SIZE, MONITORING, AND PLEA RATE: AN EXAMINATION OF UNITED STATES ATTORNEYS[1]

Richard T. Boylan and Cheryl X. Long

Washington University

St. Louis, MO 63130

(314) 935-5670

LONG@WUECONC.WUSTL.EDU

July 1999

Current version: July 10, 2000

[1]We are especially grateful to Kathleen Clark for her patience in explaining to us many of the intricacies of the legal system. We are also thankful to all the people who made data available to us and explained their meaning: John Scalia (Bureau of Justice Statistics), William Sabol (Urban Institute), Bonnie Gay, Frank Kalder, Barbara Tone (Executive Office of U.S. Attorneys). We received useful feedback from Phil Dybvig, Katherine Goldwasser, Vivian Ho, Qianbo Huai, Lee Lawess, Ted MacDonald, John Nachbar, Wilhelm Neuefeind, Jennifer Reinganum, and Paul Rothstein. Iliya Filev, Maureen Gallagher, and Meandria Tart provided cheerful and efficient research assistance.

Abstract

A general theoretical model relates monitoring to the size of an organization. The results are applied to the context of U.S. Attorney Offices where case mix, staffing, career opportunities, and monitoring are related to the likelihood of a plea settlement. Analysis of federal drug trafficking cases in fiscal years 1993 through 1996 leads to the following conclusions: There are fewer pleas in U.S. Attorney districts where caseload is lower, where private salaries are higher, and where crimes are more severe. Further, there are fewer pleas in districts with many or with few prosecutors, and there are more pleas in districts with an average number of prosecutors. The explanation for the latter results is that, unless they are closely monitored, prosecutors may take cases to trial to acquire human capital. Hence there is a lower plea rate in districts where prosecutors are not closely monitored. Further, the monitoring technology exhibits increasing returns to scale for small districts and decreasing returns to scale for large districts due to the tradeoff between gains from specialization and losses due to difficulty in coordination. Hence, small and large offices do not find monitoring as effective and adjust their monitoring expenditures accordingly; specifically, in these offices, prosecutors are not as closely monitored and the plea rates are lower.

Journal of Economic Literature Classification Numbers: H11, H40, K14, K41, K42.

1 Introduction

The role of federal prosecution in the U.S. justice system has experienced rapid and continuous expansion in the last two decades. This has led the number of prosecutors in the U.S. Attorney Offices to increase from 1386 to 3938 between 1977 and 1997. Since there are 94 offices, each in charge of the federal prosecution in one federal district, this means that the average number of prosecutors per U.S. Attorney Office has increased from 15 to 42. Faced by the rapid growth of the size of U.S. Attorney Offices, it is now more important than ever to study how efficiently the offices are operated. This paper contributes to this task by examining how effective monitoring is in the U.S. Attorney Offices.

Like other public agencies, the performance of the offices depends heavily on how well the prosecutors are monitored by the U.S. Attorneys. Monitoring is particularly important in the public sector because the heads of the public agencies have much less discretion in hiring, firing, and determining the pay for their subordinates than their private sector counterparts. To study the effectiveness of monitoring and its effects on performance in the U.S. Attorney Offices, this paper measures their performance by examining the plea rates for the following reasons.

First of all, plea rate is an index of great economic importance. Compared to pleas, trials take substantially more prosecutorial and judicial resources (Hollandar-Blumoff [18]), and it is therefore important that plea rates be high. As stated by Chief Justice Burger [7]: "The consequence of what might seem on its face a small percentage change in the rate of guilty pleas can be tremendous. A reduction from 90 per cent to 80 per cent in guilty pleas requires the assignment of twice the

judicial manpower and facilities – judges, court reporters, bailiffs, clerks, jurors and courtrooms. A reduction to 70 per cent trebles the demand." Because trials are so costly, despite the asymmetry in information, approximately 90% of all cases are disposed by plea agreement.[2]

Secondly, because U.S. Attorneys and prosecutors have different preferences over plea versus trial, it is important for chief prosecutors to monitor the plea process. Schulhofer and Nagel [31] studied in depth ten U.S. Attorney Offices and found that prosecutors did not always abide by the office plea policies, and that the extent to which prosecutors pleas were monitored varied greatly among districts.

Thirdly, there is wide variation in plea rates across districts. In 1993, for instance, plea rates varied among districts between 97% (New Hampshire) and 67% (Middle District of Alabama).[3] Further, plea rates vary with the size of a district: plea rates are higher for medium-sized districts while lower for both small and large districts. (See Table 1 on page 35.) The potential regularity relating plea rate and district size merits further exploration.

This paper explains the variation in plea rates by differences in case mix, staffing, career opportunities for prosecutors, and supervision among districts. Differences in case mix lead to differences in plea rates, because more serious offenses are more likely to lead to trials.[4] Differences in staffing lead to differences in plea rates because understaffed offices face higher opportunity costs of trial. Differences in outside career opportunities lead to differences in plea rates because prosecutors have

[2]For discussion of the asymmetry of information, see page 17.

[3]Data source: FJSRC Standard Analysis Files.

[4]This was first pointed out by Landes [19].

more incentives to accumulate trial experience when private salaries are higher. Finally, differences in supervision lead to differences in plea rates because more effective supervision makes it more difficult for prosecutors to deviate from the district plea policies.

To analyze what factors determine whether the outcome of a drug trafficking case is a plea agreement or a jury trial, this paper uses Federal Justice Statistics Resource Center data. Compared to the more commonly used National Correction Reporting Program data, the newly available FJSRC data has the following comparative advantages: First, since the information is recorded and provided by different agencies in the federal justice system and the federal judicial, every federal district is part of the program and hence the dataset covers exhaustively all the matters and cases that go through the system in relevant years; Second, the data sources include various stages of federal prosecution and a link file is provided by the Center so that information on cases or defendants from different agencies can be integrated in any analysis. In the context of our analysis, this dataset enables us to obtain detailed information on case and defendant characteristics and to analyze the information at the district level.[5]

The analysis in this paper is restricted to simple drug trafficking cases for four reasons.[6] First, the objective functions of the prosecutors seem easier to characterize for such cases (see Section 3.4). Second, there are better proxies for case mix for drug trafficking cases (see Section 4.1). Third, the

[5] In contrast, the analysis on federal prisoners can only be carried out at the state level for the NCRP data, since county information is not reported for federal prisoners.

[6] The restriction to 'simple drug trafficking cases' consists of excluding drug possession cases and Organized Crime Drug Enforcement Task Force cases.

large number of drug trafficking cases prosecuted lends statistical significance to the results in the paper. Fourth, the agency problem discussed in this paper is specific to the type of drug trafficking cases analyzed (see Section 3.4).

This paper finds that there are fewer pleas in districts where caseload is lower, where private lawyer salary is higher, and where crimes are more severe. Further, there are fewer pleas in U.S. Attorney districts with many or with few prosecutors, and there are more pleas in U.S. Attorney districts with an average number of prosecutors. In terms of monitoring, the latter results imply that the monitoring level is not always a decreasing function of size. Clearly, if monitoring is a fixed resource, then the monitoring level always decreases with district size. However, it is shown in this paper that, if an agency can allocate resources to monitoring, then the optimal level of monitoring is increasing for small offices and decreasing for large offices.

The paper proceeds as follows. The next section discusses a model on monitoring and presents the main results on the relationship between monitoring and size. Section 3 elaborates on the implications of monitoring on plea rate via a bargaining model and compares the assumptions in the model to the existing literature. In Section 4, the hypotheses derived from the monitoring model and the bargaining model are tested empirically. Section 5 concludes. Proofs of the theoretical results and all the tables are contained in the Appendix.

2 Monitoring and Size: Theory

It is widely believed that larger firms find it more difficult to monitor their employees. (For general references, see Brown and Medoff [6]; Glaeser, Kessler, and Piehl [13] discuss the difficulty in monitoring in large U.S. Attorney districts.) Clearly, in a larger firm, the owner can personally monitor a smaller fraction of the employees. But the large size also makes it a more feasible option for the owner to hire managers with specialized monitoring skills. This suggests that the relationship between firm size and the degree of difficulty in monitoring may be different from the conventional belief.

In a model discussed in Boylan and Long [5], *monitoring effectiveness* (the inverse of monitoring difficulty) is shown to be increasing in size for small firms and decreasing for large firms under very general conditions. The intuition is that there is tradeoff between gains from specialization and losses due to difficulty in coordination. The fact that some firms may find it more difficult to monitor does not necessarily mean that the *monitoring level* (the likelihood with which an employee's shirking is detected) is lower because these firms may find it worthwhile to spend a greater proportion of their resources on monitoring. However, it is further shown in [5] that, unless monitoring is a "Giffen good," the optimal allocation of resources implies a lower level of monitoring in firms that find it more expensive to provide monitoring.

The model in [5] is summarized below. The firm has N homogeneous employees and has to perform a continuum of tasks along a unit interval $[0, 1]$ to produce one good. Denote jobs by their locations. Hence, Job $\tilde{s}$ is located at $\tilde{s}$ in the interval. To maximize output, the firm allocates

n employees to production—performing the continuum of tasks—and the rest $m = N - n$ to monitoring—supervising the employees in production.

The output of the firm, U, depends on the monitoring level p. Following Becker and Murphy [1], the output also depends on the size of the production team:

$$U = (N - m)^{1+\theta} u(p), \theta \geq 0,$$

where the non-negative parameter θ captures increasing returns to specialization in production as discussed in [1], $u(p)$ captures how the monitoring level affects output, $u'(p) > 0$, and $u''(p) < 0$.

The ith of the m monitors is allocated at $\frac{2i-1}{2m}, i = 1, 2, ..., m$, and supervises tasks in $\left[\frac{i-1}{m}, \frac{i}{m}\right]$. The monitors supervise by detecting unsatisfactory performance in the interval of jobs and punishing corresponding delinquent employees. The probability with which an employee in production is monitored is $\frac{m}{N-m}$.

When monitored, the probability of a delinquency being detected decreases in the distance between the monitor and the job. This probability, denoted by $H(m)$, is shown to increase in m. In other words, with more monitors, each monitor is more effective in detecting delinquency. Monitoring effectiveness also depends on the amount of effort exerted by each monitor. Since only the owner monitors the monitors, the amount of effort exerted by each monitor decreases with the size of the monitoring team. Denote the effort level by $g(m)$ and the monitoring effectiveness by $e(m)$. Hence, $e(m) = g(m)H(m)$. It is shown in [5] that **the monitoring effectiveness, $e(m)$, is increasing for small m and decreasing for large m.** *(Result 1.)*

The level of monitoring (the probability of any delinquency being monitored AND detected)

is equal to the product of the probability with which an employee in production is monitored and the effectiveness of the monitoring. Hence, $p = \tilde{p}(m, N) \equiv \frac{m}{N-m} e(m)$. Define p^* to be the level of monitoring when the firm maximizes the output U by choosing the optimal number of monitors $m^* = \text{argmax}_{m \in [0,N]} (N-m)^{1+\theta} u(p)$. Hence, $p^\star(N) = \tilde{p}(m^*, N)$. Boylan and Long [5] give sufficient conditions under which **the optimal monitoring level $p^*(N)$ is increasing when the firm size N is small and decreasing when N is large**. (*Result 2.*) Under these conditions, the advantages of specialization are more important than difficulties due to coordination for small firms, and the losses due to difficulty of coordination are more important than gains from specialization for large firms.

3 Monitoring and Plea Bargaining: Model

The results given in Section 2 relate size to monitoring. This section presents a model on plea bargaining that provides a clear relationship between monitoring and plea rate. The model also clarifies several relevant factors, including case severity, case load, and private lawyer salary, which need to be taken into account in the subsequent analysis. Data used in the empirical analysis and results obtained are discussed in the next section.

It is well-known that prosecutors (AUSA below for Assistant U.S. Attorney) do not always act according to the wishes of their U.S. Attorney (USA below) during plea bargaining [31]. For instance an AUSA may plea a case to reduce their workload or try a case to gain expertise in a particular area of prosecution [18]. Such characterization, however, fails to distinguish the difference between

younger and older AUSAs. For younger prosecutors it is important to gain trial experience as a way of being assigned to more complex prosecutions and gaining private sector employment.[7] Based on this observation, this paper examines the entry level position in USA districts: the prosecution of drug trafficking cases that are not handled by the Organized Crime Drug Enforcement Task Force (OCDETF). For AUSAs in charge of such cases, taking a case to trial has a significant impact on the accumulation of human capital relevant for future jobs. We argue in more detail in Section 3.4 that this implies that AUSAs prefer more trials compared to the USA.

In order to simultaneously consider the bargaining problem between the defendant and the AUSA and the agency problem between the USA and the AUSA, the AUSA is assumed to possess private information on the strength of a case.[8] Thus, there are two distinct reasons for why such private information leads to trials. First, an AUSA with strong private information cannot convince the defendant to accept a long term in prison, and hence such cases end up in trials. Second, AUSA values a trial more highly than the USA because of the private benefits of trial experience, so that depending on the level of monitoring of the USA, cases the USA would have wanted to plea bargain may end up in trials.

The rest of the section is organized as follows. The preferences of the three individuals in the model – defendant, AUSA, and USA – are first discussed. After reviewing the timing of moves in the model, the relationships between monitoring, plea rates, and other relevant factors are derived, followed by some discussion on how the assumptions made in the model compare to

[7] Nelson [25, page 155], personal communication with St. Louis Circuit Attorney, Dee Joyce-Hayes.

[8] Section 3.4 provides justification for this assumption.

existing literature.

3.1 Players

There are three players in the model: the defendant, the AUSA, and the USA.

Defendant

The defendant is an individual suspected of drug trafficking and prosecuted by an AUSA. Two parameters characterize the case that the defendant is suspected of committing. The severity of the case, S, represents the length of the prison term if the defendant is convicted in case of a trial, and is distributed according to Π over $[\underline{S}, \bar{S}]$. The probability with which the defendant is convicted in case of a trial, t, is distributed according to F over $[0, 1]$, where F has a continuously differentiable density f. The preferences of the defendant are represented by the utility function $u_D = -s$, where s is the expected length of the prison term.[9]

Assistant U.S. Attorney

As in Glaeser et al [13], Landes [19], and Reinganum [29], the objective functions of the USA and the AUSA depend on the time that the defendant spends in jail. Specifically, preferences for the AUSA are represented by a utility function $u_A = (N-m)^\theta(s+iT-jP)$, where s is the expected time the defendant spends in jail, $i = 1$ indicates that there is a trial, $j = 1$ indicates that the AUSA turned down a plea the USA would have wanted the AUSA to accept, T is the AUSA's

[9] The defendant is assumed to be risk neutral. See Polinsky and Shavell [27] for a discussion of risk preferences of the defendant. Note that this paper assumes away agency problems between the defendant and the defense counsel. See Miller [23] for a theoretical study on the efficiency of different attorney compensation schemes.

personal benefit from a trial, P is the expected penalty the AUSA obtains from disobeying the USA's plea bargaining policy, and $(N-m)^{\theta}$ is the number of cases prosecuted by each AUSA.[10]

Trial experience is often thought to contribute to the private careers of former prosecutors [13]. In the same spirit, we assume that the personal benefit of a trial T is increasing in the private lawyer salary z. Trial is time consuming, hence will incur opportunity cost to the AUSA. Denote the opportunity cost by c_A. Hence, T is decreasing in c_A.

Furthermore, $P = pP_0$, where P_0 is the cost to the AUSA due to the punishment imposed by the USA when the AUSA is found to turn down a plea the USA would have wanted to accept, and p is the probability of the USA detecting such delinquency.[11]

U.S. Attorney

The utility the USA derives from an individual case prosecuted in the district is $u_U^0 = s - ic$, where s is the expected time the defendant spends in jail, c is the opportunity cost of a trial to the USA, and $i = 1$ if there is a trial. If the USA expects a trial to lead to a jail sentence of S with probability t, the USA wants the AUSA to accept any plea offer in which the defendant receives a sentence of at least $tS - c$. The utility of the USA derived from the prosecution of *all* cases is

[10] Following the discussion in Section 2, N is the number of AUSAs in the district, m is the number of supervisory AUSAs, and hence $N - m$ is the number of AUSAs who prosecute cases in the district. The non-negative parameter θ captures the increasing returns to specialization.

[11] To a large degree, the USA determines the salary raise, job assignments, and promotions of an AUSA. Further, if an AUSA wants to work in the private sector, the USA may write a letter of recommendation for the AUSA. The parameter P_0 represents the variety of ways in which a USA can punish an AUSA.

$u_U = (N - m)^{1+\theta} u_U^0$.

As long as $c > 0$ and $c_A > 0$, trial is costly for both USA and AUSA. Assume that the time spent on trial could be allocated to at least one other task that proves valuable for both. This implies that c_A is increasing in c.

3.2 Timing

Given that the potential agency problem in monitoring AUSAs studied here exists, we have $P_0 - T < c$. Throughout the paper, the parameters P_0, c, c_A, z, and T are taken to be fixed and known by all parties. The time structure in the game is as follows:

(1) USA chooses the level of monitoring, p, by allocating m AUSAs to supervisory positions, and reveals it to all parties;

(2) A case is filed, the severity S is known to all parties, but only AUSA learns the probability of conviction, t;

(3) Defendant makes a take-it-or-leave-it settlement offer x, where x denotes time in prison;

(4) AUSA accepts or rejects the offer: if AUSA accepts, the case is ended through plea bargain, otherwise, there is a trial;

(5) USA observes with probability p whether the AUSA has rejected a plea bargain that the USA would have accepted and punishes the AUSA when delinquency is detected.

Figure 1 summarizes the time structure.

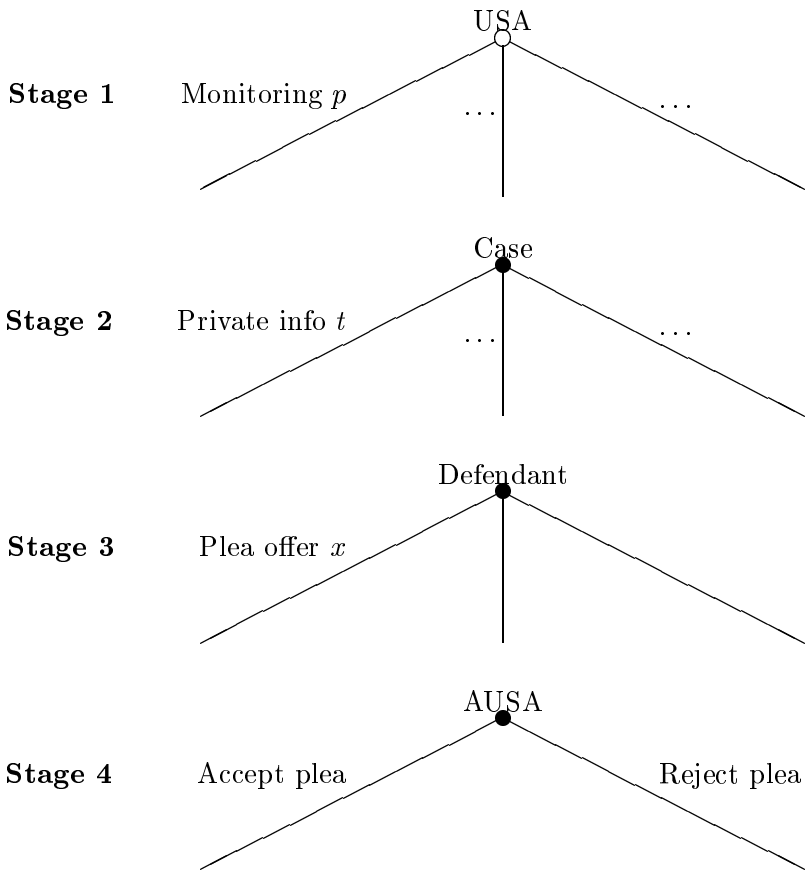

Figure 1: Model of monitoring and plea bargain. USA selects the probability p with which an AUSA who violates district policy is punished. For a particular case, AUSA has private information t over the strength of the case. Defendant makes a plea offer x that the AUSA either accepts or rejects.

3.3 Results

Proposition 1 below relates monitoring to plea rates. It establishes that higher monitoring level increases the probability of a plea. Further, cases prosecuted in an office with higher opportunity cost for trial are more likely to go to plea; cases prosecuted in an office where private lawyer salary is higher are more likely to go to trial; and more severe cases are more likely to go to trial.[12]

The following notation is introduced to state the proposition. The minimum offer acceptable to an AUSA of type t is given by $x(t) = tS - (P - T)$. Hence if the defendant offers a plea x, all AUSAs of type $t \leq \frac{x+P-T}{S} \equiv A$ accept the offer. The probability of a plea bargain settlement is thus given by $\rho = F(A)$. Further, the defendant's expected utility can be written as

$$U_D(A; p, T, S) = -F(A)x(A) - S\int_A^1 tf(t)dt,$$

where $x(A) = AS - (P - T)$.

Proposition 1 *Consider a subgame perfect equilibrium of the game where the AUSA accepts an offer if and only if* $t \leq A$. *If for all* p, T, *and* S, *the solution to* $\max_A U_D(A; p, T, S)$ *is regular and interior,*[13] *then the probability of a plea bargain* ρ *is increasing in* p, *increasing in* c, *decreasing in* z, *and decreasing in* S.

The Appendix contains the proof of Proposition 1 and provides sufficient conditions on f for the solution to the problem $max\ _A U_D(A; p, T, S)$ to be regular and interior (Lemma 1).

[12] We are most grateful to Jennifer Reinganum whose comments helped greatly improve this proposition.

[13] The solution $A^\star$ to $\max_A U_D(A; p, T, S)$ is regular and interior if $A^\star \in (0, 1)$ and $\frac{d^2U_D}{d^2A}|_{A=A^*} \neq 0$.

On the other hand, the USA's expected utility function can be written as

$$U(p;N) = E_{\{S,t\}}[u_U] = (N-m)^{1+\theta} E_{\{S,t\}}[u_U^0] = (N-m)^{1+\theta} u(p),$$

where E denotes the expectation operator, and $u(p) = \int_{\underline{S}}^{\bar{S}} \left[x(p)F(A(p)) + \int_{A(p)}^{1} (tS-c)f(t)dt \right] d\Pi(S)$. Hence, the results in Section 2 predict that both the level and the effectiveness of monitoring are increasing in size for small U.S. Attorney Offices and decreasing in size for large offices.[14] Proposition 1, therefore, predicts that the plea rate is increasing in office size for small offices and decreasing in size for large offices. The latter half of the paper tests this hypothesis empirically.

3.4 Discussion of assumptions and literature review

Before the hypotheses linking monitoring, plea rate and size are empirically tested, we provide some additional justification for the assumptions in the model and briefly review related literature in this section. Specifically, we discuss assumptions that AUSAs prefer trials more often than USAs, that prosecutors maximize the time the defendant spends in jail, that AUSAs have private information that is unverifiable during plea bargaining, that plea bargaining is a feasible option, and that the defendant (and not the AUSA) makes a take-it-or-leave-it settlement offer.

Nature of the agency problem

The nature of the agency problem is more controversial than the existence of the problem.[15] For instance, one often hears the argument that compared to USAs, AUSAs want to plea cases

[14] For the results in Section 2 to hold, it is required that $u'(p) > 0$ and $u''(p) < 0$. For an example in the context of plea bargaining where these conditions are satisfied, see the Appendix.

[15] For instance, see Schulhofer and Nagel [31].

more often because this will reduce their workload [18]. However, it is our belief that this is not the appropriate way of modeling the agents considered in this paper.

First, although the USA and AUSA both want to maximize the time that the defendant spends in jail, as the *de facto* regional head of Department of Justice, the USA is more concerned than the AUSA that trials take up prosecutorial resources that could be used more effectively in other cases, and is also more sensitive to the fact that Federal judges dislike drug trials (Little [20]).

Second, to acquire expertise in particular areas of prosecution requires going to trial [18], and this is particularly relevant to the AUSAs in our sample [25, page 155]).[16] In our paper, the cases studied are limited to drug trafficking cases that are not handled by the Organized Crime Drug Enforcement Task Force (OCDETF). Because the handling of such cases is an entry position in a USA office, these AUSAs are more likely to be interested in seeking trial experience. For these cases and the time period studied, the number of trials per prosecutor and per year is 3.14. The approximate number of trials necessary for a prosecutor to gain familiarity with such issues as jury selection, on the other hand, is 5 to 6 trials.[17] Hence, it takes on average two years for an AUSA to acquire most of the human capital available with such an entry position.

For these reasons, we argue that the nature of the agency problem is that AUSAs would like

[16] This fact is also confirmed during our personal communication with St. Louis Circuit Attorney, Dee Joyce-Hayes. Gifford [12] suggests a different motive for preferring trial. As an assistant prosecutor in Philadelphia commented: "When I get a case that looks interesting and I think I can win it, I don't want to encourage a guilty plea. I joined the district attorney's office so that I could try that kind of case to a jury."

[17] Personal communication with Lee Lawess, Federal Public Counsel.

to go to trial more often than the USA, and the seriousness of the agency problem will depend on the level of monitoring of the USA.

Glaeser, Kessler, and Piehl [13] examine the agency problem in terms of the cases prosecuted by Assistant U.S. Attorneys. In particular, they claim that AUSAs tend to prosecute high status individuals, because such prosecution increases their career capital and hence their returns in the private sector. High status individuals are taken to be white defendants because such defendants are more likely to be represented by private counsels. In order to be able to examine which cases are prosecuted at the federal vs. state level, Glaeser et al. examine the National Correction Reporting Program data. They find that the fraction of drug offenders that are white is higher in federal than state prisons. This bias is also found to be positively related to the number of AUSAs in a state. The explanation for this finding is that in large districts (i.e., with many AUSAs), the USA cannot supervise AUSAs as closely as in smaller districts.

While the National Correction Reporting Program data has the advantage of being able to provide information on which cases are prosecuted by AUSAs, it has some disadvantages in studying the effect of size on monitoring. First, only 35 states are part of the National Correction Reporting Program. Second, the data allows to identify Federal prisoners by their state and not by the USA district that prosecuted them.[18] Combining districts of a state is particularly problematic because of the particular way in which states are divided into districts. Glaeser et al measure supervision by adding the number of AUSAs in a state. This gives Alabama (divided in three districts of size:

[18]Even though the data set contains a variable for the county, for federal prisoners, the county is not known.

34, 14, 16) more AUSAs than Maryland (with only one district of size 50) even though each district in Alabama is considerably smaller than in Maryland.

Prosecutors maximize the time the defendant spends in jail

Just as in Glaeser et al [13], Landes [19], and Reinganum [29], the objective functions of the USA and the AUSA depend on the time that the defendant spends in jail. In Grossman and Katz [14] and Reinganum [28], the prosecutor also tries to minimize the likelihood of convicting an innocent defendant. Concern for the innocent defendant seems more applicable to state prosecution. Lynch states [30]: "De facto, in the real criminal justice system that operates in the U.S. Attorney's Office, there is not a presumption of innocence, there is a presumption of guilt." Similarly, Berlin [2] writes: "Prosecutors and law enforcement officials have incentives to obtain harsh sentences for offenders because of the adversarial nature of their jobs and because of public pressure to put criminals behind bars." While these statements may not be entirely accurate, they seem to be a reasonable approximation for the drug trafficking cases considered in this paper. For instance, if one looks at drug trafficking matters considered by AUSAs in fiscal year 1994, 1.3 % were disposed for lack of evidence of criminal intent, while 1 % were disposed on the basis that no federal offense was evident. For all other matters these numbers are 3 % and 2.1 %, respectively.

AUSAs have unverifiable private information

As discussed in McMunigal [22], a defendant's rights to discovery are much more limited during a plea than during a trial. For instance, in 1976 Sylvester Jones pleaded guilty to a crime because he did not know that four days prior, the only eye-witness had died. On sentencing, Jones, having

been informed of the death, was not allowed to withdraw the plea. The appeal by the defendant was denied because the Court found that "no prosecutor is obliged to share his appraisal of the weakness of his own cases (as opposed to specific exculpatory evidence) with defense counsel" (People v. Jones 44 N.Y.2d; 375 N.E.2d 41).

This implies that during plea bargaining, the AUSA has private information about the evidence that can be brought to trial. Specifically, in our model the AUSA is assumed to have private information on the probability of conviction when a case is brought to trial, t. The example above illustrates one type of the private information considered in this model: the identity and credibility of witnesses.[19]

The model also does not allow the AUSA to share the private information with the defendant. Note that the AUSA always has an incentive to claim to have the most incriminating unverifiable information (i.e., $t = 1$). Such a statement involves convincing the defendant that there are no weaknesses in the case that the defendant is not aware of. Given that a defendant's rights to discovery are much more limited during a plea than during a trial [22], it is our contention that such a statement is not verifiable at the plea bargaining stage. Since, in equilibrium, the defendant does not believe in unverifiable claims, it is without loss of generality that one assumes that unverifiable information is not revealed.[20]

[19] For different sources of private information examined by different authors in the plea bargaining literature, see Hay [16].

[20] This result does not hold if, as in Shavell [32], the defendant cannot tell whether the AUSA cannot convey information or chooses not to convey information.

Federal sentencing guidelines and plea bargaining

The sentencing guidelines were passed to ensure uniformity in sentencing. Note, however, that the guidelines do not preclude bargaining over drug sentences. As discussed in Schulhofer and Nagel [31], under the sentencing guidelines, AUSA can engage in charge bargaining, fact bargaining, and guideline factor bargaining.

Suppose an individual is charged with drug trafficking. Under charge bargaining, the drug trafficking charges are dismissed, but the defendant pleads guilty to lesser charges (such as simple possession or use of a communication facility involving drug trafficking).

Under fact bargaining, in exchange of a guilty plea, the AUSA makes a motion for reducing the sentence because of substantial assistance, even though the defendant did not assist the prosecutor at all.[21]

Under guideline factor bargaining, the plea agreement includes stipulations that yield predictable results under the guidelines.

Defendant makes settlement offer

Just as in Reinganum [29], this paper assumes that the defendant makes the plea offer. It is straightforward to solve the model where, just as in Reinganum [28], the AUSA – instead of the defendant – makes a take-it-or-leave-it offer.[22] The theoretical results used in the empirical section do not change. Specifically, a plea is more likely the higher the punishment to the AUSA in case

[21] Other ways to vary the sentence follow: putting in or leaving out from the indictment that a sale took place within 1000 feet of a school, including or excluding a gun count [8].

[22] As usual, proof of this assertion is available from authors.

of a trial and the lower the sentence for the defendant if convicted in a trial.

However, such a model requires assumptions more restrictive than the ones in the paper. First, since in equilibrium the defendant rejects (almost) all plea offers with positive probability, the insubordination is not as clear cut as in the model in the paper where the AUSA rejects a favorable plea bargain. Second, the equilibrium concepts must be stronger than subgame perfection to ensure a separating equilibrium.

4 Monitoring and Plea Bargaining: Data and Results

This section empirically studies differences across districts in the fraction of drug trafficking cases that are settled by a plea agreement. The analysis is restricted to individuals suspected of drug trafficking, because of their large number, the availability of better information on severity, and the fact that the agency problem discussed in this paper is specific to those cases.[23]

The case information is from the Central System and Central Charge files of the Executive Office of U.S. Attorneys (EOUSA) for fiscal years 1993, 1994, 1995, and 1996. The size information on the USA district offices was requested via Freedom of Information Act (FOIA) from the Executive Office of the United States Attorneys. The counsel information is from the Administrative Office of

[23] As shown in Section 4.1, for non-organized crime drug cases, the variables available in the EOUSA data files provide good measures of the severity of the crime, including the type and amount of drugs seized, whether a case involves multiple defendants, and the percentage of Organized Crime Drug Task Force cases. See Section 3.4 for a discussion of why the agency problem is specific to these cases.

U.S. Courts (AOUSC). The source of the biographical information on drug trafficking defendants is the U.S. Sentencing Commission (USSC) data files.

Two groups of defendants are studied in this paper. The first group (*Sample 1*) contains cases listed in both EOUSA and AOUSC data. The second group (*Sample 2*) contains cases that are included in all of the three data sets listed above and for which defendant biographical information is available. For cases included in the second group there is additional information, but the sample may be biased due to the fact that defendants receiving non-guilty verdicts are not included in the USSC data files.

Sample 1 includes 25076 cases and Sample 2 includes 15784. The main characteristics of these cases are shown in Tables 3 and 4. Compared to their counterparts from Sample 1, the defendants from Sample 2 on average receive a longer sentence, are involved with a larger amount of drugs, and are more likely to hire a private counsel. But these cases are less likely to involve multiple defendants. The differences, however, are all statistically insignificant.

Three empirical tests are included in this section. Section 4.1 explores the appropriateness of the proxies for case severity using a Tobit model. The empirical relationship between plea rate and district size is discussed in Section 4.2. Section 4.3 estimates the production function for monitoring.

4.1 Proxies for case severity

In order to compare plea rates, one needs to adjust for differences in case mix across districts. In the context of the model analyzed in this paper, case mix is measured by severity – the time a defendant spends in jail if convicted in a trial of the crime.

It is important to note that one cannot use information on the charges brought against the defendant to measure severity. Under a plea agreement, the prosecutor may file lower charges in exchange for a guilty plea. This results in a negative relationship between the severity of the charges and the likelihood of a plea, which needs to be distinguished from the relationship between the severity of the crime and the likelihood of a plea agreement found in the model. Further, the relationship between severity of charges and the severity of the crime varies across districts as a function of the cost of prosecution and the plea policies.

Similarly, realized sentence length cannot serve as a measure for case severity. Instead, the following variables that are available from the FJSRC dataset are used to control for case severity:

1. Weight of drugs. One of the main variables that affect the sentence is the amount of drugs in the case. The U.S. Sentencing Guidelines (USSG) provide tables that convert the severity of different drugs. For instance, one gram of cocaine equals 200 grams of marijuana. The equivalent amount of marijuana is further converted to its corresponding minimum sentence using the USSG conversion table.

2. Multiple defendants. The USSG provides higher penalties for multiple defendant cases

(see [33], §3B1.1C.).

3. Public counsel. The cost of private counsel for a federal drug trial is quite high. For this reason, for less severe offenses, even defendants with means will select public counsel.[24]

4. Percentage of cases that are Organized Crime Drug Enforcement Task Force (OCDETF) cases. Part of case severity is not observed from the data used here. For instance, whether a gun is involved in the case greatly affects the severity of the penalty but is not recorded in the EOUSA dataset. OCDETF targets high-level drug traffickers and large-scale money laundering operations [15]; hence the percentage of OCDETF cases provides information on the level of criminality in the district. In other words, a higher percentage of OCDETF cases reveals a higher average level of severity of drug cases prosecuted at the federal level in that district.

5. Biographical variables. Drug dealers, to minimize the risks in being caught by the police, tend to hire poor, illiterate youth as retailers and women to rent crack and stash-houses [34]. Hence, female, young, less educated individuals are less likely to be convicted for offenses that carry longer jail sentences (such as Continuing Criminal Enterprise cases). Non-white drug defendants are also more likely to carry a weapon, which is a significant factor in the prison sentence [21].

6. Dummy variables for the year when a case was received by the AUSA are included to take

[24] Personal communication with Lee Lawess, Federal Public Counsel.

into account the change in composition of cases across years.

The summary statistics of the proxies are found in Tables 2, 3, and 4 on pages 35 and 36.

As a crude test on how appropriate these variables are as proxies for case severity, some Tobit regressions are run with observed sentence length as the dependent variable and the variables discussed above as explanatories. The results in Table 5 on page 36 confirm that the variables discussed above provide reasonable measures of severity. Regression (1) examines Sample 1, while Regression (2) examines Sample 2.

4.2 District size and plea probability

In this subsection, a logistic regression is used to estimate the relationship between the probability with which the outcome of the prosecution is a plea (versus a jury trial verdict in District Court) and the monitoring level in that district. As discussed in Section 2, size affects monitoring level. Hence, for each year and district, the number of AUSAs (size) is used as a predictor for monitoring level. As discussed in Section 2 and Section 3.3, the effect of size on plea rates need not be monotonic. Specifically, the models predict that plea rates are increasing in size for small offices and decreasing in size for large offices (page 14). Hence both a linear term and a quadratic term of the size variable are included in the regressions. (The number of districts and years is not large enough for a non-parametric analysis.) Further, the proxies for case severity discussed in Section 4.1 are used to control for case mix. For trial cost, the number of cases per AUSA is used as a proxy.[25] Finally,

[25]The following argument provides one justification for using cases per AUSA as a proxy for trial cost. Assume there is some small drug trafficking case that the AUSA could be taking that would lead to n years in prison. The

the size of the largest law firm in the district is used as a measure for local private lawyer salary.[26]

Summarizing, the unit of observation is a defendant suspected of drug trafficking. A logistic regression estimates the likelihood with which the case is settled by a plea as a function of variables controlling for case mix, staffing, private lawyer salary, and monitoring abilities, and variables indicating the year. The results can be found on Table 6, page 37.

Regressions (1), (2) and (3) use Sample 1, while regressions (4) and (5) use Sample 2. The results of Regression (1) show that, as predicted in the model, higher severity leads to a lower likelihood of a plea. The proxies for severity that have an effect on the probability of plea bargain and are statistically significant at the 1% level are the following: weight of drugs, multiple defendants, public counsel, and percentage of OCDETF cases. Further, as predicted in the model, a higher cost of trial leads to a higher likelihood of a plea. The proxy for the cost of trial, cases per AUSA,

probability that the AUSA takes on the case, $P(x,\rho)$ is a decreasing function of both the number of cases an AUSA has to prosecute, x, and the percentage of trial cases, ρ. Furthermore, assume that $\frac{d^2P(x,\rho)}{dxd\rho} < 0$. Then, the marginal expected cost of going to trial, $-n\frac{dP(x,\rho)}{d\rho}$, is increasing in x.

[26]Boylan and Long [4] show that the size of the law firm is a good predictor for the salary level using a survey run in Chicago. For a theoretical model justifying the positive relationship between law firm size and lawyer salary, see Farrell and Scotchmer [10]. The law firm size information is from the online version of *Corporate Legal Times 1996* (http://www.Nexis-Lexis.com), which lists the number of attorneys in the U.S. for the largest 1000 law firms. We match the locations of the law firms with the federal districts and record the size of the largest law firm for each district. For 14 of the 89 districts in our study, there are no law firms ranked among the top 1000. For these districts, we use one-half the size of the smallest firm listed ($37/2 \approx 19$) as the size of the largest law firm in the district. Using different sizes for these districts (for instance, 1 or 37) does not change the results.

has a coefficient that is positive and statistically significant at the 1 % level. On the other hand, the coefficient of private salary is negative and significant at 1 % level, which is consistent with the prediction from the model that higher private lawyer salaries lead to higher trial rate. Year dummies all have positive and significant effects on plea rates. The sign and magnitude of their coefficients reflect the fact that a larger proportion of plea cases are included in the data set for later years due to the delay involved in reporting trial cases.

Finally, both size and $(\text{size})^2$ have effects on the probability of plea bargain that are statistically significant at the 1% level. Further, the signs of their coefficients confirm the prediction from the models. Specifically, plea rates are increasing in size for small offices and decreasing in size for large offices (page 14). Hence one concludes that the monitoring level is higher in average-sized districts than in small and large districts.[27]

One relevant concern for Regression (1) is that defendants involved in the same case or cases prosecuted in the same district-year might have correlated error terms that result from case effects or district-year effects not captured in the regression. Regression (2) and Regression (3) address this issue. Regression (2) considers the random effects of case and replicates all the significant results from Regression (1). Similarly, Regression (3) controls for district-year random effects and obtains significant effects for all the variables except for case load variable that has a p-value of 0.11.

Regressions (4) and (5) use Sample 2. Since biographical information is provided for these

[27] The results discussed above are robust when intra-case and intra-district correlations are corrected for.

observations, some additional proxies for severity are included in Regression (4). As predicted, cases involving young, white, females are more likely to be resolved by a plea agreement.[28] To test whether the results in Regression (4) follow from selection bias present in this sample, Regression (5) is run with the same regressors contained as in Regressions (1)–(3) but with Sample 2. Again, the results from both Regression (4) and Regression (5) are significant and of the predicted signs, except for the education variables.

4.3 District size and effectiveness

The results from Section 2 suggest that under very general conditions the monitoring technology exhibits increasing returns to scale for small districts, and decreasing returns to scale for large districts. This hypothesis is tested in this subsection.

We use the same logistic regressions run in the last subsection except that monitoring level is substituted with the product of monitoring expenditure per prosecuting attorney and monitoring effectiveness. Following the notation in Section 2, 'effective monitoring' is equal to the product of monitoring expenditure per prosecuting attorney, $\frac{m}{N-m}$, and the monitoring effectiveness, $e(m)$. To be able to test the relationship predicted in Result 1 (page 6) between m and $e(m)$, namely, $e(m)$ is increasing in m when m is small and decreasing in m when m is large, we include both a linear term and a quadratic term of m in the regressions. Again, variables controlling for case mix, case

[28] One change from the previous results in Regression (4) is that the private lawyer salary no longer has a significant effect ($p = 0.116$) on plea rates when ethnicity, gender, and age variables are included. It may be explained by the fact that the salary variable is correlated with these demographical variables.

load, and private lawyer salary are included.

The results from these regressions are shown in Table 7 on page 38. In Regression (1), observations from Sample 1 are used. Regression (2) studies case random effects, and Regression (3) studies district-year random effects. Regressions (4) and (5) use observations from Sample 2 with (4) containing biographical information and (5) excluding these variables. The results are very similar to those from the previous subsection.[29]

The results in Regression (1) should be interpreted as follows. Monitoring expenditure is $\frac{m}{N-m}$. The effectiveness of monitoring expenditure is $-3.6377+0.5403m-0.0166m^2$. Hence, average-sized districts can monitor more efficiently than larger or smaller districts. This is consistent with the prediction from the models (Result 1, page 6).

5 Conclusion

There are large differences in the fraction of cases resolved through plea bargaining across districts and over time. Table 1 on page 35 further shows that plea rates vary with the number of Assistant U.S. Attorneys in a district (denoted by 'size'); specifically, average-sized districts have the highest plea rate. This relationship between plea rate and size does not disappear when taking into consideration other variables that may be correlated with size (e.g., ethnic composition of defendants,

[29] All the results have significant coefficients for all the variables at the 10 % level, except caseload variable in Regression (3) ($p = 0.101$), salary variable in Regression (4), and education variables included in Regression (4). See Section 4.2 for interpretation on effects of variables other than size.

private lawyer salary, severity of crimes, and case load in the district).

A theoretical model explains the variation in plea rates as the result of differences in case mix, staffing level, outside career opportunities for prosecutors, and differences in supervision in U.S. Attorney districts. There are two main assumptions in the model. First, the AUSA has private information over the outcome of a trial. The private information is used to extract longer plea bargain sentences from the defendant, which leads some cases to go to trial. Second, the cost of going to trial is higher for the USA than for the AUSA. Hence, plea bargaining is more likely to fail if the USA cannot monitor the AUSA effectively. Although the bargaining model discussed in this paper is of a particular form, it is our belief that the results derive from the two assumptions discussed above, and hence will hold true in more general cases.[30]

Empirical estimates confirm the theoretical explanation. Specifically, defendants that are more likely to plea are individuals involved in cases with less severe sentences. So, a case involving a single defendant represented by public counsel where lower amount and less severe type of drugs is seized is less likely to go to trial.

If one believes that plea rates are positively correlated with monitoring, the low plea rates in large districts provide additional evidence that offices of large size find it more difficult to monitor their employees. In contrast to the existing literature, low plea rates for small districts imply

[30]See for instance Landes [19]. Nalebuff [24] constructs a game that gives the opposite result than the one in our paper. However, the results in [24] depend crucially on the assumption that, for some subset of the cases, the prosecutor receives a negative payoffs by going to trial. This assumption is not reasonable for the set of cases considered in this paper, see page 17.

that offices of small size do not monitor their employees as effectively as offices of medium size. Our explanation for this phenomenon is that small offices do not benefit from the gains from specialization in monitoring.

The paper discusses a model of monitoring that formalizes the tradeoff between gains from specialization and losses due to difficulty in coordination and proves that there will be a low level of monitoring in small and large offices. To the best of our knowledge, the model discussed in this paper is the first theoretical model to formalize this tradeoff and study its implications for monitoring effectiveness and monitoring level. Our empirical results, however, are consistent with the ones given in the sociology literature. For instance, Nolan [26] reports a non-monotonic U-shaped relationship between population size and the relative size of government that he explains by the interaction between economies of scale and increasing complexity in monitoring of organizations of various sizes. Blau [3] and Hendershot and James [17] show that there exist economies of scale in monitoring in their studies on government agencies and U.S. school districts.[31]

These results also have policy implications that are of particular importance given the rapid growth of the federal prosecution. Surprisingly, the size of a U.S. Attorney district varies greatly in ways that are hard to justify. Some states, such as Arizona, Colorado, and Massachusetts, have only one U.S. Attorney district. Other states, such as Arkansas and Iowa, are split into two different U.S. Attorney districts. Our findings indicate that redrawing the districts would lead to greater administrative efficiency.[32]

[31] See, however, Freeman [11] for a criticism of the methodology in some of these studies.

[32] The agency relationship between USAs and the Department of Justice is beyond the scope of this paper. However,

References

[1] Becker, G. S. and K. M. Murphy. 1992. The division of labor, coordination costs, and knowledge. *Quarterly Journal of Economics* 107(4):1137–1160.

[2] Berlin, E. P. 1993. Comment, the federal sentencing guidelines failure to eliminate sentencing disparity: Governmental manipulations before arrest. *Wisconsin Law Review* January/February:187–230.

[3] Blau, P. M. 1970. A formal theory of differentiation in organizations. *American Sociological Review* 35:201–218.

[4] Boylan, R. T. and C. X. Long. 1999. Incentives of public officials: an examination of United States Attorneys. Mimeo.

[5] Boylan, R. T. and C. X. Long. 2000. Firm size and optimal level of monitoring: a note, (downloadable from http://economics.wustl.edu/∼long/monitoring.pdf). Mimeo.

[6] Brown, C. and J. Medoff. 1989. The employer size-wage effect. *The Journal of Political Economy* 97(5):1027–1059.

[7] Burger, C. J. 1970. The state of the judiciary. *American Bar Association Journal* 56:929–931.

[8] Curtis, D. E. 1996. Legislating federal crime and its consequences. *The Annals of the American Academy of Political and Social Science* 543:85–96.

[9] Eisenstein, J. 1978. "Counsel for the United States: U.S. Attorneys in the political and legal systems." Baltimore: Johns Hopkins University.

[10] Farrell, J. and S. Scotchmer. 1988. Partnerships. *Quarterly Journal of Economics* 103(2):279–297.

[11] Freeman, J. H. 1973. Environment, technology, and the administrative intensity of manufacturing organizations. *American Sociological Review* 38:750–763.

[12] Gifford, D. G. 1983. Meaningful reform of plea bargaining: the control of prosecutorial discretion. *University of Illinois Law Review* 37:37–98.

[13] Glaeser, E. L., D. P. Kessler, and A. M. Piehl. 1998. What do prosecutors maximize? An analysis of drug offenders and concurrent jurisdiction. NBER Working Paper 6602.

[14] Grossman, G. M. and M. L. Katz. 1983. Plea bargaining and social welfare. *American Economic Review* 73:749–767.

[15] Guerra, S. 1995. The myth of dual sovereignty: multijurisdictional drug law enforcement and double jeopardy. *North Carolina Law Review* 73:1160–209.

[16] Hay, B. L. 1995. Effort, information, settlement, trial. *Journal of Legal Studies* 24:29–62.

[17] Hendershot, G. E. and T. F. James. 1972. Size and growth as determinants of administrative-production ratios in organizations. *American Sociological Review* 37:149–153.

in considering the optimal size of a district one should keep in mind the main finding in Eisenstein's [9] seminal work on U.S. Attorneys, that USA in large districts act more independently than in smaller districts.

[18] Hollander-Blumoff, R. 1997. Getting to "guilty": Plea bargaining as negotiation. *Harvard Negotiation Law Review* 2:115–146.

[19] Landes, W. M. 1971. An economic analysis of the courts. *Journal of Legal Studies* 14:61–107.

[20] Little, R. K. 1995. Myths and principles of federalization. *Hastings Law Journal* 46:1029–85.

[21] McDonald, D. C. and K. E. Carlson. 1993. "Sentencing in the Federal Courts: Does Race Matter?" Washington, D.C.: Bureau of Justice Statistics.

[22] McMunigal, K. C. 1989. Disclosure and accuracy in the guilty plea process. *Hastings Law Journal* 40:957–1029.

[23] Miller, G. P. 1987. Some agency problems in settlement. *Journal of Legal Studies* 16:189–215.

[24] Nalebuff, B. 1987. Credible pretrial negotiation. *Rand Journal of Economics* 18(2):198–210.

[25] Nelson, R. L. 1988. "Partners with Power: The transformation of the Large Law Firm." Berkeley: University of California Press.

[26] Nolan, P. D. 1979. Size and administrative intensity in nations. *American Sociological Review* 44:110–125.

[27] Polinsky, A. M. and S. Shavell. 1999. On the disutility and discounting of imprisonment and the theory of deterrence. *Journal of Legal Studies* 28:1–16.

[28] Reinganum, J. F. 1988. Plea bargaining and prosecutorial discretion. *American Economic Review* 78:713–728.

[29] Reinganum, J. F. 1998. Sentencing guidelines, judicial discretion and plea bargaining. Mimeo.

[30] Richman, D. C. 1999. Panel discussion: the expanding prosecutorial role from trial counsel to investigator and administrator. *Fordham Urb. Law Journal* 26:679–702.

[31] Schulhofer, S. J. and I. H. Nagel. 1997. Symposium: the federal sentencing guidelines: ten years later: plea negotiations under the federal sentencing guidelines: guideline circumvention and its dynamics in the post-Mistretta period. *Northwestern University Law Review* 91:1284–316.

[32] Shavell, S. 1989. Sharing of information prior to settlement or litigation. *RAND Journal of Economics* 20:183–195.

[33] United States Sentencing Commission. 1994. "Federal Sentencing Guideline Manual." St. Paul, Minn: West Pub. Co.

[34] United States Sentencing Commission. 1995. "Cocaine and Federal Sentencing Policy." Washington, D.C.: Special Report to Congress.

Appendix: Proofs and Tables

This section contains the proofs of the results in the paper and the tables.

Proposition 1 *Consider a subgame perfect equilibrium of the game where the AUSA accepts an offer if and only if $t \leq A$. If for all p, T, and S, the solution to $\max_A U_D(A; p, T, S)$ is regular and interior, then the probability of a plea bargain, ρ, is increasing in p, increasing in c, decreasing in z, and decreasing in S.*

Proof: The defendant chooses A to maximize $U_D(A; p, T, S)$ where

$$U_D(A; p, T, S) = -F(A)x(A) - S\int_A^1 tf(t)dt.$$

At the optimum,

$$\begin{aligned}\frac{dU_D}{dA} &= -F(A)S - f(A)[AS - (P-T)] + ASf(A) \\ &= -F(A)S + f(A)(P-T) \\ &= 0.\end{aligned}$$

Let $G(A, p, T, S) = -F(A)S + f(A)(P-T)$. Then, $\frac{\partial G}{\partial A} = \frac{d^2U_D}{dA^2} < 0$, $\frac{\partial G}{\partial p} = f(A)P_0 > 0$, $\frac{\partial G}{\partial T} = -f(A) < 0$, and $\frac{\partial G}{\partial S} = -F(A) < 0$, where the first inequality holds since the solution is regular and interior.

By the implicit function theorem, $\frac{dA}{dp} = -\frac{\frac{\partial G}{\partial p}}{\frac{\partial G}{\partial A}} > 0$, $\frac{dA}{dT} = -\frac{\frac{\partial G}{\partial T}}{\frac{\partial G}{\partial A}} < 0$, and $\frac{dA}{dS} = -\frac{\frac{\partial G}{\partial S}}{\frac{\partial G}{\partial A}} < 0$.

Hence,

$$\frac{d\rho}{dp} = \frac{d\rho}{dA}\frac{dA}{dp} = f(A)\frac{dA}{dp} > 0,$$

$$\frac{d\rho}{dc} = \frac{d\rho}{dA}\frac{dA}{dT}\frac{dT}{dc} = f(A)\frac{dA}{dT}\frac{dT}{dc} > 0,$$

$$\frac{d\rho}{dz} = \frac{d\rho}{dA}\frac{dA}{dT}\frac{dT}{dz} = f(A)\frac{dA}{dT}\frac{dT}{dz} < 0,$$

and,

$$\frac{d\rho}{dS} = \frac{d\rho}{dA}\frac{dA}{dS} = f(A)\frac{dA}{dS} < 0.$$

■

Lemma 1 *Let $f(1) < \frac{S}{P-T} \leq \frac{f(\frac{P-T}{S})}{F(\frac{P-T}{S})}$, and let $\frac{f(t)}{F(t)}$ be a monotone decreasing function. Then, the solution to $\max_A U_D(A; p, T, S)$ is regular and interior.*

Proof: It is easily seen from the assumptions that $0 < \frac{P-T}{S} < 1$.

Fix p, T, and S, from the defendant's optimization problem,

$$\frac{dU_D}{dA} = -F(A)S + f(A)(P-T).$$

At $A = \frac{P-T}{S}$, $\frac{dU_D}{dA} = -F(\frac{P-T}{S})S + f(\frac{P-T}{S})(P-T) \geq 0$, since $\frac{f(\frac{P-T}{S})}{F(\frac{P-T}{S})} \geq \frac{S}{P-T}$. At $A = 1$, $\frac{dU_D}{dA} = -S + f(1)(P-T) < 0$, since $f(1) < \frac{S}{P-T}$. Hence, $\frac{f(A)}{F(A)}$ being strictly decreasing implies that there is a unique solution $A^* \in \left[\frac{P-T}{S}, 1\right)$ and thus a unique $x^* \geq 0$ to the equation defined by the first order condition. ∎

Remark 1 *Since $\frac{f(t)}{F(t)}$ is a decreasing function of the hazard rate $\frac{f(t)}{1-F(t)}$, the monotonicity condition required on $\frac{f(t)}{F(t)}$ in Lemma 1 is equivalent to requiring an increasing hazard function. The condition $f(1) < \frac{S}{P-T}$ is needed to guarantee that there are cases where the defendant would like to go to trial, while $\frac{f(\frac{P-T}{S})}{F(\frac{P-T}{S})} \geq \frac{S}{P-T}$ is necessary to ensure that the defendant will make only non-negative offers in plea bargaining, which is consistent with the fact that drug trafficking cases are severe cases.*

Example 1 (Example where $u'(p) > 0$ and $u''(p) < 0$):

Assume that the probability of a case being convicted in trial, t, is uniformly distributed in $[0,1]$. Hence $f(t) = 1$ and $F(t) = t$. From the proof of Proposition 1, $A = \frac{P-T}{S}$ and $x = AS - (P-T)$. Hence, $x = 0$, and,

$$u(p) = \int_{\underline{S}}^{\bar{S}} \Big[\int_{\frac{P-T}{S}}^{1} (tS - c)dt \Big] d\Pi(S).$$

Therefore,

$$\begin{aligned} u'(p) &= -\int_{\underline{S}}^{\bar{S}} (P - T - c)\frac{P_0}{S} d\Pi(S) \\ &= -P_0(P - T - c)\int_{\underline{S}}^{\bar{S}} \frac{1}{S} d\Pi(S) \\ &> 0, \end{aligned}$$

since $P - T < P_0 - T < c$.

Further, $u''(p) = -P_0^2 \int_{\underline{S}}^{\bar{S}} \frac{1}{S} d\Pi(S) < 0$.

Table 1: Differences in plea rates across districts

Group	Plea rate	Average Size
Small $[9, 70]$	0.88462	27.678
Medium $[71, 140]$	0.92308	95.333
Large $[141, 219]$	0.86667	187.942

Note: The breakpoints 70 and 140 are chosen to roughly divide the range of district size into three equal intervals. The pattern shown here is robust when different breakpoints are used.

Table 2: District variables used in regressions.

Variable	Source	Range	Mean	σ	Observations
District size	FOIA	[9.11, 219.47]	44.454	41.239	353
Monitoring expenditure per prosecutor	EOUSA	[0.033, 0.336]	0.128	0.045	353
Cases per AUSA	FOIA	[0.617, 130.556]	22.2	21.438	353
% OCDETF	EOUSA	[0, 0.949]	0.254	0.227	353
Size of largest law firm	Nexis-Lexis	[43, 1053]	226.5	216.4	75

Note: The number of observations is 353 (instead of 356) since in this sample there are no (non-OCDETF) drug trafficking cases for Idaho in 1993 and 1996 and for Vermont in 1996. "Monitoring Expenditure per prosecutor" is defined as the number of prosecutor hours in management and administration divided by the number of prosecutor hours in prosecution in a district (total number of prosecutor hours - number of prosecutor hours in management and administration).

Table 3: Defendant variables used in regressions (Sample 1).

Variable	Source	Range	Mean	σ	Observations
Plea (vs jury trial)	EOUSA	0, Plea=1	0.904	0.294	25076
Prison time (months)	EOUSA	[0, 720]	69.869	81.300	25076
Weight of drugs (months in prison)	EOUSA	[0, 235]	18.864	41.510	25076
Multiple defendants	EOUSA	0, Multi=1	0.849	0.358	25076
Public Counsel	AOUSC	0, Public=1	0.322	0.467	25076
Case received in 1994	EOUSA	0, 1994=1	0.317	0.465	25076
Case received in 1995	EOUSA	0, 1995=1	0.333	0.471	25076
Case received in 1996	EOUSA	0, 1996=1	0.109	0.312	25076

Table 4: Defendant variables used in regressions (Sample 2).

Variable	Source	Range	Mean	σ	Observations
Plea (vs jury trial)	EOUSA	0, Plea=1	0.907	0.290	15784
Prison time (months)	EOUSA	[0, 720]	71.589	81.627	15784
Weight of drugs (months in prison)	EOUSA	[0, 235]	19.149	42.302	15784
Multiple defendants	EOUSA	0, Multi=1	0.819	0.385	15784
Public Counsel	AOUSC	0, Public=1	0.3	0.458	15784
White defendant	USSC	0, White=1	0.603	0.489	15784
Male defendant	USSC	0, Male=1	0.865	0.342	15784
Age of defendant	USSC	[18, 78]	32.795	9.692	15784
Years of education	USSC	[0, 20]	10.511	3.032	15784
Case received in 1994	EOUSA	0, 1994=1	0.3	0.458	15784
Case received in 1995	EOUSA	0, 1995=1	0.48	0.5	15784
Case received in 1996	EOUSA	0, 1996=1	0.156	0.363	15784

EOUSA denotes Executive Office of United States Attorneys. AOUSC denotes Administrative Office of U.S. Courts. USSC denotes U.S. Sentencing Commission. FOIA denotes Freedom of Information Act request.

Table 5: Tobit regression for the number of months in jail.

Regression	(1)	(2)
Weight of Drugs	$0.2652_{0.0001}$	$0.2282_{0.0001}$
Multiple defendants	$19.3559_{0.0001}$	$16.1408_{0.0001}$
Public Counsel	$-13.4412_{0.0001}$	$-8.8321_{0.0135}$
White defendant		$-48.0001_{0.0001}$
Male defendant		$55.7611_{0.0001}$
Age of defendant		$0.6308_{0.0002}$
Years of Education		$3.4685_{0.0814}$
(Years of Education)^2		$-0.2604_{0.0120}$
Fraction of OCDETF cases	$91.0206_{0.0001}$	$76.9649_{0.0001}$
Case received in 1994	$-4.4530_{0.2065}$	$-22.4329_{0.0012}$
Case received in 1995	$-1.2919_{0.7161}$	$-21.2225_{0.0016}$
Case received in 1996	$-3.4473_{0.4834}$	$-21.2407_{0.0055}$
Intercept	$29.9479_{0.0001}$	$11.6811_{0.4480}$
Number of Observations	25076	15784

Each cell contains the coefficient of the regression and the P-value.

Table 6: Probability with which a case leads to plea agreement.

Regression	(1)	(2)	(3)	(4)	(5)
District size	$0.01318_{0.0001}$	$0.02157_{0.0001}$	$0.01122_{0.0001}$	$0.00932_{0.0001}$	$0.01300_{0.0001}$
(District size)^2	$-0.00006_{0.0001}$	$-0.00009_{0.0001}$	$-0.00005_{0.0001}$	$-0.00004_{0.0001}$	$-0.00005_{0.0001}$
Weight of drugs	$-0.00411_{0.0001}$	$-0.00654_{0.0001}$	$-0.00437_{0.0001}$	$-0.00403_{0.0001}$	$-0.00399_{0.0001}$
Multiple defendants	$-0.45972_{0.0001}$	$-0.60619_{0.0001}$	$-0.43061_{0.0001}$	$-0.446460_{0.0001}$	$-0.45396_{0.0001}$
% OCDETF	$-0.61864_{0.0001}$	$-1.13801_{0.0001}$	$-0.69919_{0.001}$	$-0.40780_{0.020}$	$-0.53216_{0.0020}$
Public Counsel	$0.18391_{0.0001}$	$0.32678_{0.0001}$	$0.13792_{0.0090}$	$0.23633_{0.0001}$	$0.25113_{0.0001}$
White defendant				$0.64126_{0.0001}$	
Male defendant				$-0.32394_{0.0001}$	
Age of defendant				$-0.01434_{0.0001}$	
Years of education				$0.04886_{0.1560}$	
(Years of education)^2				$-0.00155_{0.387}$	
Cases per AUSA	$0.00534_{0.0001}$	$0.00754_{0.0020}$	$0.00350_{0.110}$	$0.00332_{0.018}$	$0.00352_{0.012}$
Private lawyer salary	$-0.00061_{0.0001}$	$-0.00111_{0.0001}$	$-0.00060_{0.006}$	$-0.00023_{0.116}$	$-0.00044_{0.002}$
1994	$0.15139_{0.0080}$	$0.21266_{0.0410}$	$0.17393_{0.128}$	$0.24009_{0.021}$	$0.23404_{0.023}$
1995	$0.26106_{0.0001}$	$0.37313_{0.001}$	$0.27585_{0.021}$	$0.46294_{0.0001}$	$0.45597_{0.0001}$
1996	$0.62085_{0.0001}$	$0.98150_{0.0001}$	$0.66096_{0.0001}$	$0.72661_{0.0001}$	$0.74029_{0.0001}$
Intercept	$2.12374_{0.0001}$	$3.80239_{0.0001}$	$2.38398_{0.0001}$	$2.08945_{0.0001}$	$1.89685_{0.0001}$
No. of Observations	25076	25076	25076	15784	15784

Each cell contains the coefficient and the P-value.

Table 7: Monitoring and plea probability.

Regression	(1)	(2)	(3)	(4)	(5)
$\frac{m}{n-m}$	$-3.6377_{0.0001}$	$-5.96986_{0.0001}$	$-3.63005_{0.0040}$	$-3.96057_{0.0001}$	$-4.84209_{0.0001}$
$m\frac{m}{n-m}$	$0.54033_{0.0001}$	$0.85683_{0.0001}$	$0.47984_{0.0040}$	$0.28882_{0.0690}$	$0.58748_{0.0001}$
$m^2\frac{m}{n-m}$	$-0.01661_{0.0001}$	$-0.0261_{0.0001}$	$-0.01314_{0.0110}$	$-0.00989_{0.040}$	$-0.01770_{0.0001}$
Weight of drugs	$-0.00399_{0.0001}$	$-0.00644_{0.0001}$	$-0.00440_{0.0001}$	$-0.00402_{0.0001}$	$-0.00400_{0.0001}$
Multiple defendants	$-0.46426_{0.0001}$	$-0.62075_{0.0001}$	$-0.43603_{0.0001}$	$-0.48200_{0.0001}$	$-0.47029_{0.0001}$
% OCDETF	$-0.76220_{0.0001}$	$-1.39919_{0.0001}$	$-0.66037_{0.0010}$	$-0.48913_{0.005}$	$-0.63694_{0.0001}$
Public Counsel	$0.19041_{0.0001}$	$0.33933_{0.0001}$	$0.14246_{0.0060}$	$0.24541_{0.0001}$	$0.26446_{0.0001}$
White defendant				$0.64046_{0.0001}$	
Male defendant				$-0.32783_{0.0001}$	
Age of defendant				$-0.01450_{0.0001}$	
Years of education				$0.04929_{0.153}$	
(Years of education)^2				$-0.00158_{0.379}$	
Cases per AUSA	$0.00573_{0.0001}$	$0.00829_{0.0010}$	$0.00379_{0.1010}$	$0.00315_{0.027}$	$0.00362_{0.011}$
Private lawyer salary	$-0.00057_{0.0001}$	$-0.00103_{0.0001}$	$-0.00077_{0.0001}$	$-0.00013_{0.411}$	$-0.00040_{0.011}$
1994	$0.14793_{0.0010}$	$0.20755_{0.0530}$	$0.20161_{0.050}$	$0.25503_{0.015}$	$0.23608_{0.023}$
1995	$0.27525_{0.0001}$	$0.40137_{0.0010}$	$0.32083_{0.006}$	$0.52897_{0.0001}$	$0.49781_{0.0001}$
1996	$0.62727_{0.0001}$	$0.99752_{0.0001}$	$0.66677_{0.0001}$	$0.76907_{0.0001}$	$0.76471_{0.0001}$
Intercept	$2.76026_{0.0001}$	$4.86991_{0.0001}$	$2.95339_{0.0001}$	$2.73219_{0.0001}$	$2.66359_{0.0001}$
No. of Observations	25076	25076	25076	15784	15784

Each cell contains the coefficient and the P-value.