Informal Sanctions on Prosecutors and Defendants and the Disposition of Criminal Cases*

by

Andrew F. Daughety and Jennifer F. Reinganum**
Department of Economics and Law School
Vanderbilt University

Original: May 2014
This version: February 2015

* This paper was partly-written while visiting at the Paris Center for Law and Economics (Daughety) and the University of Paris 2 (Reinganum); we especially thank Bruno Deffains for providing a supportive research environment. We also thank Scott Baker, David Bjerk, Richard Boylan, Rosa Ferrer, Francoise Forges, Luigi Franzoni, Nancy King, Lewis Kornhauser, Kathryn Spier, Christopher Slobogin, and seminar participants at the University of Bologna, The NSF Research Coordination Network on Guilty Pleas Meeting at Rutgers University, the Center for the Study of Democratic Institutions at Vanderbilt University, the Colloquium on Law, Economics, and Politics at NYU Law School, the Law and Economics Theory IV Conference, and the Law and Economics Workshop at Berkeley Law School for comments and suggestions on an earlier version.

** andrew.f.daughety@vanderbilt.edu; jennifer.f.reinganum@vanderbilt.edu

Informal Sanctions on Prosecutors and Defendants and the Disposition of Criminal Cases

Abstract

We model the strategic interaction between a prosecutor and a defendant when informal sanctions by outside observers may be imposed on both the defendant and the prosecutor. Non-strategic outside observers rationally use the disposition of the case (plea bargain, case drop, acquittal, or conviction) to impose these sanctions, but also recognize that errors in the legal process (as well as hidden information) means they may misclassify defendants and thereby erroneously impose sanctions on both defendants and prosecutors. If third parties prefer a legal system with minimal expected loss from misclassification, there is a unique prediction wherein the guilty defendant accepts the prosecutor's proposed plea offer with positive (but fractional) probability, the innocent defendant rejects the proposed offer, and the prosecutor chooses to take all defendants who reject the offer to trial. Furthermore, we show that: 1) changes in the level of the formal sanction affect the level of informal sanctions imposed by outsiders on defendants and prosecutors; and 2) increases in the informal sanction rate imposed on prosecutors results in changes in the level of informal sanctions imposed on defendants. The latter case is particularly noteworthy, as (for example) an increase in the rate associated with informally sanctioning prosecutors for convicting the innocent can result in an increase in the level of informal sanctions by third parties on innocent defendants.

We use the base model to examine two extensions of the analysis. In the first extension, we assume that a fraction of defendants are risk and/or ambiguity averse. If this response is strong enough, then some innocent defendants accept the prosecutor's plea offer. In the second extension, we consider the effect of increasing the informativeness of the jury's decision by extending the model to allow for a three-verdict outcome (not guilty, not proven, and guilty), sometimes referred to as the "Scottish" verdict. We find that: 1) guilty defendants are worse off, as plea bargains get tougher but the rejection rate does not change; 2) innocent defendants are better off; 3) the prosecutor's overall payoff goes up; and 4) the outside observers' concern over possible misapplication of informal sanctions is reduced. In this sense, the Scottish verdict is justice-improving when compared with the standard (two-outcome) verdict.

1. Introduction

All of us are familiar (if only from TV and the movies) with the fact that the criminal justice process provides formal sanctions for convicted defendants. These formal sanctions generally take two forms: incarceration (with the possibility of probation being used in some cases) and fines. In this paper we consider a third form of sanction that arises from members of society who observe pieces of the process, draw conclusions about the participants and, as a consequence of self-interested action, impose costs on the (perceived) offending party; we refer to these as informal sanctions.

Informal sanctions on convicted defendants have a long history. For example, defendants who have been convicted (that is, pled or been found guilty) and served their sentences (or paid their fines) may find it difficult to find housing and employment after release. Informal sanctions can also fall on defendants who have only been arrested, and for whom any charges have been dropped. Only one-fourth of the states actually prohibit the use of (pure) arrest information by employers when hiring (and the degree of enforcement is unclear).[1] Many states are silent on such matters (leaving the use of such information for hiring purposes entirely at the discretion of employers), while the remainder have imposed some limitations. For example, while Michigan prohibits employers asking about misdemeanor arrests that did not lead to conviction, no restrictions are placed on asking about felony arrests that did not lead to conviction. On the other hand, employers may be liable for hiring people (such as teachers or various types of care-givers) with criminal histories who ultimately harm someone else. There are a number of firms that specialize in investigating job candidates' past criminal records (which typically include arrests, even if those arrests did not lead to conviction) and provide such services to employers. A recent online development has been websites that publish booking photos ("mug shots") that are part of the public record.[2] Moreover, defendants who have

[1] The website Nolo.com provides state-level detail on the state and federal restrictions on the use by employers of information about arrest or conviction (www.nolo.com/legal-encyclopedia/state-laws-use-arrests-convictions-employment.html; accessed January 24, 2015). The Equal Opportunity Employment Commission provides guidance on what could constitute discriminatory hiring from a federal perspective, and only prohibits blanket policies of not hiring those with arrest records. The EEOC reports survey results that 92% of responding employers use criminal or background checks on all or some of job candidates (http://www.eeoc.gov/laws/guidance/arrest_conviction.cfm#IIIA; accessed January 24, 2015).

[2] See the discussion of the case of Dr. Janese Trimaldi, among others, in Segal (2013). Despite the fact that all charges against her were dropped, her booking photo (which is a public record) began to appear at online mug-shot sites. Segal estimates that there are over 80 such sites that generally charge people to remove the images; he indicates that fees for removal of information tend to run between $30 and $400 and, since multiple sites may post the picture, the cost of eliminating this information from the web can be exorbitant.

been acquitted may be greeted with the suspicion that they were actually guilty but the jury was unable to formally reach that conclusion.[3]

Informal sanctions may also be applied to officials in the system;[4] for concreteness we specifically focus on prosecutors, but others may also be subject to such sanctions. Some prosecutors may be viewed as "soft on crime" or "not up to the task" in that cases against defendants believed to be guilty are dismissed, or trials are lost. Other prosecutors may be viewed as abusing their position, railroading possibly innocent defendants into accepting plea bargains, or by winning trials that seem unfair. Informal sanctions on the prosecutor can affect her career concerns via election, appointment, promotion, or selection for judgeships, or outside opportunities in private law firms and universities.

We view formal sanctions as operating via the existing judicial system while informal sanctions come from members of society and reflect the beliefs of "outside observers." Thus, both defendants and prosecutors may experience losses due to informal sanctions applied by these same outside observers. Importantly, we show how informal sanctions may affect (including limit) the use of formal sanctions.

More precisely, informal sanctions for the defendant involve outside observers drawing an inference about how likely it is that the defendant is guilty, given the case disposition, and applying sanctions that are proportional to this belief.[5] These sanctions correspond to the outside observers withdrawing from further interactions with the defendant; for instance, they may choose not to hire him for a job or rent an apartment to him, or to avoid social interactions with him, and so on. The proportional specification is a simple way to ensure that the defendant suffers worse informal sanctions the higher is the outside observers' belief in his

---

[3] One juror in the case against Casey Anthony (acquitted of murdering her two-year-old daughter), stated: "I did not say she was innocent; I just said there was not enough evidence. If you cannot prove what the crime was, you cannot determine what the punishment should be." See Burke, et. al. (2011).

[4] The analysis abstracts from the use of formal sanctions for officials, but abuses such as prosecutorial misconduct can lead to formal sanctions. A fairly well-known example of prosecutorial misconduct involved Michael Nifong, the District Attorney for Durham County, NC, who was disbarred for his actions in the 2006 Duke University lacrosse case prosecution; see *NC State Bar v. Nifong* (June 16, 2007). Nifong also was convicted of criminal contempt of court for lying to a judge and served one day in jail and paid a $500 fine; see Associated Press (2007).

[5] Other recent papers that incorporate payoffs (representing the assessments of third parties) that are proportional to inferred type include Benabou and Tirole (2006), Daughety and Reinganum (2010), and Deffains and Fluet (2013).

guilt. Although these informal sanctions are costly to the defendant, we assume that the outside observers do not bear any net loss of imposing them; for instance, they can simply hire, or interact socially with, someone else.[6] As suggested earlier, we also assume that the outside observers impose informal sanctions on the prosecutor, and these too we take as being in proportion to their posterior belief that she has punished an innocent defendant (through conviction at trial or through a plea-bargained conviction), or failed to punish a guilty defendant (either through acquittal at trial or through dropping the case).

In the model developed in Section 3, the only type of equilibrium that can exist is a semi-separating one wherein innocent defendants reject the plea offer,[7] whereas guilty defendants mix between accepting and rejecting the plea offer. Because the prosecutor has the option to drop the case, there must be a sufficient fraction of guilty defendants among those that reject the plea offer in order to incentivize the prosecutor to go to trial following rejection.

As the equilibrium fraction of guilty defendants among those that reject the plea offer is co-determined with the fraction that outside observers expect to reject the plea offer, there is a continuum of semi-separating equilibria, indexed by this fraction. There is a smallest equilibrium fraction that is necessary to incentivize the prosecutor to go to trial following a rejection (rather than dropping the case). We show that, if outside observers prefer to minimize the expected loss from erroneously-imposed informal sanctions, then they prefer the equilibrium wherein the smallest fraction of guilty defendants reject the plea offer. That is, they prefer the equilibrium which entails the greatest amount of successful plea bargaining. This equilibrium involves the lowest expected loss from misclassification because those that accept the plea offer are revealed to be guilty types, and trial is the clearest possible signal of innocence (subject to the noise that is required to incentivize the prosecutor to go to trial following a rejected plea offer).

---

[6] In other words, we assume that an observer's perceived risk of loss from a transaction is compared to the expected match value. The higher the posterior assessment of guilt, the more matches the outside observer will choose to forego. Observers do not consider the negative externalities that their individual decisions confer on the defendant or on society at large.

[7] Innocent defendants never accepting the equilibrium plea offer is a common characteristic of many of the economic analyses of plea bargaining. In reality, some innocent defendants do accept plea offers. In Section 5 we modify the base model to account for the possibility that some innocent defendants will accept the equilibrium plea offer.

We find that informal sanctions will affect both the feasibility, and the willingness, of the prosecutor to employ plea bargaining. In particular, informal sanctions may restrict the feasibility of the equilibrium wherein at least some defendants settle, as (in equilibrium) accepting the plea bargain results in a clear inference of guilt, which results in the highest informal sanction against the defendant. If the informal sanction rate for the defendant is too high, then it will not be possible to induce a defendant to accept a plea bargain. Similarly, a prosecutor may prefer to take a case to trial rather than settling via a plea bargain if the informal sanction rate on the defendant is too high, because the prosecutor must discount the formal sanction (in the plea offer) in order to induce the defendant to accept both the plea offer and the informal sanction that results when he thereby reveals his guilt. This selected equilibrium also yields a number of interesting implications, among them that: 1) there is an induced (type-specific) correlation between the formal and informal sanctions on defendants, and 2) that increases in the informal sanction rates on prosecutors can feed back to induce seemingly-perverse adjustments in the level of informal sanctions imposed on defendants.

We then consider two extensions to the base model described above. The first is to allow for another form of unobservable heterogeneity (in attitudes towards risk and/or ambiguity) of the defendants that results in innocent defendants accepting equilibrium plea offers. Such heterogeneity can increase the equilibrium plea offer and the likelihood of its acceptance (even among guilty defendants). The second extension involves reverting to the base model but modifying the legal system so that the outside observers are able to acquire more information from the trial verdict as to the degree of potential guilt of an acquitted defendant. Specifically, we extend the model to consider the "Scottish verdict" wherein the verdict allows for three outcomes: not guilty, not proven, and guilty. The intermediate case, not proven, carries no formal sanctions (it is a form of acquittal); it represents an outcome wherein jurors felt that the prosecution's case against the defendant was insufficiently strong to meet the high evidentiary standard needed in a criminal case (beyond a reasonable doubt), but also reflects an unwillingness on the part of the jury to assert a belief that the defendant was not guilty. This finer resolution of the jury's assessment leads to increased expected costs to a truly guilty defendant, lower expected costs to a truly innocent defendant, and informal sanctions by outside

observers that are more likely to be deserved; altogether, these results suggest that the Scottish verdict is likely to be justice-enhancing relative to the standard two-outcome (convict/acquit) verdict.

Related Literature

Landes (1971) provides a complete-information model wherein the prosecutor's objective is to maximize expected sentences obtained from a collection of defendants, subject to a resource constraint; the potential for innocent defendants is not considered. Grossman and Katz (1983) and Reinganum (1988) provide screening and signaling models, respectively, of plea bargaining wherein the prosecutor (who is committed to trial following a rejection of the plea offer) maximizes a utility function that corresponds to social welfare. Absent this commitment to trial following a rejected plea, a putative equilibrium in which the guilty accept the plea offer and the innocent reject it (as occurs in Grossman and Katz) is undermined by the prosecutor's desire to drop the case rather than proceeding to trial against a defendant that (she now believes) is innocent. Franzoni (1999) and Baker and Mezzetti (2001) explicitly incorporate a credibility constraint into a screening model, which requires that a sufficiently high fraction of guilty defendants reject the plea offer.[8] We also incorporate this kind of credibility constraint; our model is closest in terms of the prosecutor's payoff functions to that of Baker and Mezzetti because in both models the prosecutor faces a risk of convicting an innocent defendant.[9] However, we will incorporate informal sanctions that fall on both the defendant and the prosecutor, depending on the disposition of the case (e.g., convicted; acquitted; plea-bargained; or dropped).

The papers discussed thus far (as well as our paper) assume that the judge or jury makes its decision based only on the evidence they observe in the course of the trial and the specified standard of proof. That

---

[8] See Nalebuff (1987) for a screening model with a possibly-binding credibility constraint in the case of a civil suit. In Franzoni's model, innocent defendants are never convicted, so the prosecutor simply maximizes the expected penalty imposed on the guilty less the cost of the effort she expends to investigate prior to trial. However, if only innocent defendants reject the plea offer, then the prosecutor is unwilling to expend any effort on investigation; thus, equilibrium must involve some guilty defendants rejecting the plea offer as well.

[9] In Baker and Mezzetti's model, a prosecutor obtains a payoff of x (resp., - x) if a guilty (resp. innocent) defendant gets a sentence of x. The prosecutor obtains a payoff of zero if she frees an innocent defendant and - αx if she frees a guilty defendant. Finally, the prosecutor does not have a cost of trial, but loses the amount c whenever she loses at trial. Thus, in their model the prosecutor has internal concern for punishing the innocent and letting the guilty go free, and they obtain a unique semi-separating equilibrium. In our model these sanctions are provided by the outside observers, and we obtain a continuum of semi-separating equilibria, and then use a selection criterion to obtain a unique (selected) equilibrium.

is, they follow their instructions rather than acting as rational Bayesian agents. Bjerk (2007) assumes a jury that acts as a rational Bayesian agent, and this undermines an equilibrium wherein the prosecutor screens defendant types perfectly. For if the prosecutor was expected to induce a guilty plea from all of the guilty defendants, then the jury would rationally infer that those coming to trial must be innocent, and the jury would acquit (but then the guilty would refuse to plead). The beliefs of the jury are self-fulfilling and thus the model has a continuum of equilibria, indexed by the evidentiary threshold needed for the jury to convict.[10] Our model also has a continuum of equilibria, but these are based on the (rational Bayesian) beliefs of the outside observers, who impose informal sanctions on both the defendant and the prosecutor. As in Franzoni (1999), Baker and Mezzetti (2001), and Bjerk (2007), our prosecutor will face a credibility constraint in that she will have to ensure that there are enough guilty defendants among those that reject the plea offer to rationalize her going to trial. Our multiple equilibria are indexed by the fraction of guilty defendants among those that reject the plea offer (again, innocent defendants always reject the plea offer in our base model). We select among the equilibria in our model by positing a preference on the part of the outside observers to minimize the expected loss due to erroneously-imposed informal sanctions.[11]

Finally, as indicated above, there are a number of different formulations for the prosecutor's objective, ranging from expected sentences to social welfare to a mixture of motivations. Prosecutors are supposed to represent society but they will clearly have personal preferences and career concerns as well. The general issue of what it is that prosecutors are maximizing is important to formulating models of plea bargaining. Glaeser, Kessler, and Piehl (2000) find evidence that some federal prosecutors are motivated by reducing crime while others are primarily motivated by career concerns. Boylan and Long (2005) find that assistant U.S. attorneys in districts with very high private salaries are more likely to take cases to trial,

[10] Bjerk assumes that both the prosecutor and the jury maximize social welfare (given their respective beliefs and information), and trials are costless. Our prosecutor's objective increases in the sentences she obtains, but is also influenced by informal sanctions imposed by outside observers based on her perceived errors. Bjerk does not argue for a particular selection from among the equilibria he identifies. We argue for a particular equilibrium to be selected based on the desire of outside observers to minimize the extent of informal sanctions that they impose in error.

[11] Defendants also prefer this equilibrium, so an alternative basis for choosing this equilibrium is that outsiders under a veil of ignorance (i.e., they recognize that they may be defendants one day) prefer the same equilibrium as that which minimizes expected loss from misclassification.

suggesting that they seek trial experience in anticipation of an eventual private-sector job. Boylan (2005) finds that the length of prison sentences obtained (but not conviction rates) is positively related to positive outcomes in the career paths of U.S. attorneys. Bandyopadhyay and McCannon (forthcoming) find that prosecutors subject to reelection pressure try to increase the number of convictions at trial (including taking more weak cases to trial).[12] McCannon (2013) finds that a prosecutor's motivation to influence the election leads to more wrongful convictions (or, at least, more reversals on appeal).

In our model prosecutors maximize the expected sentence minus the cost of trial, and minus the expected informal sanctions from outside observers arising from convicting the innocent or not convicting the guilty.[13] Thus, our prosecutor's objective reflects career concerns that are modeled as being a function of the statutory sentence length, the likelihood of conviction, the cost of trial, and the possibility of informal sanctions from outside observers.

Plan of the Paper

In Section 2, we provide the notation and formal model. In Section 3, we describe the equilibria of the model (the equilibrium concept will be Perfect Bayesian equilibrium), and a rationale for selecting among them. In Section 4 we discuss some implications of the model through a variety of comparative statics. Section 5 allows for heterogeneity in the defendant's response to risk and/or ambiguity. Section 6 extends the model to the Scottish verdict and shows that this refinement enhances justice. In Section 7, we provide a summary and suggest further extensions. The most salient technical issues are included in the Appendix while a Technical Appendix[14] contains other details of the analysis.

---

[12] Gordon and Huber (2002) argue that voters concerned with prosecutorial power (including the conviction of innocents) and who wish to impose accountability on prosecutors should follow a strategy of reelecting prosecutors who pursue cases to trial and obtain convictions. In their model P can, with effort, observe the true guilt or innocence of D and discover "unimpeachable evidence" so that truly innocent cases are dropped.

[13] One might question whether it is fair to place all the weight of getting it right on the prosecutor when there is incomplete information about the defendant and imperfect information about the evidence and the jury. We would argue that it is the prosecutor who chooses to make an offer (or not), to drop a case or to pursue the case to trial, and that the foregoing empirical studies, *in toto*, generally support a model focused around career concerns that reflect social preferences regarding criminality and the use of prosecutorial power.

[14] Available at http://www.vanderbilt.edu/econ/faculty/Daughety/DR-InformalSanctionsandCaseDispositions-TechApp.pdf

2. Modeling Preliminaries

Description of the Game

Our game commences after the police arrest the defendant on suspicion of committing a specific crime. The defendant, D, will be taken to be male, and the prosecutor, P, female. The exogenous parameters of the game include the sentence upon conviction (S_c), the evidentiary criterion used by the jury for conviction (γ_c), and the cost of trial for each agent (k^P for P and k^D for D). More detail on the notation (and the informal sanctions, which also have exogenously-determined elements) will be provided as we progress, but a basic convention will be that outcomes or actions appear as subscripts while "ownership" – that is, which agent is affected by the variable or parameter of interest – is indicated by a superscript.[15] Moreover, as this model represents the interaction between one P and one D, with respect to one crime, all parameters introduced below can be made conditional on observed characteristics of P, D, and/or the crime itself.

There are five stages in the game; note that P's payoff represents her net gains and D's payoff represents his total losses, so P maximizes her payoff while D minimizes his payoff:

Stage 1: Nature (N) draws D's type, denoted by t, and this is revealed to D only.

Stage 2: P makes a plea bargain offer of $S_b \geq 0$.

Stage 3: D chooses whether to accept (A) or reject (R) the plea bargain offer; if he accepts the offer (outcome b), the game ends and payoffs (π_b^P and π_b^D) are obtained.

Stage 4: If D has chosen R, then P now chooses whether to drop the case (outcome d) or pursue it to trial (action T). A dropped case yields of payoffs π_d^P and π_d^D.

Stage 5: If the case goes to trial, then Nature (N) draws the evidence of guilt, e, and the jury (J) uses the rule that if $e > \gamma_c$, then the outcome is conviction (outcome c), while otherwise the outcome is acquittal (outcome a). Conviction yields payoffs of π_c^P and π_c^D, while acquittal yields payoffs π_a^P and π_a^D.

---

[15] For ease of reading, when ownership is clear we omit this superscript.

Figure 1 below illustrates the extensive form for the game, with information sets indicated by "bubbles." The last move illustrated in the game tree (on the right side of the figure) is that by Nature in Stage 5, given that P chose to go to trial rather than drop the case. This trial subgame on the right is purely mechanical (and will be discussed in more detail in the narrative).

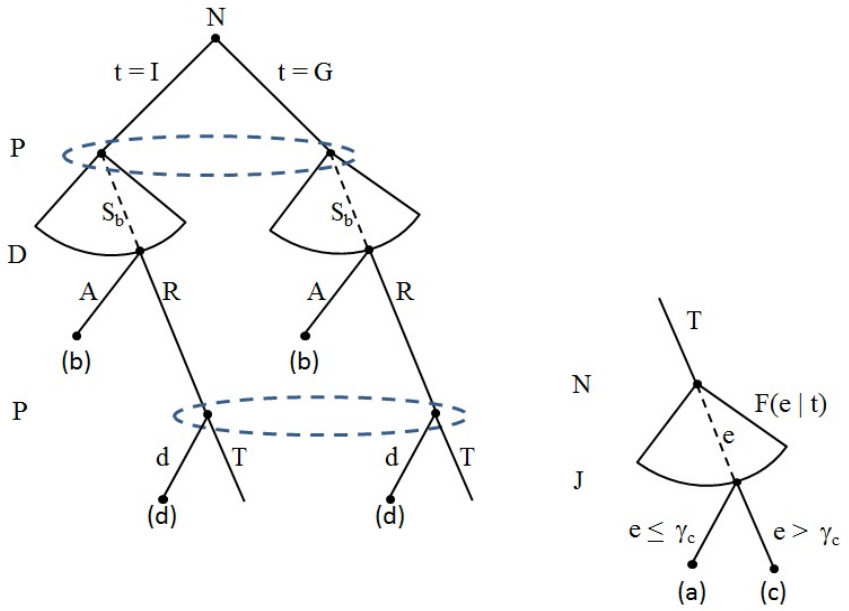

Figure 1: Game between P and D

As illustrated at the top of the tree, D's type t is either I (Innocent) or G (Guilty), and this is D's private information. Let $\lambda > 0$ denote the fraction of innocent Ds; that is $\lambda = \Pr\{t = I\}$, the probability that a D is innocent of the specific crime for which he was arrested. This parameter reflects some initial level of evidence gathered by the police. As indicated above, no further evidence is available until P and D go to trial, in which case a draw of evidence of guilt, $e \in [0, 1]$, occurs (this is shown on the right side of Figure 1). This draw is influenced by the underlying type for D and is observed only by the jury. The jury is instructed to convict if its evidence signal exceeds a threshold (γ_c) and to otherwise acquit the defendant. We denote the distribution of evidence (given D's type) as F(e | t), which we assume is continuous in e. Since for any evidentiary standard for conviction, γ_c, the jury (J) will choose outcome a when $e \leq \gamma_c$, then for either type

t, $F(\gamma_c \mid t)$ is the probability that D is acquitted and $1 - F(\gamma_c \mid t)$ is the probability that D is convicted. This motivates the assumption that at any level of evidence e, the probability of acquittal for an innocent D is higher than that for a guilty D: $F(e \mid I) > F(e \mid G)$ for all e. Finally, to aid readability in writing out payoffs, let F_t denote $F(\gamma_c \mid t)$, for t = I, G, so our assumption implies that $F_I > F_G$.

The sentences S_c and S_b are formal sanctions. Informal sanctions are penalties imposed by outside observers on both defendants and prosecutors; to reduce the verbiage, let Θ denote the outside observer(s).[16] Informal sanctions are based on Θ's beliefs, which are contingent on the case disposition (a, b, c, or d). We assume that these informal sanctions are proportional to the observers' beliefs, which depend upon the inferred type of defendant and the observed outcome of the legal process.[17] More precisely, these beliefs represent Θ's posterior probability that the defendant is type t, given the case disposition was y, and are denoted $\mu(t \mid y)$, where t = I, G and y = a, b, c, d. Note that this also means that Θ cannot directly observe the plea offer, S_b, the levels of P's and D's payoffs, or the evidence draw e.[18]

Informal sanctions imposed on D are of the form $r^D\mu(G \mid y)$, where $r^D \geq 0$ is an exogenous parameter.[19] That is, given any case disposition, Θ assesses the posterior likelihood that D is guilty, and then imposes informal sanctions at the rate r^D. These informal sanctions, which are increasing in the posterior assessment of guilt, reflect the fact that Θs may be future trading partners (broadly-construed) who decline to trade with the defendant; as discussed earlier, we assume that the observers do not suffer a net loss from avoiding these

---

[16] Since we will not be accounting for any heterogeneity among the outside observers, we will refer to both singular possessive and plural possessive Θ (e.g., we will refer to Θ's and Θs' beliefs) interchangeably. Outside observer beliefs will always be denoted by μ.

[17] For simplicity, we assume that the outside observers always observe the case disposition. However, it is trivial to allow this to occur only with positive probability. Probabilistic observation would simply re-scale the informal sanction rates by pre-multiplying these rates by the probability that the observers actually do observe the case disposition.

[18] It is very plausible that Θ would not observe a rejected plea offer. We assume that Θ does not observe the plea bargain offer S_b, even if D accepts the bargain and that acceptance is observed. We speculate that, due to the structure of the game and the fact that there are only two types of D, observing S_b if D accepts the offer would not affect Θ's out-of-equilibrium beliefs, but we leave this as an item for future research.

[19] We think of this rate as positive, but it could be negative (which the model accommodates), such as might hold if D was a gang member seeking "street cred" and the relevant outside observers are other gang members and friends. Moreover, r^D may differ from one crime to another; a heinous crime is likely to have a higher value of this parameter than a petty crime. Other characteristics of D may also affect the magnitude of r^D. For example, a repeatedly-convicted burglar charged with a new burglary may have a lower value of r^D than that of a first-time burglary defendant. On the other hand, a career burglar charged with a very different crime (e.g., child molestation) might still have a very high value of r^D.

transactions (though we will later assume that they prefer a society that minimizes the extent of erroneously-imposed informal sanctions).

As indicated earlier, Θs also impose informal sanctions on P, reflecting the notion that errors occur within the legal process; informal sanctions on the prosecutor arise when there is a belief that prosecutors should be blamed for such perceived errors; for instance, a guilty defendant may be acquitted or the case may be dropped. In these instances, P has allowed a guilty D to escape punishment. The associated informal sanctions are given by $r_G^P\mu(G \mid y)$, for $y \in \{a, d\}$. On the other hand, an innocent defendant can be convicted or may accept a plea bargain, so the prosecutor has punished an innocent defendant. The associated informal sanctions are given by $r_I^P\mu(I \mid y)$, for $y \in \{b, c\}$. We assume that r_I^P and r_G^P are non-negative. While not needed in the model, a natural assumption would be that the sanction rate for the prosecutor is higher when an innocent defendant is punished than when a guilty defendant is not punished; that is, $r_I^P > r_G^P$.[20]

D's Payoffs

We are interested in non-cooperative solutions for the game that exhibit sequential rationality by G, I, P and Θ, so we will first develop payoff functions starting from the outcomes (a, b, c, and d). Since trial ends in conviction or acquittal, D's loss on the right-hand-side of Figure 1 can be written as:

$$\pi_c^D = S_c + k^D + r^D\mu(G \mid c); \tag{1a}$$

and

$$\pi_a^D = k^D + r^D\mu(G \mid a). \tag{1b}$$

That is, going to trial costs D the amount k^D.[21] Conviction results in the formal sanction S_c plus the informal sanction $r^D\mu(G \mid c)$; since I-types may have been convicted (the evidence draw for an I-type could, conceivably, result in conviction), Θ recognizes that conviction is not a guarantee of guilt, so $\mu(G \mid c)$ will be less than one. Similarly, acquittal generally does not imply innocence, so Θ's belief $\mu(G \mid a)$ will be positive and D will bear both the cost of trial and the informal sanction $r^D\mu(G \mid a)$.

---

[20] As with r^D, r_I^P and r_G^P may vary with the crime in question or observable attributes of P (and possibly D).

[21] This cost may include the costs of legal assistance as well as the disutility of choosing to go to trial if there is pre-trial detention, so even an indigent D whose attorney fees are subsidized may face a substantial value of k^D.

We can write D's expected loss from going to trial (given his type t) as the weighted combination of the elements in equations (1a) and (1b), where the weights reflect the likely outcome at trial:

$$\pi_T^D(t) = S_c(1 - F_t) + k^D + r^D\mu(G \mid c)(1 - F_t) + r^D\mu(G \mid a)F_t,\ t \in \{I, G\}. \quad (2)$$

For instance, if t = I, then if D goes to trial, he expects to be convicted (outcome c) with probability $1 - F_I$, in which case he will receive the formal sanction S_c and the informal sanction $r^D\mu(G \mid c)$. He expects to be acquitted (outcome a) with probability F_I, in which case he will receive no formal sanction but Θ still believes there is a chance D is guilty despite his acquittal, and imposes the informal sanction $r^D\mu(G \mid a)$. Also note that D pays his trial costs k^D regardless of the trial outcome. As shown in the Technical Appendix, we obtain:

Remark 1: $\pi_T^D(I) < \pi_T^D(G)$.

That is, a D of type I has a lower expected loss from trial than does a D of type G. This follows from the assumption that an innocent type is less likely to be convicted than a guilty type ($F_I > F_G$), which implies that the posterior belief by outsiders that a D is a G-type is higher following a conviction than an acquittal ($\mu(G \mid c) > \mu(G \mid a)$). Remark 1 is important in that it suggests that the equilibrium might involve full screening (wherein, say, P makes an offer that only G-types accept and I-types reject), or partial screening (wherein, say, P makes an offer that all of one type reject, and that some of the other type reject); that is, properly-constructed offers in the plea bargaining stage may yield information about D's type. We return to this in Section 3 in our discussion of the equilibria of the game.

If P offers a plea bargain of S_b, then D can choose to accept (A) or reject (R) the offer. D's payoff from accepting a plea bargain of S_b is:

$$\pi_b^D = S_b + r^D\mu(G \mid b); \quad (3)$$

that is, he receives the formal sanction S_b plus the informal sanction that observers impose because, having accepted the plea offer (outcome b), they believe that he is guilty with probability $\mu(G \mid b)$.

Similarly, D's payoff if P drops the case is:

$$\pi_d^D = r^D\mu(G \mid d), \quad (4)$$

which reflects Θ's beliefs that D might have been guilty.

Since P may mix between going to trial and dropping the case following a rejection of the plea offer by D, let ρ^P denote the probability that P takes the case to trial following rejection by D; this occurs on the left side of Figure 1, at the second information set for P. Combining equations (2) and (4), weighted by the probability that P takes the case to trial, yields D's expected payoff following rejection (given his type) as:

$$\pi_R^D(t) = \rho^P \pi_T^D(t) + (1 - \rho^P)\pi_d^D. \tag{5}$$

P's Payoffs

Again, starting at the right of Figure 1, since trial ends in conviction or acquittal, P's payoffs on the right-hand-side of Figure 1 can be written as:

$$\pi_c^P = S_c - k^P - r_I^P \mu(I \mid c); \tag{6a}$$

$$\pi_a^P = -(k^P + r_G^P \mu(G \mid a)). \tag{6b}$$

Next we obtain P's expected payoff from going to trial; this turns out to be somewhat more complicated than D's corresponding payoff because P and Θ have different amounts of information on which to form beliefs. When the prosecutor makes the plea offer S_b, she does not know whether the defendant is guilty or innocent (she relies at this point on the prior, λ), so D's decision to accept or reject the offer will affect the prosecutor's posterior belief that he is guilty. The prosecutor's beliefs may differ from those of the observer because she observes the plea offer, whereas the observer observes only the disposition of the case. To capture this, let ν(G | R) (resp., ν(G | A)) denote the prosecutor's posterior probability that the defendant is guilty,[22] given that he rejected (resp., accepted) the plea offer S_b. Of course, in equilibrium, P's beliefs and Θ's beliefs must be the same (and must be correct).

The prosecutor's payoff from going to trial (given her beliefs following the defendant's rejection of her plea offer) can be written as:

$$\begin{aligned}\pi_T^P = {} & \nu(G \mid R)\{S_c(1 - F_G) - k^P - r_I^P \mu(I \mid c)(1 - F_G) - r_G^P \mu(G \mid a)F_G\} \\ & + \nu(I \mid R)\{S_c(1 - F_I) - k^P - r_I^P \mu(I \mid c)(1 - F_I) - r_G^P \mu(G \mid a)F_I\}.\end{aligned} \tag{7}$$

[22] P's beliefs will also depend on the plea offer S_b, but this would needlessly complicate the notation so this dependence is suppressed.

This is interpreted as follows. Given rejection of the plea offer, P believes that D is guilty with probability $\nu(G \mid R)$, in which case she expects a conviction with probability $1 - F_G$ and an acquittal with probability F_G. If D is convicted, P obtains utility from the formal sanction S_c but observers still harbor the posterior belief $\mu(I \mid c)$ that D may be innocent (despite his conviction), and impose on P the informal sanction $r_I^P \mu(I \mid c)$. If D is acquitted, then the Θs still harbor the posterior belief $\mu(G \mid a)$ that D is guilty (despite his acquittal) and impose on P the informal sanction $r_G^P \mu(G \mid a)$. Regardless of the case disposition (a or c), P pays the trial costs k^P. The second part of P's payoff, wherein she believes that D is innocent with probability $\nu(I \mid R)$, is interpreted similarly.

If P's plea offer is accepted then she obtains the following payoff:

$$\pi_b^P = S_b - r_I^P \mu(I \mid b). \tag{8}$$

Equation (8) indicates that P's payoff if the offer is accepted is the level of the plea offer minus an informal sanction imposed on P by Θ that reflects Θ's belief in the possibility that an innocent D accepted the offer.

Following rejection of the plea offer, P has the option to drop the case. If she does so, then she receives no payoff from formal sanctions, but she receives an informal sanction from Θ, who believes with probability $\mu(G \mid d)$ that D is guilty, so by dropping the case P let a guilty defendant go free. Thus, P's payoff from dropping the case is simply:

$$\pi_d^P = - r_G^P \mu(G \mid d). \tag{9}$$

As earlier, since P may mix between dropping the case and going to trial, P's expected payoff following a rejection by D is given by:

$$\pi_R^P = \rho^P \pi_T^P + (1 - \rho^P)\pi_d^P. \tag{10}$$

3. Results

In this section we provide the main results; a sketch of the derivation is in the Appendix while the Technical Appendix contains the complete analysis. We start by providing notation for the strategies for each type of D and for P, for Θ's conjectures about these strategies, and two restrictions on the parameter space

that reflect sensible behavior by P. We then describe the game's equilibria and the reasons for selecting a specific equilibrium.

First, a strategy profile consists of: 1) a plea offer (S_b) made by P in Stage 2 and, in Stage 4, a probability (ρ^P) of taking a case to trial conditional on D having rejected the plea; this pair forms P's strategy for the game; and 2) one strategy for each type of D in Stage 3, denoted as ρ_G^D and ρ_I^D. In equilibrium, each strategy will take on numerical values, so we can summarize a hypothesized equilibrium by the four-tuple: $(S_b, \rho_G^D, \rho_I^D, \rho^P) \in [0,\infty)\times[0,1]\times[0, 1]\times[0, 1]$.

Thus, suppressing the plea offer for now, a candidate equilibrium wherein both types of D always reject the offer S_b and P always goes to trial is (1, 1, 1). Notice that P has conjectures about what types of D will choose to reject the offer. D has conjectures about what P will do, conditional on D's action and, while Θ cannot observe P's and D's actions, Θ has conjectures about what P will do and about what the types of D will do. All conjectures must be correct in equilibrium, but since P's and D's conjectures are not the primary focus of the paper (and are pretty standard) we do not generate additional notation for them.

Let $(\rho_G^{D\Theta}, \rho_I^{D\Theta}, \rho^{P\Theta})$ denote Θ's conjectures about the equilibrium choices that will be made by G, I, and P (at Stage 4), respectively. For any such triple of conjectures, Θ's beliefs about D are provided in the Appendix as equations (A.1a) through (A.1d).[23] We employ Perfect Bayesian equilibrium and require that: 1) P maximizes her expected payoff by choosing her plea offer S_b, given Θ's conjectures, P's prior beliefs about D's type, and anticipating how the continuation game will play out following P's choice of plea offer; 2) each type of D minimizes his expected loss by choosing his response to the plea offer, given Θ's conjectures, and anticipating how the continuation game will play out following his decision to accept or reject the plea offer; 3) P maximizes her expected continuation payoff via her choice to pursue trial or drop the case, given Θ's conjectures, and given P's posterior beliefs about D's type (based on his decision regarding the plea offer); and 4) all conjectures and beliefs are correct in equilibrium.

[23] We only provide the beliefs that D is of type G given an outcome; the corresponding beliefs for a D of type I are readily derivable.

In what follows we provide some natural parametric restrictions that imply that, in equilibrium, $\rho_I^D = 1$. Let $\pi_T^D(G; \rho_G^D)$ be $\pi_T^D(G)$, as specified in equation (2), with Θ's beliefs evaluated at arbitrary ρ_G^D and at $\rho_I^D = 1$.[24] Furthermore, let ρ_G^{D0} be the (unique) solution to the condition that $\pi_T^P = \pi_d^P$, where both of these expressions have Θ's beliefs and P's beliefs evaluated at ρ_G^{D0}. The value of ρ_G^{D0} is given by:

$$\rho_G^{D0} = - \lambda[(S_c - r_I^P)(1 - F_I) - k^P]/(1 - \lambda)[(S_c + r_G^P)(1 - F_G) - k^P]. \quad (11)$$

which is easily shown to be a positive fraction.

There are two scenarios concerning the decision by P whether to drop the case or go to trial that restrict the parameter space. First, if P (and Θ) know (or commonly believe) that D is innocent, then P should prefer dropping the case to going to trial. Moreover, since P's plea offer is not observable by Θ, and because P's beliefs upon seeing a choice by D whether or not to accept the offer may differ (at least out of equilibrium) from Θ's beliefs, we still want parameters consistent with P choosing to drop the case if she believes D to be innocent.

Maintained Restriction 1 (hereafter, MR1):

(a) If P and Θ know (or commonly believe) that D is of type I, P strictly prefers to drop the case. Formally, this reduces to: $(S_c - r_I^P)(1 - F_I) - k^P < 0$.

(b) If Θ's beliefs are consistent with $\rho_G^{D\Theta} \geq \rho_G^{D0}$, but P believes that D is of type I, then P strictly prefers to drop the case. Formally, this reduces to:

$$(S_c - r_I^P \mu(I \mid c))(1 - F_I) - k^P + r_G^P[\mu(G \mid d) - \mu(G \mid a)F_I] < 0.$$

Intuitively, MR1(a) will hold if either k^P or r_I^P is sufficiently large, or if both together are sufficiently large. MR1(b) further suggests that r_G^P should not be too large. Note that we abstract away from a personal disutility for P associated with convicting a defendant who she beleives is innocent, but subtracting such a disutility would simply reinforce MR1.

Second, imagine that both types of D were expected to reject P's offer. Then P's (and Θ's) posterior

---

[24] Beliefs for Θ are given by equations (A.1a) - (A.1d), evaluated at arbitrary ρ_G^D and $\rho_I^D = 1$; beliefs for P are given by $\nu(G \mid R) = \rho_G^D(1 - \lambda)/[\rho_G^D(1 - \lambda) + \lambda]$, because ρ_G^D of the G-types, and all of the I-types, are expected to reject the plea offer.

beliefs about the types would be the same as the prior beliefs (that $\Pr\{t = I\} = \lambda$); this would also be true if there was no opportunity for P to make a plea offer. In this scenario, P should prefer trial over dropping the case; this will be true if the police arrest process procedure is sufficiently effective at discriminating between guilty and innocent persons; that is, if λ, the prior probability that D is of type I, is not "too large."[25] These two scenarios imply the following maintained restrictions.

Maintained Restriction 2 (hereafter, MR2): If P and Θ know (or commonly believe) that the fraction of type G among those that reject the plea offer is the same as the prior, then P should prefer to take the case to trial rather than dropping it. Formally, this restriction reduces to:

$$(1 - \lambda)[(S_c + r_G^P)(1 - F_G) - k^P] + \lambda[(S_c - r_I^P)(1 - F_I) - k^P] > 0.$$

Note that the second term in brackets in MR2 is negative by MR1(a), so the above condition is an upper bound on λ (the arrest process is sufficiently effective). Moreover, MR1(a) implies that the first term in brackets in MR2, which represents the difference in the value of going to trial versus dropping the case against a D that is known (or commonly believed) to be type G, is positive.

Using these natural parameter restrictions, we find that the only[26] equilibria of the game involve the D of type I always rejecting the equilibrium plea offer, the D of type G rejecting the equilibrium plea offer with a positive probability ρ_G^D, and P never dropping a case when D rejects an offer. We provide a sketch of the proof of the following Proposition, which formalizes the description of the game's continuum of equilibria, in the Appendix (the complete proof is in the Technical Appendix).

Proposition 1: If r^D is not "too large" then there is a unique family of Perfect Bayesian equilibria for the game wherein P's equilibrium plea offer is $S_b(\rho_G^D) = \pi_T^D(G; \rho_G^D) - r^D$, the G-type rejects the offer with probability ρ_G^D in the interval $[\rho_G^{D0}, 1]$, the I-type always rejects the plea offer ($\rho_I^D = 1$), and P always goes to trial following a rejection ($\rho^P = 1$).

---

[25] Essentially, this is the reason for requiring probable cause for an arrest (i.e., there is a reasonable basis to believe a potential D committed a specific crime).

[26] Alternative candidates for equilibria, such as fully-separating or fully-pooling candidates, or candidates involving type I accepting a plea offer, cannot be equilibria; see the Technical Appendix for details.

Note the following:

1) Both the D of type I and P use pure strategies in equilibrium: all Ds who are innocent reject P's equilibrium offer, and the equilibrium involves a sufficient fraction of Ds who are guilty that also reject the plea offer, so that P will not choose to drop any case.[27]

2) One member of this family of equilibria is $(S_b(1), 1, 1, 1)$ wherein $\rho_G^D = 1$, in which case all Ds are rejecting P's offer, meaning that P's (and Θ's) beliefs about the types that chose R is the prior, so (by MR2) P goes to trial against all Ds.[28]

3) The smallest ρ_G^D-value (ρ_G^{D0}) in the family of equilibria is in (0, 1), so in almost all of the equilibria in the Proposition, G-types accept the plea offer with positive probability. From the perspective of Θ, since I-types always reject the offer, this means that $\mu(G \mid b)$ is one: a D who accepts the offer incurs the full sanction r^D from Θ.

Limits on Informal Sanctions

What do we mean by the qualifier in Proposition 1 "If r^D is not 'too large'"? Consider those equilibria wherein $\rho_G^{D0} \leq \rho_G^D < 1$; that is, G-types <u>accept</u> P's offer with positive probability. In order for this to occur, P must: 1) choose S_b from a non-empty set and 2) not wish to defect by making a very high offer to D so as to make D reject the offer for sure. Proposition 1 indicates that $S_b(\rho_G^D) = \pi_T^D(G; \rho_G^D) - r^D$, so that the interval of feasible S_b-values capable of inducing some acceptance is $[0, \pi_T^D(G; \rho_G^D) - r^D]$. Hence, in order to have a non-empty feasible set for S_b, it must be that $r^D \leq \pi_T^D(G; \rho_G^D)$. Substituting into this inequality yields our first condition restricting the level of informal sanctions, Condition 1 (see the Appendix):

<u>Condition 1</u> (feasibility). In order for P to be able to induce a D of type G to accept a plea offer, it must be that: $r^D \leq [S_c(1 - F_G) + k^D]/[1 - \mu(G \mid c)(1 - F_G) - \mu(G \mid a)F_G]$.

---

[27] Out-of-equilibrium beliefs for Θ following an unexpected dropped case are $\mu(G \mid d) = \rho_G^D(1 - \lambda)/[\rho_G^D(1 - \lambda) + \lambda]$; since ρ_G^D of the G-types and all of the I-types are expected to reject the plea offer, Θ interprets the unexpected dropped case as an error on the part of P.

[28] Now we also need an out-of-equilibrium belief for Θ's observation of outcome b; that belief would be that the type is G, since G does worse at trial than does I, so observing b should be associated with t = G: $\mu(G \mid b) = 1$. Alternatively put, I is willing to reject the plea offer for a strictly larger probability of subsequently going to trial than is G, so selection of the "safe" option (accept) is attributed to G. This argument is an application of the Cho and Kreps (1987) refinement D1.

As noted in the Appendix, the denominator on the right-hand-side of the above condition is positive. Multiplying through both sides of the above inequality by the expression in the denominator on the right, the resulting expression $r^D[1 - \mu(G \mid c;)(1 - F_G) - \mu(G \mid a;)F_G]$ is the increment in informal sanctions that the D of type G suffers by accepting a plea (which only a true G is expected to do) rather than going to trial (where there are trial costs plus a chance of conviction and a chance of acquittal, with corresponding informal sanctions), which is the resulting term now on the right (i.e., $S_c(1 - F_G) + k^D$). If there were no positive informal sanctions for D, then Condition 1 would be satisfied automatically. Thus, positive informal sanctions on D constrain P's ability to settle cases via plea bargain. Were r^D to violate Condition 1, this would eviscerate plea bargaining. This means that beliefs by P (and Θ) would be given by the prior, and according to MR2, P would always go to trial.

The second issue of concern is that the informal sanctions may result in P defecting from her part of the hypothesized equilibrium by making a plea offer large enough to provoke both types to reject. This could occur if, despite the presence of informal sanctions in P's expected payoff from trial, $S_b(\rho_G^D) = \pi_T^D(G; \rho_G^D) - r^D$ was less than what P could expect by driving those D's that would have otherwise settled to trial. In the Appendix we derive Condition 2 as the restriction that eliminates this incentive for defection.[29]

Condition 2 (no defection). For P to find it preferable to settle with a D of type G rather than provoking trial, it must be that:

$$r^D \leq [k^P + k^D + r_I^P\mu(I \mid c)(1 - F_G) + r_G^P\mu(G \mid a)F_G]/[1 - \mu(G \mid c)(1 - F_G) - \mu(G \mid a)F_G].$$

While r^D again appears on the left (and the denominators of the expressions in the two Conditions are the same), the informal sanctions on P, at rates r_I^P and r_G^P, contribute to the magnitude of the right-hand-side, and therefore also affect the ability to conclude a successful plea bargain. In particular, higher informal sanction rates on P increase the allowable range for r^D such that P will choose not to defect. Finally, P could also defect by dropping all cases following rejection; thus, we need to verify that P prefers the hypothesized

[29] Condition 2 is not necessary in the (1, 1, 1) equilibrium (i.e., when all types reject P's offer). Note also that Condition 2 holds Θ's beliefs fixed at the equilibrium ρ_G^D because trial is on the equilibrium path.

equilibrium outcome to what she would get by dropping all cases. Condition 1, however, is sufficient to imply this preference.[30]

Since P prefers to settle via plea bargain rather than going to trial against type G when Condition 2 holds, why can't P offer a slightly lower plea offer than $S_b(\rho_G^D) = \pi_T^D(G; \rho_G^D) - r^D$ and induce type G to accept with probability 1? The reason is that, since P is not pre-committed to trial, she needs to maintain her own incentives to go to trial following rejection. If P were to offer a slightly lower plea offer designed to induce type G to accept for sure, then *G should expect the case to be dropped following rejection*. This is because subsequent to such a deviation in the plea offer, P should infer that every rejection is coming from a D of type I, whereas Θ's beliefs are unchanged, since outside observers cannot observe this deviation. Under these conditions, MR1(b) implies that P strictly refers to drop the case. Anticipating that P will drop the case following the deviation induces G to reject the plea offer, despite its apparent allure. Consequently, P cannot gain by deviating to a lower plea offer.[31]

Selecting an Equilibrium

Proposition 1 characterizes the nature of the equilibria for the game, but we are still left with a continuum of equilibria. We now propose a basis to select a unique member of that family, namely $(S_b(\rho_G^{D0}),\rho_G^{D0}, 1, 1)$, which is the equilibrium with the highest probability of plea acceptance. Notice that Θ's beliefs punish I-types by lumping them in with G-types, since the two types can't be distinguished. For example, $r^D\mu(G \mid c)$ is the sanction for a D that is convicted, whether he is truly guilty or innocent; if Θ knew that D was a wrongly-convicted I, then one would expect the sanction to be (at worst) zero. Θ erroneously punishes an I, based on observing a conviction, with probability $\lambda(1 - F_I)$, so the expected misclassification

[30] To see why, notice that in the hypothesized equilibrium, P's payoff involves some pleas and some trials. P is indifferent between trying and dropping the case for $\rho_G^D = \rho_G^{D0}$ (and strictly prefers trying to dropping for $\rho_G^D > \rho_G^{D0}$). Then Condition 1 implies that the settlement offer $S_b(\rho_G^D)$ is non-negative, whereas P's payoff from dropping the case is $- r_G^p\mu(G \mid d)$, which is strictly negative. Thus, P strictly prefers the outcome involving some accepted plea offers and some trials to defecting to dropping all cases.

[31] For the equilibrium at $\rho_G^D = \rho_G^{D0}$, there is actually a continuation equilibrium following the out-of-equilibrium plea offer $S_b < S_b(\rho_G^{D0})$ wherein type G mixes between accepting and rejecting the plea offer with probability ρ_G^{D0} and P mixes between taking the case to trial and dropping it; see the Appendix for details.

loss for this scenario is $\lambda(1 - F_I)r^D\mu(G \mid c)$. Similarly, if P obtains a conviction against D, then she will suffer an informal sanction based on Θ's beliefs that the convicted D might be innocent, in the amount of $r_I^P\mu(I \mid c)$. But if Θ knew that D was a wrongly-convicted I, then the appropriate informal sanction for P would be r_I^P. Thus the outside observer's expected misclassification loss for this scenario is $\lambda(1 - F_I)[r_I^P - r_I^P\mu(I \mid c)]$.

In the Appendix we provide an overall expected loss from misclassification, denoted as $M(\rho_G^D)$, which is shown to reduce to the following expression:

$$M(\rho_G^D) = (r^D + r_I^P)\{\lambda(1 - F_I)\mu(G \mid c) + \rho_G^D(1 - \lambda)(1 - F_G)\mu(I \mid c)\}$$
$$+ (r^D + r_G^P)\{\lambda F_I\mu(G \mid a) + \rho_G^D(1 - \lambda)F_G\mu(I \mid a)\}. \quad (12)$$

We further show that $M(\rho_G^D)$ is increasing on $[\rho_G^{D0}, 1]$. That is, if we assume that outside observers, while not bearing any net costs for imposing informal sanctions, have a preference to have a legal system that allows them to achieve the smallest expected loss from misclassification, then the observers should adopt the conjecture ρ_G^{D0}, and the associated beliefs. In Proposition 2 we adopt this notion to select the specific equilibrium $(S_b(\rho_G^{D0}), \rho_G^{D0}, 1, 1)$.

Proposition 2. If r^D is not "too large" and if the observers adopt the conjectures and beliefs that minimize the expected loss from misclassification, then the unique equilibrium for the game is that P makes the plea offer $S_b(\rho_G^{D0})$, G-types reject the offer with probability $\rho_G^{D0} < 1$, I-types always reject the offer, and P always takes all Ds that reject the plea offer to trial.

Notice that the notion of preferring a lower value of ρ_G^D is also consistent with a "veil of ignorance" argument for a Θ who realizes that they might become a D some day. Higher values of ρ_G^D than ρ_G^{D0} mean lower payoffs for both G-types and I-types. Behind a veil of ignorance as to whether or not Θ might become a D, any positive probability associated with that possibility means that Θ should prefer ρ_G^{D0} to any higher value of ρ_G^D.

Notice also that the equilibrium in Proposition 1 wherein $\rho_G^D = 1$ provides the same payoffs as if there were no plea bargaining. Thus, since the selected equilibrium involves $\rho_G^D = \rho_G^{D0} < 1$, then both types of defendant and the outside observer *all prefer that plea bargaining be possible*. This reflects the externality that even though all I-types reject the plea offer and go to trial, the fact that some G-types accept the offer

reduces the likelihood that a defendant choosing trial is guilty, thereby raising the equilibrium belief of innocence for a defendant who chooses to reject the plea offer.

A special subcase of our model involves eliminating the informal sanctions, so that $r^D = r_I^P = r_G^P = 0$. Again, the G-type mixes because P is not committed to going to trial, so some fraction of the G-types must reject the offer in order that P not defect to dropping cases (due to MR1). In this case Θ's beliefs do not affect D or P. Conditions 1 and 2 now always hold, so there is always a plea bargain that P wants to make and that results in acceptance by some G-types. Thus, we can see that it is the informal sanctions that can restrict or eliminate plea bargaining.

4. Implications of the Equilibrium: Comparative Statics

The three most important endogenous variables in the analysis are the likelihood of bargaining failure (ρ_G^{D0}), the equilibrium plea offer ($S_b(\rho_G^{D0})$), and the observer's beliefs after observing the outcome of a trial, $\mu(G \mid a)$ and $\mu(G \mid c)$, which are computed using the selected equilibrium in equations (A.1a) and (A.1c) in the Appendix. We can focus on $\mu(G \mid a)$ as it turns out that the comparative statics for $\mu(G \mid c)$ have the same sign, which are opposite in sign to those for the beliefs about I-types. Thus, in equilibrium, the posterior belief as to acquitted G-types is:

$$\mu(G \mid a) = \rho_G^{D0}(1 - \lambda)F_G/[\rho_G^{D0}(1 - \lambda)F_G + \lambda F_I]. \tag{13}$$

A small amount of algebra, after employing equation (11), yields the somewhat surprising result that $\mu(G \mid a)$ is, in equilibrium, independent of λ (and this holds for all the other equilibrium beliefs for the observer). That is, the presence of plea bargaining (and the fact that P is not pre-committed to trial) isolates the beliefs following the the trial outcomes from the arrival frequency of innocent defendants (λ) into the system. Notice that this is despite the fact that λ affects ρ_G^{D0} (that derivative is positive).

Table 1 (which uses results from the Technical Appendix) provides a summary of the effects of increases in the parameters of the model (listed at the top of the Table) on these three endogenous variables. The columns are labeled by the parameters of interest as well as the "source" of power (or weakness) in the

following sense: 1) Θ imposes the informal sanctions at rates r^D, r_I^P, and r_G^P; 2) higher S_c (or lower k^P) should make P a stronger player; and 3) higher k^D should make D a weaker player, while higher λ suggests a higher likelihood that D is innocent. In what follows we discuss these basic comparative statics results and draw out some further implications.

Table 1: Primary Comparative Statics

	Exogenous Parameters						
	Θ			P		D	
	r^D	r_I^P	r_G^P	S_c	k^P	k^D	λ
ρ_G^{D0}	0	+	–	–	+	0	+
$S_b(\rho_G^{D0})$	–	+	–	**?***	+	+	+
$\mu(G \mid a)$	0	+	–	–	+	0	0

Table note *: The direct effect is positive and the indirect effect is negative.

The Effect of Changes in the Informal Sanction Rates

From the Table we see that an increase in r^D has no effect on ρ_G^{D0}, but it reduces the equilibrium plea offer, $S_b(\rho_G^{D0})$.[32] However, an increase in r_I^P (or a decrease in r_G^P) increases both ρ_G^{D0} and $S_b(\rho_G^{D0})$. For example, consider an increase in r_I^P; it enters ρ_G^{D0} via the numerator, so according to equation (11) increasing r_I^P means that the computed value of ρ_G^{D0} should (in equilibrium) rise. The intuition for this rise in ρ_G^{D0} is that increasing r_I^P makes trial less appealing to P, since it is the errors arising from trial that can result in an I-type being convicted, resulting in a Θ-imposed penalty on P. This undermines P's incentives to take Ds who reject the offer to trial so, to maintain those incentives, more G-types must be in the mix, thereby requiring that more reject the offer. More G-types in the mix allows the equilibrium plea bargain to rise. Furthermore, a higher value of r_I^P, which causes a greater fraction of G-types to reject the offer, means that the pool of Ds at trial is richer in G-types, meaning that both conviction and acquittal are more likely to be associated with a G-type

[32] Perhaps variations in states' laws regarding the extent to which employers can (or even must) account for a potential employee's arrest and/or conviction history can be used to test this pair of predictions.

(that is, $\mu(G \mid a)$ increases, as shown in the last row of the Table).

It can also be shown that an increase in r_I^P increases both the right-hand-side of Condition 1 and the right-hand-side of Condition 2, thereby relaxing the restriction on r^D-values consistent with the possibility of plea bargaining. Further, an increase in r_G^P decreases the right-hand-side of Condition 1, so greater social opprobrium towards both defendants and towards P's who might be viewed as "soft on crime" (i.e., enabling the guilty to escape justice) means that the options for P to successfully conclude a plea bargain go down. Finally, and unfortunately, the effect of an increase in r_G^P on the right-hand-side of Condition 2 is not clear.[33]

The effect of a change in the informal sanction rates on the beliefs that outside observers hold about Ds who go to trial has a further, seemingly paradoxical effect. Notice that an increase in r_I^P increases $\mu(G \mid a)$, so a stronger negative reaction by outside observers to the possibility that P is "railroading" innocents actually leads to <u>higher</u> informal sanctions on all Ds, since there is no way for a Θ to know whether D is a G or an I. In particular, this would imply that I-types caught up in the system would suffer greater levels of informal sanctions when outsiders wish to penalize P for railroading. This seemingly perverse result reflects the lack of a feedback effect between the magnitude of the outsiders' distaste for Ps that "railroad" I-types and any effort by Ps to reduce the proportion of I-types in the original pool. While we have not incorporated effort by P on (say) monitoring the arrest process so as to reduce the frequency of innocents arrested, the effects of this are clear: if P could exert effort to reduce λ (for example, by more intensive screening of cases provided by the police) and this added effort could be brought to the outside observers' attention (as, say, "police reform"), such effort might act to reduce λ and, more importantly from P's perspective, r_I^P.

Finally, the Table indicates that an increase in r_G^P has the opposite effect on the equilibrium beliefs held by Θs. Thus, if outside observers increase their informal sanction rates on P due to a perspective that P is "weak on crime," then this reduces $\mu(G \mid a)$, leading to <u>lower</u> informal sanctions on Ds who go to trial, since they anticipate that (in equilibrium) the likelihood of G-types going to trial has gone down, making the

[33] Its effect appears to be complexly-related to the other parameter values as well as the characteristics of the F-distribution.

pool of Ds going to trial richer in I-types.

Changes in the Level of the Exogenous Formal Sanction and in the Costs of Trial

Another set of predictions concern the effect of an increase in the formal sanction, S_c. As indicated in the Table, an increase in S_c results in an increase in the likelihood of accepting the plea offer (ρ_G^{D0} falls). Intuitively, this occurs since increasing S_c increases P's expected payoff from trial. Since ρ_G^{D0} is determined by P's indifference between taking a case to trial versus dropping it, then fewer G-types need to reject a plea offer to make P just indifferent between trial and dropping the case. That is, P's choice to go to trial is credible for a greater range of possible rejection rates and therefore a lower minimal level of that range.

What if the cost of trial to P, k^P, falls? As indicated in the Table, direct computation from equation (11) above shows that ρ_G^{D0} falls as does the equilibrium plea offer, so again what would seem like a reason for P to pursue more trials (since trial is cheaper) turns out, in equilibrium, to induce fewer trials (all I-types go to trial, as before, but fewer G-types do). An increase in k^D weakens D; since this cost only appears in D's expected loss from going to trial, the equilibrium plea offer increases but the likelihood of bargaining success does not change (since this latter strategy is determined by P's incentives to drop or proceed to trial).

Another measure of potential interest is P's overall conviction rate, which is given by:

$$\lambda(1 - F_I) + (1 - \lambda)\rho_G^{D0}(1 - F_G) + (1 - \lambda)(1 - \rho_G^{D0}),$$

where the first two terms come from I-types and G-types who go to trial and the last term reflects G-types who plead guilty. This reduces to $1 - \lambda F_I - (1 - \lambda)\rho_G^{D0}F_G$, which is clearly increasing as ρ_G^{D0} falls. Thus, anything that lowers ρ_G^{D0} (holding λ and the evidence distributions constant) will raise P's conviction rate since any Ds accepting a plea are (formally) convicted of the offense. In particular, an increase in S_c (or a decrease in k^P) results in an increase in the conviction rate.

Finally, as shown in the Technical Appendix, an increase in S_c increases the overall expected level of informal sanctions that a G-type faces and decreases the overall expected level of informal sanctions that an I-type faces. Thus, the expected level of informal sanctions is positively correlated, for G-types, with the

level of the formal sanction, while it is negatively-correlated for those who are truly innocent.[34]

Police Arrest and Trial Process Effectiveness

Earlier we reflected on police arrest effectiveness as involving λ being small: the intake process into the legal system is effective if the likelihood that the police arrested an I-type is (relatively) small. From equation (11) above (and as indicated in the Table) reductions in λ lead to reductions in ρ_G^{D0} and $S_b(\rho_G^{D0})$, meaning that improvements in police arrest effectiveness reduce the plea offer and increase the likelihood of plea bargaining success, thereby reducing the expenditure of court costs, since fewer cases go to trial.

We are also interested in the effectiveness of the trial process at properly convicting the guilty and acquitting the innocent. If trials were perfect discriminators of guilt or innocence, we would have $F_I = 1$ (all innocent Ds are acquitted) and $F_G = 0$ (all guilty Ds are convicted). This thought experiment informs us about what we want investment in trial resources to do; such investments might involve better procedures for obtaining and vetting evidence, or improved procedures for the trial itself.

Let z be a level of investment in trial resources, and now extend our earlier notation $F(\gamma_c \mid t)$ to incorporate investment z. Let $F(\gamma_c \mid t, z)$, t = I, G, denote the probability that type t's evidence draw yields an acquittal if the investment level is z and the evidentiary standard for conviction is γ_c. Based on our earlier example of a perfect trial, this means that (employing a first-order stochastic dominance model of investment):

$$F_{Iz} \equiv \partial F(\gamma_c \mid I, z)/\partial z > 0 \text{ and } F_{Gz} \equiv \partial F(\gamma_c \mid G, z)/\partial z < 0.$$

Such an investment would increase trial effectiveness (we abstract from concerns about the cost of such investments). Some investments may only affect F_I or F_G, while some might affect both.

Again, returning to equation (11), an increase in z that only affects F_G thereby only increases the denominator of ρ_G^{D0}, meaning that such an investment increases the likelihood of plea-bargaining success. The

[34] We have not attempted, in this paper, to consider the question of the optimal level of the formal sanction itself; this would require the addition of a sufficiently detailed model of criminal choice and deterrence, as well as social welfare, which is beyond the scope of the current paper. Our point here is that such a model of optimal formal sanction choice must be responsive to both formal and informal sanctions, as well as the social costs (e.g., possible aggregate productivity losses) that each type of sanction may engender.

stand-alone effect of z via F_I is more complicated. This is because of the sign of $S_c - r_I^P$ is unconstrained by MR1 and MR2. If the informal sanction rate is small ("small r_I^P," meaning $S_c - r_I^P > 0$), then an increase in z leads to an increase in F_I, which leads to an increase in ρ_G^{D0}, which means less settlement, and higher plea offers as well (since $S_b(\rho_G^{D0})$ also rises in equilibrium). If the informal sanction rate is large ("large r_I^P," meaning $S_c - r_I^P < 0$), then an increase in z leads to an increase in F_I, which leads to a decrease in ρ_G^{D0}, which means more settlement, and lower plea offers as well (since $S_b(\rho_G^{D0})$ also falls in equilibrium). Putting this together with the effects on F_G, we see that the effect of an increase in z in the large-r_I^P case definitely leads to a reduction in ρ_G^{D0} (and in $S_b(\rho_G^{D0})$), as the numerator of equation (11) is falling while the denominator of equation (11) is increasing. Here, trial is becoming clearly more effective at separating I-types and G-types and providing them with the corresponding dispositions of a and c, respectively: trial is a better tool for <u>not</u> convicting the innocent and <u>for</u> convicting the guilty. Sadly, this clarity is not present in the small-r_I^P case when investment affects both F_I and F_G because it is not possible to sign the effect of z on ρ_G^{D0}, since both the numerator and the denominator of equation (11) are increasing.

5. Equilibrium Plea Acceptance by Innocent Defendants

A common result[35] for models of the law and economics of plea bargaining is that all truly innocent defendants reject the plea offer and choose to go to trial. This prediction is inconsistent with reality wherein some fraction of innocent defendants plead guilty.[36] While there may be a number of non-rational reasons for such behavior (e.g., mental illness or poor legal representation), in this section we will modify our base model to obtain an equilibrium wherein a fraction of innocent defendants rationally choose to accept a plea offer. This adjusts an important previous result since, if innocent defendants do accept a plea offer, then an

[35] To our knowledge, the only exception is Reinganum (1988). She constructs an equilibrium wherein some innocent defendants plead guilty. However, this is due to the fact that in her model, prosecutors can commit to go to trial upon a plea offer rejection and they have accurate information about the probability of a win at trial (independent of the defendant's true type), which thereby eliminates meaningful distinction between the expected trial payoffs of guilty and innocent defendants.

[36] The National Registry of Exonerations, maintained by the University of Michigan Law School, indicates that of 1465 exonerees (as of October 2014), 13% of the wrongful convictions involved false confessions (http://www.law.umich.edu/special/exoneration/Pages/learnmore.aspx, accessed January 24, 2015).

outside observer's belief about the likelihood of guilt of a defendant who accepts a plea offer must be less than one; we examine this, and other, implications below.

Modifying the Basic Model

Recall that the type space for the model described and analyzed earlier was $t \in \{G, I\}$, with the prior probability that a D was innocent as $\lambda = \Pr\{t = I\}$, and with D's type his private information. We now revise the model's type space by assuming that, independent of whether D is truly guilty or innocent, a defendant may also be "strong" or "weak," denoted as $\{S, W\}$, with ω the probability that D is a weak type and, again, this attribute is D's private information. That is, our overall type space is now $t \in \{GS, GW, IS, IW\}$, and in our re-formulated structure, $\lambda\omega = \Pr\{t = IW\}$, $\lambda(1 - \omega) = \Pr\{t = IS\}$, and so forth. In what follows, we assume that ω is sufficiently small, so that the equilibrium is of the same form as in Section 3 in the sense that it is the GS-types who will be made indifferent between trial and accepting the plea offer, and this will make P willing to go to trial against any D who rejects the offer.

By "strong" we mean a D as in Sections 2 and 3: a risk-neutral, expected loss-minimizing, agent. In particular, we know that Remark 1 will now hold for a strong D: $\pi_T^D(IS) < \pi_T^D(GS)$, where these losses are exactly as developed in Section 2 (that is, $\pi_T^D(IS)$ is precisely what was denoted in Sections 2 and 3 as $\pi_T^D(I)$ and $\pi_T^D(GS)$ is precisely what was denoted in Sections 2 and 3 as $\pi_T^D(G)$). By "weak" we will mean that such a G or I suffers a higher loss than the analogous strong D, and one that will be sufficiently high (as to be detailed shortly) so as to produce the prediction that some innocent Ds accept the equilibrium plea offer.

What might be a source of such weakness? It needs to be something that is D's private information and that cannot be costlessly, credibly revealed by D. Two important possibilities come to mind. First, a W might be a risk-averse version of an S. This is not an argument that an I is more risk averse than a G (as discussed in Becker, 1968), but rather that an IW (resp., a GW) is more risk-averse than an IS (resp., a GS). We have argued above that the equilibrium will involve a plea offer that makes a D of type GS indifferent; a D of type IS will therefore reject such an offer for sure. Risk aversion increases a D's willingness-to-pay to avoid risk, so any plea offer that makes a strong G indifferent would be accepted for sure by a weak G. The

only remaining question is whether the D of type IW is willing to accept the plea offer that makes the D of type GS indifferent. Being I rather than G makes him less willing, but being W (i.e., risk-averse) rather than S (i.e., risk-neutral) makes him more willing. In what follows, we assume that the degree of risk aversion among weak defendants is sufficient to make the D of type IW prefer this plea offer to the risks of trial.

A second possible source would be ambiguity aversion,[37] wherein a weak D would put greater weight on bad outcomes from Nature's choice of evidence than the given probability distributions $F(e \mid t)$, for $t = G$ or I would imply. For example, if we follow the results of Gilboa and Schmeidler (1989), the decision-maker places all the weight on the worst possible outcome (i.e., $e = 1$), yielding the maximum possible expected loss as the expected loss from trial. Alternatively, one could employ the model developed by Klibanoff, Marinacci, and Mukerji (2005), where the decision-maker uses his set of possible priors (relative to the ambiguity-neutral, strong D) and a function which captures his response to ambiguity, to provide an expected loss from trial, which will imply a larger expected loss than that of the analogous strong type. Again, if IW and GW are sufficiently ambiguity averse, then any plea offer that makes GS indifferent will be preferred to trial by IW and GW, but will be rejected by IS.

As shown in the Technical Appendix, the resulting equilibrium is a direct extension of that developed in Section 3: all IS-types reject the offer, the same fraction ρ_G^{D0} of GS-types reject the offer, all W-types accept the offer, and P takes all Ds who reject the offer to trial. The beliefs formed by Θ under the outcomes acquit (a), convict (c), and drop (d) are undisturbed by the modifications made above (see the Technical Appendix). However, $\mu(G \mid b)$ is no longer unity, as it was in Section 3. Rather, since an outside observer knows that there is a positive probability of an IW-type, then $\mu(G \mid b) < 1$ for all $\omega > 0$, and $\mu(G \mid b)$ is a decreasing function of ω: as the fraction of weak defendants increases, accepting the plea offer is an

[37] Crudely, a decision maker is ambiguity averse if they entertain a number of alternative models of the probability of the random events they face, and they prefer a world of one model to a world with many; such decision makers need not be risk averse, but many analyses allow for attitudes towards risk and attitudes towards ambiguity. See Machina and Siniscalchi (2014) for a comprehensive survey of this topic. The typical result in this area (and there are many) starts with the Savage axioms for subjective expected utility theory, modifies one (or more) of those (typically, the sure-thing principle), substitutes some new axioms, and then derive the form of the criterion for the ambiguity-averse decision-maker to use. For a recent application of ambiguity aversion to civil suits, see Franzoni (2014).

increasingly weak signal of guilt. The equilibrium rate of acceptance is now $\omega\lambda + [\omega + (1 - \omega)(1- \rho_G^{D0})](1 - \lambda)$, which is higher than in the base model, wherein this rate was $(1 - \rho_G^{D0})(1 - \lambda)$. P is able to obtain a plea agreement with more guilty defendants, but also unavoidably sweeps up some innocent defendants as well.

The equilibrium plea offer is also affected because a defendant accepting a plea offer is no longer inferred to be guilty for sure. The plea offer in the base model is $S_b(\rho_G^{D0}) = \pi_T^D(G; \rho_G^{D0}) - r^D$, whereas the new plea offer is $S_b(\rho_G^{D0}) = \pi_T^D(GS; \rho_G^{D0}) - r^D\mu(G \mid b; \rho_G^{D0})$. Since $\pi_T^D(G; \rho_G^{D0})$ from the base model and $\pi_T^D(GS; \rho_G^{D0})$ from the modified model are the same function, the plea offer is higher in the model with weak types.

6. Refining the Jury's Assessment: The Scottish Verdict

For almost 300 years, Scotland has used a three-outcome verdict for criminal juries; a defendant is found not guilty, or not proven, or guilty, with no formal sanction attaching to the first two outcomes.[38] Such a refinement of the jury's assessment of a defendant's guilt or innocence should provide more information to the outside observers to employ in applying informal sanctions. Does it and, if so, what else does it do?

To address this extension, we return to our base model but strengthen an earlier assumption about the distribution $F(e \mid t)$. We assume that F is differentiable in e and that the strict monotone likelihood ratio property (SMLRP) holds:

SMLRP: $f(e \mid G)/f(e \mid I)$ is strictly increasing in e, for e in (0, 1). (14)

This assumption implies (see Müller and Stoyan, 2002, p. 61): 1) $F(e \mid I) > F(e \mid G)$ (strict stochastic dominance by G); 2) $f(e \mid G)/(1 - F(e \mid G)) > f(e \mid I)/(1 - F(e \mid I))$ (strict hazard rate dominance by G); and 3) $f(e \mid G)/F(e \mid G) > f(e \mid I)/F(e \mid I)$ (strict reverse hazard rate dominance by G). We represent the three-outcome verdict by the triple {ng, np, g}, with the obvious interpretation, and assume that $\gamma_g \equiv \gamma_c$ (that is, the same evidentiary standard for a conviction under the previous two-outcome verdict is used to find a defendant

[38] See Duff (1999) for an extensive discussion of the history of the development of this institution; he indicates that not proven verdicts are reached by juries in approximately one-third of the acquittals. Bray (2005) indicates that the same three-outcome verdict was used in the 1807 trial of Aaron Burr for treason. Also, see Leipold (2000) for a discussion of the "California Alternative" wherein an acquitted defendant can petition the court for a declaration of factual innocence. While this two-stage process seems similar to the three-verdict approach in Scotland, Leipold (2000, p. 1324) indicates that California imposes "a nearly prohibitive burden of proof" on the defendant, as the defendant must prove that there was "no reasonable cause to believe that he committed the crime." As Leipold observes, this results in relatively few defendants pursuing this remedy.

"guilty" under the three-outcome verdict).[39] Further, let γ_{ng} be the cutoff for not guilty versus not proven, where $0 < \gamma_{ng} < \gamma_g$. Thus, we extend the previous notation so that $F_t(\gamma_g) \equiv \Pr\{e \leq \gamma_g \mid t\}$ and $F_t(\gamma_{ng}) \equiv \Pr\{e \leq \gamma_{ng} \mid t\}$, for t = I, G.

In the Technical Appendix we show that:

1) For any non-zero vector of strategies by D, (ρ_G^D, ρ_I^D), Θ's beliefs as to D's likelihood of being of type G, having observed one of the mutually-exclusive outcomes ng, np, or g, satisfies:

$$\mu(G \mid ng) < \mu(G \mid np) < \mu(G \mid g); \tag{15}$$

and 2) I's expected loss from proceeding to trial ($\pi_T^D(I)$) is strictly lower than G's expected loss from proceeding to trial ($\pi_T^D(G)$), where:

$$\pi_T^D(t) = S_c(1 - F_t(\gamma_g)) + k^D + r^D\mu(G \mid g)(1 - F_t(\gamma_g)) + r^D\mu(G \mid ng)F_t(\gamma_{ng}) + r^D\mu(G \mid np)(F_t(\gamma_g) - F_t(\gamma_{ng}))\}, \; t \in \{I, G\}. \tag{16}$$

The ordering of payoffs indicated above means that Proposition 1 applies to the modified game, so that the family of equilibria again involves I-types always rejecting the plea offer and P always taking any D who rejects a plea offer to trial, while G-types mix between accepting the plea offer and rejecting it with probability ρ_G^D. When P's expected payoff from trial (π_T^P) is extended to allow for the three outcomes, it turns out that this function is independent of γ_{ng}. This means that since ρ_G^D makes P indifferent between dropping and going to trial, then the equilibrium values for ρ_G^D are the same as in the two-outcome verdict regime: $\rho_G^D \in [\rho_G^{D0}, 1]$. This occurs because P's computed expected payoffs from trial simply reflect whether D is found guilty or is acquitted. Furthermore, as shown in the Technical Appendix, Conditions 1 and 2 now hold for a larger range of the parameter r^D, so the use of the three-outcome verdict means that informal sanctions are less likely to interfere with plea bargaining. Proposition 2 carries over to the modified game as well: the selected equilibrium value is still ρ_G^{D0}.

---

[39] One might wonder whether the introduction of the partitioning of acquittal into not guilty and not proven might cause juries to effectively adjust the evidentiary standard used to convict. Not surprisingly, there is not much evidence about this (we have assumed no change). Our assumption is consistent with results in the only experimental work of which we are aware; three psychologists consider just this issue in a series of laboratory experiments (see Smithson, et. al., 2007) and find that the introduction of the third verdict option does not significantly alter the likelihood of a conviction (p. 492). They also find that reported assessments of guilt follow the same montonicity as shown in our equation (15) below.

The change in the number of possible verdict outcomes affects Θ and D. The difference between the functions capturing the expected loss from misclassification for the three-outcome verdict and the two-outcome verdict is that now the term $\mu(G \mid np)(F_t(\gamma_g) - F_t(\gamma_{ng})) + \mu(G \mid ng)F_t(\gamma_{ng})$ replaces $\mu(G \mid a)F_t(\gamma_g)$ in the computation. This change also appears in D's expected cost from the three-outcome trial verdict. Using the assumption SMLRP, we show in the Technical Appendix that the following result holds:

$$\mu(G \mid np)(F_G(\gamma_g) - F_G(\gamma_{ng})) + \mu(G \mid ng)F_G(\gamma_{ng}) > \mu(G \mid a)F_G(\gamma_g), \tag{17a}$$

and $$\mu(G \mid np)(F_I(\gamma_g) - F_I(\gamma_{ng})) + \mu(G \mid ng)F_I(\gamma_{ng}) < \mu(G \mid a)F_I(\gamma_g). \tag{17b}$$

The implication of inequalities (17a) and (17b) is summarized in Proposition 3.

> Proposition 3: The expected cost for a D of type G is higher under the three-outcome verdict than under the two-outcome verdict. The expected cost for a D of type I is lower under the three-outcome verdict than under the two-outcome verdict.

This means that while a G-type still rejects the equilibrium plea offer at the same rate as before, ρ_G^{D0}, the equilibrium offer itself is larger, since $S_b(\rho_G^{D0}) = \pi_T^D(G; \rho_G^{D0}) - r^D$, but the expected loss for the G-type has increased. Thus, P's overall payoff increases (since the plea bargains are tougher and are accepted at the same rate, and P's trial payoff is unchanged). Finally, it is also straightforward to show that Θ's expected loss from misclassification is lower in the three-outcome verdict regime. We take this to mean that, overall, use of the Scottish verdict would enhance justice: I-types lose less, G-types lose more, and Θs impose fewer erroneous informal sanctions.

Further Assessment Refinement

Clearly the foregoing analysis suggests that schemes providing more precision with respect to the jury's assessment may be socially valuable. One should be cautious, however, in how this is implemented. For example, transcripts of trials are generally pubic records, but few people wish to expend the effort cost of obtaining and reading them, and the transcripts lack information about the visual or vocal cues that the jury

observed during testimony (which the jury used to assess credibility).[40] Televised trials (which occur only for a rare, selected, subset of trials) often involve extra-legal commentary, introducing expert (and/or incompetent) opinion and evidence that was not contemplated by the jury. Overall, one might expect that the social costs (i.e., to the jury as well to the outside observers) of finer resolution of the jury's verdict are likely to be strictly convex in the degree of resolution. Moreover, since a jury is not a unitary agent, but comes to a judgment via voting, aggregation of the jurors' assessments into a small number of discrete alternatives is possible, but requiring a jury to announce a specific assessment for a continuous variable is likely to fail. Thus, a prescription to "reveal e" precisely might, in reality, not be as useful as it seems.

7. Summary and Further Discussion

Our model considers the strategic interaction between a prosecutor and a defendant when informal sanctions by third parties can be imposed on both the defendant and the prosecutor. These sanctions affect the feasibility of plea bargaining, as well as the level of the bargain offered and the frequency of bargaining success. The model follows the action from choosing a plea offer up through trial, allowing for dropping of cases, thereby not relying on prosecutors being able to pre-commit to taking defendants who reject offers to trial. The defendant's private information concerns his guilt or innocence of the crime for which he was arrested; this underlying state of the world affects the evidence that is presented at trial. Third parties form beliefs about the defendants who are processed through the system, allowing for outcomes wherein a plea is accepted, or a case is dropped, or a defendant goes to trial and is convicted or acquitted. Significantly, while the third parties can observe the disposition of the cases, they cannot observe plea offers or evidence generation; of course they are rational and can construct the equilibrium of the game, but the errors in the legal process (as well as hidden information) means that they will misclassify defendants and thereby erroneously impose sanctions on both defendants and prosecutors.

[40] This is why appeals courts in the U.S. do not use the trial record to re-try the case (or incorporate new evidence) and give deference to the jury's decision; if there appears to have been a procedural error (for example, evidence was included that should have been excluded), the appeals court may order a new trial with a new jury again seeing evidence presented in court.

We show that there is a unique family of equilibria and, if third parties prefer a legal system with minimal expected loss arising from misclassification, a unique equilibrium within this family, wherein the guilty defendant accepts the prosecutor's proposed plea offer with positive (but fractional) probability, the innocent defendant rejects the proposed offer, and the prosecutor chooses to take all defendants who reject the offer to trial. The plea offer and the decisions by each agent are all a function of the informal sanctions.

We find that informal sanctions on the defendant act to constrain the set of possible plea offers the prosecutor can make, but informal sanctions on the prosecutor can work in opposite directions depending upon whether the concern is for convicting the innocent or letting the guilty escape justice. If the informal sanction rate on defendants is high enough, then plea bargaining is eviscerated. High potential sanction rates on prosecutors can also lead to distortions, with high sanction rates on prosecutors for likely conviction of innocents leading to tougher plea offers, and more guilty defendants rejecting the plea offer and going to trial, which increases the association between any trial outcome (conviction or acquittal) and guilt. This, in turn, leads to a higher level of informal sanctions on both acquitted and convicted defendants, thereby more heavily penalizing innocents. Alternatively, high potential informal sanction rates on prosecutors for perceived release of the guilty ("soft on crime") leads to greater use of plea bargaining, as offers are made less tough, leading to fewer guilty defendants going to trial, thereby leading to lower levels of informal sanctions for those who went to trial (including those who are truly guilty).

We further consider the importance of police effectiveness (that police have a solid basis for an arrest in the first place, thereby reducing the likelihood that innocents are swept up into the process) as well as trial effectiveness (the degree to which evidence acts to correctly classify defendants as to their guilt). Increased police effectiveness leads to greater use of plea bargaining to resolve offenses, and thereby saves trial costs. Increased investment in the trial process, when informal sanctions on prosecutors for possibly "railroading" innocents (that is, r_I^P) is sufficiently high, has a clear result of enhancing the effectiveness of plea bargaining; results are less clear if the informal sanctions are less important to the prosecutor than her utility for the formal sanction associated with conviction at trial.

Finally, we consider two primary extensions. First, we show that the model can readily be extended to provide the prediction that a fraction of innocent defendants will accept the plea offer; this is obtained by incorporating unobserved heterogeneity of the defendants with respect to their response to risk and/or ambiguity. Moreover, if the fraction of risk- and/or ambiguity-averse defendants is small, then while he plea acceptance rate is higher (as is the plea offer itself), the fraction of truly guilty defendants going to trial is exactly the same, and therefore the comparative statics remain as in the base model. Second, we examine the effect of extending the analysis to the "Scottish" (three-outcome) verdict, wherein juries find a defendant not guilty, not proven, or guilty. We find that the same equilibrium obtains, including the same likelihood of bargaining success, with the only modification being that the equilibrium plea offer is higher. This reflects the result that, at trial, guilty defendants do worse under this scheme while innocent defendants do better. Moreover, outside observers impose informal sanctions with a lower total expected loss from misclassification. Overall, the Scottish verdict leads to an increase in justice.

Possible Extensions

One possible direction of extension would be to allow for the presence of multiple chargeable offenses, thereby opening the door to both charge-bargaining (where prosecutors can modify the charged offense, thereby possibly affecting the informal sanctions defendants will face as well as the formal sanction at stake) and the employment of lesser included offenses for juries to consider if they find the major charged offense to be insufficiently supported by the evidence. Of course, this would require a more complete model of the trial, especially the generation of evidence and jury decision-making.

Another extension of interest is to allow the jury (in a two-outcome system) to award damages to an acquitted defendant against whom the evidence appeared to be especially weak. Would such a scheme achieve the informational advantage of the three-outcome verdict in a two-outcome system? Juries compute compensation in civil cases all the time (including complex allocations of losses between a plaintiff and a defendant). Alternatively, this particular form of compensation could be set by statute and awarded to those the jury finds are (strongly) not guilty. Moreover, if the size of such an award affects the prosecutor's career

prospects (for example, via internal reviews or public disclosure before elections), this may further incentivize her to screen defendants more carefully.

Finally, this model could be used as a foundation on which to develop a more comprehensive model that incorporates a potential offender's decision to commit a crime, the allocation of resources to law enforcement,[41] the statutory determination of formal sanctions, and an overall welfare analysis. In particular, although a potential offender might view formal and informal sanctions as perfect substitutes, a social welfare-maximizing planner may not view them this way. In particular, our outside observers are not welfare-maximizers; rather, they are self-interested agents who decline to interact with some defendants without concern about any costs they thereby externalize. But, in the aggregate, unwillingness to interact with some defendants will generate social costs (through distortions in labor and rental markets, for instance). In addition to these efficiency costs, the use of informal sanctions is a sort of vigilante justice that may be discounted in the social welfare function. The development of such a comprehensive model is a very interesting research question, but it must be postponed to future research.

[41] See Reinganum (1993) for an example of a model with criminals choosing whether to commit a crime, police potentially detecting a crime committed, and plea bargaining with a prosecutor if arrested.

References

Associated Press, "Day in Jail for Ex-Duke Prosecutor," *New York Times*, September 1, 2007. Online at http://www.nytimes.com/2007/09/01/us/01nifong.html; accessed on May 20, 2014.

Baker, Scott, and Mezzetti, Claudio, "Prosecutorial Resources, Plea Bargaining, and the Decision to Go to Trial," *Journal of Law, Economics, and Organization*, 17(1), 2001, pp. 149-167.

Bandoyopadhyay, Siddhartha, and McCannon, Bryan C., "The Effect of the Election of Prosecutors on Criminal Trials," *Public Choice*, forthcoming 2014.

Becker, Gary S., "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76(2), 1968, pp. 169-217.

Benabou, Roland, and Tirole, Jean, "Incentives and Prosocial Behavior," *American Economic Review* 96 (2006), 1652-1678.

Bjerk, David, "Guilt Shall not Escape of Innocence Suffer: The Limits of Plea Bargaining When Defendant Guilt is Uncertain," *American Law and Economics Review*, 9(2), 2007, pp. 305-329.

Boylan, Richard T., and Long, Cheryl X., "Salaries, Plea Rates, and the Career Objectives of Federal Prosecutors," *Journal of Law and Economics*, 48(2), 2005, pp. 627-652.

Boylan, Richard T., "What do Prosecutors Maximize? Evidence from the Careers of U.S. Attorneys," *American Law and Economics Review*, 7(2), 2005, pp. 379-402.

Bray, Samuel, "Not Proven: Introducing a Third Verdict," *The University of Chicago Law Review*, 72(4), 2005, pp.1299-1329.

Burke, Mary Kate; Hopper, Jessica; Francis, Enjoli; and Effron, Lauren, "Casey Anthony Juror: 'Sick to Our Stomachs' Over Not Guilty Verdict," ABC News, July 6, 2011. Online at: http://abcnews.go.com/US/casey_anthony_trial/casey-anthony-juror-jury-sick-stomach-guilty-verdict/story?id=14005609

Daughety, Andrew F., and Reinganum, Jennifer F., "Public Goods, Social Pressure, and the Choice Between Privacy and Publicity," *American Economic Journal: Microeconomics* 2 (2010), 191-221.

Deffains, Bruno, and Fluet, Claude, "Legal Liability when Individuals Have Moral Concerns," *Journal of Law, Economics, and Organization* 29 (2013), 930-955.

Duff, Peter, "The Scottish Criminal Jury: A Very Peculiar Institution," *Law and Contemporary Problems*, 62(2), 1999, pp. 173-201.

Franzoni, Luigi A., "Negotiated Enforcement and Credible Deterrence," *The Economic Journal*, 109, 1999, pp. 509-535.

Franzoni, Luigi A., "Liability Law and Uncertainty Spreading," Working Paper, Department of Economics, University of Bologna, February, 2014.

Gilboa, Itzhak, and Schmeidler, David, "Maxmin Expected Utility with Non-unique Prior," *Journal of Mathematical Economics*, 18(2), 1989, pp. 141-153.

Glaeser, Edward L., Kessler, Daniel P., and Piehl, Anne Morrison, "What do Prosecutors Maximize? An Analysis of the Federalization of Drug Crimes," *American Law and Economics Review*, 2(2), 2000, pp. 259-290.

Gordon, Sanford C., and Huber, Gregory A., "Citizen Oversight and the Electoral Incentives of Criminal Prosecutors, *American Journal of Political Science*, 46(2), 2002, pp. 334-351.

Grossman, Gene M., and Katz, Michael L., "Plea Bargaining and Social Welfare, *The American Economic Review*, 73(4), 1983, pp. 749-757.

Klibanoff, Peter, Marinacci, Massimo, and Mukerji, Sujoy, "A Smooth Model of Decision Making under Ambiguity," *Econometrica*, 73(6), 2005, pp. 1849-1892.

Landes, William M., "An Economic Analysis of the Courts," *Journal of Law and Economics*, 14(1), 1971, pp. 61-108.

Leipold, Andrew D., "The Problem of the Innocent, Acquitted Defendant," *Northwestern University Law Review*, 94(4), 2000, 1297-1356.

Machina, Mark J., and Siniscalchi, Marciano, "Ambiguity and Ambiguity Aversion," Chapter 13 in the Handbook of the Economics of Risk and Uncertainty, Vol. 1, Edited by Mark J. Machina and W. Kip Viscusi, published by Elsevier North-Holland, Amsterdam, 2014.

McCannon, Bryan C., "Prosecutor Elections, Mistakes, and Appeals," *Journal of Empirical Legal Studies*, 10(4), 2013, pp. 696-714.

Müller, Alfred, and Stoyan, Dietrich, *Comparison Methods for Stochastic Models and Risks*, John Wiley and Sons, England, 2002.

Nalebuff, Barry, "Credible Pretrial Negotiation, *The RAND Journal of Economics*, 18(2), pp. 198-210.

NC State Bar v. Michael B. Nifong, Disciplinary Hearing Commission, NC State Bar, 06 DHC 35. Online at http://www.ncbar.com/discipline/printorder.asp?id=505

Reinganum, Jennifer F., "Plea Bargaining and Prosecutorial Discretion," *The American Economic Review*, 78(4), 1988, pp. 713-728.

Reinganum, Jennifer F., "The Law Enforcement Process and Criminal Choice," *International Review of Law and Economics,* 13(2), 1993, pp. 115-134.

Smithson, Michael; Deady, Sara; and Gracik, Lavinia, "Guilty, Not Guilty, or ...? Multiple Options in Jury Verdict Choices," *Journal of Behavioral Decision Making*, 20(5), 2007, pp. 481-498.

Segal, David, "Mugged by a Mug Shot Online," *New York Times*, October 5, 2013. Online at http://www.nytimes.com/2013/10/06/business/mugged-by-a-mug-shot-online.html?pagew; accessed on May 20, 2014.

Appendix

Outside Observer Posterior Beliefs as to D's Guilt

Technically, Θ has a conjecture about S_b as well as about D's strategies, but it is not needed for the beliefs and we suppress this to avoid further clutter. Formally, the mathematical descriptions of Θ's beliefs presume that the strategy profile is fully-mixed, so that all nodes in the game are visited with positive probability, allowing us to use Bayes' Rule to provide the indicated formula. As we will see, (1, 1, 1) is an equilibrium of the game, so that in this equilibrium, the outcomes b and d are out-of-equilibrium outcomes, and the value for $\mu(G \mid b)$ and $\mu(G \mid d)$ will need to be otherwise specified, since b and d will not be visited in equilibrium. Moreover, P's strategy, ρ^P, does not affect the beliefs because it (or $1 - \rho^P$) multiplies each relevant numerator and denominator and thereby drops out of the analysis.

$$\mu(G \mid a) = \rho_G^{D\Theta}(1 - \lambda)F_G/[\rho_G^{D\Theta}(1 - \lambda)F_G + \rho_I^{D\Theta}\lambda F_I]; \quad \text{(A.1a)}$$

$$\mu(G \mid b) = (1 - \rho_G^{D\Theta})(1 - \lambda)/[(1 - \rho_G^{D\Theta})(1 - \lambda) + (1 - \rho_I^{D\Theta})\lambda]; \quad \text{(A.1b)}$$

$$\mu(G \mid c) = \rho_G^{D\Theta}(1 - \lambda)(1 - F_G)/[\rho_G^{D\Theta}(1 - \lambda)(1 - F_G) + \rho_I^{D\Theta}\lambda(1 - F_I)]; \quad \text{(A.1c)}$$

and $$\mu(G \mid d) = \rho_G^{D\Theta}(1 - \lambda)/[\rho_G^{D\Theta}(1 - \lambda) + \rho_I^{D\Theta}\lambda]. \quad \text{(A.1d)}$$

Characterizing Equilibria

The only candidates for an equilibrium involve $\rho_I^D = 1$ (I-types always reject the plea offer) and $\rho_G^D \in (0, 1]$ (G-types reject the plea offer with positive probability); see the Technical Appendix wherein other candidates are ruled out. P may also mix between taking the case to trial and dropping it following rejection.

The timing of the game is such that each type of D chooses to accept or reject the plea offer, taking as given the likelihood that P takes the case to trial following rejection; and P chooses to take the case to trial or drop it, given her beliefs about the posterior probability that D is of type G, given rejection. Both of these decisions are taken following P's choice of plea offer, S_b, so both parties must take this offer as given at subsequent decision nodes.

We first characterize the equilibrium in the continuation game, given S_b, allowing for mixed strategies for both P (ρ^P) and the D of type G (ρ_G^D). Since the observers' beliefs will depend on their conjectured value for ρ_G^D, we will augment the notation for the observers' beliefs to reflect these conjectures, $\rho_G^{D\Theta}$. Other functions that also depend on these conjectures through the observers' beliefs will be similarly augmented.

Suppose that observers conjecture that the D of type G rejects the plea offer with probability $\rho_G^{D\Theta}$. Then $\mu(G \mid c; \rho_G^{D\Theta}) = \rho_G^{D\Theta}(1 - \lambda)(1 - F_G)/[\rho_G^{D\Theta}(1 - \lambda)(1 - F_G) + \lambda(1 - F_I)]$; $\mu(G \mid a; \rho_G^{D\Theta}) = \rho_G^{D\Theta}(1 - \lambda)F_G/[\rho_G^{D\Theta}(1 - \lambda)F_G + \lambda F_I]$; $\mu(G \mid d; \rho_G^{D\Theta}) = \rho_G^{D\Theta}(1 - \lambda)/[\rho_G^{D\Theta}(1 - \lambda) + \lambda]$; and $\mu(G \mid b; \rho_G^{D\Theta}) = 1$. Moreover, suppose that the D of type G anticipates these beliefs, and also expects that P will take the case to trial following rejection with probability ρ^P. Then type G will be indifferent, and hence willing to mix, between accepting and rejecting the offer S_b, if $\pi_R^D(G; \rho_G^{D\Theta}) = \rho^P\pi_T^D(G; \rho_G^{D\Theta}) + (1 - \rho^P)\pi_d^D(\rho_G^{D\Theta}) = \pi_b^D(\rho_G^{D\Theta})$. Substitution and simplification yields the value of ρ^P that renders G indifferent:

$$\rho^P(S_b; \rho_G^{D\Theta}) = \frac{\{S_b + r^D(1 - \mu(G \mid d; \rho_G^{D\Theta}))\}}{\{S_c(1 - F_G) + k^D + r^D[\mu(G \mid c; \rho_G^{D\Theta})(1 - F_G) + \mu(G \mid a; \rho_G^{D\Theta})F_G - \mu(G \mid d; \rho_G^{D\Theta})]\}}.$$

The numerator of the expression $\rho^P(S_b; \rho_G^{D\Theta})$, which is the difference between type G's payoff from accepting the plea offer versus having his case dropped, is clearly positive, meaning that D would prefer to have his case dropped than to accept a plea offer. The denominator of the expression $\rho^P(S_b; \rho_G^{D\Theta})$ is the difference between type G's payoff from trial versus having his case dropped. This denominator is also positive (see Remark 3 in the Technical Appendix), which implies that type G would prefer that P drop the

case against him rather than take it to trial.

Since the observers' beliefs are based on their conjectures $\rho_G^{D\Theta}$ and the case disposition, and NOT on S_b, which they do not observe, the expression $\rho^P(S_b; \rho_G^{D\Theta})$ is an increasing function of S_b. That is, when S_b is higher, P must take the case to trial following rejection with a higher probability in order to make the D of type G indifferent about accepting or rejecting S_b. Notice that even a plea offer of $S_b = 0$ requires a positive probability of trial following a rejection in order to induce the D of type G to be willing to accept it; this is because acceptance of a plea offer comes with a sure informal sanction of r^D (as only a truly guilty D is expected to accept the plea).

Now consider P's decision about trying versus dropping the case. Again suppose that observers – and P – both conjecture that type G rejects the plea offer with probability $\rho_G^{D\Theta}$ in this candidate for equilibrium; thus $\nu(G \mid R; \rho_G^{D\Theta}) = \rho_G^{D\Theta}(1 - \lambda)/[\rho_G^{D\Theta}(1 - \lambda) + \lambda]$. Since these conjectures must be the same (and correct) in equilibrium, it is valid to equate them at this point in order to identify what common beliefs for P and Θ will make P indifferent, and hence willing to mix, between trying and dropping the case following a rejection. P will be indifferent between these two options if $\pi_T^P(\rho_G^{D\Theta}) = \pi_d^P(\rho_G^{D\Theta})$; that is, if:

$$\nu(G \mid R; \rho_G^{D\Theta})\{S_c(1 - F_G) - k^P - r_I^P\mu(I \mid c; \rho_G^{D\Theta})(1 - F_G) - r_G^P\mu(G \mid a; \rho_G^{D\Theta})F_G\}$$

$$+ \nu(I \mid R; \rho_G^{D\Theta})\{S_c(1 - F_I) - k^P - r_I^P\mu(I \mid c; \rho_G^{D\Theta})(1 - F_I) - r_G^P\mu(G \mid a; \rho_G^{D\Theta})F_I\} = - r_G^P\mu(G \mid d; \rho_G^{D\Theta}).$$

Substituting for the beliefs and solving for the value of $\rho_G^{D\Theta}$ that generates this equality (see the Technical Appendix for details) yields:

$$\rho_G^{D0} = - \lambda[(S_c - r_I^P)(1 - F_I) - k^P]/(1 - \lambda)[(S_c + r_G^P)(1 - F_G) - k^P],$$

where the numerator is positive by MR1(a); MR2 implies that the denominator is positive and the ratio is a fraction. For any $\rho_G^{D\Theta} > \rho_G^{D0}$, P will strictly prefer to take the case to trial following a rejection, and for any $\rho_G^{D\Theta} < \rho_G^{D0}$, P will strictly prefer to drop the case following a rejection.

To summarize, type G is willing to mix between accepting and rejecting the plea offer S_b if he anticipates that the observers' beliefs are $\rho_G^{D\Theta} = \rho_G^{D0}$ and he expects that P will take the case to trial following rejection of offer S_b with probability $\rho^P(S_b; \rho_G^{D0})$. P is indifferent between trying and dropping the case if she and the observers believe that type G rejects the plea offer with probability ρ_G^{D0}. Thus, the mixed-strategy continuation equilibrium, given S_b, is $(\rho_G^{D0}, \rho^P(S_b; \rho_G^{D0}))$.

We can now move back to the decision node at which P chooses the plea offer S_b, anticipating that it will be following by the mixed-strategy equilibrium $(\rho_G^{D0}, \rho^P(S_b; \rho_G^{D0}))$ in the continuation game. P's payoff from making the plea offer S_b is:

$$(1 - \rho_G^{D0})(1 - \lambda)S_b + (\rho_G^{D0}(1 - \lambda) + \lambda)[\rho^P(S_b; \rho_G^{D0})\pi_T^P(\rho_G^{D0}) + (1 - \rho^P(S_b; \rho_G^{D0}))\pi_d^P(\rho_G^{D0})].$$

The set of S_b values that support some plea acceptance is bounded below by 0 and above by $S_b = \pi_T^D(G; \rho_G^{D0}) - r^D$, where $\pi_T^D(G; \rho_G^{D0})$ is the expression for $\pi_T^D(G)$, evaluated at the beliefs $\mu(G \mid c; \rho_G^{D0}) = \rho_G^{D0}(1 - \lambda)(1 - F_G) /[\rho_G^{D0}(1 - \lambda)(1 - F_G) + \lambda(1 - F_I)]$; and $\mu(G \mid a; \rho_G^{D0}) = \rho_G^{D0}(1 - \lambda)F_G/[\rho_G^{D0}(1 - \lambda)F_G + \lambda F_I]$. This is because accepting the plea offer results in a combined sanction of $S_b + r^D$ (since only guilty D's accept the plea offer) and thus any plea offer higher than $\pi_T^D(G; \rho_G^{D0}) - r^D$ will be rejected for sure (rather than with probability ρ_G^{D0}). At this upper bound, the function $\rho^P(S_b; \rho_G^{D0})$ just reaches 1. In order for this range to be non-empty, we need $\pi_T^D(G; \rho_G^{D0}) - r^D \geq 0$; or, equivalently (note that the denominator of the expression below is positive):

Condition 1. In order for P to be able to induce a D of type G to accept a plea offer, it must be that:

$$r^D \leq [S_c(1 - F_G) + k^D]/[1 - \mu(G \mid c; \rho_G^{D0})(1 - F_G) - \mu(G \mid a; \rho_G^{D0})F_G].$$

Returning to P's payoff as a function of S_b, notice two things. First, since ρ_G^{D0}, which is independent of S_b, renders P indifferent between trying and dropping the case following rejection, the term in square brackets simply equals $\pi_d^P(\rho_G^{D0}) = - r_G^P\mu(G \mid d; \rho_G^{D0})$, where $\mu(G \mid d; \rho_G^{D0}) = \rho_G^{D0}(1 - \lambda)/[\rho_G^{D0}(1 - \lambda) + \lambda]$. Thus, the optimal S_b that supports some plea acceptance is $S_b(\rho_G^{D0}) = \pi_T^D(G; \rho_G^{D0}) - r^D$. This offer is rejected by type G with probability ρ_G^{D0}, and P goes to trial with certainty following a rejection. Note that a D of type I would always reject this plea offer, consistent with the hypothesized form of the equilibrium.

Every plea offer in the feasible set $[0, \pi_T^D(G; \rho_G^{D0}) - r^D]$ is consistent with a mixed-strategy equilibrium in which some D's of type G accept, and others reject, the offer. But – taking the observers' beliefs of ρ_G^{D0} as given – P could make a higher demand that would provoke certain rejection by both D-types. We need to verify that P prefers the hypothesized equilibrium described above to the "defection payoff" she would obtain if all cases went to trial.

In the hypothesized equilibrium, P settles with $(1 - \rho_G^{D0})(1 - \lambda)$ guilty defendants and goes to trial against the rest of the guilty defendants and all of the innocent defendants; if P defects and provokes rejection by all, then she will simply replace the plea offer $S_b(\rho_G^{D0}) = \pi_T^D(G; \rho_G^{D0}) - r^D$ with the expected payoff from taking a guilty defendant to trial (the observers' beliefs are fixed at ρ_G^{D0} because trial is on the equilibrium path). Thus, P prefers (at least weakly) the hypothesized equilibrium to defection as long as:

$$\pi_T^D(G; \rho_G^{D0}) - r^D = S_c(1 - F_G) + k^D + r^D\mu(G \mid c; \rho_G^{D0})(1 - F_G) + r^D\mu(G \mid a; \rho_G^{D0})F_G - r^D$$

$$\geq S_c(1 - F_G) - k^P - r_I^P\mu(I \mid c; \rho_G^{D0})(1 - F_G) - r_G^P\mu(G \mid a; \rho_G^{D0})F_G.$$

Rearranging, we can write this as:

Condition 2. For P to find it preferable to settle with a D of type G rather than provoking trial, it must be that:

$$r^D \leq [k^P + k^D + r_I^P\mu(I \mid c; \rho_G^{D0})(1 - F_G) + r_G^P\mu(G \mid a; \rho_G^{D0})F_G]/[1 - \mu(G \mid c; \rho_G^{D0})(1 - F_G) - \mu(G \mid a; \rho_G^{D0})F_G].$$

Multiple Equilibria

There can also be equilibria wherein the D of type G rejects the plea bargain with probability $\rho_G^D > \rho_G^{D0}$. The reason for this multiplicity of equilibria is that the observers' beliefs actually drive the equilibrium behavior of P and the D of type G. To see this, suppose that $\rho_G^{D\Theta} = \rho_G^{D1} > \rho_G^{D0}$, and that P and type G anticipate these beliefs. Since the observers' posterior beliefs $\mu(G \mid c; \rho_G^{D\Theta})$ and $\mu(G \mid a; \rho_G^{D\Theta})$ are increasing in $\rho_G^{D\Theta}$, both P and the D of type G expect that D will face harsher informal sanctions following either trial outcome (than they would face at ρ_G^{D0}).

This means that P can demand the higher plea sentence, $S_b(\rho_G^{D1}) = \pi_T^D(G; \rho_G^{D1}) - r^D$, which will make the D of type G indifferent about accepting and rejecting it (if he expects P to take the case to trial following a rejection); thus, type G is willing to randomize and reject the plea bargain with probability equal to ρ_G^{D1}. Since $\rho_G^{D1} > \rho_G^{D0}$, P will take the case to trial with probability 1 following a rejection (if he thinks that the D of type G is using the rejection probability ρ_G^{D1}). Thus, there is an equilibrium at any $\rho_G^D \in [\rho_G^{D0}, 1)$ as long as Conditions 1 and 2 continue to hold at that ρ_G^D.

Selecting Among Equilibria

The following terms summarize erroneously-imposed informal sanctions. Excessive sanctions imposed on type I defendants following conviction, or acquittal, at trial (ideally, there would be no sanctions):

$$\lambda(1 - F_I)r^D\mu(G \mid c) + \lambda F_I r^D\mu(G \mid a).$$

Insufficient sanctions imposed on type G defendants following conviction, or acquittal, at trial (ideally, the sanction would be r^D):

$$\rho_G^D(1 - \lambda)(1 - F_G)[r^D - r^D\mu(G \mid c)] + \rho_G^D(1 - \lambda)F_G[r^D - r^D\mu(G \mid a)].$$

Prosecutors also suffer erroneously-imposed sanctions. With respect to innocent defendants:

$$\lambda(1 - F_I)[r_I^P - r_I^P\mu(I \mid c)] + \lambda F_I[r_G^P\mu(G \mid a)].$$

Note that the first term reflects the fact that P convicted a type I and ideally would have received the sanction r_I^P, but she only received $r_I^P\mu(I \mid c)$; the second term reflects the fact that an innocent was acquitted, but P was still sanctioned because observers' beliefs admit some probability that the acquittal was an error, for which P is sanctioned (undeservedly).

With respect to guilty defendants:

$$\rho_G^D(1 - \lambda)(1 - F_G)[r_I^P\mu(I \mid c)] + \rho_G^D(1 - \lambda)F_G[r_G^P - r_G^P\mu(G \mid a)].$$

The first term reflects the fact that P convicted a guilty D, but was still sanctioned because observers' beliefs admit some probability that the conviction was an error, for which P is sanctioned (undeservedly); the second term reflects the fact that a guilty D was acquitted, so P ideally would have received the sanction r_G^P but only received $r_G^P\mu(G \mid a)$.

Let $M(\rho_G^D)$ denote the measure of expected loss from misclassification that observers experience due to these erroneous sanctions. Then (all of the observers' beliefs are evaluated at ρ_G^D):

$$\begin{aligned} M(\rho_G^D) = {} & \lambda(1 - F_I)r^D\mu(G \mid c) + \lambda F_I r^D\mu(G \mid a) + \rho_G^D(1 - \lambda)(1 - F_G)[r^D - r^D\mu(G \mid c)] \\ & + \rho_G^D(1 - \lambda)F_G[r^D - r^D\mu(G \mid a)] + \lambda(1 - F_I)[r_I^P - r_I^P\mu(I \mid c)] + \lambda F_I[r_G^P\mu(G \mid a)] \\ & + \rho_G^D(1 - \lambda)(1 - F_G)[r_I^P\mu(I \mid c)] + \rho_G^D(1 - \lambda)F_G[r_G^P - r_G^P\mu(G \mid a)]. \end{aligned}$$

Collecting terms simplifies the above expression greatly:

$$\begin{aligned} M(\rho_G^D) = {} & (r^D + r_I^P)\{\lambda(1 - F_I)\mu(G \mid c) + \rho_G^D(1 - \lambda)(1 - F_G)\mu(I \mid c)\} \\ & + (r^D + r_G^P)\{\lambda F_I\mu(G \mid a) + \rho_G^D(1 - \lambda)F_G\mu(I \mid a)\}. \end{aligned}$$

Upon recalling the definitions of $\mu(t \mid y)$, and evaluating them at ρ_G^D), it is straightforward (though tedious) to show that both of the terms in curly brackets in the expression $M(\rho_G^D)$ are increasing in ρ_G^D.