Judicial Mechanism Design

Ron Siegel and Bruno Strulovici

July 2018

Introduction

- A criminal defendant goes through a complex process
  - Arrest, plea bargaining, cross-examination, verdict, sentencing, etc.
- Existing work considers different aspects of the process
  - Grossman and Katz (1983), Reinganum (1988), Baker and Mezzetti (2001), Kaplow (2011, 2017), Daughety and Reinganum (2015)
- We investigate a broad class of processes that determine guilt and appropriate punishment from two different welfare perspectives
  - Impose little structure on the process
- Provide insights into key features of existing judicial systems
- Conduct a mechanism design analysis focused on the defendant's private information

Introduction

- Reduce judicial process to a single-agent mechanism
- Derive properties of interim and ex-ante welfare-maximizing mechanisms
  - Welfare criteria differ in their treatment of deterrence
  - Properties hold if optimize over more instruments (prevention, policing, etc)
- Similarities and differences with features of the American criminal justice system
  - Plea bargains, trials with binary verdicts, evidence threshold similar to BARD
  - Adversarial system, separation of fact-finding and sentencing as commitment devices

Today

- Judicial mechanism
- Interim welfare
- Ex-ante welfare
- Main assumption and class of mechanisms
- Interim optimal mechanism
- Ex-ante optimal mechanism
- Comparison to existing judicial systems

Judicial mechanism

- A crime has been committed and a suspect is arrested and charged
- The defendant is privately informed about his guilt, $\theta \in \{i, g\}$
  - Prior λ that the defendant is guilty

    (main results are “prior-free” can be stated as complete class theorems)
- Criminal justice machinery put into motion, leading to a judicial decision and a sentence
  - May involve multiple actors and several stages
- Model the process as an extensive-form game and an equilibrium
- Summarize the process by a signal $t \in [0,1]$ regarding the defendant’s guilt and a mapping from signals to (possibly random) sentences
- Consider the corresponding truthful mechanism: direct-revelation mechanisms in which the defendant truthfully reports his guilt $\hat{\theta} \in \{\hat{\imath}, \hat{g}\}$

Reduction to a single-agent DRM

- Take an extensive form game and an equilibrium
- Fix strategies of other players, and focus on strategy of the defendant, which is a function of his type
- Consider Direct Revelation Mechanism in which defendant reports his type, and corresponding strategy is played
- Truth-telling is optimal
- Outcome of the game: signal t capturing likelihood of guilt + sentence s
- Normalize signal t to lie in [0,1]

Summary: Judicial DRM

- A mechanism is a pair $M = (F, S)$, where
  - $F = \left(F_i^{\hat{\imath}}, F_g^{\hat{\imath}}, F_i^{\hat{g}}, F_g^{\hat{g}}\right)$ is a vector of signal distributions
  - $S\left(t, \hat{\theta}\right) \in \Delta[0, \bar{s}]$ is a sentence function
    - $\bar{s}$ Is the highest allowable sentence for the crime

- Signal has support $[0,1]$ and is ordered by its likelihood ratio (wlog)

- Signal distributions have positive densities $f_\theta^{\hat{\theta}}$, and $f_g^{\hat{\theta}}(t)/f_i^{\hat{\theta}}(t)$ is strictly increasing in t

(non atomicity can be relaxed)

Interim welfare

- Denote by $W(s,\theta)$ the welfare from imposing sentence s on defendant of type θ
  - $W(\cdot,g)$ continuous, concave, and single peaked at $\hat{s} > 0$
  - $W(\cdot,i)$ continuous, concave, and strictly decreasing, $W(0,i) = 0$
- Given prior λ, sentence s leads to welfare

$$\lambda W(s,g) + (1-\lambda) W(s,i)$$

  - ($W(\tilde{s},\theta)$ is also expected welfare from random sentence $\tilde{s}$)
- Interim welfare given a mechanism is

$$\lambda\left(E^{F_g^{\hat{g}}}\left(W(S(\cdot,\hat{g}),g)\right) - C(F_g^{\hat{g}})\right) +$$
$$(1-\lambda)\left(E^{F_i^{\hat{\imath}}}\left(W(S(\cdot,\hat{\imath}),i)\right) - C(F_i^{\hat{\imath}})\right)$$

  - $C(F_\theta^{\hat{\theta}}) \geq 0$ is expected welfare cost of generating $F_\theta^{\hat{\theta}}$

Ex-ante welfare

- Also considers the number of crimes committed
- The mechanism acts as a a deterrent
- Individuals weigh cost and benefit of committing crime
  - Benefit varies in the population
- At most one individual is apprehended and prosecuted for it
- Ex-ante social welfare given mechanism M is

$$H(M)\left(\pi_g\left(E^{F_g^{\hat{g}}}\left(W(S(\cdot,\hat{g}),g)\right)-C(F_g^{\hat{g}})\right)\right)$$

Optimal mechanisms

- Derive properties of optimal mechanisms for interim and ex-ante welfare

- Let $u(s)$ be an individual's utility from sentence s
- Assume that when defendant is innocent, social preferences over sentences agree with those of the defendant
  - $W(s, i) = u(s)$ (generalizes to $W(s,i) = \Phi(u(s))$ with Φ increasing and convex)

- **Feasible DRM** = DRM obtained from earlier reduction for some game and equilibrium. Which (truthful) mechanisms are feasible?
  - Depends on technology, unmodeled agents

- **Assumption 1**: *Replacing the sentence function in a feasible mechanism with any other sentence function that maintains truthfulness leads to a feasible mechanism*
  - Puts some structure on the set of feasible truthful mechanisms
  - Captures a notion of commitment

Interim welfare

- A mechanism is *interim optimal* if it maximizes interim welfare among all feasible mechanisms
  - Considering interim welfare allows us to disentangle the effect of deterrence from other welfare implications

- Theorem 1: *Any interim optimal mechanism has the following properties*
  - *The innocent defendant's sentence is a step function of* t*, which jumps from* 0 *to* $\bar{s}$ *at some cutoff signal* $\bar{t}$
  - *The guilty defendant's sentence is constant*

Interim welfare

- Resembles system in which plea is available before trial, and trial ends in one of two verdicts
- If defendant pleads guilty, fixed sentence and avoids trial
- Otherwise, faces a trial and is either "acquitted" and obtains a sentence of 0 or "convicted" and obtains a severe sentence
- Conviction occurs if evidence is sufficiently strong (exceeds a threshold)
- *No punishment* following an "acquittal" was not assumed
- Extreme sentence not due to deterrence (≠ Becker (1968))
- Signal not used following a "guilty" plea, even if informative
  - Not due to cost saving (but shows cost saving need not be inefficient)
  - Screening value of pleas, noted by Grossman and Katz (1983)
- Relation to Crémer-McLean: FB achievable if disutility and likelihood ratio are both unbounded, or if utility unbounded in both direction

Proof idea

- Fix a feasible mechanism. Modify sentences (only) to maximize social welfare subject to truthful reporting
- Use signal to incentivize truthful reporting
- Intuitively, binding IC: guilty pretends to be innocent
- Given utility level for innocent, choose sentence scheme that is least attractive for the guilty
- MLRP implies that it is a step function with extreme sentences
- Guilty sentence is constant because defendant and society are risk averse and innocent’s IC not binding

Proof

Ex-ante welfare

- Sentence modification may affect deterrence and thus number of crimes committed
- Affects ex-ante welfare
- If, in Theorem 1, guilty’s constant sentence is less than ex-post optimum $\hat{s}$, construction leaves guilty’s utility unchanged
  - Set of individuals who commit the crime does not change
- Corollary: *Theorem 1 characterizes mechanisms that maximize ex-ante welfare among all mechanisms in which guilty’s certainty equivalent is less than ex-post optimal sentence*

Ex-ante welfare

- In general, deterrence may optimally require a higher sentence than the ex post optimum $\hat{s}$
- Construction in Theorem 1 then reduces sentence of the guilty
- Increases interim welfare but also the utility of the guilty
- Leads to more individuals committing the crime, which may decrease ex-ante welfare

Ex-ante welfare

- A mechanism is *ex-ante optimal* if it maximizes ex-ante welfare among all feasible mechanisms
- Theorem 2: *Any ex-ante optimal mechanism (generically) has the following properties*
  - *The innocent defendant's sentence is a step function of* t*, which jumps from* 0 *to* $\bar{s}$ *at some cutoff signal* $\bar{t}$
  - *The guilty defendant's sentence is either constant or is a lottery over two sentences in* $[\hat{s}, \bar{s}]$*. The lottery can be chosen to be independent of the signal*

Ex-ante welfare

- Similar to interim optimal mechanisms, except for possibility of random guilty sentence
- May be optimal to give guilty defendant a constant sentence even when it is higher than ex-post optimal
- For random sentence to be optimal, need two things:
  - Deterrence optimally requires sentences that are higher than ex-post optimal; happens when deterrence concern dominates welfare loss from excessive punishment
  - Society must be sufficiently less risk averse, conditional on facing a guilty defendant, than individuals

Skip

Proof idea

- Modify only the sentences to increase welfare
- Similarly to Theorem 1, optimal scheme for innocent is a step function with extreme sentences
- Given a utility level for the guilty, choose threshold for step function to make the guilty indifferent
- Choose sentence scheme for guilty that maximizes welfare conditional on facing the guilty among all schemes that give him this utility level
  - No distortion because innocent does not want to mimic guilty
- This involves a concavification argument reminiscent of optimal contracting and information design
  - Here randomization concerns defendant's utility rather than belief

Proof

Similarities to the American legal system

- If a plea bargain is reached, no trial
  - Uncertain outcome for serious crimes (deterrence important)
- A trial ends with one of two outcomes: an acquittal (no punishment) or a conviction (punishment that is severe relative to the plea bargain)
  - Conviction if the evidence is sufficiently incriminating (similar to BARD)
- (Did not assume a binary verdict, no punishment following an acquittal, or availability of plea bargaining)

End

The role of evidence

- In trials, evidence is used to determine defendant's guilt

- In optimal mechanisms, evidence is used to incentivize guilty defendants to admit their guilt

- Appear similar: BARD

Commitment and Assumption 1

- Optimal mechanisms achieve full separation
- Only innocent goes to trial, punished if evidence is sufficiently incriminating
- Relies on Assumption 1
  - Feasible to punish defendant known to be innocent

- US system does try to minimize the influence of punishment severity on verdict determination
  - Separation of fact finding and sentencing
  - Keep the jury uninformed about possible punishment

Reconcile

Conclusion

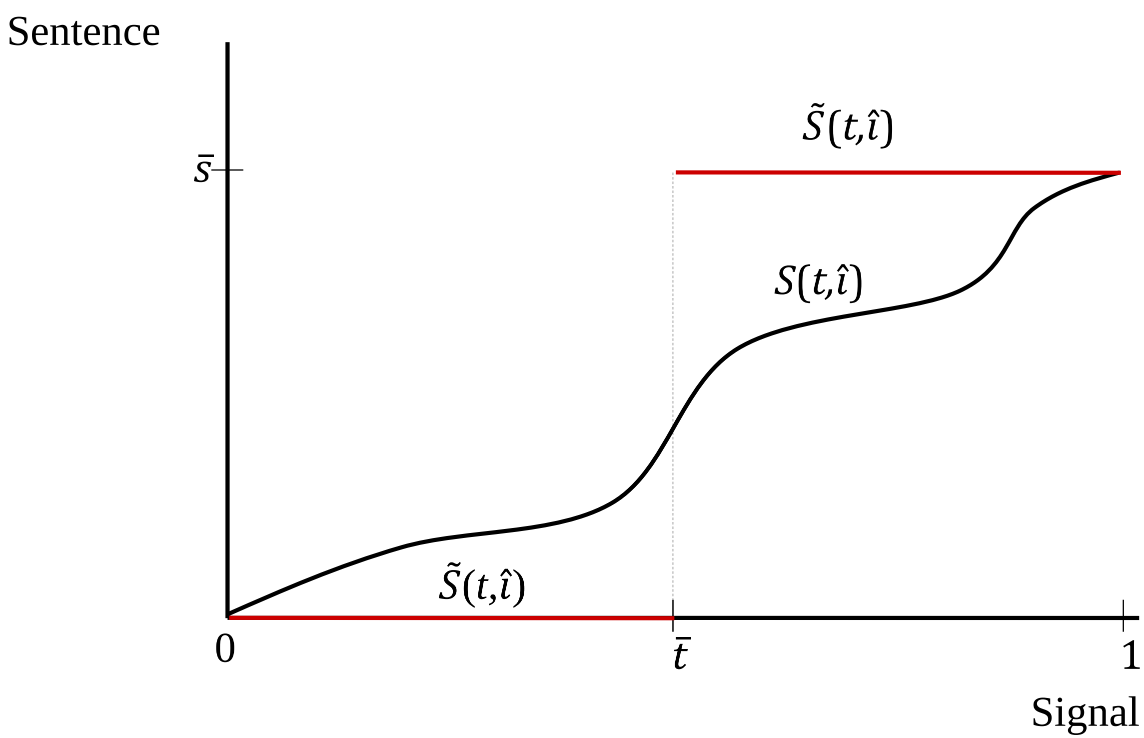
- Mechanism design approach to study optimal judicial systems
  - Reduce judicial process to single-agent mechanism
  - Formalize notion of commitment
  - Identify properties of optimal mechanisms
- Consider interim and ex-ante welfare
- Features that parallel those in the American criminal justice system
  - Plea bargains, trials with binary verdicts, adversarial, fact-finding and sentencing
- The role of evidence

Proofs

Proof

- Show that any feasible mechanism can be improved upon by a mechanism as stated in the theorem, with a strict improvement if the mechanism is not as stated in the theorem
- Consider a feasible mechanism $M = (F, S)$
- Modify S to increase interim welfare and maintain incentive compatibility
- Replace $S(\cdot, \hat{\imath})$ with step function $\tilde{S}(\cdot, \hat{\imath})$ to make an innocent defendant indifferent

Improvement with pleas



- Choose $\bar{t}$ to make the innocent indifferent
  - $(u(0)F_i^{\hat{\imath}}[0,\bar{t}] + u(\bar{s})F_i^{\hat{\imath}}[\bar{t},1]$ is continuous in $\bar{t}$)

(If distribution has atoms, may randomize at threshold.)

Proof

- Function $D(t) = u\big(S(t,\hat{\imath})\big) - u\left(\tilde{S}(t,\hat{\imath})\right)$ crosses 0 once, from below
- The ratio $f_g^{\hat{\imath}}(t)/f_i^{\hat{\imath}}(t)$ is increasing in t, by MLRP

- Lemma (Karlin 1968): *Under the two conditions above*,

$$\int_0^1 D(t) f_i^{\hat{\imath}}(t)dt \geq 0 \Rightarrow \int_0^1 D(t) f_g^{\hat{\imath}}(t)dt \geq 0$$

- So, conditional on misreporting his type, a guilty defendant prefers sentence function $S(\cdot,\hat{\imath})$ to $\tilde{S}(\cdot,\hat{\imath})$
- By truthfulness of the original mechanism, he prefers reporting truthfully with sentence function $S(\cdot,\hat{g})$ to misreporting with sentence function $S(\cdot,\hat{\imath})$
- So incentive compatibility holds when $S(\cdot,\hat{\imath})$ is replaced with $\tilde{S}(\cdot,\hat{\imath})$

Return

Proof

- Denote by s^{ce} and s^{a} the guilty defendant's certainty equivalent and expected sentence when reporting truthfully
- By concavity of $u(\cdot)$, $s^{ce} \geq s^{a}$
- Set plea sentence $s^{b} = \min\{s^{ce}, \hat{s}\}$
- This increases social welfare conditional on facing the guilty
  - If $s^{b} = \hat{s}$, then $W\left(s^{b}, g\right)$ is the highest possible utility
  - If $s^{b} < \hat{s}$, then $W\left(s^{b}, g\right) = W(s^{ce}, g) \geq W(s^{a}, g)$ and the concavity of $W(\cdot, g)$

Proof

- Because $s^b \leq s^{ce}$, truthfulness is maintained for the guilty
- Increase threshold $\bar{t}$ until the guilty is indifferent between s^b and misreporting with sentence function $\tilde{S}(\cdot, \hat{\imath})$
- This increases welfare and guarantees truthfulness by the innocent, by MLRP and the lemma

Return

Proof

- Consider a feasible mechanism $M = (F, S)$
- Replace $S(\cdot, \hat{\imath})$ with step function $\tilde{S}(\cdot, \hat{\imath})$ to make an innocent defendant indifferent
- Increase the threshold $\bar{t}$ until the guilty defendant is indifferent between reporting truthfully with $S(\cdot, \hat{g})$ and misreporting with $\tilde{S}(\cdot, \hat{\imath})$
- This increases social welfare and maintains truthfulness for the innocent

Proof

- Denote by U^g the guilty defendant's expected utility in mechanism M
- Replace sentence function $S(\cdot, \hat{g})$ with $\tilde{S}(\cdot, \hat{g})$ that the guilty is indifferent to and that maximizes ex-ante welfare

$$H(\widetilde{M})\left(\pi_g E^{F_g^{\hat{g}}}\left(W\left(\tilde{S}(\cdot, \hat{g}), g\right)\right) - C(F_g^{\hat{g}})\right.$$

Proof

- Reformulate the problem in terms of the defendant's utility
- Let $\widehat{W}(U) = W(u^{-1}(U), g)$ be the social welfare from sentencing the guilty to a sentence that gives him utility U
- Choose utility mapping $\hat{u}(t) \in \Delta[u(\bar{s}), u(0)]$ to maximize

$$E^{F_g^{\hat{g}}}\left(E\left(\widehat{W}(\hat{u}(\cdot))\right)\right) \text{ s.t. } E^{F_g^{\hat{g}}}\left(E(\hat{u}(\cdot))\right) = U^g$$

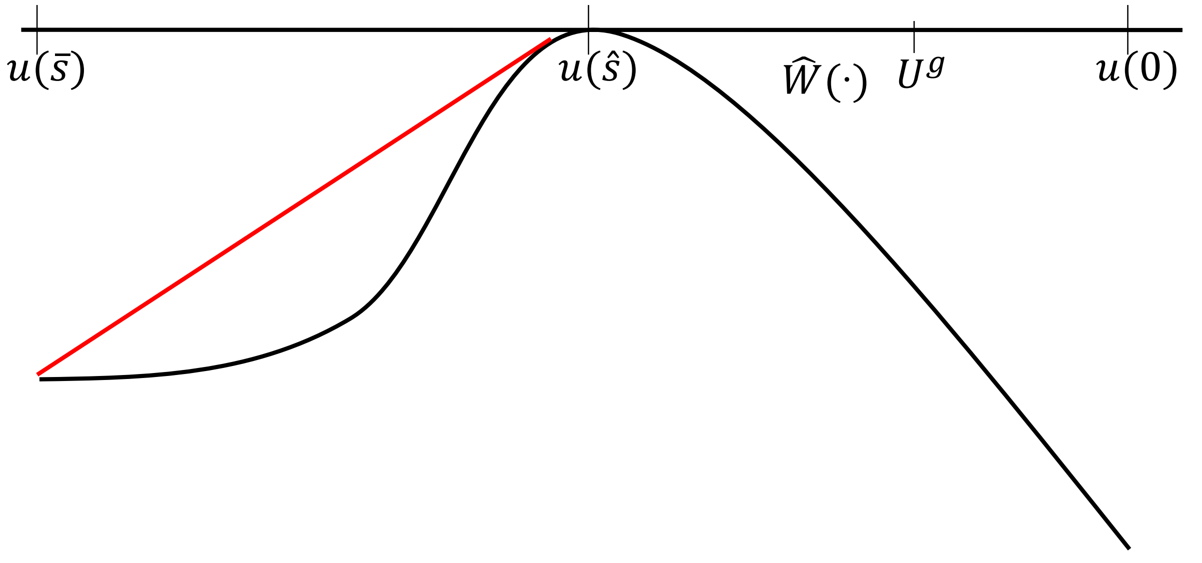
Proof

- Mapping $\hat{u}$ induces a single distribution in $\Delta[u(\bar{s}), u(0)]$
- Thus, consider choosing utility distribution $\dot{u} \in \Delta[u(\bar{s}), u(0)]$ to maximize

$$E\left(\widehat{W}(\dot{u})\right) \text{ s.t. } E(\dot{u}) = U^g$$

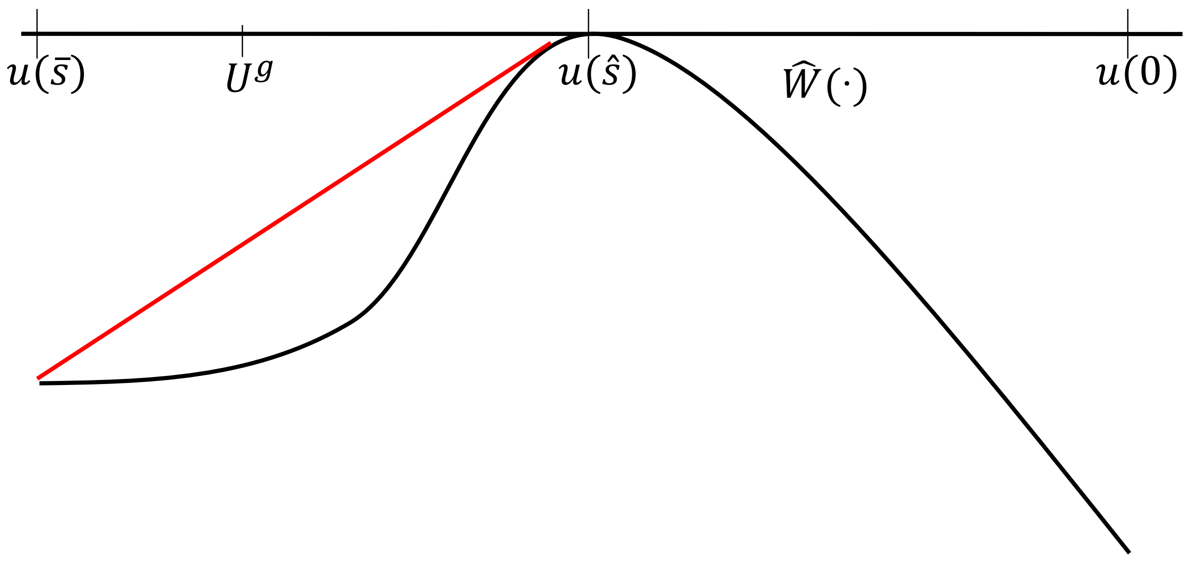
- The maximal value is $\overline{W}(U^g)$, where $\overline{W}$ is the concavification of $\widehat{W}$

Proof



- If $\overline{W}(U^g) = \widehat{W}(U^g)$, it is achieved by the constant sentence $u^{-1}(U^g)$
- $\overline{W}(\cdot) = \widehat{W}(\cdot)$ on $[u(\hat{s}), u(0)]$

Proof



- If $\overline{W}(U^g) < \widehat{W}(U^g)$, it is achieved by randomizing between two sentences
- Both sentences exceed $\hat{s}$

Return

Sending guilty defendants to trial

- In reality most convicted defendants are guilty
- Are existing trials far from optimal?

- In an optimal mechanism the guilty are indifferent between the plea bargain and going to trial
- Suppose a small fraction α of guilty defendants go to trial
- Given signal t, Bayesian updating gives guilt posterior $$p(t) = \frac{\lambda \alpha r(t)}{\lambda \alpha r(t) + (1-\lambda)}, \text{ where } r(t) = f_g^{\hat{\imath}}(t) / f_i^{\hat{\imath}}(t)$$

Sending guilty defendants to trial

- For an illustration, suppose that $r(\bar{t}) = 10$
  - the likelihood ratio at the optimal threshold is 10
  - The “Blackstone ratio”
- Suppose that $\lambda = 0.9$
  - 90% of defendants are guilty
- For $\alpha = 0.1$, the lowest posterior associated with a conviction is

$$p(\bar{t}) = \frac{9\alpha}{9\alpha+0.1} = \frac{0.9}{0.9+0.1} = 0.9$$

- “Certainty threshold is 90% when 10% of guilty defendants go to trial”

- Small welfare loss relative to the optimal mechanism when α is small

Return