.

The solution to any implementation problem is a mechanism, we first show that any mechanism has a graph representation. For any mechanism, $\Gamma = (M, g)$, the graph representation of Γ is a colored weighted graph, $G = (V, E)$, with the following properties: $V = M$; the color of a vertex m is $g(m)$; and for any two vertices $m, m' \in V$ if $m_i \neq m'_i$ and $m_j = m'_j$ for all $j \neq i$ then m and m' are connected with an edge of weight i.

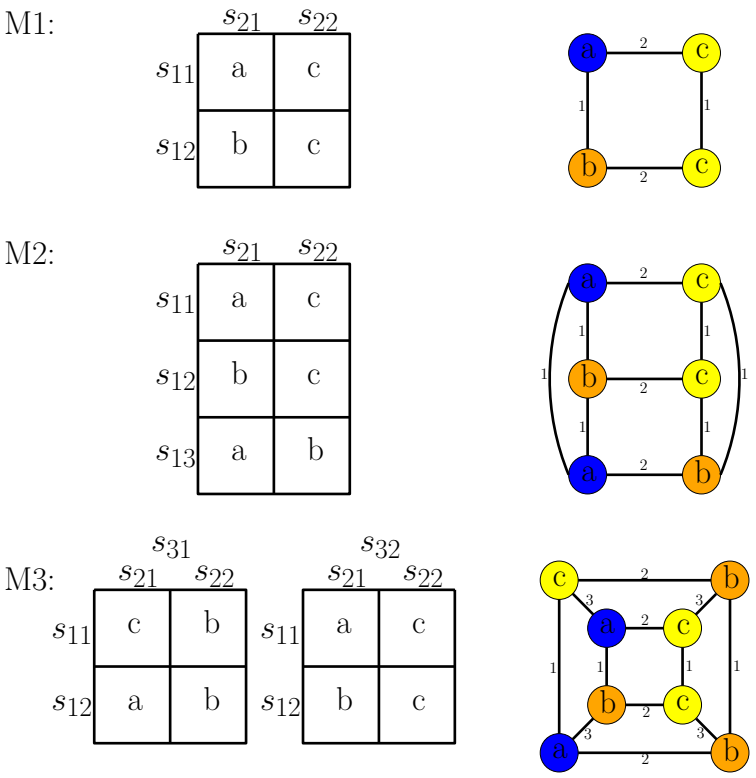

Figure 3.1: Example Mechanisms and Their Graph Representations

Figure 3.1 shows some mechanisms and their graph representations. Mechanisms $M1$ and $M2$ have two players and $M3$ has three players; s_{ij} denotes strategy j for player i. The graph preserves the essential structure of the mechanism: the color of each vertex corresponds to an outcome at a strategy profile and edges of weight i indicate outcomes player i can obtain by deviating.

Any mechanism has a corresponding graph representation, but not every graph is the graph representation of a mechanism. We denote the set of all graphs that are graph representations of some mechanism by $\mathcal{G}$. Note that, up to a renaming of players' messages, any $G \in \mathcal{G}$ has a unique associated mechanism. If (M, g) is a mechanism associated with graph $G = (G, V) \in \mathcal{G}$, then we denote by h the bijection $h : V \to M$. Note also that

$g(h(v)) = C(v)$.

Given a graph, $G \in \mathcal{G}$, we denote by $\mathcal{C}(a|G, S, R)$ the set of all colorings of graph G such that there is an equilibrium of the corresponding mechanism with outcome a under solution concept S and preference profile R. The structure of $\mathcal{C}(a|G, S, R)$ depends on the solution concept. For example, if the solution concept is Nash equilibrium then a coloring C of graph $G = (V, E)$ is contained in $\mathcal{C}(a|G, S, R)$ if there exists a vertex $v \in V$ with color a such that, for all $i \in I$, any vertex v' connected to v by an edge of weight i has color in $L_i(a, R)$, where $L_i(a, R) = \{b \in A : aR_ib\}$ (i.e. player i does not have a profitable deviation from the message corresponding to vertex v).

As we mentioned previously, a mechanism $\Gamma = (M, g)$ solves an implementation problem if and only if $g(S(\Gamma, R)) = F(R)$ for all $R \in \mathcal{R}$. This condition is equivalent to a set of coloring constraints on the graph representation of Γ, which we formalize next.

Definition 3.3.1 (**Existence Constraints**)**.** Let $(I, A, \mathcal{R}, F, S)$ be an implementation problem and G be a graph in $\mathcal{G}$ with coloring $C(G)$. The following are the existence constraints on G associated with the implementation problem: for each $R \in \mathcal{R}$ and for each $a \in F(R)$, $C(G) \in \mathcal{C}(a|G, S, R)$.

Definition 3.3.2 (**Nonexistence Constraints**)**.** Let $(I, A, \mathcal{R}, F, S)$ be an implementation problem and G be a graph in $\mathcal{G}$ with coloring $C(G)$. The following are the nonexistence constraints on G associated with the implementation problem: for each $R \in \mathcal{R}$ and for each $a \notin F(R)$, $C(G) \notin \mathcal{C}(a|G, S, R)$.

The following Theorem is the main result of the paper. It establishes that any implementation problem can be reformulated as a question about the existence of a graph that solves a graph coloring problem.

Theorem 3.3.1. Suppose $(I, A, \mathcal{R}, F, S)$ is any implementation problem. Then there exists a mechanism that solves the implementation problem if and only if there exists a graph in $\mathcal{G}$

satisfying the existence and nonexistence coloring constraints associated with the implementation problem.

Proof. Let $(I, A, \mathcal{R}, F, S)$ be an implementation problem, and let (M, g) be a mechanism that solves the implementation problem with graph representation $G = (V, E)$. If $a \in F(R)$, then implementation implies the existence of an equilibrium of the game induced by the mechanism and the preference profile R with outcome a. Thus, $C(G) \in \mathcal{C}(a|G, S, R)$. Similarly, if $a \notin F(R)$ then implementation implies that there is no equilibirum of (M, g) at R with outcome a, showing that $C(G) \notin \mathcal{C}(a|G, S, R)$.

Next, suppose $G \in \mathcal{G}$ satisfies the existence and nonexistence constraints implied by implementation problem $(I, A, \mathcal{R}, F, S)$; let $\Gamma = (M, g)$ be the mechanism corresponding to G. Then, for any $a \in F(R)$, (M, g) has an equilibrium with outcome a under solution concept S at R; also, for any $a \notin F(R)$, (M, g) has no equilibrium with outcome a under solution concept S at R. This shows that $g(S(\Gamma, R)) = F(R)$ for any R and the (M, g) solves the implementation problem. ■

Theorem 3.3.1 implies that to decide whether an implementation problem has a solution or not it is enough to answer the question of whether a graph $G \in \mathcal{G}$ exists satisfying the existence and nonexistence coloring constraints.

Definition 3.3.3 (**No-Conflict Condition**)**.** Let $(I, A, \mathcal{R}, F, S)$ be an implementation problem. There exists $G \in \mathcal{G}$ such that, for all $(a, R, a', R') \in A \times \mathcal{R} \times A \times \mathcal{R}$ with $a \in F(R)$ and $a' \notin F(R')$, $\mathcal{C}(a|G, S, R) \nsubseteq \mathcal{C}(a'|G, S, R')$.

Note: we allow for the possibilities that $a = a'$ and $R = R'$ in the no-conflict condition. The no-conflict condition simply says that the nonexistence constraints do not rule out a set of colorings, one of which is stipulated to be used to color the graph by the existence constraints. This condition is obviously necessary for implementation. It can give us an easy first check on whether an implementation problem can never have a solution, and we

show later that it is equivalent to Maskin monotonicity when the solution concept is Nash equilibrium.

Multi Implementation

In some instances it is interesting to implement social choice rules in more than one solution concept (e.g. with double implementation). In this case the existence and nonexistence constraints for the multi implementation problem are simply the combined constraints from each of the solution concepts. Theorem 3.3.1, and the rest of the paper, then proceed naturally with the combined coloring constraints.

3.4 Constraint Satisfaction Problems

Given a graph, determining whether there is a feasible coloring for the graph that solves a set of coloring constraints can be expressed as a *Constraint Satisfaction Problem* (CSP) (A review of constraint satisfaction problems can be found in Brailsford et al. (1999)). Given a set of variables and a finite domain for each variable, a (finite domain) CSP is the problem of assigning a value to each variable from its domain so that a set of constraints is satisfied. CSPs are common in Artificial Intelligence as many problems, and in particular graph coloring problems, can be represented as CSPs. We discuss next two classical examples of CSPs.

The n-Queens Problem

In chess, a queen is a piece that can move any number of squares vertically, horizontally, and diagonally. The n-Queens problem is to find a placement of n queens on a $n \times n$ chess board so that no two queens attack each other (i.e. no two queens are on the same row,

column, or diagonal). Figure 3.2 shows a solution to the 8-Queens problem on a standard 8×8 chess board.

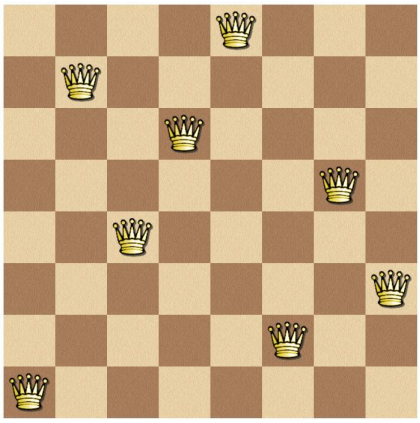

Figure 3.2: A Solution to the 8-Queens Problem

The n-Queens problem can be formulated as a CSP as follows. Because each queen must be on a different row, we know that there will be exactly one queen on each row; thus the set of variables can be $\{c_1, \ldots, c_n\}$, the column number of each of the queens. The domain of each variable can be $\{1, \ldots, n\}$, and the constraints are that no pair of variables has the same value (queens cannot be in the same column) or the same diagonal line.

(Standard) Graph Coloring Problems

Given a graph, the standard graph coloring problem is to find a coloring of the vertices such that no two adjacent vertices share the same color. This problem and its extensions have many applications, which led to their being studied extensively. For example, consider a problem of scheduling final exams in a university in the least amount of time so that no two classes have finals at the same time if a student is enrolled in both classes. Representing this problem as a graph coloring problem, each class can be represented as a vertex with edges connecting classes that have students in common. A "color" for each vertex can then

represent a time for the final. A coloring of the resulting graph with no adjacent vertices sharing the same color is then a feasible solution to the scheduling problem; a coloring that uses the minimum number of colors is an optimal solution.

Algorithms for Solving CSPs

A simple algorithm for solving CSPs that is often considerably faster than a brute-force approach of checking all the combinations of possible variable assignments is the *backtracking algorithm*. Backtracking searches in a tree of variable assignments by sequentially assigning variables an allowable value, and continually checking the constraints. If a constraint fails then the algorithm repeatedly assigns different values to the current variable; if none of the values work then the algorithm backtracks to the previous variable and tries a different value. Backtracking is faster than a brute-force approach because once a partial variable assignment is discovered to violate the constraints all the variable assignments with the same partial variable assignment are eliminated from consideration.

Backtracking algorithms only check the satisfaction of the constraints based on the current variable assignment and the previous variables, but they do not look ahead to future variables. A *forward checking* algorithm, introduced by Haralick and Elliott (1980), keeps track of allowable values for future variables and backtracks if a partial variable assignment leads to an empty set of allowable values for a future variable.

3.5 Application: Solving Arbitrary Implementation Problems with Simple Mechanisms

Given any implementation problem, Theorem 3.3.1 allows us to focus on the graph coloring problem for finding a solution. Given a graph in $\mathcal{G}$, determining whether it satisfies the existence and nonexistence coloring constraints is simply a non-standard graph coloring

problem. In this section we propose an algorithm, which uses backtracking with forward checking, to determine whether a graph in $\mathcal{G}$ satisfies the coloring constraints or not. The algorithm also gives us such coloring when it exists. We nest this algorithm in another that iteratively checks graphs in $\mathcal{G}$ starting with the simplest possible ones to give us the simplest mechanisms that solve the implementation problem, if they exist.

Because of the generality of the problems we consider, the algorithms in this section cannot be used in general to determine that a given implementation problem has no solution. However, for an arbitrary simplicity criterion and an arbitrary simplicity threshold, the algorithms we propose can tell us whether an implementation problem can be solved using a mechanism simpler than the threshold or not. In less general problems (e.g. with a fixed solution concept), it is sometimes possible to determine necessary and sufficient conditions for an implementation problem to have a solution, and algorithms for checking those conditions may be used.

In the basic algorithm, we assign the vertices an arbitrary order and sequentially color them to make sure the existence constraints are satisfied. One complication is that the existence constraints may be infeasible with a particular vertex ordering, but feasible with another. Thus, (initially) all possible orderings must be checked, but in the process of doing so we locally "explore" and eliminate all orderings that are discovered to be bad, making the process more efficient. We also forward check by keeping track of the allowable colors of all vertices and backtrack when a vertex has an empty set of allowable colors. In less general problems with a specific solution concept the algorithms can potentially be simplified considerably.

Let $F(\mathcal{R}) = \{a : a \in F(R), R \in \mathcal{R}\}$ be the set of unique outcomes in the range of F. $F(\mathcal{R})$ is the set of equilibrium outcomes that must exist in the mechanism. We assume that $F(\mathcal{R}) > 1$, otherwise the implementation problem is trivial.

For any implementation problem, $(I, A, \mathcal{R}, F, S)$, the first algorithm, Algorithm A1,

determines if a given graph $G = (V, E) \in \mathcal{G}$ with $|V| \geq 2$ can be colored in a way to satisfy the existence and nonexistence constraints of the implementation problem.

Throughout the algorithm, we refer to a vertex's set of *allowable colors*. Define $\mathcal{C}_v(a|G, S, R)$ to be the set of all colorings of graph G such that the message corresponding to vertex v in the corresponding mechanism is an equilibrium with outcome a under solution concept S and preference profile R. Suppose at any point in the algorithm we have assigned colors $c_1, \ldots, c_k \in A$ to vertices $v_1, \ldots, v_k \in V$ in steps 2, 3, or 4; then the set of allowable colors for vertex v is the set of colors assigned to v (possibly $\varnothing$) by the multi-coloring $\bigcap_{i=1}^{k} \mathcal{C}_{v_i}^{\bigcup}(c_i|G, S, R)$, where $\mathcal{C}_{v_i}^{\bigcup}(c_i|G, S, R)$ is the union of all colorings in $\mathcal{C}_{v_i}(c_i|G, S, R)$. The set of allowable colors for a vertex v is simply the set of colors for v that are consistent with vertices $(v_1, \ldots, v_k)$ being equilibria with colors $(c_1, \ldots, c_k)$.

Algorithm A1

Given An implementation problem $(I, A, \mathcal{R}, F, S)$, and a graph $G = (V, E) \in \mathcal{G}$ with $|V| \geq 2$. We remove any coloring of G, if present.

Initialization If $|F(\mathcal{R})| > |V|$ then the graph cannot satisfy the existence constraints and the algorithm ends. Label the vertices $\{v^1, \ldots, v^{|V|}\}$. Let $Q = \{(a, R) : a \in F(R), R \in \mathcal{R}\}$ and label the elements in Q as $\{q^1, \ldots, q^d\}$, where $d = |Q|$. For each $i \in \{1, \ldots, d\}$, let $q^i = (c^i, R^i)$. Let Π be the set of all d-tuples such that each $\pi = (\pi_1, \ldots, \pi_d) \in \Pi$ has the following properties: (1) $\pi_1 = 1$; (2) $\pi_i \in \{1, \ldots, |V|\}$; and (3) $\pi_i \neq \pi_j$ if $c^i \neq c^j$ for all $i, j \in \{1, \ldots, d\}$ with $i \neq j$.

Step 1 Color vertex v^1 with c^1. If there is any color c^i, $i \in \{2, \ldots, d\}$, such that no vertex in V has c^i as an allowable color then G does not satisfy the existence constraints and the algorithm ends.

Step 2 If Π is empty then the graph does not satisfy the constraints and the algorithm ends. Otherwise, pick a $\pi \in \Pi$.

Step 3 Do this step for each $1 < k \leq d$. Subtract from Π all π' with the following properties: (1) $(\pi'_1, \ldots, \pi'_{k-1}) = (\pi_1, \ldots, \pi_{k-1})$, and (2) c^k is not an allowable color for $v^{\pi'_k}$; if the current π is subtracted then go back to step 2, otherwise color v^{π_k} with c^k. If any vertex in V has an empty set of allowable colors then subtract the current π from Π, along with all π' such that $(\pi'_1, \ldots, \pi'_k) = (\pi_1, \ldots, \pi_k)$, and go back to step 2.

Step 4 The graph now satisfies the existence constraints, but there are some vertices with unassigned colors, each of which has a nonempty set of allowable colors. Use backtracking with forward checking again to check if a feasible coloring exists:

- Enumerate all the unassigned vertices in an arbitrary order.
- Sequentially color the vertices with a color from the allowable set of colors for each vertex, reducing the allowable set of colors for all the unassigned vertices based on the nonexistence constraints. After each vertex is colored, backtrack to the previously colored vertex and use a different color if any unassigned vertex has an empty set of colors.
- Use the first coloring that satisfies the nonexistence constraints; if none exist then subtract the current π from Π and go back to step 2.

The next algorithm, Algorithm A2, searches for the simplest graphs in $\mathcal{G}$ that can be colored using Algorithm A1. "Simplicity" can be defined in many ways, and we allow for fairly general choices using the following definition of a *simplicity criterion.*

Definition 3.5.1 (**Simplicity Criterion**)**.** A simplicity criterion is a function $\delta : \mathbb{N}^n \to \mathbb{R}$, such that for any $z \in \mathbb{R}$ the set $\{b \in \mathbb{N}^n : \delta(b) \leq z\}$ is bounded. We say that $b \in \mathbb{N}^n$ is δ-simpler than $b' \in \mathbb{N}^n$ if $\delta(b) < \delta(b')$.

The simplicity of mechanism (M, g) will be measured by $\delta(|M_1|, \ldots, |M_n|)$. Algorithm A2 will find the simplest mechanisms (up to an arbitrary threshold $z \in \mathbb{R}$) that implement a social choice rule, which allows for great flexibility. For example, to find a mechanism with the smallest strategy space in terms of the sum of players' strategies define $\delta(b) = \sum_{i=1}^{n} b_i$, $b = (b_1, \ldots, b_n) \in \mathbb{N}^n$; to find a mechanism with the smallest number of strategy profiles define $\delta(b) = \prod_{i=1}^{n} b_i$; and to find a mechanism that minimizes (one measure of) strategic interactions let $\delta(b) = \sum_{i=1}^{n} \mathbf{1}\{b_i > 1\}$. These examples, and many others, can be combined to yield more complicated notions of simplicity: for example, to find the mechanism with the smallest number of strategy profiles, giving preference to less strategic mechanisms at ties, define $\delta(b) = \prod_{i=1}^{n} b_i + (1/(n+1)) * \sum_{i=1}^{n} \mathbf{1}\{b_i > 1\}$.

Algorithm A2

Given: an implementation problem $(I, A, \mathcal{R}, F, S)$, and $z \in \mathbb{R}$.

Step 1 Let B be the set of vectors in $b = (b_1, \ldots, b_n) \in \mathbb{N}^n$ such that $\sum_{i=1}^{n} b_i > n$, $\prod_{i=1}^{n} b_i \geq |F(\mathcal{R})|$, and $\delta(b) \leq z$. If B is empty then the algorithm ends: the implementation problem has no solution using a mechanism with a simplicity criterion value up to z. Create an enumeration of the elements in B as $(b^1, b^2, b^3, \ldots)$ such that $\delta(b^i) \leq \delta(b^j)$ whenever $i < j$.

Step 2 Starting at $k = 1$, for $b^k \in B$, let $G \in \mathcal{G}$ be a graph representation of a mechanism with $(|M_1|, \ldots, |M_n|) = (b_1^k, \ldots, b_n^k)$. Apply Algorithm A1 to graph G; if Algorithm A1 finds a coloring of G that satisfies the existence and nonexistence constraints, then the *δ-simplest mechanism that solves the implementation problem* is the mechanism corresponding to G with the coloring assigned in Algorithm A1. If Algorithm A1 does not find a coloring for G that satisfies the constraints, increase k by 1 and repeat this step, unless $k = |B|$, in which case the algorithm ends: the implementation problem has no solution using a mechanism with a simplicity criterion value up to z.

3.6 Examples and Special Cases

3.6.1 Nash Implementation

To simplify the notation in this subsection, we let $\overset{i}{v}$ denote the set of all vertices connected to vertex v by an edge with a weight i.

Example 5. Let $A = \{a, b, c\}$ and and consider a social choice function given by $F(R) = b$ and $F(R') = c$, where R and R' are given by

R_1	R_2	R'_1	R'_2
a	c	b	c
b	b	c	a
c	a	a	b

Any mechanism that implements F must have a NE with outcome b at R and a NE with outcome c at R'; it must also have no NE at R with outcomes a or c and no NE at R' with outcomes a or b. The existence and nonexistence constrains state these constraints in terms of the mechanism.

Let $G = (V, E)$ be a graph in $\mathcal{R}$. Note that for any $(v, q) \in V \times A$, the set of all colorings that are consistent with v being an equilibrium with outcome q at R, $\mathcal{C}_v(q|G, S, R)$, is simply the set of all colorings with $C(v) = q$ and $C(\overset{i}{v})$ in the set of weakly less preferred outcomes for player i. We now state the existence and non-existence constraints.

Existence Constraints

- $\exists v \in V$ with: $C(v) = b$, $C(\overset{1}{v}) \in \{b, c\}$, and $C(\overset{2}{v}) \in \{b, a\}$
- $\exists v \in V$ with: $C(v) = c$ and $C(\overset{1}{v}) \in \{c, a\}$

Nonexistence Constraints

- $\nexists v \in V$ with: $C(v) = a$ and $C(\overset{2}{v}) \in \{a\}$
- $\nexists v \in V$ with: $C(v) = c$ and $C(\overset{1}{v}) \in \{c\}$
- $\nexists v \in V$ with: $C(v) = b$ and $C(\overset{2}{v}) \in \{b\}$
- $\nexists v \in V$ with: $C(v) = a$, $C(\overset{1}{v}) \in \{a\}$, and $C(\overset{2}{v}) \in \{a, b\}$

We can use the algorithms from the previous section to find the δ-simplest mechanism to implement the above social choice rule. We define simplicity by the number of strategy profiles in the mechanism, so $\delta(b) = \prod_{i=1}^{n} b_i$ for $b = (b_1, \ldots, b_n) \in \mathbb{N}^n$. The algorithms yield the following two mechanisms, depending on the enumeration chosen in step 4 of Algorithm A1:

Mech. 1

b	a
b	c

Mech. 2

b	a
c	c

Each of these mechanisms implement F in Nash equilibria, and they are the *smallest* (in terms of the number of strategy profiles) mechanisms that do so. In contrast, the canonical mechanism for Nash implementation (Maskin's mechanism) has twelve strategies for each player when the strategy space is reduced by using a modulo game.

We next show that Maskin monotonicity is equivalent to the no-conflict condition when the solution concept is Nash equilibrium.

Theorem 3.6.1. If Nash equilibrium is the solution concept then Maskin monotonicity holds in an implementation problem $(I, A, \mathcal{R}, F, S)$ if and only if the no-conflict condition holds.

Proof. First, suppose the no-conflict condition fails: for every $G \in \mathcal{G}$, there exists $a \in F(R)$ and $a' \notin F(R')$ such that $\mathcal{C}(a|G, S, R) \subseteq \mathcal{C}(a'|G, S, R')$. I will show that Maskin monotonicity

also fails, that is: there exists $(a, R, R') \in A \times \mathcal{R} \times \mathcal{R}$ with $a \in F(R)$, $a \notin F(R')$, but where for all $(i, b) \in I \times A$ it is true that $aR'_i b$ whenever $aR_i b$. Thus, I will show that a is chosen at R and not R', but no player has a preference reversal between a and a weakly less preferred outcome going from R to R'.

Let G be any graph in $\mathcal{G}$. $\mathcal{C}(a|G, S, R)$ is the set of all colorings such that a vertex v with color a exists and where any vertex v' connected to v with weight i has a color in $L_i(a, R)$ (i.e. $C(\overset{i}{v}) \subseteq L_i(a, R)$ for all $i \in I$), where $L_i(a, R) \equiv \{b \in A : aR_i b\}$. By assumption, there exists $a \in F(R)$ and $a' \notin F(R')$ such that $\mathcal{C}(a|G, S, R) \subseteq \mathcal{C}(a'|G, S, R')$. Suppose $a \neq a'$; then the coloring where all vertices are colored a is in $\mathcal{C}(a|G, S, R)$ but not in $\mathcal{C}(a'|G, S, R')$ showing that $\mathcal{C}(a|G, S, R) \nsubseteq \mathcal{C}(a'|G, S, R')$ and violating the assumption that the no-conflict condition fails. Thus, $a = a'$. Again, by assumption, there exists $a \in F(R)$ and $a \notin F(R')$ (with $R \neq R'$) such that $\mathcal{C}(a|G, S, R) \subseteq \mathcal{C}(a|G, S, R')$. Suppose some player has a preference reversal between a and a weakly less preferred outcome going from R to R'; that is, there exists $(i, b) \in I \times A$ with $aR_i b$, but $bP'_i a$. Then a graph with two vertices with colors a and b connected with an edge of weight i has coloring in $\mathcal{C}(a|G, S, R)$ (the vertex with color a is a Nash equilibrium in the corresponding mechanism), but its coloring is not in $\mathcal{C}(a|G, S, R')$, leading to a contradiction. Thus, no player has a preference reversal between a and a weakly less preferred outcome going from R to R' and Maskin monotonicity fails.

Next, suppose that Maskin monotonicity fails. I will show that the no-conflict condition also fails. By failure of Maskin monotonicity, there is some $a \in F(R)$ and a weakly rises in each player's preferences going from R to R', but $a \notin F(R')$. Let G be any graph in $\mathcal{G}$. Because $a \in F(R)$, the existence constraints imply that there is a vertex v with color a and where $C(\overset{i}{v}) \subseteq L_i(a, R)$ for all $i \in I$ (the set of all colorings satisfying this property is $\mathcal{C}(a|G, S, R)$). Because $a \notin F(R')$, the nonexistence constraints imply that there is no vertex v with color a and where $C(\overset{i}{v}) \subseteq L_i(a, R')$ for all $i \in I$. However, $L_i(a, R) \subseteq L_i(a, R')$ for all $i \in I$, which implies that $\mathcal{C}(a|G, S, R) \subseteq \mathcal{C}(a|G, S, R')$, and the no-conflict condition fails. ■

3.7 Conclusions and Further Work

This paper shows that any implementation problem can be formulated as a question about the existence of a graph that solves a graph coloring problem. We presented a necessary no-conflict condition for implementation, and, focusing on the special case of Nash implementation, we have shown that it is equivalent to Maskin monotonicity when the solution concept is Nash equilibrium. We also presented general algorithms that can solve arbitrary implementation problems using simple mechanisms (in an arbitrary notion of simplicity) when they exist.

Understanding the computational complexity of general implementation problems is important from a practical point of view. We have shown that coloring a graph so that its corresponding mechanism solves an implementation problem is a special case of a constraint satisfaction problem. In general, constraint satisfaction problems are NP-complete. Because computational complexity of graph coloring, and constraint satisfaction problems more generally, has been extensively studied, the connection we make to graph theory is an important step in understanding the general complexity of solving implementation problems and the possibility of obtaining different complexity results based on the environment.

Besides Maskin monotonicity, it is likely that the no-conflict condition is also equivalent to other monotonicity conditions in the literature when the appropriate solution concept is used. More generally, the graphical approach we use opens the door for applications of graph theoretic, combinatorial, and geometric results in implementation theory. Such applications would lead to a better understanding of existing results in implementation theory and potentially establish powerful general results that are independent of any solution concept.

Appendices

Appendix A

Chapter 1 Appendix

Proof of Theorem 1.3.1

Let (M, g) be a mechanism that strongly $(\bar{d}, \bar{v})$-safe implements F. To prove this theorem, we will show that if condition (1) does not hold, then condition (2) holds. Suppose $a \in F(R)$ and condition (1) doesn't hold so that $\exists R'$ with $a \notin F(R')$. Because F is strongly $(\bar{d}, \bar{v})$-safe implementable, there is a Nash equilibrium at R, say m, such that $g(m) = a$. Because $a \notin F(R')$, m cannot be a NE at R'. Hence, at least one player, say player i, has a profitable deviation from m to obtain some other outcome, say b, at R' and thus, $bP_i'a$. At R, player i must weakly prefer a to b, otherwise obtaining b would be profitable for player i at R and m would not be a NE at R. Hence, aR_ib. Now consider another profile, R'', where a rises in all players' ranks relative to R. m must then be a NE at R''. Because every NE is $(\bar{d}, \bar{v})$-safe and because player i can obtain b by deviating from m, b cannot victimize more than $\bar{v}$ players in $I\backslash\{i\}$ at R''. Hence, condition (2) holds.

Proof of Corollary 1.3.2

To show a contradiction, suppose (M, g) is a mechanism that strongly $(\bar{d}, \bar{v})$-safe implements F on $\mathcal{R}$ and $\mathcal{R}$ contains all possible strict preferences. Let $a \in F(R)$ for some $R \in \mathcal{R}$. We will first show that condition (1) of Theorem 1.3.1 holds by showing that condition (2) of that theorem does not hold. Suppose $a \notin F(R')$ for some $R' \in \mathcal{R}$. Because the set of admissible preference profiles contain all possible preference profiles, for every pair of player and outcome $(i, b) \in I \times A$, there is a preference profile where a weakly rises in every

player's preferences relative to R and b is the least preferred by all players, violating condition (2). Hence, condition (1) holds and, because a was arbitrary, F is a constant social choice rule. Consider a profile, $R^* \in \mathcal{R}$, where a is strictly the least preferred outcome by all players. By condition (1) of Theorem 1.3.1, $a \in F(R^*)$. By strong $(\bar{d}, \bar{v})$-safe implementability, there is a NE at R^*, say m^*, such that $g(m^*) = a$. However, a victimizes all players so m^* cannot be safe, leading to a contradiction.

Proof of Theorem 1.3.3

The necessity of the $\bar{v}$-safe outcome property is trivial as it follows directly from the $(\bar{d}, \bar{v})$-SE definition. To show the necessity of $(\bar{d}, \bar{v})$-safe monotonicity, suppose that F is implementable in $(\bar{d}, \bar{v})$-safe equilibrium by mechanism (M, g) and suppose that R and R' are two preference profiles such that $a \in F(R)$ and conditions (1) and (2) of the $(\bar{d}, \bar{v})$-safe monotonicity property hold. By implementability, there must be a $(\bar{d}, \bar{v})$-safe equilibrium $m \in M$ at R such that $g(m) = a$. For any $i \in I$ let $O_i(m) \subseteq A$ be the set of outcomes that player i can obtain by a (potential) deviation from m. Because m is a $(\bar{d}, \bar{v})$-safe equilibrium it must be that for any $y \in O_i(m)$, aR_iy and y victimizes at most $\bar{v}$ of the players in set $I \backslash \{i\}$ under R. By condition (1) of $(\bar{d}, \bar{v})$-safe monotonicity, this must also hold under R'. Additionally, (in the case of $\bar{d} \geq 2$), for any subset of players $B \subseteq I$ of size $\bar{d}$, let $O_B(m)$ be the attainable set by players in B from m. Because m is a $(\bar{d}, \bar{v})$-safe equilibrium, it must be that every element in $O_B(m)$ victimizes at most $\bar{v}$ of the players in $I \backslash B$ under R. Condition (2) of $(\bar{d}, \bar{v})$-safe monotonicity implies that the same holds for R'. Hence, we have shown that, under R', the profile m is such that no player can gain by deviating and no group of $\bar{d} \geq 1$ players can victimize more than $\bar{v}$ of the others by deviating. Therefore, m is a $(\bar{d}, \bar{v})$-safe equilibrium at R' and $a \in F(R')$.

Proof of Theorem 1.3.4

Suppose that F is implementable in $(\bar{d}, \bar{v})$-safe equilibrium by mechanism $\Gamma = (M, g)$ and suppose that $\bar{d} \geq \frac{n}{2}$. Let $\mathcal{R}$ be the set of preference profiles and for each $R^j \in \mathcal{R}$ and $g^j \subseteq I$ let $H(g^j, R^j) \subseteq A$ be the set of outcomes that cause $\bar{v}$ or less victims in g^j under profile R^j. Suppose, by way of contradiction, that the set of preference profiles, $\mathcal{R}$, violates the $(\bar{d}, \bar{v}, n)$-similarity property. Then, there exists a set $\{(g^1, R^1), \ldots, (g^q, R^q)\}$ with $q = \left\lfloor \frac{n}{n-\bar{d}} \right\rfloor$, $|g^i| \geq n - \bar{d}$, and $g^i \cap g^k = \varnothing$, $i, k \in \{1, \ldots, q\}$, and where each outcome in A causes more than $\bar{v}$ victims in g^j at R^j for some $j \in \{1, \ldots, q\}$; that is to say, $\bigcap_{j \in \{1,2,\ldots,q\}} H(g^j, R^j)$ is empty. For each $j \in \{1, \ldots, q\}$ let $SE^{\Gamma}(R^j)$ be the set of $(\bar{d}, \bar{v})$-SE of Γ under R^j. Suppose m is any profile where $n - \bar{d}$ or more players g^j follow the messages in some $m' \in SE^{\Gamma}(R^j)$. Then players in $I \backslash g^j$, a set of $\bar{d}$ or less players, cannot victimize more than $\bar{v}$ players in g^j. Thus, $g(m) \in H(g^j, R^j)$ by the fact that m' is a $(\bar{d}, \bar{v})$-SE at R^j. By the definition of q and because $\bar{d} \geq \frac{n}{2}$, there exists a profile, m^*, and q different groups of players of size at least $n - \bar{d}$, call them $\{g^1, \ldots, g^q\}$, such that for each $j \in \{1, \ldots, q\}$, $m^*_{g^j} = m'_{g^j}$ for some $m' \in SE^{\Gamma}(R^j)$. Hence, it must be that $g(m^*) \in \bigcap_{j \in \{1,\ldots,q\}} H(g^j, R^j)$. However, $\bigcap_{j \in \{1,2,\ldots,q\}} H(g^j, R^j)$ is empty by assumption, leading to a contradiction.

Proof of Theorem 1.3.5

Suppose first that $\bar{d} < \frac{n}{2}$ and the conditions of the Theorem are satisfied. Let $\Gamma = (\times_{i \in I} M_i, g)$ be the implementing mechanism, where the message space for each player $i \in I$ is $M_i = A \times \mathcal{R}^* \times \{1, \ldots, n\}$ and $\mathcal{R}^*$ is the set of preference profiles. A typical message is $m_i = (a^i, R^i, z^i)$. The outcome function, g, is determined as follows:

Rule 1 If $(a^i, R^i, z^i) = (a, R, z)$ for all $i \in I$ and $a \in F(R)$ then $g(m) = a$.

Rule 2 Suppose there exists i such that $(a^j, R^j, z^j) = (a, R, z)$ for all $j \neq i$, $a \in F(R)$, and $(a^i, R^i, z^i) \neq (a, R, z)$. Then $g(m) = a^i$ if $aR_i a^i$ and if a^i does not victimize more than

$\bar{v}$ players in the set $I\backslash\{i\}$ under R; otherwise, $g(m) = a$.

Rule 3 Suppose $\bar{d} \geq 2$ and there exists a set of players $B \subseteq I$ with $1 < |B| \leq \bar{d}$ such that $(a^j, R^j, z^j) = (a, R, z)$ for all $j \notin B$, $a \in F(R)$, and $(a^i, R^i, z^i) \neq (a, R, z)$ for all $i \in B$. Let B' be the set of players in B who report an outcome that does not victimize more than $\bar{v}$ players in $I\backslash B$ under R. If B' is not empty, then sort the players in B' from lowest to highest according their index, rename player $i \in B'$ by the new sorted order $q_i \in \{1, \ldots, |B'|\}$, and let $g(m) = a^{q_i^*}$, where $q_i^* = 1 + ((\sum_{i \in B'} z^i) \mod |B'|)$ is the q_i^*'th player in B'; otherwise $g(m) = a$.

Rule 4 In all other cases, $g(m) = a^j$ for $j = 1 + ((\sum_{i \in I} z^i) \mod n)$.

Let R be the true preference profile and suppose $a \in F(R)$. We first show that there exists a safe equilibrium m of the mechanism Γ such that $g(m) = a$. Let m be given by $m_i = (a, R, 1)$. By rule 2, any unilateral deviation from m cannot result in a more preferred outcome under R for the deviator, so m is a Nash equilibrium. Also under rule 2, a unilateral deviation from m by an arbitrary player i cannot cause more than $\bar{v}$ players in $I\backslash\{i\}$ to be victimized under R. A multilateral deviation from m by an arbitrary set B of $\bar{d}$ or less players falls under rule 3, which guarantees that the outcome implemented will not victimize more than $\bar{v}$ players in $I\backslash B$ under R. Finally, a does not victimize more than $\bar{v}$ players under R because F satisfies the $\bar{v}$-safe outcome property. Hence, m is a $(\bar{d}, \bar{v})$-safe equilibrium.

Next, I show that if m is a $(\bar{d}, \bar{v})$-SE of Γ then $g(m) \in F(R)$. Because of the assumption that R satisfies $(\bar{d}, \bar{v})$-exposure, no $(\bar{d}, \bar{v})$-SE of Γ can fall under rules 2-4. If there were a $(\bar{d}, \bar{v})$-SE under rules 2-4 it would be possible for a group, B, of $\bar{d}$ players to attain any outcome in A via rule 4.[1] By assumption, R satisfies $(\bar{d}, \bar{v})$-exposure so the set of outcomes

[1] Rule 4 is reached when there are more than $\bar{d}$ *different* messages reported by the players. Note that it is enough for a message to differ only on the R dimension in order to be counted "different".

that cause more than $\bar{v}$ players in $I \backslash B$ to be victims is nonempty. This implies that a group of $\bar{d}$ players can victimize more than $\bar{v}$ of the rest, violating the definition of the $(\bar{d}, \bar{v})$-SE. Hence, the only $(\bar{d}, \bar{v})$-safe equilibria of Γ fall under rule 1.

Suppose m is a $(\bar{d}, \bar{v})$-SE under rule 1 such that $m_i = (a', R', z')$ for all $i \in I$. For an arbitrary player i, let $O_i(m) \subseteq A$ be the set composed of a' and all the outcomes that player i can obtain by deviating from m. By rule 2 and the $\bar{v}$-safe outcome property, if $y \in O_i(m)$ then $a' R_i' y$ and y must not victimize more than $\bar{v}$ players in $I \backslash \{i\}$ under R'. However, because m is a $(\bar{d}, \bar{v})$-SE under the true preference profile, R, it must be that $a' R_i y$ and y cannot victimize more than $\bar{v}$ players in $I \backslash \{i\}$ under R. Additionally, rule 2 implies that $O_i(m)$ contains *all* the outcomes that are both weakly less preferred to a' under R' by player i and victimize at most $\bar{v}$ of the players in $I \backslash \{i\}$ under R'. Hence, the first condition of $(\bar{d}, \bar{v})$-safe monotonicity is satisfied.

Suppose $\bar{d} \geq 2$. Then for an arbitrary set of players, B, of size $\bar{d}$, let $O_B(m)$ be set composed of a' and all the outcomes that players in B can obtain by a (potentially multilateral) deviation from m. By rules 2 and 3 and the $\bar{v}$-safe outcome property, if $y \in O_B(m)$ then y victimizes at most $\bar{v}$ players in $I \backslash B$ under R'. However, m is a $(\bar{d}, \bar{v})$-SE under R, so it must be that y victimizes at most $\bar{v}$ players in $I \backslash B$ under R. Because $O_B(m)$ contains, via rules 2 and 3, *all* elements that victimize at most $\bar{v}$ players in $I \backslash B$ under R', the second condition of $(\bar{d}, \bar{v})$-safe monotonicity is satisfied.

By $(\bar{d}, \bar{v})$-safe monotonicity, $a' \in F(R)$, concluding the proof for the case of $\bar{d} < \frac{n}{2}$.

If $\frac{n}{2} \leq \bar{d} < n - 1$ the proof remains unchanged except that the set of possible profiles, $\mathcal{R}^*$, is assumed to be $(\bar{d}, \bar{v}, n)$-similar and rule 3 of Γ is modified to be:

Rule 3 Suppose rule 2 does not apply and suppose there are p *consensus groups*, $B_1, \ldots, B_p \subseteq I$, each of size $n - \bar{d}$ or more players with everyone in B_1 reporting $(\bar{a}^1, \bar{R}^1, \bar{z}^1)$, everyone in B_2 reporting $(\bar{a}^2, \bar{R}^2, \bar{z}^2)$, $\ldots$, and everyone in B_p reporting $(\bar{a}^p, \bar{R}^p, \bar{z}^p)$. Suppose

also that $(\bar{a}^1, \bar{R}^1, \bar{z}^1) \neq (\bar{a}^2, \bar{R}^2, \bar{z}^2) \neq \ldots \neq (\bar{a}^p, \bar{R}^p, \bar{z}^p)$, and if i is a player not in any of the consensus groups then $(a^i, R^i, z^i) \neq (\bar{a}^h, \bar{R}^h, \bar{z}^h)$ for $h \in \{1, \ldots, p\}$.

- Case of $p = 1$. Set $g(m) = a^q$ for $q = \max\{j \in I \backslash B_1\}$ if no element in the set $\{a^j : j \in I \backslash B_1\}$ victimizes more than $\bar{v}$ players in set B_1 under profile $\bar{R}^1$; otherwise $g(m) = \bar{a}^1$.
- Case of $p > 1$. Set $g(m) = b$ for some b that does not victimize more than $\bar{v}$ players in B_j under $\bar{R}^j$ for all $j \in \{1, \ldots, p\}$. Such a b exists because of the assumption that $\mathcal{R}^*$ is $(\bar{d}, \bar{v}, n)$-similar.

Note that because $\frac{n}{2} \leq \bar{d} < n - 1$, no $(\bar{d}, \bar{v})$-SE can fall under the modified rule 3. It is always true that $\bar{d}$ players can deviate from the modified rule 3 and reach rule 4. Also note that, through case $p = 1$ of the modified rule 3, a set of deviators from rule 1 can attain all outcomes that do not victimize more than $\bar{v}$ players of the rest. This is needed to satisfy $(\bar{d}, \bar{v})$-safe monotonicity.

Proof of Theorem 1.3.6

Suppose F is implementable in both $(\bar{d}, \bar{v})$-safe and Nash equilibria. If $a \in F(R)$ and safe implementation holds, then there is a safe equilibrium, call it m, at R with $g(m) = a$. If player i can deviate from m and obtain outcome b then aR_ib and b victimizes at most $\bar{v}$ players in $I \backslash \{i\}$, because m is a safe equilibrium. If all such outcomes are weakly less preferred than a for player i under preference profile R', and if this holds for all players (as $\bar{v}$-safe-Nash monotonicity states), then m is a Nash equilibrium at R' and $a \in F(R')$ by Nash implementation.

Proof of Theorem 1.3.7

Consider the mechanism used in the proof of Theorem 1.3.5 with the following transfer scheme:

Transfer Scheme 1

Rule 1 All transfers are zero under rule 1.

Rule 2 Suppose all players agree on (a, R, z) with $a \in F(R)$, except for one player i who reports $(a', R', z') \neq (a, R, z)$. If $R' \neq R$ and $z' \neq z$ then all transfers are zero; otherwise $t_i = -L$ and $t_j = 0$ for $j \neq i$.

Rule 3 Under this rule $\bar{d} \geq 2$ and all players agree on (a, R, z), except a set of players B who submit different reports and $1 < |B| \leq \bar{d}$. Set the transfer of all players in B to $-L$ and all others to zero unless the following is true: $|B| = 2$, call them players i and j, with player i reporting (a', R', z') and player j reporting (a'', R'', z''), with $R' \neq R$, $z' \neq z$, $R'' = R$, and $z'' = z'$, then set $t_i = -L$, $t_j = L$, and all other transfers to zero.

Rule 4 Set $t_i = -L \sum_{j \neq i} \mathbf{1}\{z^j = 1 + (z^i \mod n)\} + L \sum_{j \neq i} \mathbf{1}\{z^i = 1 + (z^j \mod n)\}$.

We will prove that the mechanism used in Theorem 1.3.5, along with the transfers in Transfer Scheme 1, double implement F. Let R be the true preference profile and suppose $a \in F(R)$.

First note that, following the same argument in the proof of Theorem 1.3.5, the message profile where all players report $(a, R, 1)$ is a $(\bar{d}, \bar{v})$-safe equilibrium with outcome a. Thus, all outcomes in F can be obtained via a $(\bar{d}, \bar{v})$-safe equilibrium (i.e. a Nash equilibrium) of the mechanism with transfers.

Second, we show that if m is a Nash equilibrium of the mechanism with transfers then $g(m) \in F(R)$ (this also implies that all $(\bar{d}, \bar{v})$-safe equilibria lead to outcomes in $F(R)$).

Suppose m is a Nash equilibrium under rule 1 with every player reporting (a', R', z'). For an arbitrary player i, let $O_i(m) \subseteq A$ be the set composed of a' and all the outcomes that player i can obtain by deviating from m. By rule 2 and the $\bar{v}$-safe outcome property, if $b \in O_i(m)$ then $a' R'_i b$ and b does not victimize more than $\bar{v}$ players in $I \backslash \{i\}$ under R'. However, because m is a Nash equilibrium under the true preference profile R, and because player i can obtain any outcome in $O_i(m)$ without getting fined (by reporting a different preference profile and integer than other players), it must be that $a' R_i b$. Additionally, rule 2 implies that $O_i(m)$ contains all the outcomes that are both weakly less preferred to a' under R' by player i and victimize at most $\bar{v}$ of the players in $I \backslash \{i\}$ under R'. Hence, $a' \in F(R)$ by $(\bar{d}, \bar{v})$-safe-Nash monotonicity.

Next, we show that no Nash equilibria exist under rules 2,3, and 4. Suppose m is a Nash equilibrium under rule 2 with player i reporting (a', R', z') and all others reporting $(\overline{a}, \overline{R}, \overline{z})$. Player i cannot be fined at a Nash equilibrium under rule 2, because he can report $R' \neq \overline{R}$ and $z' \neq \overline{z}$ and avoid the fine without affecting the outcome chosen. Thus, it must be true that $R' \neq \overline{R}$ and $z' \neq \overline{z}$. However, any other player can deviate and report $(\overline{a}, \overline{R}, z')$ and obtain a transfer L from player i, via rule 3, and this deviation will be profitable, contradicting m being a Nash equilibrium.

Suppose m is a Nash equilibrium under rule 3 with all players agreeing on $(\overline{a}, \overline{R}, \overline{z})$, except a set of players B with different reports. At least one player in B has transfer $-L$, and any such player has an incentive to deviate to $(\overline{a}, \overline{R}, \overline{z})$ and avoid the fine, contradicting m being a Nash equilibrium.

Finally, suppose m is a Nash equilibrium under rule 4. For the rest of the proof we assume modular n arithmetic, so that if $z = n$ then $z + 1 = 1$, $z + 2 = 2$, etc... There must be an integer that no player is reporting, otherwise each integer in $\{1, \ldots, n\}$ is being reported by exactly one player and all players would have an incentive to deviate (e.g. the player reporting integer n can deviate to reporting $n - 1$ and have transfer L instead of

0). Let z^* be the smallest element in $\{1, \ldots, n\}$ such that no player reports z^* and at least one player reports $z^* - 1$. Because no player reports z^*, it must be that no player has a negative transfer, otherwise a player with negative transfer can deviate to reporting $z^* - 1$ and pay no fine. Because the transfers among the agents are zero-sum, each agent's transfer is zero. At least 2 players must be reporting $z^* + 1$, otherwise at least one player (possibly one reporting $z^* + 1$) would have an incentive to deviate and report z^* to obtain a positive transfer. Because players reporting $z^* + 1$ have zero net transfers, and because no players report z^*, it must be that no players report $z^* + 2$. Following the same argument, at least two players report $z^* + 3$ and no players report $z^* + 4$, at least two players report $z^* + 5$ and no players report $z^* + 6$, etc... Such reporting behavior is only possible if exactly two players are reporting each of $z^* + 1$, $z^* + 3$, $z^* + 4$, etc... But then a player reporting $z^* + 3$ has an incentive to deviate and report $z^* + 2$, obtaining a transfer of L. This leads to a contradiction and m cannot be a Nash equilibrium.

Proof of Corollary 1.3.8

Suppose the preference profile is R. In the proof of Theorem 1.3.7, we showed the following: for any outcome $a \in F(R)$, there exists a safe equilibrium, m, with $g(m) = a$; for any Nash equilibrium under rule 1 where all players report (a', R', z'), $a' \in F(R)$; and that there are no Nash equilibria under rules 2, 3, or 4. Thus, any unsafe Nash equilibrium must fall under rule 1, where all players agree on a report. If the preference profile being reported by all players is R under rule 1, then the message profile is safe by the design of the mechanism (in the same way that we showed the profile where each player reports $(a, R, 1)$ is safe in the proof of Theorem 1.3.5). Hence, it must be that all players report a false preference profile at any unsafe Nash equilibrium.

Suppose m is an unsafe Nash equilibrium. Double implementation implies that $g(m) \in F(R)$. Thus, the (truthful) profile where each player reports $(g(m), R, 1)$ falls under

rule 1, is safe by the design of the mechanism, and yields the same outcome as m.

Proof of Theorem 1.3.9

Suppose F is double implementable in $(\bar{d}, \bar{v})$-safe and Nash equilibria, but $\bar{v}$-safe-Nash monotonicity fails. Thus, there are two preference profiles R and R' with: $a \in F(R)$, and, for all $i \in I$, aR_ib and b victimizes at most $\bar{v}$ players in $I\backslash\{i\}$ implies aR'_ib, but $a \notin F(R')$. There must be a safe equilibrium at R with outcome a, call it m. $a \notin F(R')$ by assumption, so m cannot be a Nash equilibrium at R', otherwise $g(m) = a$ must be in $F(R')$ by Nash implementation. Thus, some player, call it i, has a profitable deviation from m to another profile, call it m', at R'. The outcome at m' must be different than a, otherwise the transfer required to induce a deviation from m to m' at R' would also induce the same deviation at R and m would not be a Nash equilibrium at R; let the outcome at m' be b. Because m is a safe equilibrium at R, aR_ib and b victimizes at most $\bar{v}$ players in $i\backslash\{i\}$ at R, and – by our assumption about the failure of $\bar{v}$-safe-Nash monotonicity between R and R' – aR'_ib; thus, $(a,0)R_i(b,0)$ and $(a,0)R'_i(b,0)$. However, player i must have an incentive to deviate to m' at R', and this is only possible if the designer can set transfers $t(m), t(m') \in \mathbb{R}$ such that: (1) $(a,t(m))R_i(b,t(m'))$, because m is a Nash equilibrium at R; and (2) $(b,t(m'))P_i(a,t(m))$, because player i has an incentive to deviate from m to m' at R'. To set such transfers, or determine that none exist, is only possible if the designer has precise cardinal information about preferences for player i and outcomes $\{a,b\}$ at preference profiles $\{R,R'\}$ as defined in definition 1.3.10.

Proof of Lemma 1.4.1

For the purpose of deriving a contradiction assume that there exists a mechanism, $(\times_{i\in I}M_i, g)$, that implements a social choice function f in a $(n-1,0)$-SE and that for some i, $\exists D \subseteq A$ such that $|D| = K$ and $\forall m_i \in M_i\ \exists m_{-i} \in M_{-i}$ with $g(m_i, m_{-i}) \in D$.

Since implementation must hold for every possible preference profile, consider the preference profile where D is player i's least favorite K outcomes. This profile is one of the possible ones because the set of preference profiles contains all strict preferences. Let m be a $(n-1,0)$-SE. By assumption, $\exists m'_{-i} \in M_{-i}$ such that $g(m_i, m'_{-i}) \in D$ violating the definition of a $(n-1,0)$-SE.

Proof of Theorem 1.4.2

Suppose $x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$ and let f be any SCF.

Case 1: $x \leq nK$. Let $R \in \mathcal{R}$ be the preference profile where the least preferred K elements for player 1 are elements $a_1, ..., a_K$, the least preferred for player 2 are $a_{K+1}, ..., a_{2K}$, etc... Since $x \leq nK$ then at some point each element in A will be among the least preferred K ones for some $i \in I$. Hence, any element picked by f in this case will result in at least 1 victim and f is not implementable in a $(n-1,0)$-SE.

Case 2:[2] $nK < x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$. By Lemma 1.4.1, any mechanism, $(\times_{i\in I} M_i, g)$, that implements a social choice function in this case must allow every player to rule out their least favorite K outcomes. Furthermore, in a $(n-1,0)$-SE every player must be following a message that rules out their least favorite K outcomes because if not then $\exists m_{-i} \in M_{-i}$ so that the outcome is among the K least favorite for player i violating the definition of a $(n-1,0)$-SE. Let $R \in \mathcal{R}$ be any preference profile with the following properties:

- The second least favorite K outcomes for player i are $\{a_{(i-1)K+1}, \ldots, a_{iK}\}$ $\forall i \in I$.
- The least favorite K outcomes for players 1 and 2 are $\{a_{nK+1}, \ldots, a_{nK+K}\}$, for players 3 and 4 are $\{a_{nK+K+1}, \ldots, a_{nK+2K}\}$, etc..., until a_x is among the least favorite K outcomes for some player, call it player j, then sequentially use $\{a_1, \ldots, a_{K-1}\}$ to fill

[2] See example 2

the rest of the K least favorite outcomes for player j. The least favorite K outcomes for players $j+1,\ldots,n$ are the same as player j's.

Hence, R, is as follows:

R_1	R_2	R_3	R_4	$\ldots$	R_j	R_{j+1}	$\ldots$	R_n
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_1	a_{K+1}	a_{2K+1}	a_{3K+1}	$\ldots$	$a_{(j-1)K+1}$	a_{jK+1}	$\ldots$	$a_{(n-1)K+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_K	a_{2K}	a_{3K}	a_{4K}	$\ldots$	a_{jK}	$a_{(j+1)K}$	$\ldots$	a_{nK}
a_{nK+1}	a_{nK+1}	a_{nK+K+1}	a_{nK+K+1}	$\ldots$	$a_{nK+cK+1}$	$a_{nK+cK+1}$	$\ldots$	$a_{nK+cK+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	a_x	a_x	$\ldots$	a_x
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	a_1	a_1	$\vdots$	a_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{nK+K}	a_{nK+K}	a_{nK+2K}	a_{nK+2K}	$\ldots$	a_y	a_y	$\ldots$	a_y

where c is the integer quotient from dividing $j-1$ by 2 (and ignoring the remainder), and $y \in \{1,\ldots,K-1\}$. The set of outcomes below the dashed line are the least favorite K outcomes for each player.

Since we know that every player must pick a message that rules out their least favorite K outcomes, then any outcome picked by the social choice function will be in $\{a_1,\ldots,a_{nK}\}$. Hence, at any equilibrium $(m_1,...,m_n)$ of the game induced by the mechanism under R, the outcome is the among $\{a_{(i-1)K+1},\ldots,a_{iK}\}$ for some player i. However, since the least favorite K outcomes of player i are being ruled out by at least one other player ($x \leq nK + K\left\lfloor \frac{n}{2} \right\rfloor$), player i has an incentive to deviate from this equilibrium and rule out $\{a_{(i-1)K+1},\ldots,a_{iK}\}$ in order to get a choice that is more preferred for player i under R_i. By Lemma 1.4.1 we know that player i has a message that rules out $\{a_{(i-1)K+1},\ldots,a_{iK}\}$. Hence, $(m_1,...,m_n)$ is not a $(n-1,0)$-SE because it is not a Nash equilibrium.

Proof of Theorem 1.6.1

Suppose, by way of contradiction, that the conditions of the Theorem hold but F is double implementable in $(\bar{d}, \bar{v})$-safe and Nash equilibria. $F(R) = a$, because a is a Condorcet winner at R. Thus, there exists a $(\bar{d}, \bar{v})$-safe equilibrium, say m, with $g(m) = a$. No player can obtain b by deviating from m, because b victimizes at least $\bar{v} + 2$ players at R and m is a $(\bar{d}, \bar{v})$-safe equilibrium at R. Condition 4 of the theorem implies that any outcome that a player can obtain by deviating from m is weakly less preferred than a for that player under R', so m is a Nash equilibrium at R'. Thus, $g(m) = a \in F(R')$ by the assumption that F is implementable in Nash equilibria, but $F(R') = b$ because b is a Condorcet winner at R', leading to a contradiction.

Discussion of Section 1.4.1

Suppose first that $\bar{d} = 1$, the preference profile is R and consider a mechanism that $(\bar{d}, \bar{v})$- safe implements a SCR F. A message profile, m, of the mechanism is a $(\bar{d}, \bar{v})$-safe equilibrium if and only if $O_i(m) \subseteq L_i(a, R_i) \cap G_i(R)$ for all i, where a is the outcome at m. It follows that if at another preference profile, R', we have $L_i(a, R_i) \cap G_i(R) \subseteq L_i(a, R_i') \cap G_i(R')$ then it must be true that $O_i(m) \subseteq L_i(a, R_i') \cap G_i(R')$, m is a $(\bar{d}, \bar{v})$-SE at R', and $a \in F(R')$. This is precisely the statement of $(\bar{d}, \bar{v})$-safe monotonicity when $\bar{d} = 1$. Figure A.1 illustrates what must happen for each player between preference profiles R and R' for $(\bar{d}, \bar{v})$-safe monotonicity to impose the restriction that $a \in F(R')$.

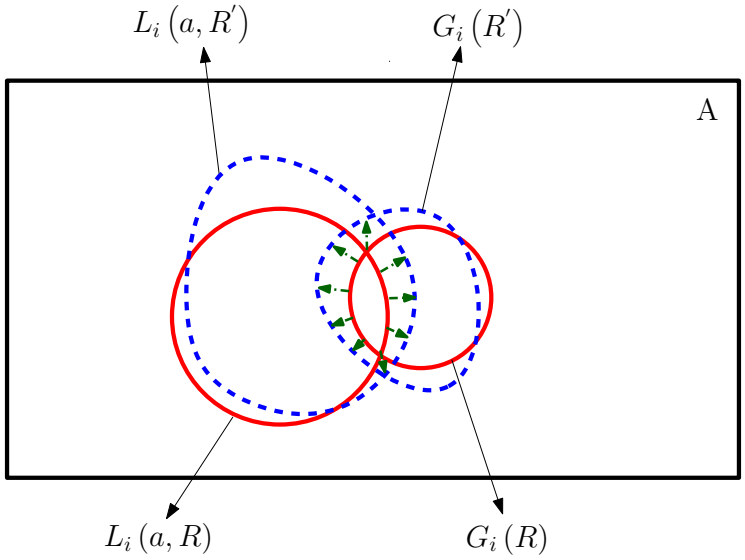

Figure A.1: Condition (1) of $(\bar{d}, \bar{v})$-Safe Monotonicity

The figure shows that the intersections of the sets $L_i(a, \cdot)$ and $G_i(\cdot)$ must (weakly) increase moving from R to R'. Note, as the figure illustrates, the lack of restrictions on the relationship between the lower contour sets $L_i(a, R)$ and $L_i(a, R')$; this contrasts with Maskin monotonicity where the only restriction is on the inclusion of $L_i(a, R)$ in $L_i(a, R')$

Figure A.1 also illustrates why $(\bar{d}, \bar{v})$-safe monotonicity does not imply Maskin monotonicity and vise versa: $L_i(a, R) \cap G_i(R) \subseteq L_i(a, R') \cap G_i(R')$ (condition (1) of $(\bar{d}, \bar{v})$-safe monotonicity) does not imply $L_i(a, R) \subseteq L_i(a, R')$ (the Maskin monotonicity condition). Similarly, if $L_i(a, R) \subseteq L_i(a, R')$ were true it would not imply $L_i(a, R) \cap G_i(R) \subseteq L_i(a, R') \cap G_i(R')$. However, it is certainly true that both conditions may hold at the same time.

When $\bar{d} \geq 2$ then a second condition must be added to the definition of $(\bar{d}, \bar{v})$-safe monotonicity to take into account the possible presence of more than one deviator. However, the first condition, which is described above and illustrated in Figure A.1, must hold for $(\bar{d}, \bar{v})$-safe monotonicity to impose a restriction on the SCR regardless of $\bar{d}$.

Appendix B

Chapter 2 Appendix

B.1 Goods and Codes Assigned

The following is a list of goods that were assigned codes. Some thefts were assigned more than one code. This occurs if more than one good was stolen in one incident, or if a good can be classified under two or more categories (e.g. sneakers get a code for athletic clothing and a code for shoes). Goods were assigned the most specific code possible. For example, shrimp was assigned the code for shrimp, not seafood. If a good was classified under a broad category (e.g. Miscellaneous), but a code for it exists in another category, it was often assigned the code from the other category as well as the recorded broad category code. For example, if the two fields describing a theft were: "Miscellaneous" and "Dog food", the codes assigned were 1200 and 802.

Alcohol	**100**
Beer	101
Spirits	102
Wine	103
Auto/Parts	**200**
Antifreeze	201
Batteries	202
Fuel	203
Oil	204
Tires	205
Wheels/rims	206
Building/Industrial	**300**
Air Conditioners/HVAC	301
Aluminum	302
Brass	303
Copper	304
Diesel/Fuel	305
Lumber/wood	306
Metal	307
Plastics	308
Shingles	309
Steel	310
Tools	311
Titanium	312
Zinc	313
Lead	314
Paint	315
Electronics	**700**
Air Conditioners/HVAC	701
Audio	702
Batteries	703
Cameras/Camcorders	704
Cell Phones and accessories	705
Computer accessories/parts	706
Computer Boards	707
Computer Monitors	708
Computers	709
Copiers/Scanners/Printers/Cartridges	710
Gaming Consoles	711
Hard Drives	712
Laptops/tablets	713
Memory	714
Microprocessors	715
TVs	716

Clothing/Shoes	**400**
Athletic	401
Boots/Shoes	402
Children's clothing	403
Fabric	404
Handbags/Purses	405
Jackets/Coats	406
Jeans	407
Women's clothing	408
Men's clothing	409
Consumer Care	**500**
Cologne/Perfume	501
Cosmetics	502
Soap/Body Wash	503
Toilet Paper	504
Toothpaste	505
Shampoo	506
Cosmetics/Personal Care	**600**
Cologne/Perfume	601
Cosmetics	602
Soap/Body Wash	603
Shampoo	604
Metals	**1100**
Aluminum	1101
Brass	1102
Bronze	1103
Cobalt	1104
Copper	1105
Lead	1106
Nickel	1107
Steel	1108
Titanium	1109
Tin	1110

Food/Drinks	**800**
Almonds	801
Animal Food	802
Apple	803
Avocado	804
Baby Food	805
Banana	806
Beans	807
Beef	808
Beverages	809
Bread	810
Butter	811
Candy/gum	812
Cereal	813
Cheese	814
Chicken	815
Chocolate	816
Coffee	817
Cooking Oil	818
Dairy	819
Energy Drinks	820
Fish	821
Fruit	822
Juice	823
Lemon/Citrus	824
Lettuce	825
Lobster	826
Meat	827
Milk	828
Nuts	829
Orange	830
Pork	831
Potatoes	832
Produce	833
Rice	834
Seafood	835
Shrimp	836
Soda	837
Sport Drinks	838
Sugar	839
Tea	840
Tomato	841
Tuna	842
Turkey	843
Water	844
Vegetables	845
Poultry	846

Home/Garden	**900**
Air Conditioners/HVAC	901
Air Freshners	902
Appliances	903
Bedding	904
Carpets/Rugs	905
Cleaning	906
Consumer Goods	907
Detergent	908
Furniture	909
Lawn Mowers/Weed Eaters	910
Lighting	911
Mattresses	912
Power Tools/Tools	913
Refrigerators	914
Soap	915
Toilet Paper	916
Vacuums	917
Washers/Dryers	918
Wood	919
Paint	920
Jewelry/Accessories	**1000**
Miscellaneous	**1200**
ATVs	1201
Bicycles	1202
Paper/Paper Products	1203
Toys	1204
Pharmaceuticals	**1300**
Baby formula	1301
Gloves	1302
Equipment (not inc gloves)	1303
OTC (not inc vitamins)	1304
Rx	1305
Vitamins	1306
Tobacco	**1400**
Cigarettes	1401
Cigars	1402

B.2 Price Series and Data Matches

We were able to make 41 matches between price series from the BLS and goods in the data. The following is a list of the matches made. For each good, a theft was counted if it had any of the "Matching Data Codes" associated with it.

Good ID	BLS Price Series Name	BLS Price Series ID	Matching Data Codes				
1	Pet Food	CUUR0000SS61031	802				
2	Apples	CUUR0000SEFK01	803				
3	Baby Food	CUUR0000SEFT05	805	1301			
4	Bananas	CUUR0000SEFK02	806				
5	Beef and veal	CUUR0000SEFC	808				
6	Bread	CUUR0000SEFB01	810				
7	Butter	CUUR0000SS10011	811				
8	Candy and chewing gum	CUUR0000SEFR02	812				
9	Breakfast cereal	CUUR0000SEFA02	813				
10	Poultry	CUUR0000SEFF	815	843	846		
11	Coffee	CUUR0000SEFP01	817				
12	Dairy and related products	CUUR0000SEFJ	819	814	828		
13	Fish and seafood	CUUR0000SEFG	821	826	835	836	842
14	Juices and nonalcoholic drinks	CUUR0000SEFN	823	838			
15	Oranges, including tangerines	CUUR0000SS11031	830				
16	Pork	CUUR0000SEFD	831				
17	Potatoes	CUUR0000SEFL01	832				
18	Tomatoes	CUUR0000SEFL03	841				
19	Carbonated drinks	CUUR0000SEFN01	837				
20	Sugar and artificial sweeteners	CUUR0000SEFR01	839				
21	Beer, ale, and other malt beverages at home	CUUR0000SEFW01	101				
22	Distilled spirits at home	CUUR0000SEFW02	102				
23	Wine at home	CUUR0000SEFW03	103				
24	Iron and steel	WPU101	310	1108			
25	Copper base scrap	WPU102301	304	1105			
26	Aluminum base scrap	WPU102302	302	1101			
27	Men's apparel	CUUR0000SEAA	409				
28	Women's apparel	CUUR0000SEAC	408				
29	Infants' and toddlers' apparel	CUUR0000SEAF	403				
30	Footwear	CUUR0000SEAE	402				
31	Tobacco and smoking products	CUUR0000SEGA	1400	1401	1402		
32	Personal care products	CUUR0000SEGB	500	-	604		
33	Fuel oil and other fuels	CUUR0000SEHE	203	305			
34	Tires	CUUR0000SETC01	205				
35	Furniture and bedding	CUUR0000SEHJ	909	904			
36	Appliances	CUUR0000SEHK	903	914	918		
37	Tools, hardware, outdoor equipment and supplies	CUUR0000SEHM	913	910	311		
38	Televisions	CUUR0000SERA01	716				
39	Audio equipment	CUUR0000SERA05	702				
40	Personal computers and peripheral equipment	CUUR0000SEEE01	706	708	709	713	
41	Toys	CUUR0000SERE01	1204				

The following is a list of other matches for major product categories that were made throughout the paper (e.g. in plots or in good-specific regressions).

BLS Price Series Name	BLS Price Series ID	Matching Data Codes										
Metals and metal products	WPU10	1100	-	1110	302	303	304	307	310	312	313	314
Food and beverages	CUUR0000SAF	800	-	846								
Apparel	CUUR0000SAA	400	-	409								
Alcoholic beverages	CUUR0000SAF116	100	-	103								

B.3 Theft Maps

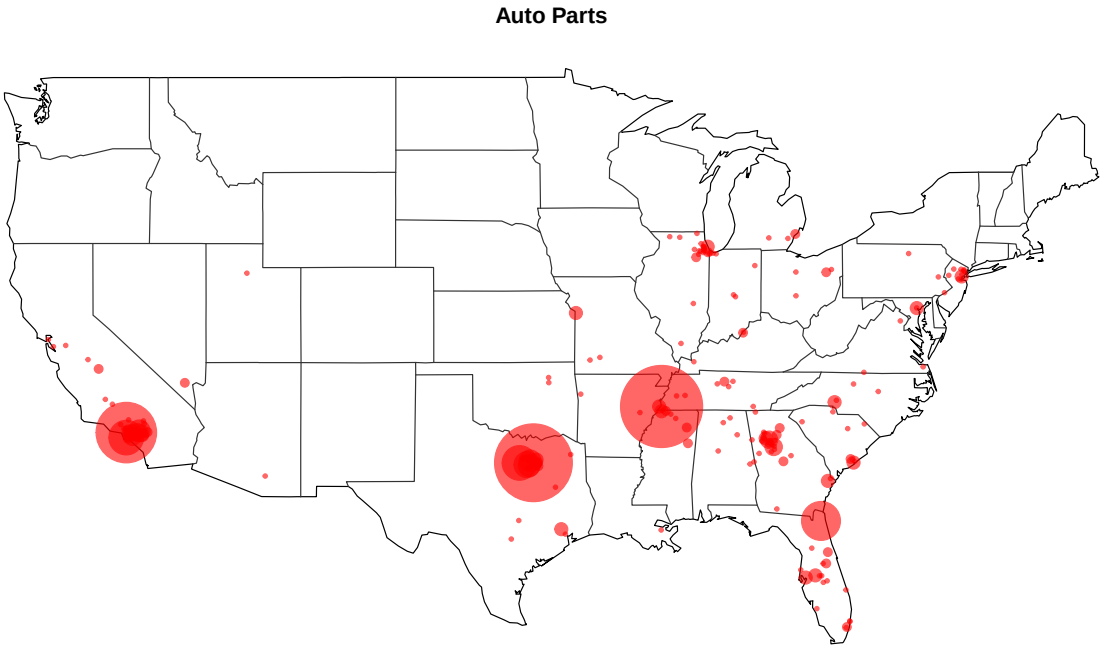

Figure B.1: Map of Auto Parts Thefts

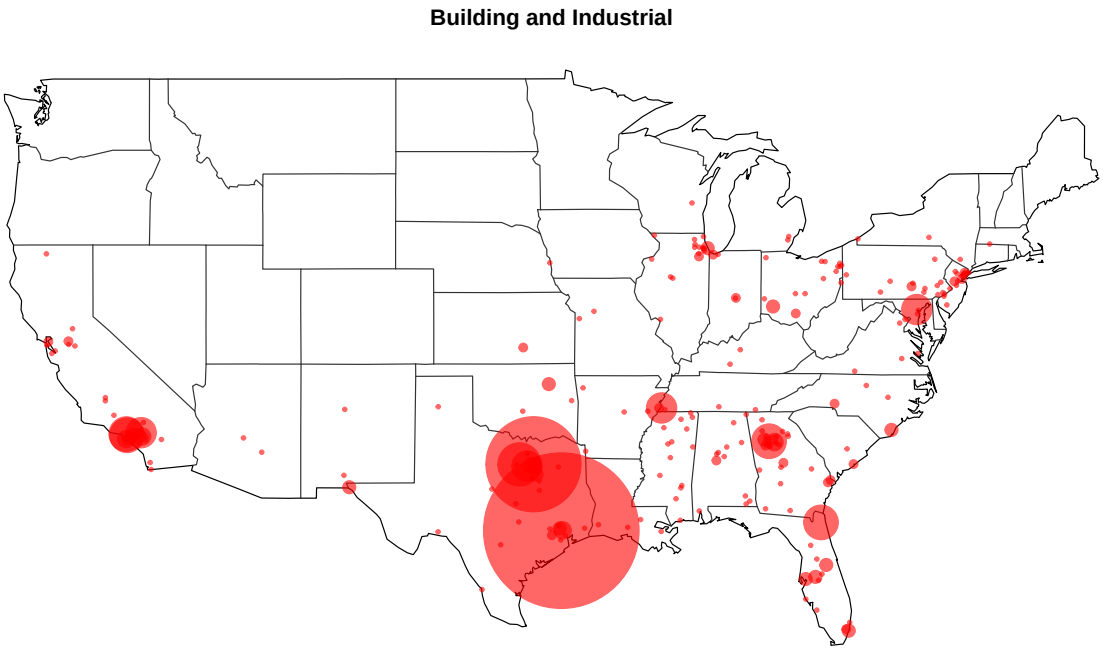

Figure B.2: Map of Building and Industrial Thefts

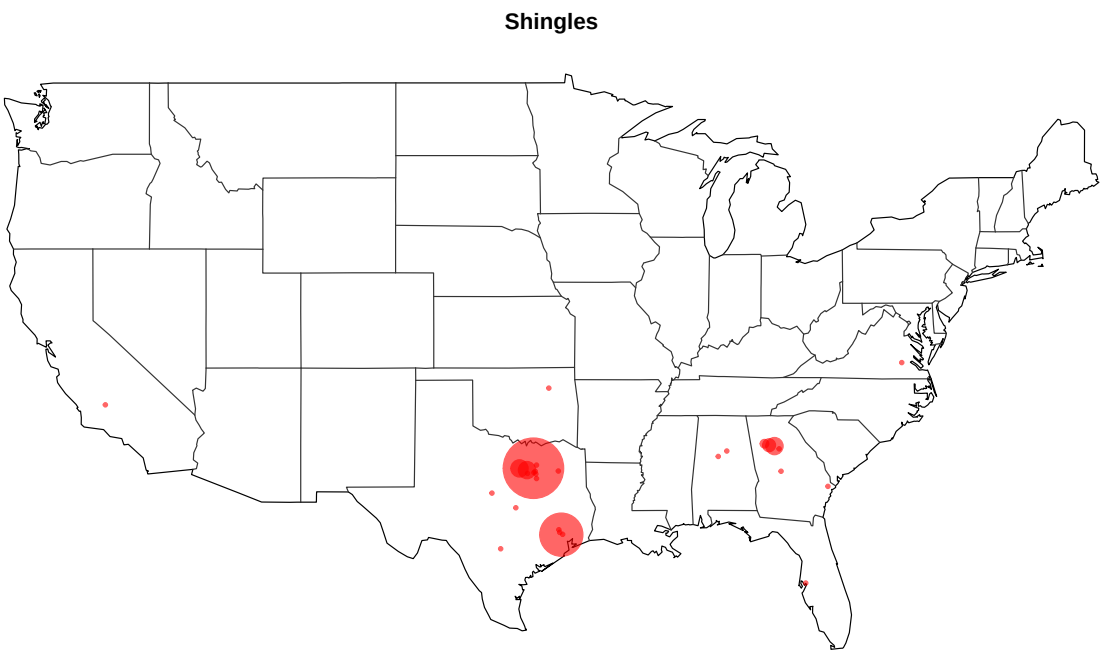

Figure B.3: Map of Shingles Thefts

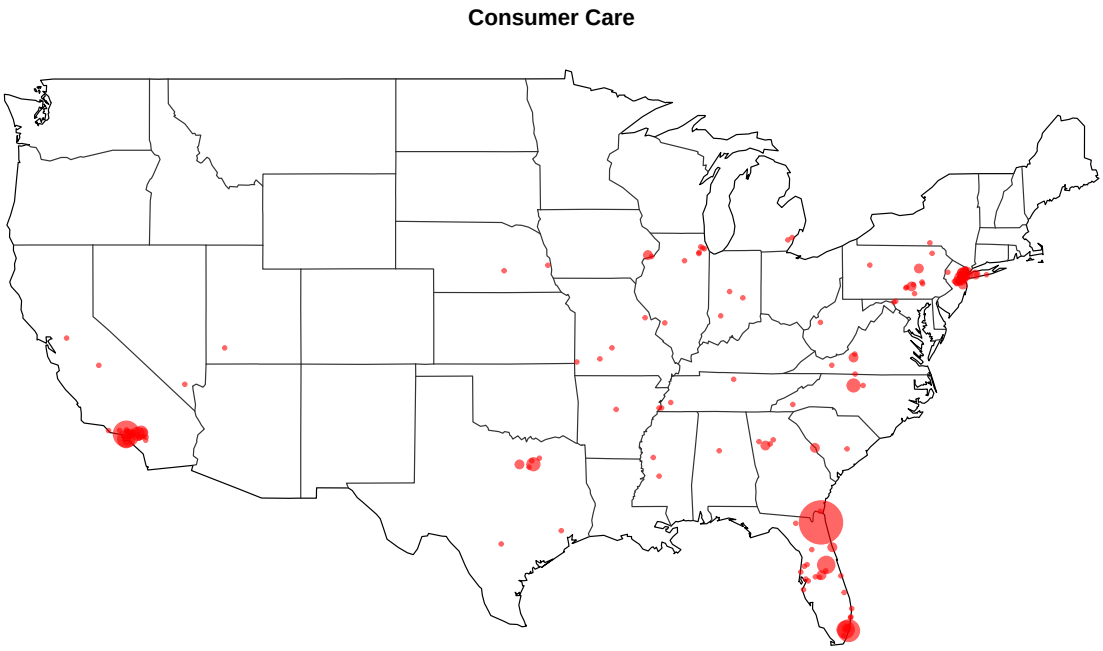

Figure B.4: Map of Consumer Care Products Thefts

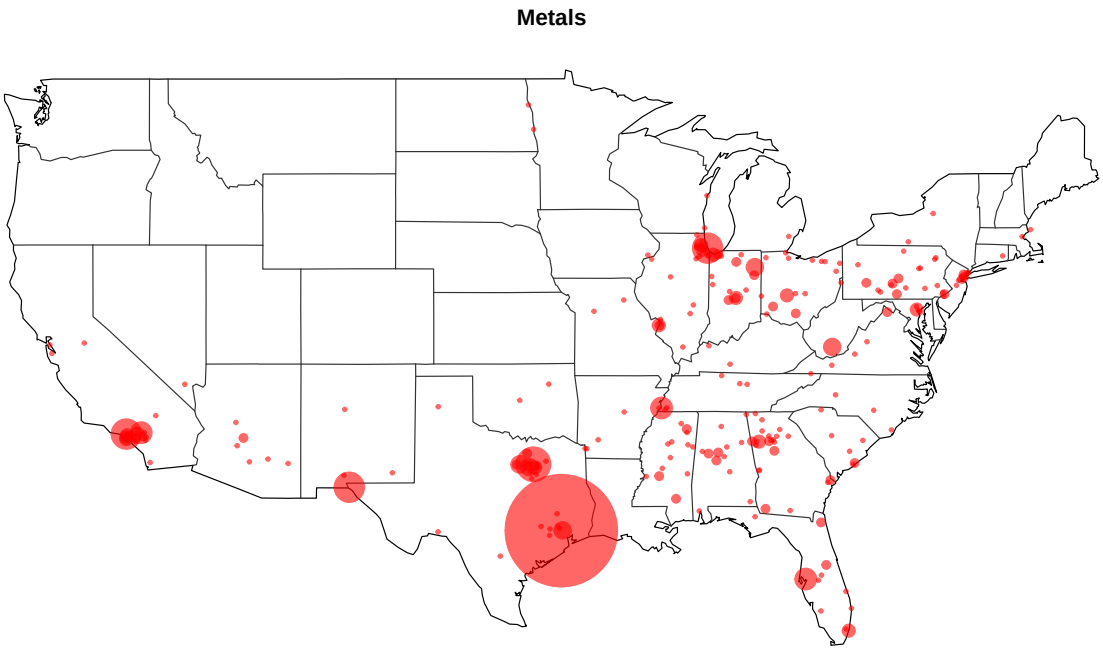

Figure B.5: Map of Metals Thefts

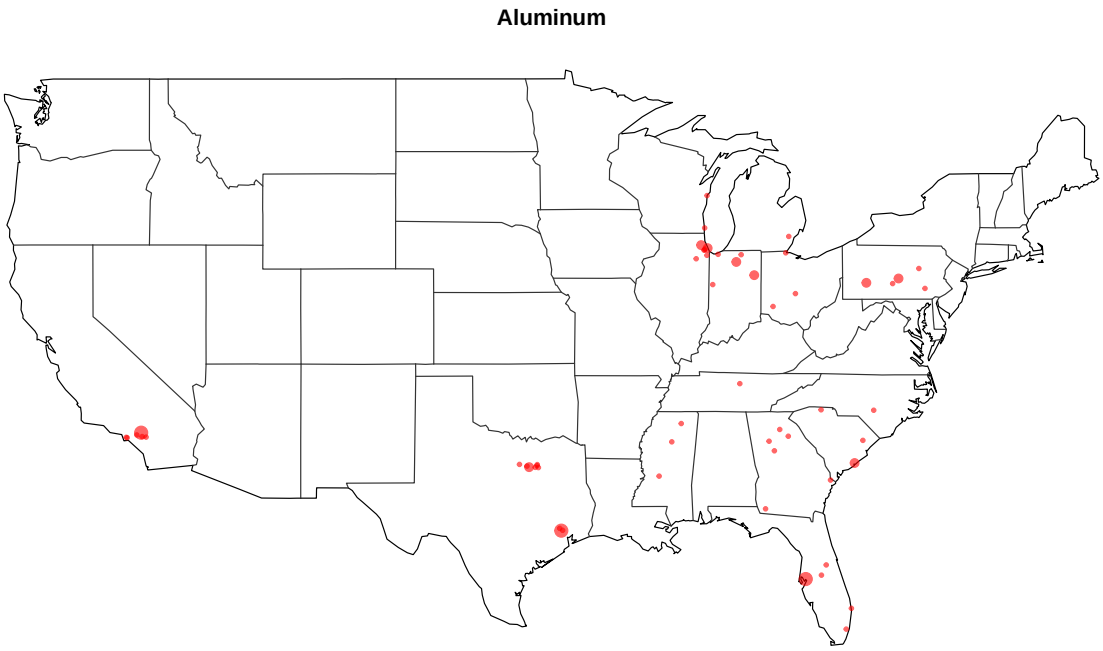

Figure B.6: Map of Aluminum Thefts

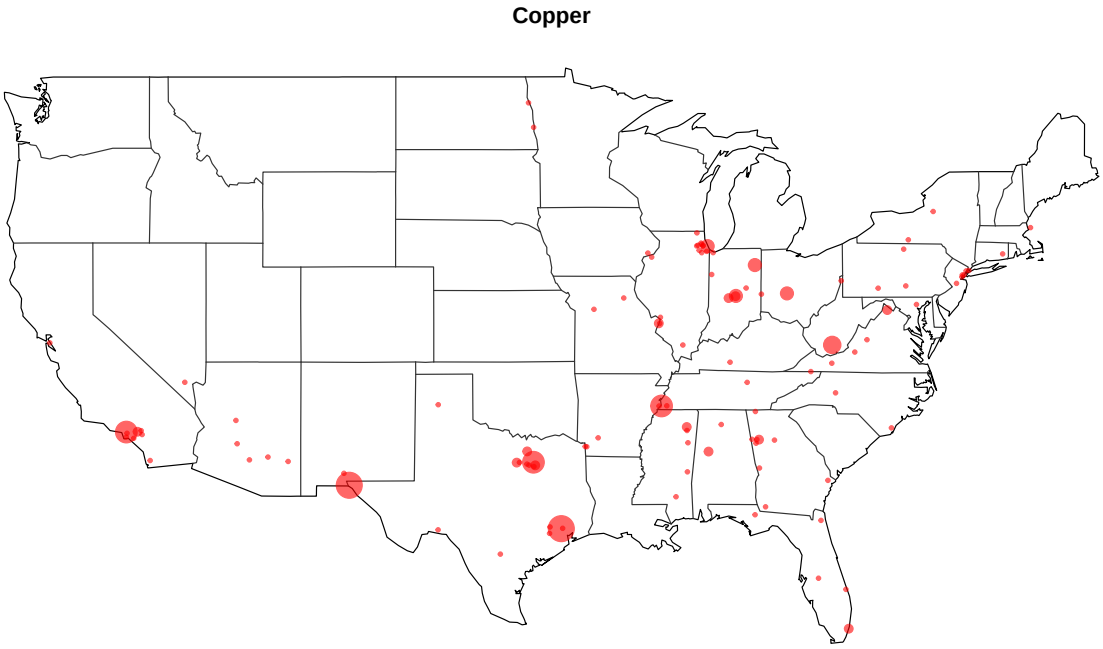

Figure B.7: Map of Copper Thefts

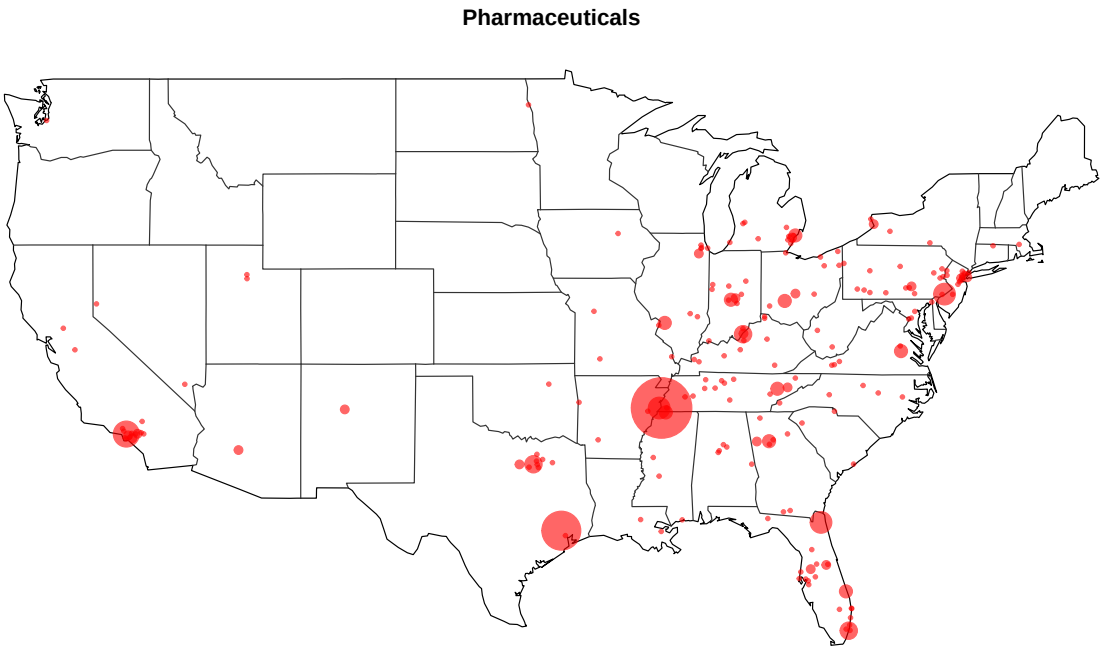

Figure B.8: Map of Pharmaceuticals Thefts

B.4 Ports Information

Port/waterway name	Rank	Total Trade (TEUS)
Los Angeles, CA	1	4919.2
Long Beach, CA	2	4063.8
New York (NY and NJ)	3	3761.3
Savannah, GA	4	1898.7
Oakland, CA	5	1542.9
Norfolk Harbor, VA	6	1413.2
Houston, TX	7	1262.8
Seattle, WA	8	1219.3
Tacoma, WA	9	1150.7
Charleston, SC	10	941.1
Honolulu, HI	12	686.2
Jacksonville, FL	13	631.4
Miami, FL	14	622.6
Port Everglades, FL	15	531.5
Baltimore, MD	16	453.1
Anchorage, AK	17	254
Wilmington, NC	19	184.3
Wilmington, DE	20	164
Portland, OR	21	162.1
Boston, MA	22	158.8

Table B.1: Largest Twenty U.S. Ports (excluding PR and LA)

Appendix C

Chapter 3 Appendix

Mechanism Design vs. Implementation Theory

The difference between mechanism design and implementation theory is easily illustrated in our environment: implementation theory requires both the existence and nonexistence constraints to be satisfied, whereas mechanism design requires only the existence constraints to hold. Mechanism design can be much easier to achieve than implementation, and in this section we illustrate this graphically for a class of solution concepts that include Nash equilibrium.

Solution concepts often impose restrictions on what players can achieve by deviating from a particular message profile. We let $\bar{d}$ denote the maximum number of deviations that a solution concept imposes restrictions on. For example, Nash equilibrium specifies that no one player can have a profitable unilateral deviation, so $\bar{d} = 1$ for Nash equilibrium. In this section we restrict our attention to a class of solution concepts we call *simple first-order* solution concepts.

Definition C.0.1 (**Simple First-Order Solution Concept**)**.** Given an implementation problem, S is a simple first-order solution concept if:

1. There is a positive integer $\bar{d}$ and, for each $R \in \mathcal{R}$ and any message profile m in any mechanism, S can determine whether m is an equilibrium at R or not based on the outcome at m and the outcomes obtainable by deviations of up to $\bar{d}$ players.

2. For any $(a, R) \in A \times \mathcal{R}$, if m is any message profile with outcome a in any mechanism, then we can always assign outcomes to message profiles that are deviations from m by up to $\bar{d}$ players to make m an equilibrium under S at R.

Nash equilibrium is a simple first-order solution concept: it satisfies (1) with $\bar{d} = 1$ and satisfies (2) by making assigning all unilateral deviations from m the same outcome as m. In the rest of this section we show that, with enough players, mechanism design is always possible in any simple first-order solution concept.

To illustrate, first consider an implementation problem with $n = 2$ in a solution concept with $\bar{d} = 1$ (e.g. Nash equilibrium). Suppose the range of F contains at least two distinct outcomes (true of any non-trivial social choice rule), a and b. The existence constraints imply the existence of two message profiles, with outcomes a and b, and where unilateral deviations satisfy certain coloring constraints. We illustrate this in the following figure.

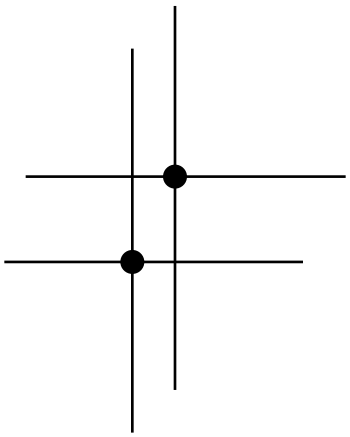

Figure C.1: Implementation with $n = 2$ and $\bar{d} = 1$

The two circles represent the two equilibrium message profiles, and the lines represent the constraints on unilateral deviations at each equilibrium. Note that, no matter where the

equilibria are placed in a mechanism, the lines emanating from the two equilibria must intersect. At each intersection point, the designer must find an outcome that satisfies the constraints arising from each of the equilibria simultaneously. This may or may not be possible, depending on the problem. However, consider the same setup with three players ($n = 3$):

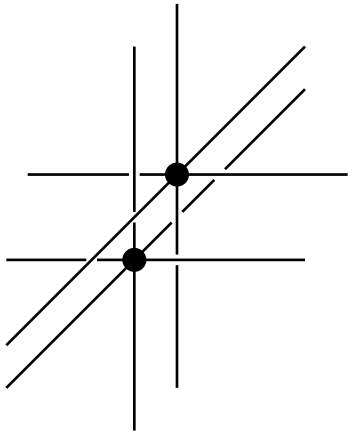

Figure C.2: Implementation with $n = 3$ and $\bar{d} = 1$

Figure C.2 shows how the equilibria can be placed so that no intersections happen. With $n = 3$ and $\bar{d} = 1$ it is *always* possible to find a mechanism that satisfies the existence constraints. This can be generalized, as the following theorem shows.

Theorem C.0.1. Suppose S is any simple first-order solution concept completely characterized by restrictions on deviations by up to $\bar{d}$ players. Then mechanism design is always possible for any social choice rule if $n \geq 2\bar{d} + 1$.

Proof. We need to show that, under the conditions of the theorem, a mechanism always exists that satisfies the existence constraints. Consider a mechanism that, for each $a \in F(R)$, gives to each player i a message $m_i^{(a,R)}$ such that: (1) $g(m^{(a,R)}) = a$, and (2) outcomes at deviations by up to $\bar{d}$ players from $m^{(a,R)}$ are consistent with $m^{(a,R)}$ being an equilibrium

with outcome a at R. This is always possible for simple first-order solution concepts, as long as the constraints imposed on deviations at one equilibrium do not conflict with constraints imposed on deviations from another equilibrium. This is always true when $n \geq 2\bar{d} + 1$. To see this, consider two outcomes, a and b, such that $a \in F(R)$ and $b \in F(R')$. A deviation by up to $\bar{d}$ players from $m^{(a,R)}$ must have a strict majority of the players following the strategy in $m^{(a,R)}$, because $n \geq 2\bar{d} + 1$. Thus, such a deviation is never also a deviation from $m^{(b,R')}$, where a strict majority of players follow the strategy in $m^{(b,R')}$. Therefore, the existence constraints are satisfied and mechanism design is achieved. ■

Bibliography

Abreu, Dilip and Hitoshi Matsushima (1992), "Virtual implementation in iteratively undominated strategies: complete information." *Econometrica: Journal of the Econometric Society*, 993–1008.

Aumann, Robert J (1960), "Acceptable points in games of perfect information." *Pacific Journal of Mathematics*, 10, 381–417.

Bar-Ilan, Avner and Bruce Sacerdote (2004), "The response of criminals and noncriminals to fines*." *Journal of Law and Economics*, 47, 1–17.

Becker, Gary S. (1968), "Crime and punishment: An economic approach." *Journal of Political Economy*, 76, 169–217.

Bergemann, Dirk and Stephen Morris (2005), "Robust mechanism design." *Econometrica*, 73, 1771–1813.

Bernheim, B Douglas, Bezalel Peleg, and Michael D Whinston (1987), "Coalition-proof nash equilibria i. concepts." *Journal of Economic Theory*, 42, 1–12.

Brailsford, Sally C, Chris N Potts, and Barbara M Smith (1999), "Constraint satisfaction problems: Algorithms and applications." *European Journal of Operational Research*, 119, 557–581.

Burges, Dan (2012), *Cargo theft, loss prevention, and supply chain security*. Butterworth-Heinemann.

Cabrales, Antonio and Roberto Serrano (2011), "Implementation in adaptive better-response dynamics: Towards a general theory of bounded rationality in mechanisms." *Games and Economic Behavior*, 73, 360–374.

Cameron, Samuel (1988), "The economics of crime deterrence: a survey of theory and evidence." *Kyklos*, 41, 301–323.

Carter, Ned (1995), "Increased theft as a side effect of sales promotion activities: An exploratory study." *Journal of Business and Psychology*, 10, 57–64.

Corman, Hope and H Naci Mocan (2000), "A time-series analysis of crime, deterrence, and drug abuse in new york city." *American Economic Review*, 90, 584–604.

Di Tella, Rafael and Ernesto Schargrodsky (2004), "Do police reduce crime? estimates using the allocation of police forces after a terrorist attack." *American Economic Review*, 115–133.

Donohue, John and Steven D Levitt (2001), "The impact of legalized abortion on crime." *Quarterly Journal of Economics*, 116.

Drago, Francesco, Roberto Galbiati, and Pietro Vertova (2009), "The deterrent effects of prison: Evidence from a natural experiment." *Journal of Political Economy*, 117, 257–280.

Dutta, Bhaskar and Arunava Sen (2012), "Nash implementation with partially honest individuals." *Games and Economic Behavior*, 74, 154–169.

Eliaz, Kfir (2002), "Fault tolerant implementation." *The Review of Economic Studies*, 69, 589–610.

Featherstone, Clayton R (2011), "Rank efficiency: Investigating a widespread ordinal welfare criterion." *Job Market Paper*.

Foley, C Fritz (2011), "Welfare payments and crime." *The Review of Economics and Statistics*, 93, 97–112.

Freeman, Richard B (1999), "The economics of crime." *Handbook of labor economics*, 3, 3529–3571.

Gibbard, Allan (1973), "Manipulation of voting schemes: a general result." *Econometrica: journal of the Econometric Society*, 587–601.

Glaze, Lauren E. (2011), "Correctional populations in the united states, 2010." Bureau of Justice Statistics. http://www.bjs.gov/content/pub/pdf/cpus10.pdf, last accessed: 8/25/2013.

Grogger, Jeffrey (1990), "The deterrent effect of capital punishment: an analysis of daily homicide counts." *Journal of the American Statistical Association*, 85, 295–303.

Halpern, Joseph Y (2008), "Beyond nash equilibrium: Solution concepts for the 21st century." In *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, 1–10, ACM.

Haralick, Robert M and Gordon L Elliott (1980), "Increasing tree search efficiency for constraint satisfaction problems." *Artificial intelligence*, 14, 263–313.

Harsanyi, John C and Reinhard Selten (1988), "A general theory of equilibrium selection in games." *MIT Press Books*, 1.

Hausman, Jerry A, Bronwyn H Hall, and Zvi Griliches (1984), "Econometric models for count data with an application to the patents-r&d relationship." *Econometrica*, 52, 909–938.

Healy, Paul J and Michael Peress (2013), "Preference domains and the monotonicity of condorcet extensions."

Heckelman, Jac C and Andrew J Yates (2003), "And a hockey game broke out: Crime and punishment in the nhl." *Economic Inquiry*, 41, 705–712.

Hutchinson, Kevin P and Andrew J Yates (2007), "Crime on the court: A correction." *Journal of Political Economy*, 115, 515–519.

Jackson, Matthew O (2001), "A crash course in implementation theory." *Social choice and welfare*, 18, 655–708.

Keteyian, Armen (2010), "Inside "grand theft cargo" - meet "the trucker" and "the broker"." CBS News. http://www.cbsnews.com/news/inside-grand-theft-cargo-meet-the-trucker-and-the-broker/, last accessed: 11/25/2013.

Kyckelhahn, Tracey and Tara Martin (2013), "Justice expenditure and employment extracts, 2010 - preliminary." Bureau of Justice Statistics. http://www.bjs.gov/index.cfm?ty=pbdetail&iid=4679, last accessed: 8/25/2013.

Lee, David S and Justin McCrary (2009), *The deterrence effect of prison: Dynamic theory and evidence*. Industrial Relations Section, Princeton University.

Levitt, Steven D (1998), "Why do increased arrest rates appear to reduce crime: deterrence, incapacitation, or measurement error?" *Economic inquiry*, 36, 353–372.

Levitt, Steven D (2002), "Testing the economic model of crime: The national hockey league's two-referee experiment." *Contributions in Economic Analysis & Policy*, 1.

Levitt, Steven D (2004), "Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not." *The Journal of Economic Perspectives*, 18, 163–190.

Maskin, Eric (1999), "Nash equilibrium and welfare optimality." *The Review of Economic Studies*, 66, 23–38.

McCormick, Robert E and Robert D Tollison (1984), "Crime on the court." *The Journal of Political Economy*, 223–235.

McNees, M Patrick, Marcia Kennon, John F Schnelle, Robert E Kirchner, and Murphy M Thomas (1980), "An experimental analysis of a program to reduce retail theft." *American Journal of Community Psychology*, 8, 379–385.

Muller, Eitan and Mark A Satterthwaite (1977), "The equivalence of strong positive association and strategy-proofness." *Journal of Economic Theory*, 14, 412–418.

Palfrey, Thomas R and Sanjay Srivastava (1991), "Nash implementation using undominated strategies." *Econometrica*, 479–501.

Raphael, Steven and Rudolf Winter-Ebmer (2001), "Identifying the effect of unemployment on crime." *Journal of Law and Economics*, 44, 259–283.

Reilly, Barry and Robert Witt (2008), "Domestic burglaries and the real price of audio-visual goods: Some time series evidence for britain." *Economics Letters*, 100, 96–100.

Satterthwaite, Mark Allen (1975), "Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions." *Journal of economic theory*, 10, 187–217.

Selten, Reinhard (1975), "Reexamination of the perfectness concept for equilibrium points in extensive games." *International journal of game theory*, 4, 25–55.

Sidebottom, Aiden, Jyoti Belur, Kate Bowers, Lisa Tompson, and Shane D Johnson (2011), "Theft in price-volatile markets: on the relationship between copper price and copper theft." *Journal of Research in Crime and Delinquency*, 48, 396–418.

Tsebelis, George (1990), "Penalty has no impact on crime: A game-theoretic analysis." *Rationality and Society*, 2, 255–286.

Vohra, Rakesh V (2011), *Mechanism Design: a linear programming approach*. 47, Cambridge University Press.

Wooldridge, Jeffrey M (1997), "Quasi-likelihood methods for count data." *Handbook of applied econometrics*, 2, 352–406.

Wooldridge, Jeffrey M (2002), *Econometric analysis of cross section and panel data*. The MIT press.

Wu, Wen-Tsun and Jia-he Jiang (1962), "Essential equilibrium points of n-person noncooperative games." *Scientia Sinica*, 11, 1307–1322.

Zhang, Junsen (1997), "The effect of welfare programs on criminal behavior: A theoretical and empirical analysis." *Economic Inquiry*, 35, 120–137.