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

Andrew F. Daughety

Vanderbilt University


Jennifer F. Reinganum

Vanderbilt University

jennifer.f.reinganum@vanderbilt.edu

## Abstract

We model the strategic interaction between a prosecutor and a defendant when non-strategic outside observers rationally use the case disposition (plea bargain, dropped case, acquittal, or conviction) to impose informal sanctions on both parties. Outside observers recognize that error in the legal process (as well as hidden information) means they may misclassify defendants and thereby impose sanctions erroneously. We show that: 1) changes in the level of the formal sanction affect the level of informal sanctions imposed; and 2) increases in the informal sanction rates imposed on prosecutors result in changes in the level of informal sanctions imposed on defendants. We also extend the model to allow for a three-outcome verdict (not guilty, not proven, and guilty), sometimes referred to as the "Scottish" verdict. We find that the Scottish verdict is justice-improving in that it benefits innocent defendants, outside observers and prosecutors in comparison with the standard (two-outcome) verdict.

1. INTRODUCTION

All of us are familiar 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 limited aspects 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 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; 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" when cases against defendants believed to be guilty are dismissed, or trials are lost. Other prosecutors may be viewed as overzealous (and possibly abusing their position) when cases against defendants believed to be innocent are resolved by plea bargain or are pursued to trial and won. 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.[4]

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 influence (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 guilt.[6] 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 rent to, or interact socially with, someone else. As suggested earlier, we also assume that 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).

We develop a model that incorporates informal sanctions and obtain several results. In Section 3, we show that an innocent defendant rejects the plea offer,[7] whereas a guilty defendant sometimes accepts and sometimes rejects the plea offer. Because the prosecutor has the option to drop the case, there must be a sufficiently high likelihood that a guilty defendant rejects the plea offer in order to incentivize the prosecutor to go to trial following a rejection. That is, following Nalebuff (1987), who first addressed this issue in the context of a lawsuit, we require that the prosecutor's threat of trial be *credible*. Since only the defendant knows whether he is guilty or innocent, and since the trial process is not perfect at distinguishing between the guilty and the innocent, outside observers will impose excessive informal sanctions on an innocent defendant and

insufficient informal sanctions on a guilty defendant. Similarly, informal sanctions on the prosecutor may also be imposed erroneously. We construct a measure of the expected loss from misclassification that reflects the outside observers' assessments of D's guilt and show that this measure is increasing in the guilty defendant's plea rejection rate.

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 a plea offer that at least some defendants accept, because accepting the plea offer 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. The equilibrium also yields a number of interesting implications, among them that: 1) an increase in the formal sanction upon conviction increases (decreases) the equilibrium expected informal sanctions on truly guilty (truly innocent) 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 of the defendants (in attitudes towards risk and/or ambiguity) that results in some innocent defendants accepting plea offers in equilibrium. Such heterogeneity can increase both 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 higher expected costs to a truly guilty defendant, lower expected costs to a truly innocent defendant, and lower expected loss due to misclassification by the observers; altogether, these results suggest that the Scottish verdict is likely to be justice-enhancing relative to the standard two-outcome (convict/acquit) verdict.

## 1.1. Related Literature

Landes (1971) provides a complete-information model wherein the prosecutor maximizes 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 proceed to trial against a defendant that (she now believes) is innocent. Franzoni (1999) and Baker and Mezzetti (2001) explicitly incorporate Nalebuff's credibility constraint into a screening model, requiring that a sufficiently high fraction of guilty defendants reject the plea offer.[8] As indicated earlier, we also incorporate such a 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 jury makes its decision based only on the evidence it observes in the course of the trial and the specified standard of proof. That is, it follows instructions rather than acting as a rational Bayesian agent.[10] Bjerk (2007) considers 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; anticipating this, the guilty defendant 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.

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, 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 (2014) find that prosecutors subject to reelection pressure try to increase the number of convictions at trial (including taking more weak cases to trial).[11] 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 concerns about convicting the innocent or not convicting the guilty.[12] 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.

1.2. Plan of the Paper

In Section 2, we provide the notation and the formal model. In Section 3, we describe the equilibrium of the model (the equilibrium concept will be Perfect Bayesian equilibrium). 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 in Section 3 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[13] contains other details of the analysis.

2. MODELING PRELIMINARIES

2.1. 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 ($\gamma_c$), 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 (which is omitted when ownership is obvious). 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 ($\pi_b^P$ and $\pi_b^D$) 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 payoffs $\pi_d^P$ and $\pi_d^D$.

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 $\pi_c^P$ and $\pi_c^D$, while acquittal yields payoffs $\pi_a^P$ and $\pi_a^D$.

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).

<<COMP: Place Fig. 1 about here>>

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.[14] Let $\lambda > 0$ denote the likelihood that $D$ is innocent; that is, $\lambda = \Pr\{t = I\}$ is 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 ($\gamma_c$) and to otherwise acquit the defendant. We denote the distribution of evidence (given $D$'s type) as $F(e \mid t)$, which we assume is continuous in $e$. Since for any evidentiary standard for conviction, $\gamma_c$, 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 $\Theta$ denote the outside observer(s).[15]  Informal sanctions are based on $\Theta$'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.[16]  More precisely, these beliefs represent $\Theta$'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 $\Theta$ cannot directly observe the plea offer, $S_b$, the levels of $P$'s and $D$'s payoffs, or the evidence draw $e$.[17]

Informal sanctions on $D$ are $r^D\mu(G \mid y)$, where $r^D \geq 0$ is an exogenous parameter.[18]  That is, given any case disposition, $\Theta$ 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 observers 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 transactions.

As indicated earlier, $\Theta$ also imposes 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 held accountable 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 $r^P_I \mu(I \mid y)$, for $y \in \{b, c\}$. We assume that $r^P_I$ and $r^P_G$ are non-negative.[19]

2.2. *D*'s Payoffs

We are interested in a non-cooperative solution for the game that exhibits 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:

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

and

$$\pi^D_a = k^D + r^D \mu(G \mid a). \tag{2}$$

That is, going to trial costs *D* the amount $k^D$.[20] Conviction results in the formal sanction $S_c$ plus the informal sanction $r^D \mu(G \mid c)$; since an innocent *D* may have been convicted (the evidence draw for an innocent *D* 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 $r^D \mu(G \mid a)$.

We can write *D*'s expected loss from going to trial (given his type *t*) as the weighted combination of the elements in equations (1) and (2), where the weights reflect the likely outcome at trial:

$$\pi^D_T(t) = S_c(1 - F_t) + k^D + r^D \mu(G \mid c)(1 - F_t) + r^D \mu(G \mid a)F_t, \ t \in \{I, G\}. \tag{3}$$

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 $\Theta$ 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:

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

That is, an innocent $D$ has a lower expected loss from trial than does a guilty $D$. This follows from the assumption that an innocent $D$ is less likely to be convicted than a guilty $D$ ($F_I > F_G$), which implies that the posterior belief by outsiders that a $D$ is truly guilty 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 a guilty $D$ accepts and an innocent $D$ rejects), or partial screening (wherein, say, $P$ makes an offer that an innocent $D$ rejects and that a guilty $D$ rejects with fractional probability); 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 equilibrium of the game.

$\quad\quad$ 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:

$$\pi_b^D = S_b + r^D\mu(G \mid b);\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (4)$$

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:

$$\pi_d^D = r^D \mu(G \mid d), \tag{5}$$

which reflects $\Theta$'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 $\rho^P$ 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 (3) and (5), weighted by the probability that *P* takes the case to trial, yields *D*'s expected payoff following rejection (given his type) as:

$$\pi_R^D(t) = \rho^P \pi_T^D(t) + (1 - \rho^P)\pi_d^D. \tag{6}$$

## 2.3. *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:

$$\pi_c^P = S_c - k^P - r_I^P \mu(I \mid c); \tag{7}$$

and

$$\pi_a^P = - (k^P + r_G^P \mu(G \mid a)). \tag{8}$$

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 $\Theta$ 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, $\lambda$), 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 $v(G \mid R)$ (resp., $v(G \mid A)$) denote the prosecutor's posterior probability that the defendant is guilty,[21] given that he rejected (resp., accepted) the plea offer $S_b$. Of course, in equilibrium, $P$'s beliefs and $\Theta$'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:

$$\pi_T^P = v(G \mid R)\{S_c(1 - F_G) - k^P - r_I^P\mu(I \mid c)(1 - F_G) - r_G^P\mu(G \mid a)F_G\}$$
$$+ v(I \mid R)\{S_c(1 - F_I) - k^P - r_I^P\mu(I \mid c)(1 - F_I) - r_G^P\mu(G \mid a)F_I\}. \qquad (9)$$

This is interpreted as follows. Given rejection of the plea offer, $P$ believes that $D$ is guilty with probability $v(G \mid R)$, 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 observers 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 $v(I \mid R)$, is interpreted similarly.

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

$$\pi_b^P = S_b - r_I^P\mu(I \mid b). \qquad (10)$$

Equation (10) 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 $\Theta$ that reflects $\Theta$'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 $\Theta$, who believes with probability $\mu(G \mid d)$ that $D$ is guilty, so by dropping the case $P$ may have let a guilty defendant go free. Thus, $P$'s payoff from dropping the case is simply:

$$\pi_d^P = - r_G^P \mu(G \mid d). \tag{11}$$

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:

$$\pi_R^P = \rho^P \pi_T^P + (1 - \rho^P)\pi_d^P. \tag{12}$$

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 $\Theta$'s conjectures about these strategies, and some natural restrictions on the parameter space that reflect sensible behavior by $P$. We then describe the game's 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 ($\rho^P$) 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 $\rho_G^D$ and $\rho_I^D$. In equilibrium, each strategy will take on numerical values, so we can summarize a hypothesized equilibrium by the four-tuple ($S_b, \rho_G^D, \rho_I^D, \rho^P$) $\in [0,\infty) \times [0,1] \times [0, 1] \times [0, 1]$.

Notice that $P$ has conjectures about which types of $D$ will choose to reject the offer. $D$ has conjectures about what $P$ will do, conditional on $D$'s action and, while $\Theta$ cannot observe $P$'s and $D$'s actions, $\Theta$ 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.

Suppressing the plea offer for now, let $(\rho_G^{D\Theta}, \rho_I^{D\Theta}, \rho^{P\Theta})$ denote $\Theta$'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, $\Theta$'s beliefs about $D$ are provided in the Appendix as equations (A1) through (A4).[22] We employ Perfect Bayesian equilibrium and require that: 1) $P$ maximizes her expected payoff by choosing her plea offer $S_b$, given $\Theta$'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 $\Theta$'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 $\Theta$'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.

Below we provide parametric restrictions that reflect reasonable preferences for $P$. Consider the terms in braces in equation (9) above, evaluated with all informal sanction rates set to zero. Since $F_G < F_I$, then it is straightforward that the resulting

payoff to $P$ from taking a guilty $D$ to trial is higher than the payoff she obtains from taking an innocent $D$ to trial: $S_c(1 - F_G) - k^P > S_c(1 - F_I) - k^P$. We extend this preference to the case of non-zero levels of the informal sanction rates via the following restriction on the parameter space.

> *Maintained Restriction 0* (hereafter, MR0): $P$ strictly prefers to go to trial against a $D$ she believes to be guilty in comparison with one she believes to be innocent. Formally, this reduces to assuming that (given $\rho_G^{D\Theta}$, $\rho_I^{D\Theta}$, and the corresponding associated beliefs for $\Theta$): $S_c - r_I^P \mu(I \mid c) + r_G^P \mu(G \mid a) > 0$. Thus, a sufficient condition for MR0 is $S_c - r_I^P > 0$.

There are two scenarios concerning the decision by $P$ whether to drop the case or go to trial that also restrict the parameter space. First, if $P$ (and $\Theta$) know (or commonly believe) that $D$ is innocent, then $P$ should prefer dropping the case to going to trial.

> *Maintained Restriction 1* (hereafter, MR1): If $P$ and $\Theta$ know (or commonly believe) that $D$ is innocent, $P$ strictly prefers to drop the case. Formally, this reduces to: $(S_c - r_I^P)(1 - F_I) - k^P < 0$.

Intuitively, MR1 will hold if either $k^P$ or $r_I^P$ is sufficiently large, or if both together are sufficiently large.

Second, imagine that both types of $D$ were expected to reject $P$'s offer. Then $P$'s (and $\Theta$'s) posterior beliefs about the types would be the same as the prior beliefs (that is, $\Pr\{t = I\} = \lambda$); 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 is sufficiently effective at discriminating between guilty and

innocent persons; that is, if $\lambda$, the prior probability that $D$ is innocent, is not "too large."[23]

   *Maintained Restriction 2* (hereafter, MR2):   If $P$ and $\Theta$ know (or commonly believe) that the likelihood of a guilty $D$ 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, so the above condition is an upper bound on $\lambda$ (the arrest process is sufficiently effective).  Moreover, for MR2 to hold, it must be 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 guilty, is positive.

   Using these natural parameter restrictions, we find that the only[24] equilibrium of the game involves an innocent $D$ always rejecting the equilibrium plea offer, a guilty $D$ rejecting the equilibrium plea offer with a positive probability $\rho_G^D$, and $P$ never dropping a case when $D$ rejects an offer.[25]  To streamline the exposition, let $\pi_T^D(G; \rho_G^D)$ be $\pi_T^D(G)$, as specified in equation (3), with $\Theta$'s beliefs evaluated at arbitrary $\rho_G^D$ and at $\rho_I^D = 1$.[26] Furthermore, let $\rho_G^{D0}$ be the (unique) solution to the condition that $\pi_T^P = \pi_d^P$, where both of these expressions have $\Theta$'s beliefs and $P$'s beliefs evaluated at $\rho_G^{D0}$.  The value of $\rho_G^{D0}$ is given by:

$$\rho_G^{D0} = - \lambda[(S_c - r_I^P)(1 - F_I) - k^P]/(1 - \lambda)[(S_c + r_G^P)(1 - F_G) - k^P]. \tag{13}$$

which is easily shown to be a positive fraction under MR1 and MR2.  We provide a sketch of the proof of the following Proposition (as well as detail on the equilibrium plea

offer via equation (A6)) 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 Perfect Bayesian equilibrium for the game wherein $P$'s equilibrium plea offer is $S_b(\rho_G^{D0}) =$ $\pi_T^D(G; \rho_G^{D0}) - r^D$, the guilty $D$ rejects the offer with probability $\rho_G^{D0}$, the innocent $D$ always rejects the plea offer ($\rho_I^D = 1$), and $P$ always goes to trial following a rejection ($\rho^P = 1$).

We observe two properties of the equilibrium. First, both an innocent $D$ and $P$ use pure strategies in equilibrium: an innocent $D$ always rejects $P$'s equilibrium offer. Furthermore, the likelihood that it is a guilty $D$ who is rejecting the plea offer is sufficiently high that $P$ will never choose to drop the case. Second, from the perspective of $\Theta$, since an innocent $D$ always rejects the offer, this means that $\mu(G \mid b)$ is one: a $D$ who accepts the offer must be guilty and therefore incurs the full sanction $r^D$ from $\Theta$.

3.1. Limits on Informal Sanctions

What do we mean by the qualifier in Proposition 1 "If $r^D$ is not 'too large'"? Since $\rho_G^{D0} < 1$, then a guilty $D$ accepts $P$'s offer with positive (but fractional) 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 $S_b(\rho_G^D) = \pi_T^D(G; \rho_G^{D0}) - r^D$, so that the interval of feasible $S_b$-values capable of inducing some acceptance is $[0, \pi_T^D(G; \rho_G^{D0}) - r^D]$. Hence, in order to have a non-empty feasible set for $S_b$, it must be that $r^D \leq \pi_T^D(G; \rho_G^{D0})$. In equation (A5) in the Appendix, we provide the formal statement of this as Condition 1. 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, $S_b(\rho_G^{D0}) = \pi_T^D(G; \rho_G^{D0}) - r^D$ was less than what $P$ could expect by driving those $D$'s that would have otherwise settled to trial. In equation (A7) in the Appendix we characterize the restriction that eliminates this incentive for defection as Condition 2. Both of these Conditions provide positive upper bounds on $r^D$, so this is the sense in which $r^D$ is not "too large." Finally, $P$ could also defect by dropping all cases following rejection; thus, we need to verify that $P$ prefers the hypothesized equilibrium outcome to what she would get by dropping all cases; however, Condition 1 is sufficient to imply this preference.[27]

Since $P$ prefers to settle via plea bargain rather than going to trial against a guilty $D$, why can't $P$ offer a slightly lower plea offer than $S_b(\rho_G^{D0}) = \pi_T^D(G; \rho_G^{D0}) - r^D$ and induce a guilty $D$ to accept more frequently? The reason is that, since $P$ is not pre-committed to trial, she needs to maintain the credibility of her threat to go to trial following rejection, which requires a guilty $D$ to reject with at least probability $\rho_G^{D0}$.[28]

## 3.2. Outside Observers' Expected Loss from Misclassification

Notice that $\Theta$'s beliefs punish an innocent $D$ by lumping them in with a guilty $D$, since the two types can't be distinguished. For example, $r^D\mu(G \mid c)$ is the sanction for a convicted $D$, whether he is guilty or innocent; if $\Theta$ knew that $D$ was wrongly-convicted, then one would expect the sanction to be zero. $\Theta$ erroneously punishes an innocent $D$, based on observing a conviction, with probability $\lambda(1 - F_I)$, so the expected

misclassification 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 $\Theta$'s beliefs that the convicted $D$ might be innocent, in the amount of $r^P_I\mu(I \mid c)$. But if $\Theta$ knew that $D$ was wrongly-convicted, then the appropriate informal sanction for $P$ would be $r^P_I$. The outside observer's expected misclassification loss for this scenario is $\lambda(1 - F_I)[r^P_I - r^P_I\mu(I \mid c)]$.

In the Appendix we provide an overall expected loss from misclassification, denoted as $M(\rho^D_G)$, which is an increasing function of $\rho^D_G$. Evaluating the function (including all the beliefs) at $\rho^{D0}_G$ yields:

$$M(\rho^{D0}_G) = (r^D + r^P_I)\{\lambda(1 - F_I)\mu(G \mid c) + \rho^{D0}_G(1 - \lambda)(1 - F_G)\mu(I \mid c)\}$$

$$+ (r^D + r^P_G)\{\lambda F_I\mu(G \mid a) + \rho^{D0}_G(1 - \lambda)F_G\mu(I \mid a)\}. \tag{14}$$

We assume that outside observers, while not bearing any net costs for imposing informal sanctions, prefer a legal system with a lower expected loss from misclassification. If there were no plea bargaining, this would correspond to $M(\rho^D_G = 1)$. Thus the observers prefer that plea bargaining be possible. Moreover, it is straightforward, using equation (3) and the posterior beliefs specified in (A1) - (A4) in the Appendix, to show that $D$'s expected loss from trial is increasing in $\rho^D_G$, regardless of type. Since $\rho^{D0}_G < 1$, this means that both types of defendant prefer that plea bargaining be possible. This reflects the externality that even though an innocent $D$ rejects the plea offer and goes to trial, the fact that a guilty $D$ accepts the offer with positive probability 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. Notice that the fact that $D$'s loss is increasing in $\rho^D_G$ means that there is an alternative basis for outside observers to prefer a

legal system with a lower expected loss from misclassification: under a veil of ignorance, $\Theta$ recognizes that they might someday be a defendant in a criminal action. Finally, we observe that the foregoing is an informational argument supporting plea bargaining, separate from the standard cost-savings argument for plea bargaining.

A special subcase of our model involves eliminating the informal sanctions, so that $r^D = r_I^P = r_G^P = 0$. Again, a guilty $D$ mixes because $P$ is not committed to going to trial, so a guilty $D$ must reject the offer with a sufficiently large probability in order that $P$ not defect to dropping cases (due to MR1). In this case $\Theta$'s beliefs do not affect $D$ or $P$, and there is always a plea bargain that $P$ wants to make and that results in acceptance with positive probability by a guilty $D$. 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 ($\rho_G^{D0}$), 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 equilibrium in equations (A1) and (A3) 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 an innocent $D$. Thus, in equilibrium, the posterior belief that an acquitted $D$ is guilty is:

$$\mu(G \mid a) = \rho_G^{D0}(1 - \lambda)F_G/[\rho_G^{D0}(1 - \lambda)F_G + \lambda F_I]. \tag{15}$$

A small amount of algebra, after employing equation (13), yields the somewhat surprising result that $\mu(G \mid a)$ is, in equilibrium, independent of $\lambda$ (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 trial outcome from the arrival frequency of innocent defendants ($\lambda$) into the system. Notice that this is despite the fact that $\lambda$ affects $\rho_G^{D0}$ (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 Table 1) 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) $\Theta$ 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 $\lambda$ 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

| Agents and Relevant Exogenous Parameters | | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\Theta$ | | | $P$ | | $D$ | |
| | $r^D$ | $r_I^P$ | $r_G^P$ | $S_c$ | $k^P$ | $k^D$ | $\lambda$ |
| $\rho_G^{D0}$ | 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.

4.1. The Effect of Changes in the Informal Sanction Rates

From Table 1 we see that an increase in $r^D$ has no effect on $\rho_G^{D0}$, but it reduces the equilibrium plea offer, $S_b(\rho_G^{D0})$.[29]  However, an increase in $r_I^P$ (or a decrease in $r_G^P$) increases both $\rho_G^{D0}$ and $S_b(\rho_G^{D0})$.  For example, consider an increase in $r_I^P$; it enters $\rho_G^{D0}$ via the numerator, so according to equation (13) increasing $r_I^P$ means that the computed value of $\rho_G^{D0}$ should (in equilibrium) rise.  The intuition for this rise in $\rho_G^{D0}$ 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 innocent $D$ being convicted, resulting in a $\Theta$-imposed penalty on $P$.  This undermines the credibility of $P$'s threat to take a $D$ who rejects the offer to trial so, to maintain those incentives, a guilty $D$ must reject the offer with a higher probability.[30]  Moreover, since the expected loss from trial for a guilty $D$ thereby increases, this allows the equilibrium plea bargain to rise.  Furthermore, a higher value of $r_I^P$, which causes a guilty $D$ to reject the offer with higher probability, means that the mix of $D$-types at trial puts higher weight on $D$ being truly guilty, meaning that both conviction and acquittal are more likely to be associated with a guilty $D$ (that is, $\mu(G \mid a)$ increases, as shown in the last row of Table 1).  An alternative intuitive explanation is that if $r_I^P$ has increased, $P$'s best response is to increase the likelihood that a truly guilty $D$ chooses to go to trial.  This is because, given the nature of the trial process and the higher likelihood that a guilty $D$ is convicted in comparison with an innocent $D$, by so altering the likelihood of the type of $D$ who is going to trial in the direction of the defendant actually being guilty, $P$ effectively enlists $J$ in reducing the likelihood that the convicted $D$ was innocent, thereby reducing the impact of the increase in $r_I^P$ on the informal sanctions imposed on $P$.

The effect of a change in the informal sanction rates on the beliefs that outside observers hold about a $D$ that goes 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 *higher* informal sanctions on both types of $D$, since there is no way for a $\Theta$ to know whether $D$ is a guilty $D$ or an innocent $D$. In particular, this would imply that an innocent $D$ caught up in the system would suffer a greater level of informal sanctions when outsiders wish to penalize $P$ for railroading innocents.[31]

Table 1 also indicates that an increase in $r_G^P$ has the opposite effect on the equilibrium beliefs held by $\Theta$. 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 *lower* informal sanctions on a $D$ that goes to trial, since they anticipate that (in equilibrium) the likelihood of a guilty $D$ going to trial has gone down, making a $D$ going to trial more likely to be truly innocent. Moreover, the expected loss from misclassification depends directly (as well as indirectly, via $\rho_G^{D0}$) on $r^D$, $r_I^P$, and $r_G^P$. Taking account of both effects, this expected loss is increasing in $r^D$ and $r_I^P$; that is, higher informal sanction rates on $D$ and on $P$ for convicting innocents lead to a greater expected loss from misclassification of $D$ by $\Theta$.[32]

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

Another set of predictions concerns the effect of increases in the formal sanction, $S_c$, and the costs of trial $k^P$ and $k^D$. Because these parameters do not affect $M(\rho_G^{D0})$ directly, increases in these parameters affect $M(\rho_G^{D0})$ with the same sign as their effect on

$\rho_G^{D0}$. As indicated in Table 1, an increase in $S_c$ results in an increase in the likelihood of accepting the plea offer (that is, $\rho_G^{D0}$ falls). Intuitively, this occurs because increasing $S_c$ increases $P$'s expected payoff from trial. Since $\rho_G^{D0}$ is determined by $P$'s indifference between taking a case to trial versus dropping it, a guilty $D$ needs to reject a plea offer less often in order to restore $P$'s indifference between trial and dropping the case. Furthermore, as shown in the Technical Appendix, an increase in $S_c$ decreases the overall expected informal sanctions facing (respectively) $P$ and an innocent $D$, but increases the overall expected informal sanctions facing a guilty $D$.[33] This result follows from the fact that increasing $S_c$ results in more plea acceptance by a guilty $D$, which self-labels him as guilty and lowers the posterior assessment of guilt following a trial (which helps an innocent $D$).

What if the cost of trial to $P$, $k^P$, falls? As indicated in Table 1, direct computation from equation (13) above shows that $\rho_G^{D0}$ 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 (an innocent $D$ goes to trial, as before, but a guilty $D$ is less likely to do so). An increase in $k^D$ weakens $D$; since this cost only appears in $D$'s expected loss from trial, the equilibrium plea offer increases but the likelihood of acceptance does not change (since this latter strategy is determined by $P$'s incentives to drop or proceed to trial). As mentioned earlier, the use of pretrial detention acts to extract a longer sentence from a $D$ who accepts the plea offer.

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 an

innocent $D$ and a guilty $D$ who goes to trial and the last term reflects a guilty $D$ who pleads guilty. This reduces to $1 - \lambda F_I - (1 - \lambda)\rho_G^{D0}F_G$, which is clearly increasing as $\rho_G^{D0}$ falls. Thus, anything that lowers $\rho_G^{D0}$ (holding $\lambda$ and the evidence distributions constant) will raise $P$'s conviction rate since any $D$ accepting a plea is (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.

## 4.3. Police Arrest and Trial Process Effectiveness

Earlier we reflected on police arrest effectiveness as involving $\lambda$ being small: the intake process into the legal system is effective if the likelihood that the police arrested an innocent $D$ is (relatively) small. From equation (13) above (and as indicated in Table 1) reductions in $\lambda$ lead to reductions in $\rho_G^{D0}$ 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.[34]

We are also interested in the effectiveness of the trial process as to 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 $D$s are acquitted) and $F_G = 0$ (all guilty $D$s 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.

Returning to equation (13), an increase in resources applied to trial process whose only effect is to reduce $F_G$ simply increases the denominator of $\rho_G^{D0}$. This means that such

an investment increases the likelihood of plea-bargaining success. The impact of increased trial resources that only affect $F_I$ is more complicated. This is because of the sign of $S_c - r_I^P$ is unconstrained by MR1 and MR2.[35] If the informal sanction rate is small ("small $r_I^P$," meaning $S_c - r_I^P > 0$), then an increase in $F_I$ leads to an increase in $\rho_G^{D0}$, 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 $F_I$ leads to a decrease in $\rho_G^{D0}$. This 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 trial resources in the large-$r_I^P$ case definitely leads to a reduction in $\rho_G^{D0}$ (and in $S_b(\rho_G^{D0})$), as the numerator of equation (13) is falling while the denominator of equation (13) is increasing. Here, trial is becoming more effective at separating an innocent from a guilty $D$ and providing them with the corresponding dispositions of $a$ and $c$, respectively: trial is a better tool for *not* convicting the innocent and *for* 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 both the numerator and the denominator of equation (13) are increasing.

5. EQUILIBRIUM PLEA ACCEPTANCE BY INNOCENT DEFENDANTS

A common result for plea bargaining models with risk-neutral defendants is that all truly innocent defendants reject the plea offer and choose to go to trial.[36] This prediction is inconsistent with reality wherein a fraction of innocent defendants do plead guilty.[37] 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 incorporate a second dimension of unobservable heterogeneity among defendants in order to obtain an equilibrium wherein some innocent defendants rationally choose to accept a plea offer.[38]   This adjusts an important previous result since, if innocent defendants do accept a plea offer, then an outside observer's belief about the likelihood of guilt of a defendant who accepts a plea offer must be less than one.

5.1.  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 $\lambda = \Pr\{t = I\}$, where $D$'s type is 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 $\omega$ 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 $\omega$ 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.

A "strong" $D$ is as described 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 $\pi_T^D(G)$).  A "weak" $D$ is one that is sufficiently risk-

averse[39] and/or ambiguity-averse[40] to be willing to accept a plea offer that is designed to render a strong guilty $D$ indifferent between accepting the plea offer and going to trial.

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 $\rho_G^{D0}$ of $GS$-types reject the offer, all $W$-types accept the offer, and $P$ takes all $D$s who reject the offer to trial. The beliefs formed by $\Theta$ under the outcomes acquit ($a$), convict ($c$), and drop ($d$) are undisturbed by the modifications made above. 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 $\omega$: as the likelihood of a weak defendant increases, accepting the plea offer is an increasingly noisy 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 unavoidably sweeps up some innocent defendants as well.

The equilibrium plea offer is also affected because a defendant that accepts a plea offer is no longer inferred to be guilty for sure. The plea offer in the original 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 some 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.[41]  Such a refinement of the jury's assessment of a

defendant's true guilt or innocence[42] should provide more information to the outside

observers to employ in applying informal sanctions.

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 $(0, 1)$.            (16)

This assumption implies (see Müller and Stoyan, 2002: 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

"guilty" under the three-outcome verdict).[43]  Further, let $\gamma_{ng}$ 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 conjectured strategies, $(\rho_G^{D\Theta}, \rho_I^{D\Theta})$, $\Theta$'s

beliefs as to $D$ being truly guilty, having observed one of the outcomes $ng$,

$np$, or $g$, satisfies:

$\mu(G \mid ng) < \mu(G \mid np) < \mu(G \mid g)$;            (17)

and            2)  the expected loss from proceeding to trial for type $I$ ($\pi_T^D(I)$) is strictly

lower than the expected loss from proceeding to trial for type $G$ ($\pi_T^D(G)$),

where:

$$\pi_T^D(t) = S_c(1 - F_t(\gamma_g)) + k^D + r^D\mu(G \mid g)(1 - F_t(\gamma_g)) + r^D\mu(G \mid ng)F_t(\gamma_{ng})$$

$$+ r^D\mu(G \mid np)(F_t(\gamma_g) - F_t(\gamma_{ng}))\}, \, t \in \{I, G\}. \tag{18}$$

The ordering of payoffs indicated above means that Proposition 1 applies to the modified

game, so the equilibrium again involves an innocent $D$ always rejecting the plea offer and

$P$ always taking any $D$ who rejects a plea offer to trial, while a truly guilty $D$ mixes

between accepting the plea offer and rejecting it with positive probability $\rho_G^D$. When $P$'s

expected payoff from trial ($\pi_T^P$) is extended to allow for the three outcomes, it turns out

that this function is independent of $\gamma_{ng}$ when $P$ and $\Theta$ hold the same conjecture about $\rho_G^D$

(which, in equilibrium, they must). This means that $P$ is made indifferent between

dropping and going to trial by the same $\rho_G^D$ as in the two-outcome regime: $\rho_G^D = \rho_G^{D0}$.

The change in the number of possible trial outcomes affects $\Theta$ and $D$. The

difference between the functions capturing the expected loss from misclassification for

the three-outcome verdict and for 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 results hold:

$$\mu(G \mid np)(F_G(\gamma_g) - F_G(\gamma_{ng})) + \mu(G \mid ng)F_G(\gamma_{ng}) > \mu(G \mid a)F_G(\gamma_g), \tag{19}$$

and

$$\mu(G \mid np)(F_I(\gamma_g) - F_I(\gamma_{ng})) + \mu(G \mid ng)F_I(\gamma_{ng}) < \mu(G \mid a)F_I(\gamma_g). \tag{20}$$

Intuitively, equations (19) and (20) reflect the underlying ordering of the evidence

distribution, wherein a truly guilty $D$ gets a stochastically worse evidence draw than a truly innocent $D$. Equation (19) indicates that subdivision of acquittal for a truly guilty $D$ yields a higher expected belief of guilt, while equation (20) indicates the reverse for a truly innocent $D$: a finer partition of the outcomes benefits a truly innocent $D$ and hurts a truly guilty $D$. The implication for $D$'s trial loss is summarized in Proposition 2.

> *Proposition 2*: The expected loss for a truly guilty $D$ is higher under the three-outcome verdict than under the two-outcome verdict. The expected loss for a truly innocent $D$ is lower under the three-outcome verdict than under the two-outcome verdict.

This means that while a truly guilty $D$ still rejects the equilibrium plea offer at the same rate as before, $\rho_G^{D0}$, the equilibrium offer itself is larger, since $S_b(\rho_G^{D0}) = \pi_T^D(G; \rho_G^{D0}) - r^D$ and the expected loss from trial for a truly guilty $D$ 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 $\Theta$'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: a truly innocent $D$ expects to lose less, a truly guilty $D$ expects to lose more, and $\Theta$ imposes (on average) lower erroneous informal sanctions.

6.1. 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).[44] 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" will not, in reality, be as useful as it seems.

## 7. SUMMARY AND FURTHER DISCUSSION

Our model considers the strategic interaction between a prosecutor ($P$) and a defendant ($D$) when informal sanctions by third parties can be imposed on either or both of them. Informal sanctions imposed by outside observers are modeled as the product of a rate (one for the defendant and two for the prosecutor) and the outside observer's Bayesian beliefs as to the underlying guilt or innocence of $D$, conditional on the disposition of the case. There are two possible sanction rates for prosecutors since $P$s may "railroad" innocents into accepting a plea or being wrongly convicted at trial, or allow guilty $D$s to go free by dropping the case or losing at trial. We first show that there is a unique equilibrium for the game, wherein $P$ makes a plea offer that adjusts for $D$'s

informal sanction rate. This offer is rejected by an innocent $D$ but is accepted with positive probability by a guilty $D$; this rejection rate by a guilty $D$ is just sufficient to make $P$'s threat of trial credible.

We describe a number of relevant comparative statics results. Very high informal sanctions on a defendant can eviscerate plea bargaining. Higher informal sanctions on a prosecutor for railroading innocents can result in higher informal sanctions on $D$ (both innocent and guilty $D$s) while higher informal sanctions on $P$ for being "soft on crime" attenuate the informal sanctions applied to $D$. Increasing the formal sanction imposed upon a $D$ convicted at trial (or decreasing $P$'s cost of trial) acts to increase expected informal sanctions for a truly guilty $D$ and decrease them for a truly innocent $D$. Actions that reduce the likelihood that an innocent $D$ is arrested reduce the equilibrium plea offer and increase the equilibrium plea acceptance rate, thereby reducing trial costs. We also show that the use of plea bargaining lowers the expected loss from misclassification by the outside observers; this is a new argument for allowing plea bargaining that is separate from the standard appeal to cost savings arising from reduced trial costs. All of the above results continue to hold if a small fraction of defendants may be sufficiently risk- or ambiguity-averse (i.e., sufficiently "weak" in contrast with the risk-neutral $D$) so that a weak innocent $D$ will choose to take the plea offer rather than reject it and go to trial. This occurs because the same equilibrium plea rejection rate is used by guilty (strong) defendants as in the original model.

In the penultimate section of the paper we consider the effect on our initial analysis if the jury's verdict could be refined, as occurs in the Scottish verdict, wherein

acquittals are further subdivided into those the jury labels as "not guilty" and those whose case is labeled as "not proven." This informational refinement leads to the same equilibrium plea rejection rate as before, but a higher equilibrium plea offer because an innocent $D$ has a lower expected loss from trial under the Scottish verdict while a guilty $D$ has a higher expected loss from trial under the Scottish verdict, when compared with the more traditional two-outcome verdict. Most significantly, it leads to a lower expected loss from misclassification by an outside observer.

7.1. Possible Extensions

One possible 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 primary 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, on which we have a recent companion paper (Daughety and Reinganum, forthcoming), is to allow the jury (in a two-outcome system) to award state-provided damages to an acquitted defendant against whom the evidence appears (to it) to be especially weak. This scheme partitions acquittals into compensated acquittals and uncompensated acquittals, and acts to replicate aspects of the Scottish verdict (the informational gains are the same, with the same plea rejection rate, but plea offers may be higher or lower, depending upon the size of the compensation). However, we also show

that if such an award is viewed as a cost to $P$, then this may raise both the plea offer and the truly guilty $D$'s plea rejection rate (since this cost would weaken $P$ and the credibility of the threat of trial will require a higher likelihood of plea rejection by a guilty $D$). Thus, full implementation via such a compensation scheme will cause overall costs (compensation plus trial costs) to rise.

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,[45] 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.

APPENDIX

A.1.  Outside Observer Posterior Beliefs as to $D$'s Guilt

Technically, $\Theta$ 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 $\Theta$'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, $\rho^P = 1$ is part of an equilibrium of the game, so that the outcome $d$ is out-of-equilibrium outcome, and the value for $\mu(G \mid d)$ will need to be otherwise specified, since $d$ will not be visited in equilibrium.  Moreover, $P$'s strategy, $\rho^P$, does not affect the beliefs because it (or $1 - \rho^P$) multiplies each relevant numerator and denominator and thereby drops out of the analysis.

$$\mu(G \mid a) = \rho_G^{D\Theta}(1 - \lambda)F_G/[\rho_G^{D\Theta}(1 - \lambda)F_G + \rho_I^{D\Theta}\lambda F_I]; \tag{A1}$$

$$\mu(G \mid b) = (1 - \rho_G^{D\Theta})(1 - \lambda)/[(1 - \rho_G^{D\Theta})(1 - \lambda) + (1 - \rho_I^{D\Theta})\lambda]; \tag{A2}$$

$$\mu(G \mid c) = \rho_G^{D\Theta}(1 - \lambda)(1 - F_G)/[\rho_G^{D\Theta}(1 - \lambda)(1 - F_G) + \rho_I^{D\Theta}\lambda(1 - F_I)]; \tag{A3}$$

and $\quad \mu(G \mid d) = \rho_G^{D\Theta}(1 - \lambda)/[\rho_G^{D\Theta}(1 - \lambda) + \rho_I^{D\Theta}\lambda]. \tag{A4}$

A.2.  Characterizing Equilibrium

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 $P$ $(\rho^P)$ and the $D$ of type $G$ $(\rho_G^D)$. Since the observers' beliefs will depend on their conjectured value for $\rho_G^D$, 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 $\rho^P$. 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 $\rho^P$ that renders $G$ indifferent:

$$\rho^P(S_b; \rho_G^{D\Theta}) = \{S_b + r^D(1 - \mu(G \mid d; \rho_G^{D\Theta}))\} \times$$

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

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 $v(G \mid R; \rho_G^{D\Theta}) = \rho_G^{D\Theta}(1 - \lambda)/[\rho_G^{D\Theta}(1 - \lambda) + \lambda]$. 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 $\Theta$ 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:

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

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

$$= - r_G^P \mu(G \mid d; \rho_G^{D\Theta}).$$

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; 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 $\rho_G^{D0}$. 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 $S_b = \pi_T^D(G; \rho_G^{D0}) - r^P$, where $\pi_T^D(G; \rho_G^{D0})$ is the expression $\pi_T^D(G)$, 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 $\rho_G^{D0}$). 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:

$$r^D \leq [S_c(1 - F_G) + k^D]/[1 - \mu(G \mid c; \rho_G^{D0})(1 - F_G) - \mu(G \mid a; \rho_G^{D0})F_G]. \tag{A5}$$

Returning to $P$'s payoff as a function of $S_b$, notice two things. First, since $\rho_G^{D0}$, 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 $\rho_G^{D0}$, 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 $G$-types accept, and others reject, the offer. But – taking the observers' beliefs of $\rho_G^{D0}$ as given – $P$ could make a higher demand that would provoke certain rejection by both 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 $S_b(\rho_G^{D0}) = \pi_T^P(G; \rho_G^{D0}) - r^D$ with the expected payoff from taking a guilty defendant to trial (the observers' beliefs are fixed at $\rho_G^{D0}$ because trial is on the equilibrium path). Thus, $P$ prefers (at least weakly) the hypothesized equilibrium to defection as long as:

$$S_b(\rho_G^{D0}) = \pi_T^P(G; \rho_G^{D0}) - r^D$$

$$= S_c(1 - F_G) + k^D + r^D\mu(G \mid c; \rho_G^{D0})(1 - F_G) + r^D\mu(G \mid a; \rho_G^{D0})F_G - r^D \qquad \text{(A6)}$$

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

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:

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

$$[1 - \mu(G \mid c; \rho_G^{D0})(1 - F_G) - \mu(G \mid a; \rho_G^{D0})F_G]^{-1}. \qquad \text{(A7)}$$

Since the right-hand-sides of Conditions 1 and 2 are evaluated at $\rho_G^{D0}$, these are purely-parametric constraints. The right-hand side of Condition 1 is always at least $S_c(1 - F_G) + k^D$, so the set of parameters satisfying that constraint is non-empty. The right-hand-side of Condition 2 is always at least $k^P + k^D$, so the set of parameters satisfying that constraint is also non-empty. Moreover, both right-hand-sides are independent of $r^D$ (since $\rho_G^{D0}$ does not depend on $r^D$). Thus, both Conditions provide upper bounds on $r^D$ but we are unable to determine which one is the least upper bound. In addition, we are

unable to provide a complete characterization of what an equilibrium would look like if one or both of these Conditions fails to hold. However, it is clear that $r^D$ can be high enough that no $D$ will accept even a plea offer of $S_b = 0$ as it is accompanied by the informal sanction $r^D$. A sufficient condition for this to occur is: $r^D > [S_c(1 - F_G) + k^D]/[1 - \mu(G \mid c; \rho_G^D = 1)(1 - F_G) - \mu(G \mid a; \rho_G^D = 1)F_G]$. In this sense, high informal sanctions on $D$ can eviscerate plea bargaining.

A.3. Uniqueness

There cannot be an equilibrium wherein the $D$ of type $G$ rejects the plea bargain with probability $\rho_G^D < \rho_G^{D0}$ as $P$'s threat to go to trial is not credible. Moreover, there cannot be an equilibrium wherein the $D$ of type $G$ rejects the plea bargain with probability $\rho_G^D > \rho_G^{D0}$; see the Technical Appendix for the details.

A.4. Expected Loss from Misclassification

The following terms summarize erroneously-imposed informal sanctions, for arbitrary $\rho_G^D$, with $\rho_I^D = 1$. Excessive sanctions imposed on type $I$ defendants following conviction, or acquittal, at trial (ideally, there would be no sanctions) are:

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

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

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

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

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

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 these are:

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

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 beliefs are evaluated at $\rho_G^D$):

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

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

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

Collecting terms simplifies the above expression greatly:

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

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

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

REFERENCES

Baker, S., and C. Mezzetti. 2001. "Prosecutorial Resources, Plea Bargaining, and the Decision to Go to Trial," 17 *Journal of Law, Economics, and Organization* 149-167.

Bandoyopadhyay, S., and B.C. McCannon. 2014. "The Effect of the Election of Prosecutors on Criminal Trials," 161 *Public Choice* 141-156.

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

Benabou, R., and J. Tirole. 2006. "Incentives and Prosocial Behavior," 96 *American Economic Review* 1652-1678.

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

Boylan, R.T., and C.X. Long. 2005. "Salaries, Plea Rates, and the Career Objectives of Federal Prosecutors," 48 *Journal of Law and Economics* 627-652.

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

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

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

Daughety, A.F., and J.F. Reinganum. 2000. "Appealing Judgments," 31 *RAND Journal of Economics* 502-525.

-------------. 2010. "Public Goods, Social Pressure, and the Choice Between Privacy and Publicity," 2 *American Economic Journal: Microeconomics* 191-221.

-------------. Forthcoming. "Selecting Among Acquitted Defendants: Procedural Choice vs. Selective Compensation," *Journal of Institutional and Theoretical Economics*.

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

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

Flegenheimer, Matt. 2015. "What to Expect in New York in 2013." *New York Times*, February 16. Online at:

http://www.nytimes.com/interactive/2012/12/30/nyregion/new-york-in-2013.html?_r=0#/#courts.

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

-------------. 2014. "Liability Law and Uncertainty Spreading." University of Bologna Department of Economics, Working paper.

Gilboa, I., and D. Schmeidler. 1989. "Maxmin Expected Utility with Non-unique Prior," 18 *Journal of Mathematical Economics* 141-153.

Glaeser, E.L., D.P. Kessler, and A.M. Piehl. 2000. "What do Prosecutors Maximize? An Analysis of the Federalization of Drug Crimes," 2 *American Law and Economics*

*Review* 259-290.

Gordon, S.C., and G.A. Huber. 2002. "Citizen Oversight and the Electoral Incentives of Criminal Prosecutors," 46 *American Journal of Political Science* 334-351.

Grossman, G.M., and M.L. Katz. 1983. "Plea Bargaining and Social Welfare," 73 *The American Economic Review* 749-757.

Iacobucci, Edward M. 2014. "On the Interaction between Legal and Reputational Sanctions," 43 *Journal of Legal Studies* 189-207.

Klibanoff, P., M. Marinacci, and S. Mukerji. 2005. "A Smooth Model of Decision Making under Ambiguity," 73 *Econometrica* 1849-1892.

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

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

Levy, Gilat. 2005. "Careerist Judges and the Appeals Process," 36 *RAND Journal of Economics* 275-297.

--------------. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules," 97 *American Economic Review* 150-168.

Machina, M.J., and M. Siniscalchi. 2014. "Ambiguity and Ambiguity Aversion," in M.J. Machina and W.K. Viscusi, eds., *Handbook of the Economics of Risk and Uncertainty, Volume 1*. Amsterdam: Elsevier North-Holland.

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

Müller, A., and D. Stoyan. 2002. *Comparison Methods for Stochastic Models and Risks*. Chichester, England: John Wiley and Sons.

Nalebuff, Barry. 1987. "Credible Pretrial Negotiation, 18 *RAND Journal of Economics* 198-210.

Rasmusen, Eric. 1996. "Stigma and Self-Fulfilling Expectations of Criminality," 39 *Journal of Law and Economics* 519-543.

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

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

Smithson, M., S. Deady, and L. Gracik. 2007. "Guilty, Not Guilty, or ...? Multiple Options in Jury Verdict Choices," 20 *Journal of Behavioral Decision Making* 481-498.

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

FIGURE LEGENDS/CAPTIONS

Figure 1:  Game Between *P* and *D*

FOOTNOTES

---

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.

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.   As an example of informal sanctions on prosecutors, Charles "Joe" Hynes lost reelection as District Attorney for Brooklyn, NY, due to voter dissatisfaction with both his failure to pursue child sexual abuse complaints in Brooklyn's  Orthodox Jewish community as well as perception that he was wrongfully convicting (innocent) defendants in other cases; see Flegenheimer (2012).  Hynes, who had served as DA for 20 years, lost the Democratic primary and then, upon running in the general election as a Republican, lost that election as well.  Our analysis abstracts from the use of formal sanctions for officials, but abuses such as prosecutorial misconduct can lead to formal

sanctions.

5. Other recent papers that incorporate payoffs (representing the assessments of third parties) that are proportional to inferred type include Levy (2005, 2007), Benabou and Tirole (2006), Daughety and Reinganum (2010), Deffains and Fluet (2013), and Iacobucci (2014).

6. For instance, suppose that an outside observer and the defendant have a random match value from a particular transaction (a job, a rental unit, a college admission etc.). The observer compares the realized match value to the perceived risk of dealing with the defendant. Across a collection of observers and transactions, the higher the posterior assessment of guilt, the more matches the outside observers will choose to forego. Outside observers do not consider the negative externalities 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.

8. 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 zero if she frees an innocent defendant and -$\alpha x$ 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.

10. In Daughety and Reinganum (2000) we argue that the rules of evidence and procedure used in trials are inconsistent with a fully Bayesian model of a jury; there we use axiomatic methods to model a jury's decision problem.

11. 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.

12. 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.

13. Available at: www.vanderbilt.edu/econ/faculty/Daughety/DR-InformalSanctionsandCaseDispositions-TechApp.pdf

14. The descriptors "innocent" or "guilty" will refer to $D$'s type, whereas the trial outcomes are "acquitted" and "convicted."

15. For ease of exposition we will refer to $\Theta$, but this encompasses the collection of outside observers that $D$ and/or $P$ may encounter.

16. 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.

17. It is very plausible that $\Theta$ would not observe a rejected plea offer. We assume that $\Theta$ 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 $\Theta$'s out-of-equilibrium beliefs, but we leave this as an item for future research.

18. 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 $r^D$.

19. As with $r^D$, $r^P_I$ and $r^P_G$ may vary with the crime in question or observable attributes of $P$ (and possibly $D$).

20. 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 $k^D$.

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

22. 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.

23. 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).

24. Alternative candidates for equilibria, such as fully-separating or fully-pooling candidates, or candidates involving an innocent $D$ accepting a plea offer, cannot be equilibria; see the Technical Appendix for details.

25. Recall that there are only two places in the game wherein evidence about $D$ is realized: at the beginning (the arrest) and at the end (the trial). While we have not included the possibility that evidence arises after the plea bargain but before the decision

by $P$ as to dropping or going to trial, if publicly-observable exculpatory evidence arose at this point in the process, then a drop might be consistent with equilibrium play.

26. Beliefs for $\Theta$ are given by equations (A1) - (A4), evaluated at arbitrary $\rho_G^D$ and $\rho_I^D = 1$; beliefs for $P$ are given by $v(G \mid R) = \rho_G^D(1 - \lambda)/[\rho_G^D(1 - \lambda) + \lambda]$, because a guilty $D$ rejects with probability $\rho_G^D$, and an innocent $D$ always rejects, the plea offer.

27. 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 $\rho_G^D = \rho_G^{D0}$. Condition 1 implies that $S_b(\rho_G^{D0})$ is non-negative, whereas $P$'s payoff from dropping the case is $-r_G^P\mu(G \mid d)$, which is negative. Thus, $P$ strictly prefers the outcome involving some accepted plea offers and some trials to defecting to dropping all cases.

28. There is a continuation equilibrium following an out-of-equilibrium plea offer $S_b < S_b(\rho_G^{D0})$, wherein a guilty $D$ mixes between accepting and rejecting the plea offer with probability $\rho_G^{D0}$ and $P$ mixes between taking the case to trial and dropping it. $P$ cannot gain from such a deviation; see the Appendix for details.

29. 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.

30. Recall that we abstract from a personal disutility for $P$ associated with prosecuting and/or convicting a defendant she believes is innocent, but incorporating such a disutility would simply reinforce (appropriately revised versions of) MR0 and MR1. Moreover, much like an increase in $r_I^P$, such a disutility would undermine $P$'s incentive to take the case to trial following a rejected plea offer; in order to re-establish the credibility of $P$'s

trial threat, a guilty defendant would (in equilibrium) reject the plea offer with a higher probability.

31.  This seemingly perverse result reflects the lack of a feedback effect between the magnitude of the outsiders' distaste for a $P$ that "railroads" an innocent $D$ and any effort that $P$ could exert to reduce the likelihood of an innocent $D$ in the original pool.  While we have not formally 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 $\lambda$ (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 $\lambda$ and, more importantly from $P$'s perspective, $r_I^P$.

32.  The change in the expected loss from misclassification is indeterminate with respect to a change in $r_G^P$ (the direct effect is positive and the indirect effect is negative).

33.  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.  We return to this in Section 7.

34.  The effect of $\lambda$ on $M(\rho_G^{D0})$ is complex as $\lambda$ enters directly (including via the beliefs) and indirectly (via $\rho_G^{D0}$).  Nevertheless, $\lambda < \frac{1}{2}$ (which seems a reasonable requirement of a

rational police system) is sufficient to ensure that the expected loss from misclassification is increasing in $\lambda$.

35. If MR0 is to hold for *every* possible conjecture by $\Theta$ about $D$'s strategies, then MR0 implies that $S_c - r_I^P > 0$. MR1 and MR2 jointly rule out an equilibrium wherein $\rho_G^D < \rho_G^{D0}$. For $\rho_G^D \geq \rho_G^{D0}$, MR0 allows the case wherein $S_c - r_I^P < 0$, which is of interest in the discussion to follow.

36. To our knowledge, the only exception is Reinganum (1988). She constructs an equilibrium wherein some risk-neutral innocent defendants plead guilty. However, this is due to the fact that in her model, prosecutors have accurate (private) information about the probability of a win at trial, which is determined by the prosecutor's evidence rather than the defendant's true type. Thus, conditional on their inferences about P's evidence, both guilty and innocent defendants anticipate the same expected trial payoff.

37. The National Registry of Exonerations, maintained by the University of Michigan Law School, indicates that of 1555 exonerees (as of March, 2015), 13% of the wrongful convictions involved false confessions; this rate is indicated as being highest in homicide cases (21%) (http://www.law.umich.edu/special/exoneration/Pages/learnmore.aspx, accessed January 24, 2015).

38. The law accounts for a $D$ that wishes to claim to be innocent and yet be allowed to plead guilty (asserting that he believes that P has sufficient evidence to convict him) by allowing $D$ to make an "Alford plea." Such a plea nonetheless results in conviction.

39. 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 $GS$).

Since the equilibrium plea offer makes a *GS* indifferent, an *IS* will 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 *GS* indifferent would be accepted for sure by a *GW*. The only remaining question is whether an *IW* is willing to accept the plea offer that makes a *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 an *IW* prefer this plea offer to trial. Grossman and Katz (1983) also invoke risk aversion to obtain plea acceptance by innocent defendants.

40. For a recent application of ambiguity aversion to civil suits, see Franzoni (2014). Crudely, a decision maker is ambiguity-averse if he is uncertain about the distribution he faces, and prefers a world with a known distribution; for a comprehensive survey, see Machina and Siniscalchi (2014). The typical approach starts with the Savage axioms for subjective expected utility, substitutes some new axioms, and then derives the ambiguity-averse decision-maker's objective function. For example, following Gilboa and Schmeidler (1989), the defendant (who, say, imagines the possibility of a sure conviction at trial) places all the weight on the worst possible outcome (i.e., $e = 1$), yielding the maximum possible expected loss from trial. Alternatively, following Klibanoff, Marinacci, and Mukerji (2005), the *GW* type uses his set of possible priors and a function that captures his response to ambiguity; this will provide an expected loss from trial that is larger that of the *GS* 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*.

41. 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 (2000:1324) 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.

42. Since the Scottish verdict refers to "guilty" as an outcome, we adopt that language and now refer to a *D* whose type is *G* as being "truly guilty," or "type *G*." In a parallel construction, a truly innocent *D* will sometimes be referred to as type *I*.

43. 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 and find that the introduction of the third verdict option does not significantly alter the likelihood of a conviction (Smithson, et. al., 2007:492). They also find that reported assessments of guilt follow the same monotonicity as shown in our

equation (17).

44. 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.

45. 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 if an arrest is made. Also, see Rasmusen (1996) for a model of criminal choice with stigma.

Figure 1