The Emergent Organization of Plea Bargaining

John F. Padgett
University of Chicago

This article analyzes four systems of plea bargaining within American criminal courts. These systems are examples of centralized authority, centralized contract, delegated authority, and delegated contract. The guiding questions are: Why do informal role structures emerge in organizations? How are such role adaptations shaped by the formal structures they subvert? Stochastic models based on the existing research literature are developed to derive numerous empirical predictions about why different U.S. cities have different plea bargaining systems.

Ever since the Hawthorne experiments, the concept of informal organization has been a central heuristic in organization theory. Planned formal control is one thing, the argument goes, but the aggregate behavior of organizations cannot be understood without taking into account the localized subversions of individuals pursuing their own self-interest. If such tactical maneuverings cumulate into a patterned set of reinforcing expectations that are continually reenacted across diverse circumstances, we may speak of an informal role structure that emerges crescively from within the constraints of a formal organization.

Plea bargaining in American felony courts is an almost ideal example of this general phenomenon. No one consciously designed plea bargaining, yet it is both widespread and stable. With some geographical variation, roughly 90% of felony convictions today are obtained through pleas of guilty rather than through trial (Newman 1966; Miller, McDonald, and Cramer 1978). Plea bargaining was not declared legal by the Supreme

1 I am indebted to Alberm W. Alschuler for his detailed comments on an earlier draft of this paper which built directly on his seminal contributions. I would also like to thank Robert Axelrod, Michael Cohen, Wendy Griswold, Anfis Hagee, Mary Vogel, and Harrison White for their helpful comments. Steve Garry did the programming and graphics. The research was supported by grant SES80-08658 from the National Science Foundation. Requests for reprints should be sent to John F. Padgett, 502 Pick Hall, Departments of Political Science and Sociology, University of Chicago, Chicago, Illinois 60637.

© 1985 by The University of Chicago. All rights reserved.
0002-9602/85/0604-0003$01.50
Court until 1970, yet in most of the jurisdictions that have been studied to date, plea bargaining in one form or another has been dominant since the early 20th century (Moley 1929; Heumann 1978; Friedman 1979).

What makes plea bargaining particularly intriguing from an informal role structure perspective is the fact that relationships created through interaction in plea bargaining are not uniform in spite of the presumed constancy of participant motivations. There are four distinct "pure" forms of plea bargaining, each of which is analyzed below: implicit plea bargaining, judicial plea bargaining, charge reduction plea bargaining, and sentence recommendation plea bargaining. These alternative informal role structures both represent different allocations of discretion and benefit and reflect distinct behavioral frames within which participants interpret and socially construct one another. Different forms of plea bargaining predominate in the criminal courts of different American cities (Alschuler 1968, 1976; Miller et al. 1978). The empirical question that motivates this paper is, Why do different cities have different plea bargaining systems? Stochastic models of the four types of plea bargaining will be derived, and their aggregate properties will be analyzed under varying environmental conditions.

Unfortunately, within the extant organization theory literature, determinants of alternative informal role structures have not been investigated with the same systematic care that has characterized the study of formal organizational structure. On the side of formal structure, we have a full menu of contemporary theoretical approaches to choose from—structural theory (e.g., Blau and Schoenherr 1971), contingency theory (e.g., Lawrence and Lorsch 1967), institutional theory (e.g., Meyer and Rowan 1977), population ecology (e.g., Hannan and Freeman 1977), and resource dependency (e.g., Pfeffer and Salancik 1978). With the possible exception of the work of Crozier (1964, 1976) and Williamson (1975), however, analyses of behavioral process in organizations (e.g., March and Simon 1958; Cyert and March 1963) have not been assembled into comparative structural conclusions.

The primary difficulty in developing comparative analyses of informal role structures in organizations lies in the necessary merger of two traditionally distinct levels of analysis—the more micro utilitarian level of analysis and the more aggregate functionalist level of analysis. Virtually all researchers of plea bargaining would agree that the utilitarian self-interests of judges, prosecutors, defense attorneys, and defendants are the crucial starting points for any analysis of plea bargaining. The stochastic models below illustrate how the sentencing terms of exchange in a felony court are established in equilibrium through an aggregation of strategic expectations and calculations at this individual level of analysis. Taken by themselves, however, participant self-interests even in interaction are

754
Plea Bargaining

insufficient for deriving organizational variation in plea bargaining systems. The essence of the concept of a role structure is not the terms of exchange but rather the terms within which exchange is defined. The models of alternative role structures will differ not in assumptions about individual self-interest but in assumptions about how individuals construe one another.

These models will be used to address essentially functionalist questions about adaptation of organizational structure to environmental constraints. Because the role structure in question is informal organization, however, a reinterpretation both of the mechanism of adaptation and of the concept of "environment" is required.

Traditional functionalist mechanisms for explaining organizational adaptation are irrelevant to plea bargaining. Differential death rates are irrelevant because courts do not die, and extrinsic legitimacy is irrelevant because plea bargaining has been illegitimate throughout most of its history. Environmental constraint in plea bargaining does not imply a set of fixed behaviors necessary for organizational survival. Instead, what is problematic in plea bargaining is the normative frame through which judges and prosecutors interpret their own actions. The problem of substantive versus formal justice exists in the first place, I will argue, because there is a contradiction between sentencing and procedural norms. Substantive justice implies a preoccupation with sentencing results rather than with the procedure of decision making. On strictly normative grounds, therefore, I will hypothesize that the microdynamics of any individual utilitarian plea bargaining "game" are embedded in a higher-order struggle to preserve the integrity of internalized sentencing standards in the face of guilty plea pressure to abandon them. In particular, "sentence discounts" measure in equilibrium the average deviation of plea bargained sentences from culturally variable sentencing ideals.

Likewise, the concept of environment requires modification when the dependent variable is informal role structure. From the perspective of informal organization, the "environment" includes the formal structure of the organization itself as well as an array of exogenous constraints. The concept of formal structure is used in this paper in two distinct ways. On the one hand, formal structure means the official and externally imposed control system of criminal courts—not only the organization chart of judge, prosecutor, defense attorney, and defendant with their respectively prescribed duties and rights, but also the legal apparatus of sentencing codes, evidentiary rules, and constitutional limitations. On the other hand, formal structure means the officially sanctioned temporal sequence of adjudicatory procedure. Temporal sequence is the behavioral reflection of certain procedural norms. The form of plea bargaining that emerges in a criminal court, I will argue, is not independent of the formal
structure it subverts. Plea bargaining is a systematic transformation in the temporal formal structure caused by problems induced by the control formal structure. Variation in control structure and in exogenous constraints shapes the distinctive problems to be overcome and thereby the informal adaptations that are provoked.

Thus, although utilitarian self-interest and functional adaptation establish the outer bounds of this analysis of plea bargaining, the core of the approach to informal role structure in this article is a feedback between two levels of normative interpretation—the role terms within which individual defendants are construed, and the sentencing terms within which an aggregate pool of defendants is construed. Informal role interpretations of defendants by judges and prosecutors shape the aggregate sentence discount schedules that emerge in equilibrium, and substantive justice sentencing problems experienced by judges and prosecutors shape the informal role interpretations they are willing to adopt. Within this perspective, the task is to demonstrate how this informal social feedback process is affected by structural contradictions within the formal structure, as these play themselves out under varying "environmental" conditions.

The arguments in this paper will be developed in a traditional format: description, theoretical background, models, analysis, and discussion.

Alternative Structures of Plea Bargaining

Four types of plea bargaining within criminal courts at the felony level have been identified in the literature (see, e.g., Alschuler 1968, 1975, 1976):

1. In implicit plea bargaining, the defendant simply "throws himself on the mercy of the court" by pleading guilty to the original charge under the expectation of receiving a more lenient sentence thereby.

2. In charge reduction plea bargaining, the prosecutor downgrades or eliminates charges in exchange for a guilty plea to the reduced charge(s).

3. In judicial plea bargaining, the judge, after consultation with prosecutor and defense counsel in conference, offers the defendant a specific guilty plea sentence.

4. In sentence recommendation plea bargaining, the prosecutor in exchange for a guilty plea recommends a particular disposition to the judge, who then (usually) imposes the sentence recommended. One or another of these "pure" types usually predominates within a single court, but combinations (particularly of 2 and 4) also are frequently observed (see Miller et al. 1978).

Implicit plea bargaining is the theoretical baseline because it represents no change in officially sanctioned pretrial role behaviors. No overt com-
Plea Bargaining

Communication takes place between the defendant (or his attorney) and any court official before the plea. The judge allocates punishment only after the issue of guilt has been decided. The prosecutor does not participate in sentencing, either directly or indirectly.

These official roles, which remain unchanged in implicit plea bargaining, reflect two procedural norms that together imply strict, temporal segregation between the trial and the sentence hearing. The judge maintains a decorous above-it-all stance because the defendant should not be punished until after the issue of guilt is resolved. And the prosecutor's subjective evaluation of the evidence should not impinge on final sentencing because the defendant categorically is either guilty or not.

The reason a defendant may plead guilty in the implicit system is the sentence discount—a differential in expected or average sentence between conviction by plea and conviction by trial. Implicit plea bargaining confronts the defendant with two lotteries. The defendant is asked to give up his probability of acquittal at trial in exchange for a sentence lottery with lower mean. Because no sentence outcome is guaranteed in this system, the defendant risks being sorely disappointed at sentence hearing.

Charge reduction plea bargaining alters the role of the prosecutor but not that of the judge. Through the discretionary evaluation of case evidence as a labeled crime, the prosecutor acts as an advisor who frames both the sentencing alternatives and the legal context within which the judge can act. The judge still makes an independent sentencing decision after the plea, within a legal framework which has been manipulated by the prosecutor. In charge reduction plea bargaining, the defendant also is confronted with two sentence lotteries—a trial lottery based on the original charge with some probability of acquittal, and a guilty plea lottery based on the reduced charge with certain conviction. Besides sentence discount differences in means, an added benefit is reduced risk in the limited form of lower statutory maxima.

Judicial plea bargaining alters the role of the judge but not that of the prosecutor. The judge descends from his or her august stance to intervene actively with the defendant by selecting a tentative sentence before the plea. This judicial action is usually presented thus: “If the defendant is really guilty, and if the facts are as I currently understand them, then my sentence would most likely be this.” The prosecutor and defense counsel in this structure act primarily as couriers who lay out before the judge in offsettingly biased presentations the relevant sentencing information, which includes facts about the defendant’s background and the circumstances of the crime. Holding aside the possibility of a judge’s reneging on his or her word, the defendant in judicial plea bargaining is presented not with a guilty plea lottery in sentences but instead with certain knowledge about the consequences of a guilty plea.
In sentence recommendation plea bargaining, the prosecutor takes over directly the judge's sentencing functions. The judge and the prosecutor are in a principal-agent relation, with the judge's role being the comparatively passive one of assuring that the prosecutor's decisions are consistent with the judge's own sentencing policies. The justifications for a judge's acceding to most of the explicit suggestions of the prosecutor are that, in the absence of trial, the prosecutor is closer to the facts of the case and that the prosecutor's sentencing standards are identical in any event. If the judge always rubber stamps the recommendations, the defendant has certain knowledge about the consequences of pleading guilty. In sentence recommendation plea bargaining, both the discretionary evaluation of case evidence and the discretionary judgment about individual sentencing have been delegated to the prosecutor.

These four plea bargaining structures emerge through transposing temporally, at an informal level, the officially prescribed sequence of procedural adjudication. In effect, charge reduction involves folding the trial but not the sentence hearing back across the boundary of the plea. Matters of conviction on a charge are explicitly discussed and settled, even though the resulting sentence is left open. Judicial plea bargaining in effect folds the sentence hearing but not the trial back across the boundary of the plea. Strength of state's evidence (i.e., probability of trial conviction) is not systematically assessed because the judge in an American framework does not have direct access to such evidence before the plea, but provisional sentence is meted out nonetheless under the tentative assumption that the defendant is guilty as charged. Implicit plea bargaining folds neither stage back across the plea, relying instead on tacit global cues. And sentence recommendation folds both stages back across the plea, since the prosecutor both evaluates his or her own evidence and makes sentence commitments for the judge.

These temporal inversions imply quite distinct informational bases within which defendants are construed. Charge reduction involves interpreting the defendant as a case. Before the plea, evidentiary considerations are primary, with historical and contextual background of the defendant relegated to a later evaluation by the judge. Judicial plea bargaining involves interpreting the defendant as a criminal. The historical and contextual circumstances of the person take precedence, with matters of factual and legal guilt left open to the discretionary judgment of the defendant. Implicit plea bargaining before the plea is blind to all idiosyncrasies of individual defendants: the judge perceives only an undifferentiated pool of defendants who clear themselves voluntarily in response to aggregate rewards. Sentence recommendation involves the most complete informational base, differentiating pre-plea defendants
both in terms of evidence as cases and in terms of personal background as criminals.

The structure of role relationships among plea bargaining participants follows directly from these underlying perceptual constructions. In the boundary relationship between defendants and court officials, what is at stake in the perception of pre-plea defendants as individually distinct criminals is the ability of the system to reach out with explicit sentence commitments. Judicial and sentence recommendation plea bargaining are structured as contract relations between the defendant and the judge, either directly or through agency, respectively. Interdependent actions of direct consequence to each party (in this case, sentence for the defendant and guilty plea for the judge or prosecutor) are “presentiated,” or spelled out in advance of commitment by either party (Macneil 1974). The procedural norm that “a defendant should not be punished until after the issue of guilt is resolved” is effectively abandoned.

Implicit and charge reduction plea bargaining, in contrast, are structured as authority relations between the defendant and the judge, in the sense of Simon (1957). One party voluntarily but bindingly submits (here, by pleading guilty) to the future discretionary choice of that party’s actions by another, which is predictable in advance only probabilistically. Except for the nonrecurrent character of defendants’ participation, this distinction between contract and authority is identical to that emphasized by Williamson (1975).

In the vertical role relationship between judge and prosecutor, what is at issue in the perception of pre-plea defendants as individually distinct evidentiary cases is the judge’s willingness to permit the final sentence outcome, either directly or indirectly, to become a function of the strength of the state’s case. In charge reduction plea bargaining, the judge has delegated to the prosecutor a discretionary capacity to interpret and manipulate the legal framework of the case. In sentence recommendation plea bargaining, the judge has delegated sentencing discretion as well. Either way, the prosecutor will bring to bear a distinctive interpretative framework, which emphasizes (as will be explained below) certainty of punishment and hence strength of the state’s case. Because calculation in terms of strength of state’s case entails a probabilistic conception of guilt, the procedural norm that “a defendant categorically is either guilty or not” is effectively abandoned.

In implicit and judicial plea bargaining, in contrast, active manipulation of sentencing discretion remains centralized in the hands of the judge, and the prosecutor’s self-interest is excluded.

These alternative role relations are summarized schematically in table I. Implicit plea bargaining is centralized authority. Judicial plea bargain-
TABLE 1
ALTERNATIVE ROLE STRUCTURES OF PLEA BARGAINING

<table>
<thead>
<tr>
<th>JUDGE-PROSECUTOR RELATIONSHIP</th>
<th>JUDGE-DEFENDANT RELATIONSHIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralized</td>
<td>Implicit plea bargaining</td>
</tr>
<tr>
<td>Delegated</td>
<td>Charge reduction plea bargaining</td>
</tr>
</tbody>
</table>

ing is centralized contract. Charge reduction plea bargaining is delegated authority. And sentence recommendation plea bargaining is delegated contract. Underlying these various informal role structures are perceptions of pre-plea defendants as differentiated cases, as differentiated criminals, as neither, or as both.

THEORETICAL BACKGROUND

Under what conditions will each of these plea bargaining systems emerge? Previous research, while cognizant of the empirical varieties of plea bargaining, has concentrated instead on trying to explain why plea bargaining should exist in any form. Competing theories addressed to this more globally defined question have been posed at different levels of analysis. Most contemporary studies are utilitarian: they emphasize the self-interest of one or more of the primary participants. Historical arguments tend to be functionalist, concentrating instead on adaptation by trial courts to one of a variety of changing environmental constraints. In spite of these level-of-analysis complications, most existing theories are variations on one of the following three themes: (1) plea bargaining as a response to the pressures of a heavy caseload; (2) plea bargaining as a response to the divergence between factual and legal guilt; and (3) plea bargaining as an effort to implement flexible sentencing standards within a constricting statutory framework. Each theme contains both a utilitarian and a functionalist component.

In this section I summarize the existing literature on plea bargaining in order to establish building blocks for the stochastic models that follow. My strategy is not to reject any of the "competing" theoretical arguments. It is rather to construct a synthetic modeling framework that incorporates arguments from all of the primary lines of thought. To meld operationally the utilitarian and the functionalist sides of these arguments, however,
Plea Bargaining

requires explicit attention to the aggregate terms of exchange within which myriad environmental constraints are experienced by participants.

Three Existing Explanatory Themes

The first explanatory image in the literature is of plea bargaining as a response to the pressures of caseload. Trial judges and district attorneys face a far greater volume of felony cases than could possibly be processed with limited resources through the more legitimate but time-consuming route of jury trials. Judges, prosecutors, and public defenders are united by their common interest in keeping the system from being overwhelmed. The motivational consequence, according to Alschuler (1976, p. 1099), is that "most trial judges look for guilty pleas the way that salesmen look for orders." Blumberg (1967) argues that this leads to the routinized treatment of idiosyncratic defendants, whereas Eisenstein and Jacob (1977) stress the informal rules of thumb and understandings which develop in small work groups. Either way, a certain percentage of undismissed cases must be channeled through guilty pleas rather than through trials if the caseload backlog is not to grow without bounds.

The two incentives for a defendant to plead guilty, in the shared view of all three schools of thought, are the sentence discount and the reduced uncertainty that a guilty plea can produce. Defendants plead guilty, in other words, both because they expect a more lenient sentence and because trial outcomes are unpredictable and involve the greater risk of high or even maximum statutory penalties (Alschuler 1976).

This "caseload hypothesis" has recently come under heavy attack because of an apparent lack of empirical correlation between fluctuations in caseload and aggregate guilty plea rates (Heumann 1978; Feeley 1979). Apart from issues of long-run equilibrium, this attack overlooks the facts that (a) internal adjustment in strategies (either sentence discount or plea bargaining mode) may be made specifically in order to sustain constant guilty plea rates and (b) more is involved than raw volume of inputs. Structural "carrying capacity" is the other constraining half of the argument.

Historically, the administrative capacity of criminal courts to cope with even a fixed volume of cases has declined dramatically for at least two reasons. In the nineteenth century, because of the rise of autonomous police and prisons on either side of the court, "the courts lost their dominant position in the system and instead became processing agencies standing between an expanding caseload over which they exercised minimal

\footnote{Omitted here is the transaction cost argument of Feeley (1979), which applies more forcefully at the misdemeanor court level. See p. 6 below.}
control and a relatively rigid prison system that could not admit all who might be eligible" (Haller 1979, p. 275). In a decentralized system with autonomous police, prosecutors, and prisons, in other words, judges could cope with criminal developments only within the confines of their narrow jurisdiction, the trial.

Within the trial itself, the average length of time necessary to complete a jury trial has increased steadily. In the 18th century, English juries processed 12–20 felony trials a day (Beattie 1977; Langbein 1978), and a Tallahassee felony court as late as 1890 conducted as many as six jury trials daily (Friedman 1979). While systematic statistics are not available for comparison, this “contrasts dramatically with the 7.2 days that an average felony jury trial required in Los Angeles in 1968” (Alschuler 1979, p. 240). The causes of this increase in length of jury trial include the increased participation of adversary lawyers and the development of procedural due process controls (Langbein 1979). Both explanations for the diminished carrying capacity of criminal courts stress that increased professionalization has inhibited the ability of the official system to function expeditiously.

The second theme in explaining plea bargaining stresses divergence between factual and legal guilt. Prosecutors and even defense attorneys, after some experience in the system, tend to believe that the vast bulk of felony defendants are in fact guilty of something, or else they would not be where they are (Heumann 1978). Except perhaps for cases of high public saliency, the argument goes, most felony charges with substantial doubt about factual guilt are dismissed. In this context, the issue for the prosecutor is not substantive guilt; the issue is how to steer the case through to legal guilt, which allows at least some punishment for the crime.

In this view, prosecutors are motivated during plea bargaining by their desire to gain greater control over dispositions than the vagaries and uncertainties of jury trials permit (Heumann 1978). The political urge to maintain high conviction rates is not inconsistent with the normative desire to impose outcomes that are “just.” If plea bargaining leads to more certain lower sentences, at least that is preferable to the risk of no sentence at all.

The key variable that structures plea bargaining microdynamics in this theme is “strength of state’s case.” The concessions that prosecutors are willing to entertain for a guilty plea depend heavily on their estimate of

---

3 Rapid bench trials, however, provide one alternative to plea bargaining in some jurisdictions. These adjudications, which resemble sentence hearings more than full trials, are often called “slow pleas” because of their uncontestable character (Greenwood et al. 1972).
probability of conviction at trial (Mather 1974). Alschuler (1968, p. 60) adds that "the universal rule is that the sentence differential between guilty plea and trial defendants increases in direct proportion to the likelihood of acquittal."

This prosecutorial self-interest argument can be embedded in environmental context by noting that the aggregate flow of case strengths that confront participants is affected, in turn, by a number of factors. Strength of state's case is ultimately a matter of evidence and admissible facts. Thus, the average strength of a case that makes it to plea bargaining is a product of earlier organizational filters—the arrest and investigation practices of police and the dismissal policies of the prosecutor. The fewer the arrests, the more thorough the investigations, and the more wholesale the dismissals, the stronger on average will be the evidence underlying plea bargaining dynamics.

On a more historical time frame, this theme would also imply by extension that the development of procedural due process controls on police investigations and jury trials increased the pressure for some adaptation like plea bargaining. Exclusionary rules of evidence and restricted powers of judicial summation increase the inherent unpredictability of juries, thereby intensifying the prosecutor's desire for extra-trial control. Constitutional limitations on police investigations (including right to defense counsel, restrictions on pretrial detention, search and seizure rules, and limitations on induced confessions) weaken on average the strength of the state's case, thereby undermining the built-in leverage of the prosecutor.

A third theme discernible in the plea bargaining literature is that plea bargaining is an effort to substitute flexible sentencing standards, which remain sensitive to the idiosyncratic background of the crime or criminal, for the rigid and often harsh provisions of the statutory code. As Feesley (1979, p. 197) has put it: "Although the relationship between prosecutor and defense is somewhat adversarial, they are actually engaging in a decision process to reach a consensus about the 'worth of the case.' . . . Much of what passes for plea bargaining is really negotiation over the meaning of facts." Equitable disposition rather than guilt is the focus of this argument, and substantive rather than formal justice is the basis for evaluating "worth of case" and "seriousness of offense."

This third theme highlights the role of sentencing norms. Even if legal guilt is attained, the sentence implied by the criminal code may be inconsistent with participants' context-specific conceptions of "equity." The determinants of "equity" are as variable as the concept of "mitigating circumstances" and hence are impossible to standardize. But plea bargaining, in this view, serves two related functions: (a) it provides an informal forum, uninhibited by evidentiary constraints, for participants to develop normative consensus about "proper" disposition; and (b) it is a
vehicle for individualizing sentence, either directly through sentence offers or indirectly through manipulation of charge.

The macro side of this theme highlights the systematic tensions that may exist between the internalized sentencing standards of trial court participants and the exogenous standards implicit in statutory codes and regulations. Historians have stressed that the emergence of plea bargaining in the 19th century was associated with the rise of urban political machines (Haller 1970). "Criminal courts comprised a distinct subculture at the turn of the century and after. The criminal bar was composed of lawyers who had attended less prestigious law schools, who usually did not join the bar associations, and who typically were members of ethnic minorities" (Alschuler 1979, p. 229). Particularistic trial judges and prosecutors, who may be generated by a machine culture, are more concerned with the criminal than with the crime and hence tend to generate more lenient sentences. Levin (1971) has pointed out that today this relative judicial orientation to the criminal rather than to the crime is itself a cultural variable. Pittsburgh trial judges, who emerge through a partisan political framework, give lower sentences for the same crimes than do their harsher, deterrence-oriented counterparts in Minneapolis, whose background is rooted in bar associations and reform governments.

Counterposed against this variation in internalized sentence standards is the variation in statutory criminal codes across American states (Note, Columbia Law Review, 1960). Criminal codes differ (a) in the number of refined charges they itemize, (b) in the severity of maximum sentence per charge, and (c) in the degree of discretion they grant judges in imposing a sentence. Alschuler (1976) has argued that the more refined the charging options, the easier it is to implement charge reduction plea bargaining. Hermann (1974) has hypothesized that plea bargaining is American because criminal codes on average are much harsher in the United States than in Western Europe. And Newman (1966) found that implicit plea bargaining dominated in Wisconsin, where broad sentencing discretion was delegated to the judge, whereas charge reduction was prevalent in Michigan, where mandatory sentence minima greatly restricted the judge's freedom to sentence within charge.

Sentence Discount Schedules

Even if we grant, as I do, the validity of all three of these schools of thought, two major puzzles must be solved before the issue of structural variation in plea bargaining can be addressed: (1) How do these "competing" causal arguments interrelate? (2) How are the levels of analysis to be spanned? The first puzzle will be solved in the course of constructing the stochastic models. The second requires an interfacing mechanism that
Plea Bargaining

simultaneously has two analytical properties. From the utilitarian side, it must be a behavioral frame within which the aggregate interaction of court participants is expressed. From the functionalist side it must be an ecological frame within which environmental and structural constraints are reflected.

The mechanism I propose is the sentence discount schedule, which is an operationalization, at the organizational level of analysis, of the common observation that trial judges (and prosecutors) are concerned with substantive rather than formal justice. I hypothesize that judges, consistent with their micro concern with the equity of individual dispositions, are concerned at the aggregate level with maintaining the integrity of their own internalized sentencing standards (whatever these might be). The sentence discount schedule traces out in equilibrium the global sentencing policies required to generate varying aggregate guilty plea rates. Each point on the schedule defines the degree to which guilty plea sentences must on average diverge from trial conviction sentences in order to induce heterogeneous defendants to plead guilty at a fixed aggregate rate.

Plea bargaining, I argue, is structured by the interaction of two levels of strategic "game" simultaneously. At the individual level, the game is between the defendant (through his attorney) and either the judge or the prosecutor and is based on the self-interest motivations outlined in the literature above. At the system level, however, what is at stake is maintaining the integrity of global sentencing policies, with the judge and/or prosecutor "playing" not the individual defendant but the entire heterogeneous pool of defendants all at once. To the extent to which a plea bargaining structure forces judges and prosecutors to generate guilty plea sentences that are on average at variance with their internalized sentencing norms, judges and prosecutors will be tempted to abandon that structure. Assuming (holding aside statutory code restrictions) that conviction at trial offers an unconstrained opportunity to apply these internalized standards directly, the sentence discount is a direct measure of this normative inconsistency.

As will be derived below, equilibrium sentence discount schedules are shaped by the formal control structure and by the environmental constraints described above. Judges and prosecutors need not perceive these underlying constraints directly; what they perceive is the response of an aggregate pool of defendants to variation in sentencing policies. In their struggle to preserve their own internalized sentencing standards, however, they are indirectly adapting to environmental exigencies.

This normative concern with keeping sentence discounts within bounds is not, it should be emphasized, a cultural universal. In Great Britain, the guilty plea is legally construed as a mitigating factor presumptively indicating an element of remorse, and sentences have actually
American Journal of Sociology

been lowered on appeal on the grounds that the sentence discount originally given was insufficient (Heberling 1973). No claim is made for the relevance of the following models outside an American context. In America, however, sentence discounts lie at the heart of most criticisms of plea bargaining (Alschuler 1981). Conservatives, for example, argue that granting lenient benefits to guilty plea defendants undermines the capacity of the court system to give criminals their just deserts. Liberals argue that the additional punishment meted out to defendants convicted at trial imposes a cost on the exercise of their constitutional rights. Plea bargaining, others have argued, subverts the process of establishing guilt accurately, since sentence discounts may induce even factually innocent defendants to cop a plea because of their fear of conviction at trial. For a variety of reasons, therefore, sentence discounts lie at the normative heart of U.S. plea bargaining, even though there is no inevitability to a negative evaluation of sentence discounts.

MODELS

Stochastic models of the four plea bargaining role structures (plus mixtures), which predict both aggregate guilty plea rates and equilibrium sentence discounts, are presented in this section. The models are based on the description in the introduction: participant motivations are held constant; the key source of variation across role structures is the terms within which pre-plea defendants are interpreted.

Implicit Plea Bargaining

In implicit plea bargaining, the prosecutor is irrelevant. The judge does not perceive individual pre-plea defendants and manipulates the system from afar only via global sentencing policies. The defendant, in consequence, is confronted with two lotteries—one sentence distribution contingent on his own choice of guilty plea, and another sentence distribution contingent on a conviction at trial. The defendant's incentives in this structure are to achieve a more lenient expected sentence and to reduce expected risk. Both are subject to manipulation by the trial judge.

Define the two sentence distributions, for guilty plea and for trial conviction, respectively, as \( x_{pij} \sim p_{pij} (\mu_{pij}, \sigma_{pij}) \) and \( x_{rij} \sim p_{rij} (\mu_{rij}, \sigma_{rij}) \). Defendant \( i \) confronted with charge \( j \), in other words, receives a particular sentence \( x_{ij} \) drawn from system-wide sentence distributions which are contingent on plea. Underlying this particular sentence, of course, are the context-specific attributes of the crime and the criminal. In implicit plea bargaining, the defendant through his attorney knows at the time of his plea only aggregate judicial sentencing policies (i.e., \( \mu_{pij}, \sigma_{pij}, \mu_{rij}, \sigma_{rij} \), and
Plea Bargaining

$\sigma_{T_j}$, but not the exact outcome for his specific case. Conversely, before the plea the judge does not know any of the specific attributes of the defendant. The sentence discount controlled by the judge is $d_{ji} \equiv (\mu_{F_j} / \mu_{T_j})$.

If the defendant pleads guilty, then with probability 1.0 he receives whatever sample $x_{P_j}$ the judge selects out of $p_{P_j}(\mu_{P_j}, \sigma_{P_j})$. But if the defendant chooses trial, he receives $x_{T_j}$ only with the probability that he in fact will be convicted at trial. Call this probability $s_{ij}$, which is identical from the prosecutor's perspective to the strength of the state's case. The defendant will plead guilty whenever $^2$

$$s_{ij}(\mu_{T_j} + \sigma_{T_j}) > (\mu_{P_j} + \sigma_{P_j})$$

$$s_{ij} > \left( \frac{\mu_{P_j} + \sigma_{P_j}}{\mu_{T_j} + \sigma_{T_j}} \right)$$

Conversely, the defendant will go to trial whenever $^6$

$$s_{ij}(\mu_{T_j} + \sigma_{T_j}) < (\mu_{P_j} + \sigma_{P_j})$$

$$s_{ij} < \left( \frac{\mu_{P_j} + \sigma_{P_j}}{\mu_{T_j} + \sigma_{T_j}} \right)$$

$^4$ Throughout, I assume that defendants obtain accurate information about judicial sentencing policies from their counsel. This assumption is most valid for experienced defense attorneys (like public defenders), who make it part of their job to monitor judges very closely (Aisculer 1975). I do not model any defense attorney self-interests that are at variance with those of clients. This is not because such interests do not exist, but because such interests (e.g., in the clearing of workloads) are held in common with other participants in plea bargaining, who are explicitly modeled. The case of the private defense counsel interested in maximizing his fee lies outside the bounds of this analysis because the typical felony defendant is poor.

$^5$ Obviously, more complex utility functions could be presumed here. I have chosen the linear form, in the absence of any information to the contrary, because it captures the essentials of risk aversion in the simplest manner. One commonly used alternative, derived from $U(x) = -x_1 \sigma_1^2$, which is quite similar in spirit to the representation selected, is the quadratic form, $s_{ij}(\mu_{T_j}^2 + \sigma_{T_j}^2) > (\mu_{P_j}^2 + \sigma_{P_j}^2)$. I prefer the simpler form because it is easier to conceive of defendants (or their attorneys) thinking literally in terms of expected sentence and expected uncertainty (e.g., $s_{ij}(\mu_{T_j}^2 + \sigma_{T_j}^2)$ than it is to conceive of defendants thinking literally in quadratic terms. I have, however, worked through all models under this alternative quadratic assumption and have verified that the conclusions about relative performance of plea bargaining systems are not altered. All that is involved is a minor scaling shift in the sentence discount schedules which applies uniformly to all systems. Additionally, one could easily augment these utilities with weights on the $s_1$'s, if one believed that expected risk is not as important as expected sentence. In the particular sentence distributions which will be hypothesized below, however, $s$ is invariably lower than $\mu$ in any event.

$^6$ In this framework, my earlier dismissal of the importance of Feeley's (1979) transaction cost argument in the case of felony courts can be justified further. Transaction costs can be included in this utility function as follows: $s_{ij} \left( \mu_{T_j} + \sigma_{T_j} \right) > (\mu_{P_j} + (\mu_{P_j}$

767
American Journal of Sociology

This binary decision rule models the behavior of a single risk-averse defendant confronted with two risky options. At the organizational level of analysis, however, the issue is the aggregate flow of such decisions (the guilty plea rate) across all \( n \) incoming defendants, each of whom has an idiosyncratic \( s_i \) associated with his case.

For this aggregation, a distribution of strengths of state case across the entire flow of undismissed cases is necessary—namely, \( p(s) \). One flexible distribution which seems capable of approximating all realistic possibilities is the power function density, which is graphed in figure 1:

\[
p(s) = \nu s_i^{\nu - 1}.
\]

The peak of this density at \( s_i = 1 \) reflects the empirical observation that most cases are “dead bang.” Namely, it is more likely than not that most defendants are in fact guilty of something or else they would not be where they are.

The degree to which this statement is true may vary considerably from city to city. The parameter \( \nu \) controls the average strength of case because \( E(s_i) = \nu(\nu + 1) \). The higher the \( \nu \), the greater the proportion of strong cases and conversely. Causally, as explained above, \( \nu \) is the consequence of earlier stages in the system: (a) The more efficient police arrest and

\[\text{FIG. 1.—The distribution of strengths of case and implicit plea bargaining}\]

\[\text{+ \( \sigma_{p_i} \), where presumably \( t_{p_i} > t_{p_j} \). Solving yields \( s_i > \left[ (t_{p_j} - t_{p_i})(\mu_{p_j} + \sigma_{p_j}) + ((\mu_{p_i} + \sigma_{p_i})/\sigma_{p_i}) \right]. \) To the extent that transaction costs are important, the guilty plea rate will be higher than it otherwise would be. The relative importance of the first term vanishes, however, as \( \mu_{p_i} \) and \( \sigma_{p_i} \) increase, as they do with more serious offenses. For misdemeanor courts where \( \mu_{p_i} \) and \( \sigma_{p_i} \) are low, the role of transaction costs remains central.} \]
Plea Bargaining

Investigation practices are, for whatever the reasons, the higher the \( v \). (b) The more thoroughly prosecutors dismiss weak cases, the higher the \( v \). (c) The more restrictive are evidentiary barriers to legal conviction, the lower the \( v \).

The aggregate flow of guilty pleas is

\[
P_j(\text{plea}) = \text{prob}\left[s_{ij} > \left( \frac{\mu_{P_j} + \sigma_{P_j}}{\mu_{T_j} + \sigma_{T_j}} \right)^v \right].
\]

For the power function density, this can be easily solved to yield

\[
P_j(\text{plea}) = 1 - \left( \frac{\mu_{P_j} + \sigma_{P_j}}{\mu_{T_j} + \sigma_{T_j}} \right)^v,
\]

with the corresponding trial rate being

\[
P_j(\text{trial}) = \left( \frac{\mu_{P_j} + \sigma_{P_j}}{\mu_{T_j} + \sigma_{T_j}} \right)^v.
\]

Graphically, as is apparent in the second panel of figure 1, these two expressions correspond to relative areas under the \( \rho(s_{ij}) \) curve. Substantively, the aggregate guilty plea rate generated by an implicit plea bargaining system is the joint outcome of sentencing policies, which are under the direct control of the judge, and of average strength of case, which is ecologically defined.

For implicit plea bargaining (and for charge reduction plea bargaining), a closed-form solution is possible without further specification of the sentence distributions. For the contract modes of judicial and sentence recommendation plea bargaining, however, a specific functional form will prove necessary for numerical calculations. As before, a plausible distribution is the power density, which is illustrated in figure 2:

\[
\rho(x_{P_j}) = (\beta_T/M_j)(1 - x_{T_j}/M_j)^{\beta_T - 1},
\]

and likewise for \( \rho(x_{T_j}) \). The term \( M_j \) represents the maximum sentence allowed in the statutory criminal code. Since \( E(x_{T_j}) = [1/(\beta_T + 1)]M_j \), \( \beta_T \) is a measure of leniency of internalized judicial sentencing norms. Low \( \beta_T \) is the case of harsh sentencing norms, and high \( \beta_T \) is the case of lenient sentencing norms. According to the findings of Levin (1971) discussed above, \( \beta_T \) is rooted in local legal culture. Partisan judges will tend to generate a higher \( \beta_T \) than will their more deterrence-oriented reformist counterparts. The term \( \beta_P \) is the guilty plea distribution analogue, the magnitude of which will be determined through sentence discount equilibrium.

This sentence distribution is defined over a finite sample space \([0, M_j]\)

769
because of the restrictions in the statutory criminal code. The particular monotonically declining shape of this density was selected because of my empirical observation of that pattern in published statistics for federal U.S. District Court (U.S. Department of Justice 1978, p. 564). A plausibly rationalized version of this empirical pattern is an underlying monotonically declining distribution of prior records (which is one of the consistently demonstrated "historical background" factors in sentencing).

---

This sentence distribution could be modified to incorporate even greater realism. As shown in fig. 2, the low point of the distribution could be redefined (by substituting \( x_{P_{m}} \) in place of \( x_{P_{m}} \) in the formula) to incorporate mandatory minimum sentences as well as statutory maxima in the codes. American states vary in their inclusion of statutory minima in the codes. In addition, assuming a common metric that might be difficult to operationalize, probation could be incorporated into the distribution by defining a fixed \( x_{p} \) point on the axis, below which the sample sentence outcome is presumed to be probation (of varying lengths) rather than actual "hard time." Newman (1966) has shown the former adjustment to be relevant to the relative attractiveness of charge reduction. The latter adjustment, being constant across all modes of plea bargaining, is irrelevant to the relative comparison of plea systems.
With this functional form, the trial rate solution above simplifies to
\[ P_j(\text{trial}) = (\overline{\beta}_r/\overline{\beta}_r)^\mu. \]

The notation is
\[ \overline{\beta}_r = \left( \frac{1}{\beta_r + 1} \right)[1 + (\beta_r/\beta_r + 2)^\mu] \]
and
\[ \overline{\beta}_r = \left( \frac{1}{\beta_r + 1} \right)[1 + (\beta_r/\beta_r + 2)^\mu]. \]

The sentence discount is \( d_{\mu} = (\beta_T + 1)/(\beta_T + 1) \).

This equation represents the behavioral responsiveness of an aggregate pool of defendants under implicit plea bargaining without the presence of any caseload constraints. Trial judges, however, need to equilibrate this \( P_j \) (trial) so that it does not exceed the carrying capacity of their own courts. Trial carrying capacity (\( K \)) is simply the number of trials which can be processed in one year. For one judge, this equals the inverse of the average length of time for each trial (\( l \)), defined in appropriate yearly units. Thus, for a city court composed of \( k \) judges, \( K = kl/\mu \). The caseload constraint, therefore, is \( n \cdot P(\text{trial}) \leq K = kl/\mu \) or \( P(\text{trial}) \leq kl/n \). Here, as before, \( n \) is the raw volume of undischmissed case input.

If we define the equilibrium \( P_{\mu}(\text{trial}) \) to be exact (i.e., if we do not grant trial judges a positive preference for free time), then \( P_{\mu}(\text{trial}) = (kl/n) \),

which yields a final equilibrium solution of
\[ (\overline{\beta}_r/\overline{\beta}_r) = (kl/n)^{1/\mu}. \]

Translation of this expression into sentence discount terms \( d_{\mu} \) is accomplished numerically.

Equilibrium here is a fixed relationship, defined by the sentence discount, between two parallel sentence distributions—a fixed trial distribution, controlled by \( \overline{\beta}_r \) and \( M_j \), and a variable guilty plea distribution, which adjusts to the trial distribution in accordance with this schedule. This sentence discount schedule is molded directly by the environmental parameters \( n, l, \nu, \) and \( \overline{\beta}_r \). The greater the raw volume of undischmissed

\[ ^* \text{Obviously, full "carrying capacity" is the summation of these expressions over all } J \text{ charge classes. This straightforward (but messy) generalization is tangential to my primary concern here; so I oversimplify by assuming that sentence discounts and guilty plea rates within courts are constant across original charge classes. To the extent to which Feeley's (1979) transaction costs are important, however, guilty plea rates are higher the less serious the original charge.} \]
American Journal of Sociology

cases, the longer the average length of trial, the weaker the average strength of case, and the more severe the judicial sentencing norms, the wider must be the sentence differential between guilty plea and trial defendants in an implicit plea bargaining system for caseload backlog not to grow (without bounds over multiple years). The sentence discount schedule represents a compacted view of the aggregate pool of defendants, the terms of which are shaped (with or without the knowledge of participants) by the exogenous structural constraints in which the court is embedded.

None of these aggregate solutions, it should be reiterated, implies either that defendants in the same plea class receive the same sentence or that each defendant receives a constant sentence discount. The stochastic sentence distributions reflect individualized evaluations.

Judicial Plea Bargaining

In the pure form of judicial plea bargaining, the trial judge, after a brief conference in chambers with prosecutor and defense attorney, communicates to the defendant (through counsel) the contractual sentence he or she will impose in exchange for the defendant pleading guilty (Alschuler 1976). Just as in a sentence hearing, the judge bases this contingent pre-plea offer on the individual attributes of the alleged criminal—the contextual circumstances surrounding the crime and the historical background of the defendant. Substantive justice considerations are central, but the judge is not affected (as is the prosecutor) by the strength of the state's case or the probability of conviction at trial. The judge does not share the adversarial position of the prosecutor or have the same detailed pre-plea access to the evidence. Given the provisional assumption that the defendant is really guilty, sentencing (excluding the sentence discount) is based solely "on the facts as I now see them."

Judges may also preserve their legal right to change their minds "if the facts turn out otherwise than I now understand them to be" by reneging on this pre-plea offer, but this mixed case will be considered in a subsequent subsection.

The main behavioral difference between judicial and implicit plea bargaining is temporal order. The judge selects his or her sample $x_{FY}$ out of

---

5 This behavioral assumption presumes that the judge does in fact act like a judge. If, on the other hand, the judge occasionally adopts a prosecutorial perspective (as some would argue), then the model developed here would require modification. In particular, the revised judicial model would be a mixture of the "pure" judicial and sentence recommendation models developed herein.
Plea Bargaining

$p(\mu_{ij}, \sigma_{ij})$ before rather than after the defendant chooses his plea. The defendant, then, will go to trial whenever

$$s_{ij}(\mu_{ij} + \sigma_{ij}) < x_{p_{ij}},$$

and the aggregate flow of trials will be

$$P_j(\text{trial}) = \text{prob} \{ x_{p_{ij}} > s_{ij}(\mu_{ij} + \sigma_{ij}) \}.$$

From the defendant's point of view, judicial plea bargaining has the advantage of greater certainty over implicit plea bargaining, but whether this leads to a higher probability of pleading guilty depends on the equilibrium sentence cost discount involved.

This expression can be solved for the case of the power density in the following manner:

$$P_j(\text{trial}|s_{ij}) = \text{prob} \{ x_{p_{ij}} > s_{ij}\bar{\beta}_T M_j \}
= \int_{s_{ij} \bar{\beta}_T M_j}^{M_j} (\beta_P/M_j)(1 - \frac{x_{p_{ij}}}{M_j})^{\beta_P - 1} dx_{p_{ij}}
= (1 - s_{ij} \bar{\beta}_T)^{\beta_P},$$

where $\bar{\beta}_T$ is defined as before. Thus,

$$P_j(\text{trial}) = \int_0^{\frac{M_j}{s_{ij} \bar{\beta}_T}} v u^{-1} (1 - s_{ij} \bar{\beta}_T)^{\beta_P} dv
= \left( \frac{\beta_P}{(\nu + \beta_P)!} \right) \sum_{i=0}^{\beta_P} \left( \frac{(\nu - 1 + i)!}{(\nu - 1)!i!} \right) (1 - \bar{\beta}_T)^i.$$

Caseload constraints are imposed graphically in the section on analysis. Equilibrium sentence discount schedules must be derived numerically.

Charge Reduction Plea Bargaining

In charge reduction plea bargaining, the prosecutor offers as concession to the defendant not an explicit sentence for his plea, but rather an alteration in charge from the original indictment. Like implicit plea bargaining, charge reduction is an authority relation between the defendant and the judge, in the sense that even after the charge has been altered the defendant still faces a discretionary directive selected for him out of the plea sentence distribution. A reduced charge is no guarantee of a reduced sentence; final disposition still depends on the evaluation of the judge.

Charge reduction differs from implicit plea bargaining in the sense that the defendant, in general, does not plead guilty "on the nose." Instead of facing a fixed $p(x_{p_{ij}})$ defined by the original charge $j$, the defendant faces an adjusted $p(x_{p_{ij}^r})$ defined by the reduced charge $j'$. In essence, the judge

773
American Journal of Sociology

poses an array of guilty plea sentence distributions, controlled by a policy set \( \{ \mu_{P_j}, \sigma_{P_j} \} \) over all possible charges, and the prosecutor selects the distribution out of this set from which the judge will choose.

The defendant will plead guilty, as before, whenever

\[
    s_{ij} > \left( \frac{\mu_{P_j} \pm \sigma_{P_j}}{\mu_{T_j} + \sigma_{T_j}} \right)
\]

But on what basis will the prosecutor select this reduced charge?

The "divergence of factual from legal guilt" school cited above argues that the prosecutor is motivated to maintain control over the case in the face of heterogeneous strengths of case. Some punishment is better than the risk of no punishment at all, or, as Alschuler's interviewed prosecutor has put it (1968, p. 59), "When we have a weak case [low \( s_{ij} \)] for any reason, we'll reduce it to almost anything rather than lose." The "flexible sentencing standards" school cited above argues that prosecutors try to generate final sentences which are as close as is realistically possible to what they regard as equitable outcomes. Under the additional assumption that the prosecutor's global sentencing standards are identical to the judge's, these two schools of thought can be combined by hypothesizing that the prosecutor will have as an implicit target disposition

\[
    E(x_{P_j}) = s_{ij}\mu_{P_j}
\]

Prosecutors under charge reduction do not have full control over the individual sentence, and idiosyncratic criminal facts considered by the judge at sentence hearing will drive any particular sentence away from this stochastic mean. But the prosecutor, I argue, will choose that reduced charge which on average drives the judge's own sentencing distribution closest to the prosecutor's own target disposition. This target is customized by the case's probability of trial conviction. Prosecutors under this hypothesis "go for what they can get" in the sense of expected values, even if such a "realistic" approach is less than the ideal of \( s_{ij} = 1 \).

To operationalize this hypothesis, it is necessary to posit structure to the charge array \( \{ \mu_{P_j}, \sigma_{P_j} \} \). Perhaps the most realistic idealization is a geometric relationship across charge means: \( \mu_{P, j - k} = \mu^{*}\mu_{P_j} \), which is equivalent for the power sentence density to the same relationship across statutory maxima, \( M_{j - k} = \mu^{*}M_j \). Here \( k \) is the number of charges the prosecutor can downgrade, and \( \mu \) is the attribute of the sentencing code which Alschuler (1976) has highlighted. Criminal codes differ in how numerous and in how refined (e.g., first degree, second degree) are the charging options available to the prosecutor. The higher the \( \mu \), the larger the number of charging options.

In other words, as is illustrated in figure 3, I take it as an empirical fact that the difference between average sentence for murder and average

774
sentence for manslaughter is greater than the difference between average sentence for manslaughter and that for armed assault, which in turn is greater than the difference between average sentence for armed assault and for simple assault and so on down the line. Put another way, there are more charges which on average generate less severe punishment than there are charges whose penalties are extreme. In the case of power
sentence densities, these geometric relations among averages are induced by geometric relations among statutory maxima.

Given this charge array structure, the prosecutor's charge reduction strategy can be operationalized as is illustrated in figure 3. Namely,

(a) $j' = j$ whenever

$\forall i (\mu_{ij} + \mu_{j-1}) < s_i \mu_{ij} < \mu_{j-1}$

$\forall i (1 + \bar{\mu}) < s_i < \forall (1 + \bar{\mu})$;

(b) $j' = j - 1$ whenever

$\forall i (\mu_{ij-1} + \mu_{j-1}) < s_i \mu_{ij} < \forall (\mu_{ij-1} + \mu_{j-1})$

$\forall i (1 + \bar{\mu}) < s_i < \forall (1 + \bar{\mu})$;

(c) $j' = j - 2$ whenever

$\forall i (\mu_{ij-2} + \mu_{j-1}) < s_i \mu_{ij} < \forall (\mu_{ij-2} + \mu_{j-1})$

$\forall i (1 + \bar{\mu}) < s_i < \forall (1 + \bar{\mu})$;

et cetera. The degree to which the prosecutor will downgrade the charge is a direct function of the strength of the state's case.

Even with charge reduction, not all defendants will plead guilty. Some may find it advantageous to take their chances at trial on the original charge. In particular, a defendant will go to trial

(a) if $j' = j$, whenever

$s_i < \frac{\mu_{ij} + \sigma_{Tj}}{\mu_{ij} + \sigma_{Tj}}$;

(b) if $j' = j - 1$, whenever

$s_i < \frac{\mu_{ij} + \sigma_{Tj}}{\mu_{ij} + \sigma_{Tj}}$;

(c) if $j' = j - 2$, whenever

$s_i < \frac{\mu_{ij} + \sigma_{Tj}}{\mu_{ij} + \sigma_{Tj}}$;

and so forth.

These two strategic sides can be assembled to generate an aggregate trial flow as follows:

$$P_{i}(\text{trial}) = \int_{\gamma(1 + \bar{\mu})}^{\gamma(1 + \bar{\mu})} \rho(s_{ij}) ds_{ij} + \int_{\gamma(1 + \bar{\mu})}^{\gamma(1 + \bar{\mu})} \rho(s_{ij}) ds_{ij} + \ldots$$

$$= (\frac{1}{1 - \bar{\mu}})^{\nu} \sum_{i=0}^{\infty} \frac{\bar{\mu}^{i}}{(1 + \bar{\mu})^{\nu}}$$

Once again, translation of this expression into sentence discount terms is accomplished numerically.

In charge reduction plea bargaining, guilty plea rates are affected directly not only by sentence discounts, average strength of case, and judicial severity in sentencing, as was the case in implicit plea bargaining.
Plea Bargaining

The charging structure in the statutory criminal code also has an impact on the viability of this mode. In particular, the more numerous are the charging options available to the prosecutor (i.e., the closer \( \mu \) is to one), the less “effort” in the form of sentence discounts the judge has to make in sentence hearing in order to equilibrate the system. In the extreme case, the judge can drive the guilty plea rate to 100% by choosing a \( \beta_p \) that makes \( (\beta_p/\beta_T) = \frac{1}{2}(1 + \pi) \). In the implicit and judicial plea structures, in contrast, this could only be done by letting everyone off scot-free (\( d_\mu = 0 \)).

Sentence Recommendation Plea Bargaining

In sentence recommendation plea bargaining, the prosecutor recommends to the judge an explicit disposition in exchange for the defendant’s plea of guilty. In theory the judge is free to accept or reject this recommendation, but in the pure form of this mode the judge always rubber stamps the request. Cases in which the judge occasionally reneges on the plea bargained agreement are examined in the next subsection.

Sentence recommendation plea bargaining is a combination of the judicial and the charge reduction modes. Sentence distribution predispositions are customized by case, as in charge reduction plea bargaining. But an explicit pre-plea sentence commitment is made on the basis of the individual attributes of the alleged criminal, as in judicial plea bargaining.

The defendant, as in judicial plea bargaining, will plead guilty whenever

\[
s_i(\mu_T + \sigma_T) > x_{Pij}.
\]

Because of the prosecutor’s incentives discussed above, these pre-plea sentence decisions are affected by strength of state’s case in addition to global sentencing policies. In particular, the prosecutor will operate within a customized sentencing distribution with mean

\[
E(x_{Pij}) = s_i\mu_{Pij},
\]

just as was the case in charge reduction plea bargaining. In sentence recommendation plea bargaining, however, the prosecutor has direct access to this distribution and thus will go on to select his or her own \( x_{Pij} \), based on the same individualized attributes as would underlie the judge’s choice.

Implicit in this model of prosecutorial sentence recommendation is an empirical observation often cited by judges who delegate their sentencing discretion to prosecutors (McDonald 1979). Judges and prosecutors in the same court system in general share identical substantive justice norms.
American Journal of Sociology

and global sentencing predispositions ($\beta$). Over the long run, one reason for this convergence may be the fact that judges and prosecutors are embedded in the same local culture. More directly in the short run, without such convergence, this mode of plea bargaining is unstable because of the judge's rejection of many recommendations.

Direct comparison of the average severity of the pre-plea sentence offers generated by the judge and by the prosecutor is not obvious. Discounting by strength of case implies that the prosecutor's punishments will be lower (but presumably more certain) than those of the judge, the average magnitude of this difference being controlled by $\nu$. The larger global sentence discounts which will be required in the judicial plea bargaining system, however, provide an offsetting effect, the magnitude of which is determined in equilibrium.

For the power sentence density with mean $s_d \left[1/(\beta P + 1)\right]$ $M_j$, the model can be solved as follows:

$$P_{j}(\text{trial}) = \text{prob} \left[ x_{P_{ij}} > s_d (\mu_T j + \sigma_T j) \right]$$

$$= \left[ \frac{(\beta P/s_d M_j)(1 - x_{P_{ij}}/s_d M_j)^{\beta P - 1}}{\mu_T j + \sigma_T j} \right] ds_{ij}$$

$$= (1 - \beta P)^{\beta P}.$$ 

The final solution is independent of $\nu$. Equilibrium sentence discount schedules must be derived numerically.

Mixtures

All of the foregoing models are ideal types in the sense that plea bargained arrangements are always adhered to. In judicial plea bargaining, the judge never "goes back on his or her word" to give a defendant a sentence other than was promised. In sentence recommendation plea bargaining, the judge always accepts the recommendation worked out between the prosecutor and the defendant. In charge reduction plea bargaining, the judge's sentencing predispositions are in fact changed by the alteration in charge.

It should be apparent, even without formal solutions, that there is pressure on trial judges to cooperate with these systems. If the trial judge is "independent minded" and does not ratify plea bargaining agreements, the benefits and viability of plea bargaining are undercut because guilty plea rates will decline in response to this additional element of uncertainty in the system.

The models nonetheless can be extended to cope with this possible
source of friction in any realistic plea bargaining system. If we call the probability that a judge will renege on a plea bargained agreement $p_r$, then the defendant will plead guilty in a mixed judicial or sentence recommendation structure whenever

$$s_{ij}((\mu \tau_j + \sigma \tau_j) > (1 - p_r)s_{ij} + p_r(\mu \tau_j + \sigma \tau_j).$$

The analogue for a mixed charge reduction structure, in which the judge sometimes sentences a defendant as if he were guilty of the original charge, is

$$s_{ij}((\mu \tau_j + \sigma \tau_j) > (1 - p_r)(\mu \tau_j + \sigma \tau_j) + p_r(\mu \tau_j + \sigma \tau_j).$$

If the judge reneges, in other words, he or she returns to the original role structure and selects another “independent” evaluation out of the guilty plea distribution. Implicit plea bargaining is the null to which the judge will return with probability $p_r$.

I do not present here the final solutions for these extensions, since these solutions are rather involved mathematically. However, it should be apparent intuitively that any mixture's numerical solution will usually lie between the two corresponding ideal types.

ANALYSIS

With tools now in hand, we can return to the primary question of this paper: When will which informal role structure of plea bargaining exist? The stochastic models themselves have been developed at the level of interlocking strategic choice and then aggregated to yield the global trial rate (or, equivalently, the guilty plea rate) characteristics of each system. Environmental constraints embodying each of the three themes in the plea bargaining literature are also embedded in the models presented above:

1. The “caseload pressure” argument is represented by $P^*(\text{trial}) = (k/n)$. Judges manipulate sentence policies and/or the structure of plea bargaining in order to drive the aggregate choices of defendants, $P(\text{trial})$, down to this equilibrium target, $P^*(\text{trial})$. Underlying this trial rate target are raw undismissed caseload ($n$), the number of trial judges in the court ($k$), and the average length of trial ($l$).

---

10 In the case of judicial plea bargaining, Alschuler (1976) refers to this mixed system as “hinds, indirectness and cajolery.” Instead of overt sentence promises, the judge gives off ritualized signals and hints (e.g., “willingness to consider” or “current inclinations”) about the sentence he or she intends to impose, the interpretation of which is well known among experienced attorneys. Judges thereby explicitly preserve their legal right to change their minds (and exercise it with $p_r$).
2. The “divergence of factual and legal guilt” argument is represented at the individual level by strength of state’s case \( (s_q) \) and at the system level by the average flow of case strengths \( (v) \) with which participants are confronted. Underlying this heterogeneous flow of “quality” of cases are police arrest and investigation constraints and practices, prosecutorial dismissal policies, and exclusionary rules of evidence.

3. The “flexible sentencing standards” argument is represented at the micro level by the fact that all final sentences are individualized draws from an overall systematic sentencing distribution and at the system level by internalized sentencing severity norms \( (\beta_r) \). Local legal culture underlies these sentencing attitudes. Statutory criminal codes are modeled by \( M_p \), the statutory maximum penalty, and by \( \bar{p} \), which models the number of charging options available to the prosecutor.

The question to be addressed in this section therefore is, How does variation in these specific attributes of “environmental context” affect the relative attractiveness, to judges and prosecutors, of the four different role structures of plea bargaining (plus mixtures)?

The sentence discount adaptation mechanism emphasized above implicitly assumes that in the final analysis it is judges who permit plea bargaining systems to crystallize beneath them. Individual-level benefits to defendants and prosecutors are the proximate causes of overt plea bargaining negotiation, but an unsympathetic judge can always unravel any plea bargaining structure simply by refusing to ratify plea bargained agreements (to the extent \( \rho_t \)). This is true even in the case of charge reduction, since a consistent judicial policy of sentencing at or near reduced charge maxima would eventually force compensatory prosecutors to reduce charges to such ludicrous levels that convictions would be overturned at appeal on the grounds of no “factual basis.”

I will begin my analysis therefore by examining each plea bargaining system from the perspective of the judge: When will the judge be most likely to intervene explicitly with defendants through judicial sentence offers? When will the judge be most likely to delegate discretionary manipulation of the charge to the prosecutor? And when will the judge be most tempted to delegate his or her own sentencing discretion as well? I will address these questions by focusing on the distinctive judicial problems that unfold as environmental circumstances change. Implicit plea bargaining is always the baseline to which the judge will retreat if he or she refuses to cooperate with more overt forms of plea bargaining.

Judicial versus Implicit Plea Bargaining
The role transformation at issue in this comparison is the switch by the judge from authority to contracting modes of relating to defendants.
Procedurally what is involved in this behavioral change is the judge's interpretation of pre-plea defendants as differentiated criminals. Judges in judicial plea bargaining look behind the veil of the decision on guilt to the sentence hearing stage, at which time the issue is the appropriateness of final disposition. In neither implicit nor judicial plea bargaining, however, do judges perceive pre-plea defendants as differentiated cases.

Figure 4 presents the equilibrium sentence discount schedules for implicit plea bargaining, $p_r = 1$, judicial plea bargaining, $p_r = 0$, and sample intermediate systems, $p_r = .3, .7$. Each graph describes, for a fixed $\beta_T$ and $v$ court environment, the array of sentence discounts, $d_\mu = \ldots$

![Graphs showing sentence discount schedules for different values of $\nu$.](image)

**Fig. 4.** Sentence discount schedules: judicial vs. implicit plea bargaining. (Intermediate sentencing severity: $\beta_T = 3$.)
\( \mu_{\mu}/\mu_{\mu} \), required by various trial or guilty plea rate targets, \( P^*(\text{trial}) = 1 - P^*(\text{guilty plea}) \). Particular criminal courts are represented by different locations within these graphs. Varying levels of caseload pressure are arrayed along the horizontal axes, since \( P^*(\text{trial}) = (k/n) \). Varying levels of average strength of case are arrayed in the different panels of the figure, since \( E(s_i) = \text{(v/v + 1)} \). And varying levels of average sentence severity are arrayed in different figures. Figure 4 presents the intermediate case of \( \beta_T = 3 \), in which \( E(x_{Tiy}) = .25 M_j \), and figures 7 and 8 present the cases of severe and lenient sentencing, respectively. Detailed discussion of these latter figures is postponed to a subsequent subsection.

To interpret these graphs from the perspective of the judge, simply focus on the vertical gap between the judicial and the implicit schedules. The closer \( d_n \) is to one, the better from the perspective of the judge; hence the vertical distance between schedules measures the discounting gain to be realized by a switch to the role set represented by the higher schedule. The larger the vertical gap between the judicial and the implicit schedules, the greater is the incentive for a judge to plea bargain overtly.

Before I explain what underlies the phenomena revealed in figure 4, let me describe the model’s predictions about the relative attractiveness of judicial plea bargaining to a “substantive justice” judge. The qualitative character of these predictions varies radically depending on whether average strength of case is low or high.

1. When a substantial proportion of cases is legally weak (\( v = 1, 2, 3 \)) judicial plea bargaining is superior in terms of discounts to implicit plea bargaining under all caseload pressure circumstances. Through the simple expedient of announcing the sentence before the plea, in other words, a judge can punish “comparable” criminals more equally. The magnitude of this judicial incentive to contract varies with caseload pressure. When caseload pressure is extremely light (owing either to light caseload or to rapid bench trials), there is not much discounting difference between systems. But when caseload pressure is moderate, \( .3 < P^*(\text{trial}) < .4 \) for \( v = 2, 3 \), sentence discount incentives for a judge to intervene are maximal. Discounting can be eliminated entirely, \( d_n = 1 \), under these circumstances. As caseload pressure increases beyond this range, the discounting performance of judicial plea bargaining deteriorates rapidly. The comparative advantage of intervention is correspondingly reduced.

2. When the overwhelming bulk of cases is legally “dead bang” (\( v = 5, 7, 10 \)), the relative attractiveness of judicial and implicit plea bargaining depends on the level of caseload pressure. For low to moderate caseload pressure, \( P^*(\text{trial}) > .15 \), the story is as before. Judicial plea bargaining requires less extensive sentence discounting to equilibrate the system than does implicit plea bargaining, even though the magnitude of this relative superiority may not necessarily be great. But as caseload pressure
Plea Bargaining

increases beyond the $0.10 < P^*(\text{trial}) < 0.15$ range, relative superiority inverts: implicit plea bargaining emerges to dominate judicial plea bargaining in the eyes of the judge. The preconditions for this inversion of centralized authority over centralized contracting are a strong strength of case flow and a heavy caseload pressure. It remains to be explained why these patterns are observed.

It is easy to understand why judicial plea bargaining can frequently equilibrate caseload with less intense discounting than can implicit plea bargaining. Judicial plea bargaining trades in against defendant's risk aversion whereas implicit plea bargaining does not. Pre-plea sentence offers are customized to suit each individually differentiated criminal, with the result that increased certainty compensates for decreased benefits in expected values.

Given this straightforward logic, it is surprising that implicit plea bargaining is superior under any conditions. To understand this counterintuitive finding, it is necessary to reconsider who it is that is pleading guilty under these two systems. A given sentence discount generates a different compositional mix of defendants who insist on trial in these two systems. This fact has sharp consequences for the relative responsiveness of this residual pool to further inducements.

As is apparent from figure 1, in implicit plea bargaining defendants with the weakest evidentiary cases against them are likely to insist on trial whatever the sentence discount difference in lotteries. Conversely, defendants with the strongest cases against them are the most likely to respond to even small discount benefits. Increased sentence discounting in the implicit plea bargaining system thus cuts deeper and deeper into a residual pool of progressively weaker cases. When $\nu$ is low and hence the percentage of such cases is high, this mix poses a problem for the judge trying to manage a heavy caseload. As caseloads and guilty plea rates rise, judges increasingly are left with defendants who require a substantial discount to overcome their natural inclination for trial. Since judges are reluctant to engage in such heavy discounting, implicit plea bargaining can remain viable only when a strong strength of case flow mitigates this problem.

This case compositional problem also confronts judicial plea bargaining, as is apparent from the shift in judicial discount schedules as a function of $\nu$. In judicial plea bargaining, however, another compositional problem potentially exists as well. Not all defendants will be delighted with the sentences a judge offers "based on the facts as I see them." Defendants will respond differentially to explicit sentence offers depending on whether the judge evaluates them as "run-of-the-mill" criminals, with consequent lenient sentence offers, or as "hard-core" criminals, with consequent stiff sentence offers. Large numbers of defen-
dants in the first category can be creamed off with no discounting whatsoever ($d_k = 1$) simply by trading against their understandable risk aversion. Typically, however, defendants in the latter category will go to trial under the accurate calculation that they have nothing to lose. As guilty plea rates increase, therefore, judges in a judicial plea bargaining system are increasingly left with a recalcitrant residual pool of hard-core criminals (arrayed in the tail of the sentencing distribution) who require substantial discounts to overcome their knowledge of the judge's opinion of them. As sentence discounting intensifies (indicated by the sharp decline in the judicial schedule after the $d_k = 1$ plateau), judges lose control over just those defendants they most want to control.

This criminal composition problem is magnified somewhat under conditions of low $\beta_T$. If judges hold harsh internalized sentencing standards, they will perceive more defendants as hard-core criminals, thereby increasing the percentage of defendants who are reluctant to cop a plea. The aggregate consequence, as will be illustrated in figures 7 and 8, is that sentence discount schedules begin to deteriorate from $d_k = 1$ at lower caseload pressure levels when sentencing standards are harsh, $\beta_T = 1$, than they do when sentencing standards are lenient, $\beta_T = \gamma$.

Implicit plea bargaining is not faced with this problem of residual hard-core criminals because no idiosyncratic evaluative information is conveyed before the plea. Of course, defendants may have exogenous information such as prior record on which to base a prediction of the judge's sentence, but such information affects estimated trial and guilty plea sentence distributions equally, making it essentially irrelevant to the ratio-based plea choice. In judicial plea bargaining, in contrast, defendants are given precise information with no uncertainty about the consequences of one of their plea options. This makes defendants either more or less cooperative depending on the content of the information.

The puzzling inversion of implicit over judicial plea bargaining can now be explained. Increasing caseload pressure causes the case composition problem latent in implicit plea bargaining and the criminal composition problem latent in judicial plea bargaining to become manifest at different rates which depend on $\nu$ and $\beta_T$. When caseload pressure is light or moderate, the discounting benefits that judicial plea bargaining can reap through creaming run-of-the-mill criminals are decisive. When caseload pressure becomes very heavy, however, judicial plea bargaining is confronted with a problem that implicit plea bargaining does not share—a residual pool of hard-core criminals who now know the score against them. Implicit plea bargaining has a weak evidentiary case problem of its own, but this problem is either decisive or trivial depending on the aggregate proportion of such cases.

None of these comparisons between centralized forms of plea bargain-
Plea Bargaining

ing implies that either one of them is satisfactory. I turn now to comparing implicit plea bargaining with the two delegated forms of prosecutorial plea bargaining—first charge reduction and then sentence recommendation.

Charge Reduction versus Implicit Plea Bargaining

The role transformation at issue in this comparison is delegation within an authority relationship. Pre-plea defendants are construed as differentiated cases, and prosecutors customize individual sentence distributions to variation in strength of state's case. Both charge reduction and implicit plea bargaining, however, remain authority relations between the defendant and the judge since the defendant still must voluntarily submit himself to the future discretionary directive of the judge. One may anticipate that the case composition problem will be ameliorated through delegation and that the criminal composition problem will not exist.

Figure 5 presents, for this comparison, the same information contained in figure 4. Analysis in this comparison is much simpler. In terms of discounting, charge reduction is superior to implicit plea bargaining under all circumstances, period. If a judge is interested in strictly minimizing sentence discounts, it will always be advantageous to delegate to the prosecutor the ability to select the judge's own sentence distribution. Given the discussion above, the reason should be obvious: even though sentence outcomes remain uncertain under charge reduction (hence eliminating the criminal composition problem which confronts contracting systems), environmental variation in strength of state's case has been absorbed into sentencing thereby matching more closely the terms in which defendants think. The procedural norm that a defendant categorically either is or is not guilty is effectively abandoned.

Charge reduction plea bargaining is indeed the most stable of the four role structures examined, in the specific sense that its equilibrium discount schedules are least sensitive to environmental variation in \( \nu, \beta_T \), and \((\phi/\bar{\mu})\). Charge reduction is insensitive to \( \nu \) because heterogeneity in \( s_{ij} \) has been absorbed into a customized charge. Variation in internalized sentencing norms (\( \beta_T \)) is essentially irrelevant to comparisons between authority systems because in these systems defendants' choices are structured as relative comparisons of analogous uncertainties. Caseload pressure is relevant but not markedly so because charge reduction plea bargaining, unlike any other system, can generate \( P^*(\text{trial}) = 0 \) for \( d_{\mu} > 0 \). In contrast, the statutory charge structure in the criminal code is vital to the viability of charge reduction, since the corner of the discounting schedule is given by \( (\bar{\beta}_{\rho}/\beta_T) = \frac{1}{2} (1 + \bar{\mu}) \). (In fig. 5, I illustrate with \( \bar{\mu} = .6 \).) The more numerous and "fine-grained" are the statutory charges in
the code ($\mu$ near 1), the less sentence discounting is required on the part of the judge to equilibrate the system to caseload constraints.

Figure 3 shows why charge reduction is capable of driving the guilty plea rate to 100% with $d_\mu > 0$. Sentence discounts control the relative proportion of defendants who plead guilty in each reduced charge class. As sentence discounting increases, each of these proportions is driven up in "venetian blind" manner regardless of $s_{ij}$, since reduced charges have placed all evidentiary cases on a more or less equivalent footing with regard to plea choice. The authority-based power of uncertain averages is reflected: hard-core criminals plead guilty with the same frequency as do
all others, even though at sentence hearing they may be sorely disappointed. This is possible only because personal evaluations are systematically withheld from defendants. Charge reduction thus derives its robust power from combining discrimination in the labeling of crime (which overcomes the case composition problem) with secrecy in the evaluation of criminals (which suppresses the criminal composition problem).

The magnitude of this superiority of charge reduction over implicit plea bargaining is decisively affected by \( \nu \) because of the sensitivity of implicit plea bargaining to variation along that dimension. When virtually all cases are “dead bang” (high \( \nu \)), the quality of case flow is fairly homogeneous. The relative benefit of customization is correspondingly decreased (as is apparent in the graphs from the convergence of implicit plea bargaining with charge reduction plea bargaining as \( \nu \) becomes infinitely large).

In the case of low \( \nu \), these conclusions apply only to the pure form of charge reduction. For the real benefits of this system to become manifest, the judge has to back up the prosecutor fully. When the quality of case flow is not high, the judge who even occasionally ignores the prosecutor’s relabeling and punishes the defendant as if the original indictment were true (insofar as maxima permit) will witness a dramatic decline in guilty plea rates, which he or she will then have to combat with increased sentence discounting—thereby vitiating the comparative advantage of the system. A subsidiary conclusion of this analysis, therefore, is that administrative pressure for a rigid division of labor between the judge’s and the prosecutor’s informal roles is greatest when quality of case flow is weak.

Sentence Recommendation versus Implicit Plea Bargaining

Sentence recommendation represents the most radical role transformation from both implicit plea bargaining and the official system because delegation and contracting are involved simultaneously. Sentence recommendation shares with charge reduction plea bargaining the feature that sentences are customized by strength of state’s case. And sentence recommendation shares with judicial plea bargaining the feature that pre-plea sentence commitments are offered explicitly. When both features are in place simultaneously, the net result is a principal-agent structure in which the prosecutor selects the judge’s actions, subject only to post hoc approval. It can be anticipated that sentence recommendation plea bargaining shares the costs and benefits of these alternative role systems; the main issue is the relative magnitudes of the various effects.

Figure 6 presents the relevant equilibrium sentence discount schedules, again for an intermediate level of \( \beta_r \). The observed inversion pattern is
Fig. 5.—Sentence discount schedules: sentence recommendation vs. implicit plea bargaining. (Intermediate sentencing severity: $\beta_T = 3$.)

qualitatively similar to that observed in judicial plea bargaining (fig. 4), with one important difference. Pure sentence recommendation plea bargaining is impervious to variation in $\nu$, whereas pure judicial plea bargaining is not. The reason is the same as in charge reduction plea bargaining: problems of quality of case flow can be compensated for by customizing sentences in terms of $s_{ij}$.

Like judicial plea bargaining, sentence recommendation plea bargaining can induce large numbers of defendants to plead guilty with no sentence discounting whatsoever because of the risk aversion effect. The models predict, indeed, that pure sentence recommendation can generate
Plea Bargaining

a guilty plea rate in the neighborhood of 85% with \( d_\mu = 1 \). This exact figure depends on underlying assumptions about \( \beta_T \) and about the defendant's utility function, but the general order of magnitude remains constant under reasonable permutations (see n. 4). Effects such as this may underlie the empirical puzzle that sometimes sentence discounts can be statistically discovered and sometimes not, even though the logic of sentence discounts remains pervasive.

As before, none of this implies that all defendants will be pleased with the sentence offers they receive. Sentence recommendation, like judicial plea bargaining, is confronted with the criminal composition problem as caseload pressure becomes severe. Hard-core defendants who now know the score against them are understandably reluctant to cop a plea, whatever the strength of the state's case against them. The same flip in relative performance of implicit over sentence recommendation plea bargaining is therefore observed under conditions of very heavy caseload pressure and high average strength of case.

In terms of quantitative estimates, the greatest pressure on the judge to devolve sentencing responsibilities onto the shoulders of the prosecutor occurs in the caseload pressure range \( 10 < P^*(\text{trial}) < 20 \). When quality of case flow is weak, the judge in addition to delegation must hardly ever overturn the recommendations of the prosecutor. A high average strength of case flow, on the other hand, gives the judge more freedom to exercise his or her own independent judgment. Indeed, under quite delimited caseload pressure and quality of case conditions, the occasional reassertion of judicial autonomy paradoxically improves the discounting performance of the system. Not only alternative role structures but also relative divisions of labor, therefore, are affected by environmental context.

All Four Systems Compared

Figures 7 and 8 suppress mixed cases \( \rho \neq 0, 1 \) and present the sentence discount schedules of all four pure plea bargaining systems simultaneously, for two extreme cases of internalized sentencing norms. The first, \( \beta_T = 1 \), equivalent to \( E(x_{Tij}) = .5M_j \), represents a very harsh court system, which Levin (1971) has argued is often associated with middle-class, deterrence-oriented, bar association judges. The second, \( \beta_T = 7 \), equivalent to \( E(x_{Tij}) = .125M_j \), represents a fairly lenient court system, which Levin has argued is often associated with ethnic judges of partisan backgrounds. The intermediate \( \beta_T = 3 \) case, equivalent to \( E(x_{Tij}) = .25M_j \), has already been presented in figures 4–6.

The format of figures 7 and 8 highlights the effect of variation in \( \beta_T \). The figures reveal that the sentence discount performance of judicial plea bargaining is affected rather markedly, the performance of sentence rec-
ommendation plea bargaining is affected somewhat, and the performances of implicit and charge reduction plea bargaining essentially are not affected. In the contracting cases, the effects are in the expected direction: more severe sentencing standards require greater sentence discounting in order to equilibrate the system. The reason has already been given: when sentences are on average severe, more defendants are aware of the extremity of their specific plights (whereas authority structures merely pose two unpleasant lotteries, both of which shift in parallel to variation in $\beta_T$). To the extent to which this effect is important, one would expect to see authority structures associated with bar association legal
cultures, and one would expect to see contracting structures associated with partisan legal cultures.

On the whole, however, the graphs reveal that this effect of legal culture on alternative role structures is comparatively minor. If one examines only the relative "shift points" defined by intersections among these schedules, hardly any difference will be observed between figures 7 and 8. Variation in internalized sentencing norms may be crucial for other interesting aspects of criminal courts, but caseload pressure and average strength of case are more important determinants of role structure adaptations in plea bargaining.
Figures 7 and 8 are also useful for comparing the prosecutorial forms of plea bargaining apart from the logical baseline of implicit plea bargaining. Compare, for example, the two authority structures (implicit and charge reduction) and the two contracting structures (judicial and sentence recommendation) in isolation from each other. For any fixed set of environmental circumstances, the prosecutorial mode of plea bargaining is superior. If one is seeking an answer only to the question why most U.S. felony courts are dominated by prosecutors, one need go no further than this observation. Prosecutors begin with the advantage that they are closer to the facts of each case ($s_i$), but this advantage becomes translated into dominance only because judges frequently have a strong sentencing incentive to delegate away their own discretion.

The absolute magnitude of this incentive to delegate downward, however, is strongly a function of $v$. It is apparent from the graphs (and can be demonstrated analytically) that as $v$ goes to infinity the judicial models of each pair converge into their prosecutorial counterparts.

Now compare only the two prosecutorial forms of plea bargaining. Both forms of plea bargaining are essentially independent of $v$. Charge reduction is superior to sentence recommendation whenever caseload pressure is approximately $P^*(\text{trial}) < .1$. And sentence recommendation is superior to charge reduction whenever caseload pressure is approximately $P^*(\text{trial}) > .1$. Again, the specific number depends somewhat on underlying assumptions, but it is nonetheless true across permutations that as caseload pressure increases delegated authority emerges over delegated contracting as the dominant role structure of plea bargaining.

This prediction is unambiguous for either extreme of caseload pressure (assuming, as always, that judges want to maximize rather than just loosely satisfice on sentence discounts). Many observers have told us, however, that guilty plea rates of approximately 90%, $P^*(\text{trial}) = .1$, are empirically quite common (except for those jurisdictions with rapid uncontested bench trials). Mixtures of prosecutorial plea bargaining, in which some proportion of cases involves charge reduction and the other proportion involves sentence recommendation, have not been derived analytically. However it is apparent that the "envelope" of such equilibrium schedules, across varying such mixtures, will trace out a convex curve linking the two corners of each pure sentence discount schedule. The actual prediction for this modal empirical situation therefore is a mixture of charge reduction and sentence recommendation plea bargaining, with delegated authority emerging only gradually as caseload pressure starts to mount. Preliminary surveys of felony courts in the United States indicate that mixtures of prosecutorial forms of plea bargaining are a very frequently observed type.
DISCUSSION

The mathematical analyses above have yielded the following cross-sectional conclusions:

1. Prosecutorial forms of plea bargaining are more likely when average strength of case is low. Prosecutors begin with the advantage that they are closer to the facts of each case, but this advantage becomes translated into dominance only when judges have a strong sentencing incentive to delegate away their own discretion.

2. Authority and contracting forms of plea bargaining alternate in likelihood as caseload pressure increases. When caseload pressure is light, there is little sentencing reason for judges to abandon their official authority role. When caseload pressure is moderate to heavy, contracting of explicit sentences allows run-of-the-mill criminals to be creamed off with no sentence discount. When caseload pressure is extremely heavy, both the centralized and the delegated forms of authority allow judges to gain control over hard-core criminals. Of the four plea bargaining systems examined, charge reduction plea bargaining is the most stable across environmental circumstances.

3. Statutory and local standards of sentencing severity have little effect on alternate forms of plea bargaining, because defendants' plea choices are structured as relative comparisons. There exists a slight tendency for authority systems to be associated with harsh courts and for contracting systems to be associated with lenient courts, but the magnitude of this effect is small.

Projections of dynamics from analyses of comparative statics are precarious because convergence to new equilibria may not be quick. The implications of these results for the emergence and transformation of plea bargaining deserve to be mentioned, however, in order to clarify the connections between the above models and the more comparative and historical literature on plea bargaining. I will conclude with a few suggestions about the structural determinants of \( \nu, (k/n), \) and \( \beta_\tau, \) in that order. Only thus will it become clear how the various local problems examined above are created by formal control structures in the first place.

The models reveal that the delegation relation between the judge and the prosecutor is an adaptation to variation in \( \nu. \) The higher the proportion of weak cases which flow through the felony court, the greater is the discounting incentive for the judge to delegate discretionary judgment to the prosecutor. What are the sources of variation in \( \nu? \)

"Weak cases" are not necessarily equivalent to shoddy investigations. Felony courts are caught between contradictory standards of legal and factual guilt, both of which are enforced by external organizations which impinge on felony courts. Appellate courts are the source for legal stan-
American Journal of Sociology

dards of guilt, which involve due process, constitutional, and "beyond reasonable doubt" issues above and beyond a concern with maximizing probabilities of accurate factual guilt determination (Tribe 1971). The police who actually make arrests, on the other hand, are motivated by public order considerations which often imply immediate intervention on the part of police officers irrespective of "legal niceties" (Skolnick 1966). What is a "weak case" from the legal guilt perspective of an appellate court may well be a rational arrest from the factual guilt perspective of the police.

Ultimately, therefore, the flow of weak cases which confronts a felony court is grounded in a contradiction between appellate courts and police in the interpretation of the meaning of guilt. The wider the divergence the lower the v, regardless of the source of the divergence. When appellate courts add to exclusionary rules of evidence (as they did in the Warren court and as they did in the late 19th century), the proportion of weak cases confronting criminal courts increases. And when police can be proactive (e.g., the FBI) rather than reactive (e.g., most municipal forces) in their investigation of alleged crimes, the proportion of weak cases confronting criminal courts decreases (Hagan and Bernstein 1979).

In America, there are particular historical reasons for this contradiction which can help to explain the timing of the emergence of plea bargaining. The mid-19th-century Jacksonian movement resulted in the establishment of election procedures for trial judges and prosecutors and in the organization of police on local rather than on state lines (Friedman 1973; Miller 1973). Both developments helped to institutionalize factual standards of guilt. The late-19th-century consolidation of a professionally oriented legal elite created the social base for an appellate imposition of strict evidentiary standards to control from afar what were perceived to be lax standards at lower court levels (Friedman 1973). These twin developments indeed were not independent, since the institutionalization of factual standards of guilt evoked the imposition of legalistic controls which in turn provoked more radical informal subversions. The analysis in this paper suggests that the comparatively unusual role of the prosecutor in American criminal courts is rooted, at least in part, in this historical divergence of a legalistic center from a democratic periphery.

This broad structural context does not imply, however, that local court personnel are passive in their response to the flow of weak cases. Contemporary cities vary in their rates of dismissal, either by prosecutors at arraignments or by judges at preliminary hearings. When weak cases are dismissed in large volumes, the average strength of remaining cases is increased and the raw caseload is decreased. This paper has not examined explicitly the trade-off between plea bargaining and dismissals, but im-
Plea Bargaining

Plicitly the models predict a greater reliance on prosecutorial forms of plea bargaining when dismissal rates are low. Dismissal of cases before they reach felony courts and delegation of judicial discretion to the prosecutor within felony courts are alternative adaptations to the structural contradiction between appellate courts and police.

On the role relationship of authority versus contract, the models predict that the primary determinant is caseload pressure. In view of the controversy surrounding this variable, interpretation of this caseload pressure effect needs to be precise. Raw undischarged caseload should not be interpreted as the direct cause of role transformations in plea bargaining. Caseload is simply a constraint that all plea bargaining systems can satisfy. The role system consequences of any level of caseload depend on how the structural context of each court has shaped the substantive justice problems to be overcome. Instead of thinking of raw caseload as a direct cause of anything, it is more fruitful to think of caseload as a catalyst that induces latent problems to become manifest successively.

The particular structural variable which sets the stage for the impact of caseload in the above models is average length of trial. Langbein (1978, 1979) has itemized the historical developments that have transformed the American jury trial from a summary judicial proceeding to a highly ritualized contest between adversarial lawyers. Some of these changes involve the elaboration of the due process rights of defendants: (1) the creation of the right to defense counsel, (2) the institutionalization of public defender offices, (3) the interpretation of self-incrimination to exclude defendants as viable witnesses, and (4) the elaboration of exclusionary rules of evidence, partly on constitutional grounds. Other changes involve restrictions on judicial discretion in trials: (1) the introduction of voir dire adversarial selection of juries, (2) the shift from judge to prosecutor as the agent of cross-examination, (3) the restriction of judicial powers of summation to matters of law rather than fact, and (4) the increased access of defendants to procedural appeal. Together, these two types of change amount to a radical segregation, within trials, of the judge from both the defendant and the jury. The image of a judge actively pursuing justice within trials has been replaced by the image of a passive umpire.

The American judge's passivity in felony jury trials is striking when compared with other national systems. In Germany, judges participate in lay deliberations about both guilt and sentence (Casper and Zeisel 1972). In France, background sentencing information is introduced and discussed at trial (Pugh 1962). And in all Continental systems, judges actively interrogate defendants on the basis of pre-trial access to dossiers. In Great Britain, judges have greater discretionary control over juries.
through summation and over witnesses through questioning. Exclusionary rules are less elaborately developed, and defendants can choose to have all but the most severe felony charges tried at the misdemeanor court level (Heberling 1973). The net consequences of these different trial structures, compared with the American system, are that judges have greater influence over trial process and outcomes and that felony trials are on average shorter.

In America, a consequence of the strict legal segregation of judge from defendant is that routine processing of run-of-the-mill criminals is inhibited. Procedurally rigid jury trials simultaneously intensify judges' administrative need to reach out for the cooperation of such defendants and inhibit their official ability to do so. Before trial, judges can neither distinguish run-of-the-mill criminals from the others nor single them out for routine treatment without the defendants' consent. Contracting emerges as a locally created and case-specific bridge to overcome this externally imposed segregation. In their procedural effort to minimize judicial discretion within trials, appellate courts and legislatures have minimized trials.

In Continental systems, in contrast, judicial manipulation of particular run-of-the-mill criminals can be accomplished within trials because of the greater pre-trial and trial judicial access to defendants (Goldstein and Marcus 1977). And in Great Britain, most routine felony criminals can be induced to clear themselves at the misdemeanor court level through legitimately heavy sentence discounts, both within and across court levels (Heberling 1973).

The significance of these formal trial considerations for extra-trial adaptations like contracting is reinforced through consideration of those few American jurisdictions where overt plea bargaining is relied on less heavily than elsewhere. Baltimore, Los Angeles, Pittsburgh, and Philadelphia offer comparatively speedy judicial bench trials at the felony level, in which the role of lawyers is greatly reduced, as a structural alternative to overt plea bargaining. A titular plea of not guilty to such a summary bench trial is often considered in practice to be a "slow plea" of guilty because of the resemblance of such a proceeding to a sentence hearing, in which the salient considerations are the prior history of the criminal and the mitigating circumstances of the crime (Greenwood et al. 1973; Eisenstein and Jacob 1977). If proportionate reliance on summary bench trials is only moderate, sentence discounts may be reduced but overt forms of plea bargaining will not vanish. But if bench trials are relied on extensively enough (and if those trials are speedy enough), implicit plea bargaining with little sentencing difference between bench trial and guilty plea is possible because caseload pressure is so drastically
reduced. A sentence differential between bench trials and jury trials must remain, however, for such a system to be viable. Whereas dismissals can be a functional alternative to reliance on prosecutors, summary bench trials can be a functional alternative to contracting.

The findings about the effect of sentencing norms have been largely negative, but a few concluding remarks are necessary in order to clarify disagreements with existing literature. I will discuss the effect of statutory criminal codes first and that of local legal culture second.

Herrmann (1974) has hypothesized that plea bargaining is uniquely American because of the comparatively harsh criminal codes in America (relative to Western Europe). The analysis in this paper does not support this particular argument, since $M_j$ canceled out all $P_{\text{trial}}$ formulae. The reason is straightforward: apart from the unexamined effect of transaction costs, defendants' plea choices are based on the relative, not absolute, comparison of two unpleasant alternatives.

The analysis is more consistent with statutory code arguments that focus on the variance in sentencing options available to court personnel. A large number of charging options (nearly 1) is crucial for the viability of charge reduction plea bargaining, just as Alschuler (1976) has argued. And while mandatory sentence minima have not been explicitly incorporated into the models' sentence distributions (see n. 6), it seems both well established and consistent with the logic of this analysis that charge reduction plea bargaining can circumvent such statutory restrictions on judicial discretion (Newman 1966). The contribution of this paper to debates about mandatory sentencing is only to point out that the impact of criminal codes on alternative forms of plea bargaining is confined mostly to the charge reduction mode.

The largely negative finding about the relative effect of internalized sentencing norms assumes significance in the light of the common historical argument that plea bargaining originated in urban political machines (Haller 1970). Insofar as machines were one element in the institutionalization of factual standards of guilt, the analysis in this paper is not inconsistent with this argument. But insofar as machines are thought to generate plea bargaining because of their inclination to construe defendants particularistically and to punish them leniently as a result, the analysis implies that plea bargaining should not be historically unique to machines. A wide variety of local legal cultures are capable of generating plea bargaining systems, with partisan legal cultures only marginally more inclined to contracting forms. This conclusion is given indirect empirical support by Moley's finding (1929) that 19th-century guilty plea percentages rose at approximately the same rates in upstate New York as they did in New York City.
CONCLUSION

This paper has been structured by three related questions about plea bargaining: How do informal role structures emerge in criminal courts? How are these role structures shaped by the structural context in which they are embedded? Why do U.S. cities have different plea bargaining systems?

The first two questions have received clear answers in this paper. Informal role structures in plea bargaining emerge through temporal transformations in the sequence of official procedures, which shape the terms in which pre-plea defendants are construed. And alternative role structures in plea bargaining are embedded in environmental context through sentence discount schedules, which reflect the impact of structural constraints on the achievement of locally defined substantive justice.

The third question has been answered abstractly: the equilibrium likelihood of different plea bargaining systems depends on the average strength of case, the caseload pressure, and (to a lesser extent) the local sentencing norms that characterize each city. To apply these predictions to specific cities, however, will require empirical testing. In principle, testing is feasible since most of the independent variables, \( n, l, k, \beta_f, \mu, \) and \( P^*(\text{trial}) \), are directly observable. Measurement of \( v \) is more problematic, but trial acquittal rates are available as long as the models presented above can be used to help eliminate selection biases.

Real difficulties exist, however, on the dependent variable side. Officially published statistics are virtually worthless for testing because crucial elements, such as sentence discounts and the character of plea bargained inducements, are systematically not reported. A few more refined data sets do exist (e.g., Eisenstein and Jacob 1977; Vera Institute 1977; Rhodes 1978), but they do not span many jurisdictions. Future empirical evaluation will no doubt require a difficult interlacing of sensitive case accounts with what aggregate statistical measures exist.

Such empirical problems are inevitable in any study of informal circumventions of official control. However, if organization theorists are stimulated into a renewed interest in one of the classic questions of the subfield, a minimal goal of this paper will have been achieved. Informal roles do indeed emerge locally and crescively within organizations, but the structure of such informal adaptations is not independent of the formal control structure they subvert.

REFERENCES

Plea Bargaining


American Journal of Sociology


800