Why are vulnerable regimes stable? Regime change games with an active defender*

Ole Jann — CERGE-EI

Christoph Schottmüller — University of Cologne and TILEC

February 19, 2020

Abstract

We analyze a regime change game in which an active defender can invest in costly, unobservable defenses. We show that if there are sufficiently many potential attackers, the game has a unique Nash equilibrium in which the defender chooses to have almost no defenses and attacks almost never occur. This provides a new perspective on coordination problems and the necessity of refinements in coordination games. We suggest that our result can explain why regimes that appear highly vulnerable to coordinated attacks can actually remain stable.

Keywords: panopticon, coordination, regime change games, global games, transparency

JEL classification: D23 (Organizational Behavior), D74 (Conflict, Revolutions), D82 (Asymmetric and Private Information), Z13 (Economic Sociology)

*Jann: CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences; ole.jann@cerge-ei.cz. Schottmüller: Department of Economics, University of Cologne; c.schottmueller@uni-koeln.de. We are grateful for helpful comments by Alberto Alesina, Hans Carlsson, Eric van Damme, Eddie Dekel, Jeff Ely, Nicola Gennaioli, Ian Jewitt, Heidi Kaila, Paul Klemperer, Nenad Kos, Pablo Kurlat, Meg Meyer, Stephen Morris, Marco Ottaviani, Alessandro Pavan, Jens Prüfer, David Ronayne, Karl Schlag, Tomas Sjöström, Joel Sobel, Peter Norman Sørensen, Jakub Steiner, Adrien Vigier and Jan Zápal as well as from audiences at Bocconi University, the Universities of Bonn, Copenhagen, Hamburg, Lund and Oxford, Tilec, SING 2016 (Odense), GAMES 2016 (Maastricht) and EEA 2017 (Lisbon). This research has been supported by PRIMUS project 20/HUM/019.

1. Introduction

Regime change games describe situations in which a group of individuals must work together to overcome a common opponent. Examples include speculative attacks, bank runs, revolutions or prison riots. For the theorist, such games pose two puzzles: First, predictions are often not clear-cut, as the attackers' actions are strategic complements and hence there are multiple, extreme Nash equilibria (in which everybody attacks or nobody does). Second, any result will crucially depend on the relative strength of the opponent compared to the combined strength of the attackers – yet what determines this strength? A common modeling approach has been to answer both questions in one, by assuming that the opponent's strength is a "state of nature" which individuals learn imperfectly – see, for example, the "global games" approach which refines the set of equilibria and yields a unique prediction.[1]

Yet in many of the situations which regime change games model, the defense strength is not a random caprice of nature, but the conscious choice of an active defender. A central bank could build up currency reserves in anticipation of having to counter a speculative attack, a company may raise capital to deter a take-over, and prison wardens or dictators can invest in security forces. How does the strategic agency of such a defender affect the thinking of the participants and the outcome of the game?

In this paper, we consider a regime change game in which there is an active defender who can invest in costly defenses. In particular, we consider a model in which the defender can make his choice in secret, as this maximizes the importance of his agency.[2] We show that introducing this assumption completely changes the structure of the conflict between attackers and defender, and leads to drastically different predictions from canonical models: If the number of potential attackers is large enough, there exists a unique Nash equilibrium in which the defender chooses to make almost no investment in defenses, yet attacks almost never occur.

We think that this result is interesting for four reasons. First, it shows that models with strategic complements can yield unique Nash equilibrium predictions if we consider a rich enough player set, without the need for any refinement methods. Since the assumption of a stochastic state of nature has usually been seen as a simplifying one, imposed for solvability, it is a striking implication of our model that it may in fact complicate matters. Second, our main result contains a discontinuity: It only applies for a large enough number of potential

[1] Cf Rubinstein (1989) and Carlsson and van Damme (1993) for ground-breaking work on global games, Morris and Shin (1998) and Goldstein and Pauzner (2005) for applications to speculative attacks and bank runs respectively, and Morris and Shin (2003) for an early survey.

[2] If his choice were observable to the attackers, we would be back to models in which the higher-order beliefs of players determine equilibrium selection – we compare our results to such models in the supplementary material.

attackers – intuitively, their coordination problem has to be severe enough for our result to hold. Indeed, we will provide (and discuss in detail) a novel understanding of what a "coordination problem" is in this context: Not the problem of trying to predict what is in others' minds, but of trying to be collectively unpredictable – a related, but not identical problem. This is in contrast to many of the models of coordination problems, which have qualitatively the same result for a game involving two as one involving millions of players. Third, it shows how the law of large numbers can be used to examine the existence of mixed Nash equilibria: In our main proof, we will make use of the fact that the aggregate behavior of a group of players in a mixed equilibrium will become relatively more predictable as the size of the group increases. This is, to our knowledge, the first time that such a proof technique has been used to establish (non-) existence of mixed Nash equilibria. Fourth, and finally, we believe that our result can explain phenomena observed in the real world, in which we often see that systems which may ostensibly appear vulnerable in fact maintain stability at relatively low cost.

In the remainder of this introduction, we will describe our main result in more detail and give an intuitive explanation for it. Within the main text of the paper we describe our model and explain the proof technique in sections 2 and 3; the full proof is in appendix A. Despite the complexity of the formal argument, we argue in section 4.3 that the result is intuitive enough to matter for the real-world situations we have in mind; we also show how it is very robust to a number of modifications and extensions (section 4.1) and relate it to the literature on information design (section 4.2).

An Intuitive Explanation of Our Result

While our model has many applications, we will consider the example of a prison to fix ideas. Imagine that a prison warden chooses how many guards he wants to hire; guards are costly. Afterwards a number of prisoners decide whether to revolt or not. If the number of prisoners who revolt is higher than the number of guards, these prisoners win and escape from the prison. Otherwise, the prisoners who revolted get punished. Prisoners who do not revolt will neither escape nor be punished. We assume throughout that the warden can make his choice in secret, so that the choices of the warden and the prisoners are in effect simultaneous. Given the similarity of our model to Jeremy Bentham's (1787) plan of how to build the perfect prison (and, as we will see, the similarities of our solution with his prediction), we call our model the "panopticon".

Our main result (Theorem 1) is that there is a unique Nash equilibrium in which the warden hires almost no guards and prisoners almost never revolt. Despite the fact that there is at most one guard (and sometimes none), a successful breakout almost never occurs. This

result arises only if the number of prisoners is sufficiently high; for low numbers of prisoners there are many more Nash equilibria.

How does the equilibrium uniqueness arise, and why does it only arise for large numbers of prisoners? We begin by noting that the situation between the warden and the group of prisoners is one of conflict: After the game has taken place and a successful breakout has happened (or not), either the warden or at least one prisoner must always regret their action.[3] This is a feature which our model shares with more simple simultaneous-move games of conflict such as *matching pennies* or *rock-paper-scissors*.

In such games, a player is weakened by being predictable: It is easy to win a game of *rock-paper-scissors* against someone who always plays *rock*, and still possible to beat them more than half the time if for example they are unable play *scissors* more than 10% of the time. In simple two-player games, players gain unpredictability by playing a mixed strategy – indeed, that is the only equilibrium of *rock-paper-scissors*. In our model, the warden can be as unpredictable as he likes by similarly randomizing between guard levels. Each prisoner can also be unpredictable by randomizing between revolting and not revolting. But, crucially, a group of randomizing prisoners will always be more predictable in aggregate than any single one of them. If, for example, 20 prisoners were to flip a coin to make their choice, the number of attacking prisoners will be between 9 and 11 more than a third of the time. As the number of prisoners grows, their unpredictability (relative to their number) falls. This means that, if there are many prisoners, there cannot be any equilibria in which breakouts happen frequently, since the warden would then know – more or less exactly – how many guards he would have to hire to prevent them. This means that the only Nash Equilibrium that exists is one in which prisoners almost never attack, the warden needs almost no guards, and successful attacks almost never occur. To be sure: This equilibrium also exists in slightly different models, such as can be found in the canonical literature on speculative attacks.[4] But our result is that in the situation we describe, this equilibrium is unique, and that this uniqueness is quite robust to the introduction of different types of stochasticity, richer payoff functions or heterogeneity between players.

The coordination problem of the prisoners, then, is the inability to make one's actions dependent on that of others, and in such a way becoming unpredictable as a group. More specifically, in order to become successfully unpredictable the prisoners would need a coordination device that would not be accessible to the warden. This is in stark contrast to other models of regime change games, which have always seen the coordination problem in the

[3]Note that we mean "action" in the sense that a player can play a mixed strategy which then picks an action; he might regret the action without regretting the strategy.

[4]For example, in a textbook bank run model it is an equilibrium that no one runs on the bank. However, there is another equilibrium in which everyone does.

inability of players to guess each other's information.

We do not claim that our model "shows" that prison wardens or governments can deter attacks with minimal (or no) defense strength; see the discussion at the end of section 3.2 for details. Instead, we think of our result as giving an insight – at a high level of abstraction – into how systems and regimes can be stable even if they look vulnerable to a coordinated attack. Reasoning about the strength of police forces in the Western world, and their problems at countering large-scale riots if they occur, can perhaps convince us of the intuition. It would clearly not be an equilibrium if there were frequent riots, as policymakers would react by strengthening the police. Neither, however, could it be an equilibrium to have so many police officers that no riots would ever be possible (as there would be pressure to save money by downsizing the police forces). We live, therefore, in a world in which there are quite few police officers per population, and large-scale riots are rare.

Jeremy Bentham and the "Panopticon"

Our main result has a striking parallel to the ideas of Jeremy Bentham (1787), who thought about how to build the perfect prison. In a series of letters, he proposed the "panopticon": A prison in which prisoners are not only kept separate from each other, but also (by an intricate construction) unable to see who is guarding them.[5] His prediction was that this construction would make revolts impossible at a low cost.[6]

Our result is identical to Bentham's prediction, and it relies on similar central assumptions: That prisoners are unable to centrally coordinate their behavior, that they are unable to observe how many guards there are, and that there are many prisoners. We comment in more detail about the connection between our result and Bentham's ideas in section 4.3. Bentham's idea has also been enormously influential in political philosophy and sociology; we draw some parallels to related literature in the same section.

Relation to Other Literature

Our model is directly related to the rich literature on regime change games and their applications to speculative attacks and bank runs. The problem of a central bank defending against speculators has received much attention (e.g. Flood and Garber, 1984; Obstfeld, 1986). Such

[5]Bentham's plans ensured that prisoners could not see into the guards' "lodge": "To the windows of the lodge there are blinds, as high up as the eyes of the prisoners in their cells can, by any means they can employ, be made to reach." He also emphasized the lack of communication possibilities: "These cells are divided from one another, and the prisoners by that means secluded from all communication with each other, by partitions in the form of radii issuing from the circumference towards the center ... "

[6]Bentham (p. 46): "Overpowering the guard requires an union of hands, and a concert among minds. But what union, or what concert, can there be among persons, no one of whom will have set eyes on any other from the first moment of his entrance?

models predict that the self-fulfilling nature of attacks leads to multiple equilibria. One of these equilibria is usually the one that emerges as unique Nash Equilibrium in our model.

The equilibrium multiplicity of such models has often been seen as unsatisfying, or at least calling for an explanation of the attackers' higher-order beliefs. Morris and Shin (1998), building on results by Rubinstein (1989) and Carlsson and van Damme (1993), show how a refinement that introduces minimal noise selects a unique equilibrium prediction, in which the probability of an attack is monotonic in the strength of the defender. More recently, Weinstein and Yildiz (2007) have shown that while equilibrium uniqueness is a robust feature of selection methods based on higher order perturbations, the identity of the selected equilibrium is not robust and depends heavily on the chosen perturbation.

Our main result, in contrast to this literature, shows that if we take the defender seriously as a player (and assume that he has the will and the possibility to defend himself), a unique equilibrium emerges naturally from the interaction of the attackers' and the defender's beliefs. This uniqueness does not require any ad-hoc assumptions about higher-order beliefs or an information structure that creates a particular structure of higher-order beliefs.

In our main result, we make use of the fact that as the number of prisoners gets larger, their overall behavior becomes relatively more predictable regardless of which strategy they each play. In describing this increase in relative predictability, we show how to use *concentration inequalities* to prove uniqueness of mixed equilibria; we will cite the relevant mathematical results directly as we make use of them.

Problems with a similar structure to ours have also been analyzed with a focus on signaling and information manipulation (Edmond, 2013), signaling through defensive measures (Angeletos and Pavan, 2013), reputation (Huang, 2017) and the optimal stopping problem when under attack (Kurlat, 2015). The main contrast between these papers and our analysis is that we consider the defender as a strategic player. We consider a simple one-shot game, and we are not concerned with the ability of the defender to distort information or signal.

Since we examine the consequences of different information structures, our work is also related to the literature on information design (e.g. Bergemann and Morris, 2017). In section 4.2, we comment on how our paper diverges from this literature, and why we consider our approach to be a more natural way of approaching our research question.

2. Model

Players and Strategies We consider a game played between one *warden* and N *prisoners*. The warden chooses a guard level $\gamma \in \mathbb{R}_+$. The prisoners decide simultaneously and independently whether to revolt (r) or not revolt (n). All revolting prisoners break out if the

number of revolting prisoners is strictly larger than γ. Otherwise, no prisoner breaks out.

Payoffs Each prisoner values breaking out by $b > 0$. If the prisoner revolts but there is no breakout, he bears a cost $-q < 0$. This cost can be interpreted in two ways: It could either represent a punishment for prisoners who unsuccessfully try to escape or it could denote a cost of effort (in the latter case b should be interpreted as the benefit of breaking out net of this effort cost). If a prisoner does not revolt, his utility is 0; see table 1 for a summary of these payoffs.

	breaks out	does not break out
r	b	$-q$
n	0	0

Table 1: A prisoner's payoff conditional on breaking out or not

The warden experiences a disutility denoted by $-B < 0$ whenever a breakout occurs; apart from that he only cares about the costs of the guards. The costs of the guards are linear in γ with slope normalized to 1, i.e. guard costs are $-\gamma$. Consequently, the utility of the warden is $-B - \gamma$ if a breakout occurs and $-\gamma$ otherwise. Each player maximizes his expected utility. Finally, we make an assumption on the size of the disutility B:

Assumption 1. $B \geq N + 1$.

The assumption implies that the warden would prevent a revolt (by setting $\gamma = N$) if he knew that all prisoners play r for sure. We mainly make this assumption to make the game interesting: If $B < N$, there is – independent of any information structure – a very robust equilibrium in which the guard level is zero and all prisoners revolt. This is a somewhat uninteresting case that we want to neglect. For concreteness, we will assume $B \geq N + 1$ (instead of $B > N$), which significantly simplifies the analysis.

Information Structure For our main result, we assume that the warden's choice of γ cannot be observed by the prisoners, and prisoners cannot coordinate their actions. This setting closely resembles Jeremy Bentham's idea of the "panopticon", which is how we call this model. In appendix B, we consider other information structures for our model and derive standard results for comparison with our result.

3. Analysis of the Panopticon

3.1. Preliminary Analysis

We begin by showing that there exist only equilibria in which all prisoners play r with the same probability p in equilibrium. Intuitively, this follows from the strategic complementarity between prisoners' actions. If $p_i < p_j$, then i would view the probability that "others" revolt higher than j. But this would imply that i has higher incentives to revolt than j which contradicts $p_i < p_j$.

Lemma 1. ***(All equilibria are prisoner symmetric)*** *There are no equilibria in which prisoners revolt with prisoner dependent probabilities p_i and $p_j \neq p_i$ for some prisoners i and j.*

We can quickly see that equilibria can only exist in mixed strategies: If the prisoners revolted for sure, the warden would best respond by setting the guard level to $\gamma = N$. Consequently, the revolt is unsuccessful and revolting is not a best response for the prisoners. Alternatively, the warden would best respond with $\gamma = 0$ if the prisoners played n for sure. But in this case revolting is a best response. Consequently, the prisoners (and possibly also the warden) will mix and revolts will succeed with some probability in equilibrium.

In any mixed equilibrium, the number of prisoners playing r follows a binomial distribution as every prisoner plays r with probability p and the prisoners' choices are independent. Call this distribution G and its probability mass function g. More precisely, $g(m) = \binom{N}{m} p^m (1-p)^{N-m}$ is the probability that m prisoners revolt if each prisoner revolts with probability p.

Clearly, the warden's best response puts positive probability only on integers between 0 and N. Therefore, the warden's maximization problem is

$$\max_{\gamma \in \{0,1,\ldots,N\}} -(1 - G(\gamma))B - \gamma. \tag{1}$$

Denote the warden's (mixed) strategy by the distribution F with probability mass function f. The warden has to be indifferent between any two γ_0 and γ_1 in the support of F which means that the following equation has to hold

$$B\left(G(\gamma_0) - G(\gamma_1)\right) = \gamma_0 - \gamma_1 \tag{2}$$

for any γ_0 and γ_1 in the support of F. Note that G is S-shaped because it is a binomial distribution, i.e. g is first strictly increasing (up to the mode of G) and then strictly decreasing. This property leads – together with assumption 1 – to the following result.

Lemma 2. ***(Support of the warden's equilibrium strategy)*** *In any mixed strategy equilibrium, the support of F consists of at most two elements and these two elements are adjacent, i.e. the warden mixes between γ_1 and $\gamma_1 + 1$ with $\gamma_1 \in \{0, \ldots, N-1\}$. For any $\gamma_1 \in \{0, \ldots, N-1\}$, there exists a unique $p \in (0, (\gamma_1 + 1)/N)$ such that γ_1 and $\gamma_1 + 1$ are the two global maxima of the warden's utility.*

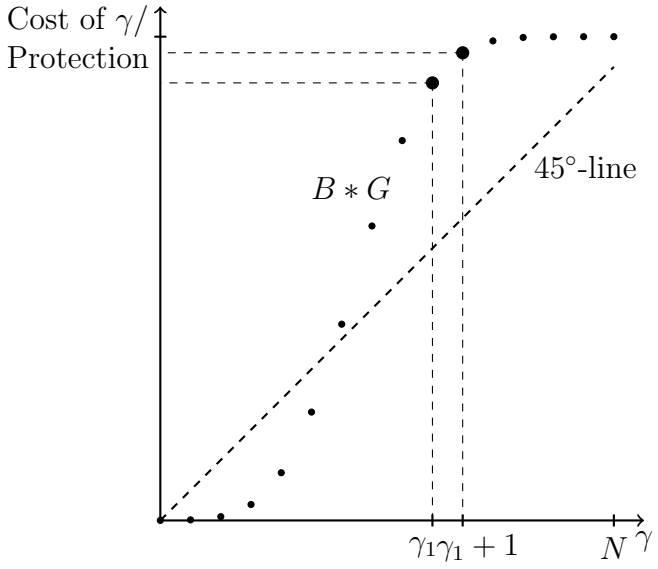


Figure 1: Equilibrium in the panopticon model.

We illustrate the lemma using figure 1. For every individual revolt probability p, we get a cumulative density function $G(m)$ that gives the probability that m or fewer prisoners revolt – in other words, the probability that a guard level $\gamma = m$ successfully prevents a breakout. This function G is (multiplied by B) given by the dots (we concentrate on values at integers). The dashed line gives the cost of setting a guard level γ, which is simply γ. The warden optimally mixes between guard levels that maximize the difference between $B * G(\gamma)$ and γ. Intuitively, he trades off the additional cost of increasing the guard strength with reducing the probability of a breakout. Choosing a higher γ than $\gamma_1 + 1$, for example, would increase the cost by much more than the probability of preventing breakouts (weighted by the disutility of a breakout), and is therefore not optimal. If there are several guard levels where the difference is equivalent, the warden is indifferent between them. The example illustrates our two intermediate results: (a) The warden will never mix between more than two guard levels, since the concavity of G (above the mode) means that the difference between $B * G$ and cost cannot be equal in three or more points. (b) For every γ_1, $\gamma_1 + 1$ we can find a p such that the warden is indifferent between the two guard levels, by finding a p such that the resulting G has the maximum distance from the 45-degree line at γ_1 and $\gamma_1 + 1$. The

condition $p < (\gamma_1+1)/N$ is equivalent to saying that γ_1 is weakly above the mode of G. That is, the optimal guard level will be in the concave part of G which is again in line with figure 1.

In equilibrium, each prisoner must be indifferent between revolting and not revolting. This indifference condition is given by

$$\mathbb{E}_\gamma \left[-qG_{N-1}(\gamma-1) + b(1-G_{N-1}(\gamma-1))\right] = 0 \tag{3}$$

where the expectation over γ is taken with respect to the warden's optimal strategy F and G_{N-1} is the binomial distribution with $N-1$ prisoners, i.e. $g_{N-1}(m) = \binom{N-1}{m} p^m (1-p)^{N-1-m}$. Note that the probability of revolting p and the guard level γ_1 of a mixed equilibrium are determined simultaneously by (1) and (2) as the warden's own mixing probability does not play a role in these conditions. Given these two values, (3) will determine the equilibrium mixing probability of the warden.

We now turn to the question which guard levels can be chosen in equilibrium. Lemma 2 stated that we can concentrate on equilibria where the warden mixes over γ_1 and γ_1+1 for $\gamma_1 \in \{0,\dots,N-1\}$. Furthermore, the warden's incentives do not pose an obstacle for the existence of such an equilibrium for any $\gamma_1 \in \{0,\dots,N-1\}$ as there is always a p for which γ_1 and γ_1+1 are optimal. Whether an equilibrium exists for $\gamma_1 \in \{0,\dots,N-1\}$ is therefore determined by the prisoner's indifference condition. More precisely, a mixed strategy equilibrium where the warden mixes over γ_1 and γ_1+1 exists if and only if a prisoner strictly preferred to revolt if the warden played γ_1 for sure and strictly preferred not to revolt if the warden played γ_1+1 for sure (holding fixed the probability p with which the other prisoners revolt). Defining

$$\Delta(\gamma) = -qG_{N-1}(\gamma-1) + b(1-G_{N-1}(\gamma-1)) \tag{4}$$

as the utility difference of a prisoner between playing revolt and no revolt if the warden uses γ guards for sure, this can be expressed as follows: An equilibrium in which the warden mixes between γ_1 and γ_1+1 exists if and only if $\Delta(\gamma_1) > 0 > \Delta(\gamma_1+1)$. In this case, the equilibrium mixing probability with which the warden plays γ_1 is

$$z = \frac{-\Delta(\gamma_1+1)}{\Delta(\gamma_1)-\Delta(\gamma_1+1)}. \tag{5}$$

Note that several equilibria can exist because Δ is not necessarily monotone: While both terms in (4) are directly decreasing in γ, there is an indirect effect through p: A higher γ is only optimal for the warden if the revolt probability p is higher. This, however, implies that Δ increases. Which of the two effects dominates (direct effect through γ or indirect effect

through p) is a priori unclear. However, $\Delta(0) > 0$ as revolting is dominant if the guard level is zero and $\Delta(N) < 0$ as not revolting is dominant when the guard level is N. Consequently, at least one equilibrium exists.

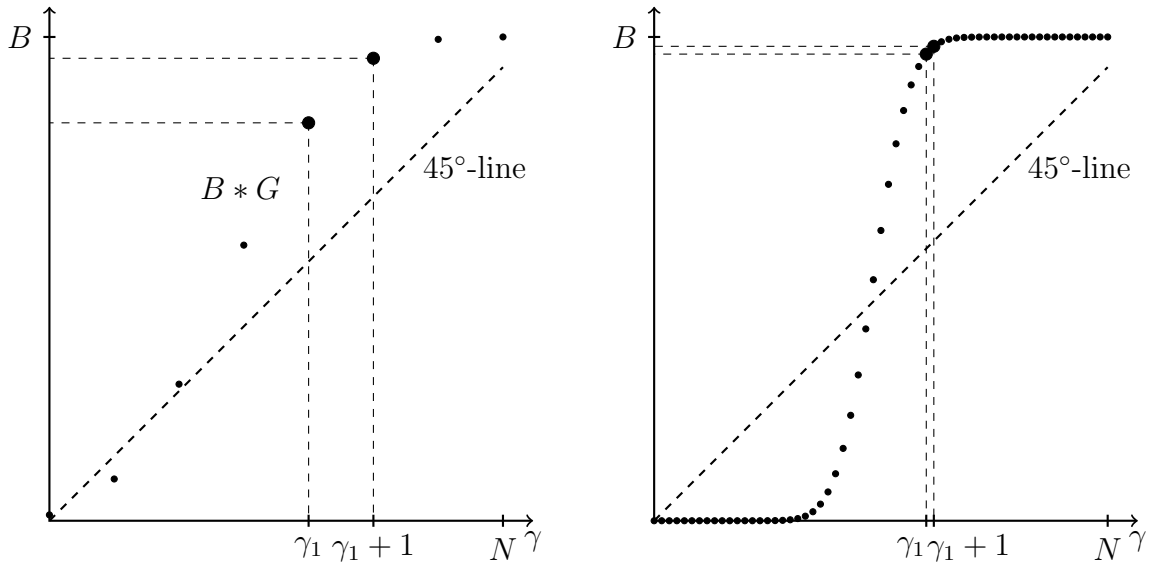
We have therefore established the following for the panopticon model:

Result 1. ***(Panopticon)*** *In every equilibrium, each prisoner mixes over r and n and all prisoners use the same mixed strategy. The warden mixes between at most two guard levels γ_1 and $\gamma_1 + 1$.*

3.2. Unique Equilibrium for large N

Making use of the preliminary results from the previous section, we can now show our main theorem:

Theorem 1. ***(Unique equilibrium for large N)*** *Take b and q as given. Let N be sufficiently large and B such that assumption 1 is satisfied.*[7] *Then, the warden mixes between 0 and 1 in the unique equilibrium of the panopticon model. The probability of a breakout is arbitrarily close to zero and $G_{N-1}(0)$ is arbitrarily close to one for sufficiently high N. The warden's payoff is bounded from below by a constant.*



(A) small N (B) large N

Figure 2: An illustration of theorem 1.

After having derived the intermediate results about the panopticon model in section 3.1, we can extend the intuition for theorem 1 that we gave in the introduction. Recall that there

[7] Assumption 1 links B and N. The theorem should be understood in the following way: Take b and q as given; then there is an $\bar{N}$ such that for all $N \geq \bar{N}$ and all B satisfying assumption 1, the results hold.

are three requirements for an equilibrium where the warden mixes between guard levels γ_1 and $\gamma_1 + 1$: (i) The warden must be indifferent between the guard levels, (ii) both guard levels must be better than all other guard levels, and (iii) the prisoners must be indifferent between revolting and not revolting. Figure 2 shows, similar to figure 1, a distribution G of attacking prisoners so that the first two requirements are fulfilled. In particular, by (2), a line through the points $(\gamma_1, BG(\gamma_1))$ and $(\gamma_1 + 1, BG(\gamma_1 + 1))$ would be parallel to the 45° line.

The third requirement can only be fulfilled if the probability of a successful revolt is sufficiently high, since it is otherwise optimal for the prisoners to never revolt. In panel (A), where N is relatively small, this is possible: There is a positive probability that the number of revolting prisoners is larger than γ_1. This can be seen on the vertical axis as $B * G(\gamma_1)$ is well below B. Hence we can find a mixing probability for the warden that makes prisoners indifferent between revolting and not revolting. But if N gets larger (panel B), the probability of a successful revolt converges to 0 for both γ_1 and $\gamma_1 + 1$ since the binomial distribution G becomes more concentrated – and therefore steeper – around its mode (which is always smaller than γ_1) for large N. Then there exists no mixing between these two guard levels that would actually make the prisoners indifferent, and thus requirement (iii) cannot be fulfilled for large N and $\gamma_1 > 0$. The only equilibrium for large N is the one where $\gamma_1 = 0$. Then each prisoner has the possibility of successfully revolting on his own, and therefore no longer cares about the probability with which others revolt.

The result that $G_{N-1}(0)$ is close to one if N is large states that every prisoner expects all other prisoners to not revolt. This is in line with Bentham's idea that prisoners would not even think about a coordinated attack in a panopticon. Given $G_{N-1}(0) \approx 1$, the equilibrium is in fact similar to a game where each prisoner faces the warden one-on-one without any prospects of support by his fellow inmates. The panopticon exploits, in this sense, the prisoners' coordination problem maximally.

While we derived our result for completely mixed equilibria, this does not necessarily mean that it is only relevant for situations with full mixing and the permanent possibility of a successful attack. We would like to make two points as to how our result can be understood.

Our preferred interpretation of the mixed equilibrium in theorem 1 is in terms of Harsanyi's (1973) purification. According to this interpretation, we can view a mixed strategy equilibrium as the limit of pure strategy Bayesian equilibria in which prisoners have private information about, say, how much they are punished in case of an unsuccessful revolt. In the panopticon equilibrium, only those prisoners who fear punishment the very least will revolt. For every other prisoner, not revolting is the unique best response in the Bayesian game (and a non-unique best response in the limit).

We would also like to point out that our result mainly means that for large N, there is a unique equilibrium in which the breakout possibility approaches zero. If we use slightly different strategy sets, this equilibrium need not necessarily be mixed. Assume that model parameters are such that a unique equilibrium would exist according to theorem 1, and let z_{eq} be the probability of playing $\gamma = 1$ in this equilibrium. Then we can analogously derive our result in a modified model in which there is some minimum guard level that the warden cannot go below:

Corollary 1. ***(Minimal guard requirement)*** *Suppose the warden has to set a guard level of at least $\gamma_{min} \geq 1$ with probability of at least $z_{min} > z_{eq}$. Then there is a unique equilibrium in the panopticon in which the warden sets γ_{min} with probability z_{min} (and $\gamma = 0$ with probability $1 - z_{min}$) and prisoners choose $p = 0$.*

This corollary clarifies the right interpretation of theorem 1: The main result is not that the warden uses zero (or one) guards for large N – which might indeed seem unrealistic in some applications. Instead the main insights are that (i) equilibrium multiplicity reduces to uniqueness and the probability of a breakout approaches zero for large N, and that (ii) this is hugely advantageous for the warden.

4. Discussion

4.1. Extensions and Robustness

In the supplementary material to this paper, we consider several extensions and generalizations of our model and show that the fundamental property of large populations upon which our proof relies is robust to such changes. We consider the following extensions:

- We consider payoff functions such that the payoff $-q$ of an unsuccessfully revolting prisoner depends on the guard level γ. While the precise equilibrium mixing probabilities change, the analysis remains qualitatively the same and the equilibrium probability of revolting remains arbitrarily close to zero.

- We consider payoff structures such that the payoff of a non-revolting prisoner depends on whether a breakout occurs or not. Our results remain unchanged unless this free riding possibility destroys the supermodularity of the game among prisoners.

- We provide a generalization of our model in which the probability of a breakout increases monotonically in m, the number of revolting prisoners.

- We consider an extension in which attackers differ in their size. The warden will then not mix between 0 and 1, but between 0 and the size of the largest attacker.

4.2. Why we don't do information design

The warden in our model benefits from the fact that his choice is unobservable, and would therefore want to keep his choice secret if he could. Since prisoners almost always get payoff 0, secrecy is also efficient. One could therefore think that an information designer should choose the information structure in which the warden's choice remains unobservable. However, we diverge from the literature on information design by not solving for the warden optimal information structure. The standard method for doing so would be to employ a revelation principle for information design which permits the formulation of the problem as a maximization over the prisoner's strategies subject to obedience constraints (c.f. Bergemann and Morris, 2016, 2017).

In our model, if prisoners can perfectly observe the guard level, there exists an equilibrium in which the warden employs one guard and no one revolts. Equilibrium strategies will clearly satisfy the aforementioned obedience constraints as no one has an incentive to deviate. But this equilibrium illustrates a problem with the information design approach: Of course, the subgame in which the prisoners have to act (after observing the guard level) has another Nash equilibrium in which everyone revolts. The common equilibrium selection method for coordination games, i.e. the global game approach, would select the latter equilibrium and not the warden's preferred equilibrium in which no one revolts. The usual information design approach, however, implicitly assumes that the warden can determine which equilibrium is played if there are multiple equilibria in the prisoners' subgame.[8] But since the coordination problem between prisoners and its interaction with the warden's choice is at the core of our paper, we find it intellectually unsatisfying to not only ignore this coordination problem, but solve it to the *warden's* advantage. Our main result is, instead, built on Nash equilibrium uniqueness and hence does not require a choice between different equilibria.

4.3. Bentham and the Literature that Followed

Our main result in theorem 1 has an interesting analogy in Bentham's ideas. Bentham explicitly stated that a single guard, i.e. a minimal guard level, would be sufficient: "[...] so far from it, that a greater multitude than ever were yet lodged in one house might be inspected by a single person." He envisioned the impossibility of a "concert among minds" to such a degree that prisoners would not even think about revolting together with other prisoners.[9]

In the 230 years since Bentham, many scholars have interpreted it as a metaphor for

[8] A similar problem is, of course, well known in the mechanism design literature where the same assumption is implicitly made when using the revelation principle.

[9] Chwe (2003) provides an interesting discussion of the panopticon and higher-order knowledge that is mostly orthogonal to ours.

modern society. Most prominently, Foucault (1975) points out that panopticism, a system in which individuals self-discipline because of the omnipresent possibility of being disciplined, has made modern society possible. Order is no longer maintained in a "contest of violence" between those opposing and those defending it. Instead, the docility of individuals allows for cost-saving minimal enforcement: There is neither waste of resources through unused guard capacity nor fruitless attempts at revolting.[10] This was a prerequisite for the establishment of organizations, firms, schools in which individuals have internalized the rules and behave in the desired way without constant supervision. Our result captures the intuition on how and why panopticism would work in a formal, game-theoretical model.

Others (e.g. Zuboff, 1988) have suggested that modern computers and indeed the internet are panoptica, where everyone can at any time be under surveillance by a small group – an idea that has gained credence by recent revelations of mass surveillance by intelligence agencies.

5. Conclusion

In this paper, we have shown how the character of coordination games changes drastically if we take the defender seriously as a strategic player. Canonical regime change games use all kinds of modelling techniques to avoid multiplicity of equilibria, or impose some order on it – whereas in our model, a unique Nash equilibrium emerges from the interplay of attackers' and defender's beliefs. Similarly, many models understand "coordination problem" to mean a problem of higher-order knowledge. Our paper suggests a more instrumental approach in which the real problem is the inability of players to condition their actions on those of others.

This suggests that in situations where it *is* possible to condition one's behavior on that of others, be it through contracts or through sequential and observable play, the coordination problem weakens or disappears. We can see examples of this in the real world. Prisons, oppressive regimes and (adversarial) employers make more use of surveillance than actual "guards". But once people are on the street, that doesn't cut it anymore: Police forces at violent demonstrations or football riots are not much smaller than the crowds they are sent to control. This is because protesters (and hooligans) can condition their behavior on each other: by being in the same place, being able to observe each other make sequential choices, and possibly even having some hierarchy among themselves. In fact, Chenoweth and Stephan (2011) have suggested that once a small minority of 3.5% of the population openly engage in

[10] "Hence the major effect of the Panopticon: to induce in the inmate a state of conscious and permanent visibility that assures the automatic functioning of power. So to arrange things ... that the perfection of power should tend to render its actual exercise unnecessary, ... that the inmates should be caught up in a power situation of which they are themselves the bearers." (Foucault, 1975)

anti-regime behavior, the regime's days are almost always numbered.

This paper is about what happens before: Why, quite often, it takes decades for opposition to turn into large-scale protest. Why regimes, political or otherwise, only rarely rely on heavy-handed enforcement and usually get away with an abstract threat that is not backed up by huge numbers of actual enforcers. And why the minds of protesters are just as much engaged by predicting the behavior of the regime as by the canonical problem of predicting each other's behavior.

In all of these situations, there are many more effects than the one we describe. But we suggest that, on a high level of abstraction, they all share one feature: that a group who is opposing a single actor may lack the unpredictability to successfully do so.

Appendix

A. Proofs

Proof of lemma 1: When analyzing the panopticon model, we restricted attention to symmetric equilibria, i.e. equilibria in which all prisoners revolt with the same probability p. We will now show that this is without loss of generality, i.e. there are no equilibria in which prisoners revolt with prisoner dependent probabilities p_i and $p_i \neq p_j$ for some prisoners i and j.

In the main text, we already argued that equilibria cannot be pure, i.e. there has to be at least one prisoner who uses a mixed strategy p_i with $0 < p_i < 1$. The argument is simple: If all prisoners used a pure strategy in equilibrium, the warden would be certain of the number of revolting prisoners, say k. In this case, the warden best responds by setting $\gamma = k$ which prevents a breakout for sure while any lower guard level would lead to a breakout with probability 1. If $k > 0$, the revolting prisoners could profitably deviate to not revolting. If, however, $\gamma = k = 0$, then each prisoner could profitably deviate by revolting. Since at least one prisoner has a profitable deviation, we can conclude that there is no equilibrium in which all prisoners use pure strategies. Without loss of generality, let us therefore assume that prisoner 1 uses a completely mixed strategy, i.e. $0 < p_1 < 1$.

First, we will show the following: Take any equilibrium in the panopticon model. If $0 < p_i \leq p_j < 1$ holds for two prisoners i and j, then $p_i = p_j$. To see this, note that both i and j have to be indifferent between revolting and not revolting because both use a completely mixed strategy. If $p_j > p_i$ and j is indifferent between revolting and not revolting, then i would strictly prefer to revolt: For any $\gamma > 0$, the probability that at least $\lfloor \gamma \rfloor$ other prisoners revolt is higher for i than for j if $p_j > p_i$. Since j was indifferent, i will then strictly prefer to revolt. This contradicts that i is indifferent (because he plays a completely mixed strategy) and we must therefore have $p_i = p_j$.

Note that the previous argument actually says that if two players are indifferent between revolting and not revolting, then they must play revolt with the same probability. This is a bit stronger than what we said before because it rules out the possibility that some prisoner plays revolt with probability 0 or 1 while being indifferent between the two actions. (Recall that prisoner 1 uses a completely mixed strategy.)

What remains to be shown is that no prisoner strictly prefers one of the two actions in equilibrium. Suppose to the contrary that prisoner j strictly preferred to revolt and therefore plays revolt with probability 1 in equilibrium. Now consider prisoner 1: Since $p_1 < p_j = 1$, the probability that at least $\lfloor \gamma \rfloor$ other prisoners revolt is higher from prisoner 1's perspective

than from prisoner j's perspective. Therefore, prisoner 1 strictly prefers to revolt given that prisoner j strictly prefers to revolt. This contradicts that prisoner 1 plays a completely mixed strategy in equilibrium. Consequently, there cannot be a prisoner j who strictly prefers to revolt.

An analogous argument yields that there is no prisoner who strictly prefers not revolt. This completes the proof. □

Proof of lemma 2: We start with the first part of the lemma. As a first step, we show a weaker result: The support of the warden can consist of at most three elements. Denote the mode of G by γ^m (for a given p).[11] The binomial distribution G has the property that G is convex on $\{0, \ldots, \gamma^m\}$ and G is concave on $\{\gamma^m, \ldots, N\}$. Therefore, the maximization problem of the warden over the domain $\{0, \ldots, \gamma^m\}$ is convex and consequently only the boundary values 0 and γ^m can be local maxima (on this restricted domain). If we take $\{\gamma^m, \ldots, N\}$ as domain of the warden's maximization problem, the problem is concave and therefore (because γ takes integer values) this problem can have at most two local maxima γ_1 and γ_2 such that $\gamma_2 = \gamma_1 + 1$ (clearly, it could have only one local maximizer as well in which case we are already done). This implies that (1) has (at most) three local maxima: one at $\gamma_0 = 0$, γ_1 weakly above γ^m and possibly $\gamma_2 = \gamma_1 + 1$. Therefore, f's support will contain at most three elements.

Next we will show that the case where the warden is indifferent between $\gamma_0 = 0$, $\gamma_1 \geq \gamma^m$ and $\gamma_2 = \gamma_1 + 1$ is impossible. To see this, note that the fact that the warden is indifferent between γ_1 and $\gamma_1 + 1$ implies that $g(\gamma_1 + 1) = 1/B$. The warden is indifferent between γ_1 and γ_0 if and only if $(G(\gamma_1) - G(0))/\gamma = 1/B$. This is equivalent to saying that the average $g(\gamma)$ for $\gamma \in \{1, \ldots, \gamma_1\}$ equals $1/B$. Since $\gamma_2 - 1 \geq \gamma^m$ and as $g(\gamma_2) = 1/B$, we know that $g(\gamma) < 1/B$ for all $\gamma > \gamma_2$ (because g is strictly decreasing above the mode). Since $\sum_{\gamma=0}^{N} g(\gamma) = 1 \geq (N+1)/B$ by assumption 1 (i.e. the average $g(\gamma)$ is at least $1/B$), this implies that $g(0) \geq 1/B$. But then the single peakedness of g implies that $g(\gamma) > 1/B$ for all $\gamma \in \{1, \ldots, \gamma_1\}$ (recall that $g(\gamma_1 + 1) = 1/B$) which contradicts our earlier result that the average $g(\gamma)$ for $\gamma \in \{1, \ldots, \gamma_1\}$ is at most $1/B$.[12]

Last we reuse the argument of the previous paragraph to show that there cannot be an equilibrium in which the warden mixes between $\gamma_0 = 0$ and $\gamma_1 > 1$. Suppose there was such an equilibrium. Since the warden prefers γ_1 to $\gamma_1 + 1$, we must have $g(\gamma_1 + 1) \leq 1/B$.[13] As γ_1 has to be at least as high as the mode γ^m, we know that $g(\gamma) \leq g(\gamma_1 + 1)$ for all $\gamma \geq \gamma_1 + 1$.

[11] In the non-generic case that G has two modes, let γ^m be the smaller one.

[12] This last argument can be easily extended using inequalities to show that whenever there are γ_1 and $\gamma_2 = \gamma_1 + 1$ forming a local maximum of the warden's profit this local maximum must be the global maximum; i.e. is preferred to $\gamma_0 = 0$.

[13] For $\gamma_1 = N$, this step can be skipped and the rest of the argument works analogously.

The warden prefers γ_1 to $\gamma_1 - 1$ which implies $g(\gamma_1) \geq 1/B$. Furthermore, the warden has to be indifferent between γ_0 and γ_1 which implies that the average $g(\gamma)$ for $\gamma \in \{1, \ldots, \gamma_1\}$ equals $1/B$. As $\sum_{\gamma=0}^{N} g(\gamma) = 1 \geq (N+1)/B$, we obtain that $g(0) \geq 1/B$. But the single peakedness of g and the fact that $g(\gamma_1) \geq 1/B$ would then imply that the average $g(\gamma)$ for $\gamma \in \{1, \ldots, \gamma_1\}$ is strictly above $1/B$ contradicting that the warden is indifferent between γ_0 and γ_1. Taking the last three paragraphs together, the warden's equilibrium support can consist of at most two elements and these two elements have to be adjacent.

Finally, we turn to the second part of the lemma. Note that $\pi(\gamma_1) = \pi(\gamma_1 + 1)$ holds iff

$$g(\gamma_1 + 1) = 1/B.$$

This equation (viewed as an equation in p which indirectly determines g) has a solution $p < (\gamma_1+1)/N$: To see this note that $g(\gamma_1+1) = \binom{N}{\gamma_1+1} p^{\gamma_1+1}(1-p)^{N-\gamma_1-1}$ viewed as a function of p is 0 for $p = 0$ and single peaked with its maximum at $p = (\gamma_1 + 1)/N$. Furthermore, $g(\gamma_1+1)$ is continuous in p. Hence, it is sufficient to show that $g(\gamma_1 + 1)|_{p=(\gamma_1+1)/N} > 1/(N+1)$ as $1/(N+1) \geq 1/B$ by assumption 1. Note that for $p = (\gamma_1 + 1)/N$, $\gamma_1 + 1$ is the mode and therefore the maximum of g (viewed as function over γ). If $g(\gamma_1 + 1)|_{p=(\gamma_1+1)/N} \leq 1/(N+1)$, then $g(\gamma) \leq 1/(N+1)$ for all γ (with strict inequality for some) which contradicts that g is a probability mass function (it cannot sum to 1!). Hence, $g(\gamma_1 + 1)|_{p=(\gamma_1+1)/N} > 1/(N+1)$ which proves that there is a $p < (\gamma_1 + 1)/N$ such that $g(\gamma_1 + 1) = 1/B$.

The fact that $p < (\gamma_1 + 1)/N$ implies that $\gamma_1 + 1$ will be above the mode. As π is concave on $\{\gamma^m, \ldots, N\}$, $g(\gamma_1+1) = 1/B$ implies that γ_1 and γ_1+1 yield a higher warden payoff than any other γ weakly above the mode. Since π is convex on $\{0, \ldots, \gamma^m\}$, it follows that γ_1 and $\gamma_1 + 1$ are global maximizer of π iff $\pi(0) \leq \pi(\gamma_1 + 1)$. This last inequality can be written as

$$\frac{G(\gamma_1 + 1) - G(0)}{\gamma_1 + 1} \geq \frac{1}{B} \tag{6}$$

(where G is the cumulated binomial distribution for the $p < (\gamma_1 + 1)/N$ solving $g(\gamma_1 + 1) = 1/B$). The same argument as above shows that (6) holds: Suppose it did not. Then the average $g(\gamma)$ for $\gamma \in \{1, \ldots, \gamma_1 + 1\}$ would be strictly less than $1/B$ and as $\gamma_1 + 1$ is above the mode and $g(\gamma_1 + 1) = 1/B$, the same holds for $\gamma > \gamma_1 + 1$. Using the assumption $B \geq N + 1$ and the fact that $g(\gamma)$ has to sum to 1 over all $\gamma \in \{0, \ldots, N\}$, it follows that $g(0) \geq 1/B$. But then the single peakedness of g and $g(\gamma_1 + 1) = 1/B$ contradict that the average $g(\gamma)$ over $\{1, \ldots, \gamma_1 + 1\}$ is less than $1/B$. □

Proof of theorem 1: We will first show that an equilibrium in which the warden mixes over 0 and 1 exists in the panopticon for N sufficiently high. Second, we will derive

a lower bound on the warden payoff for this 0-1 mixed equilibrium. Finally, we will show uniqueness of the equilibrium for N sufficiently high. The other results in the theorem appear as intermediate results of the uniqueness proof.

It will be convenient to denote $B = \alpha(N+1)$ for some $\alpha \geq 1$ which can be done by assumption 1. In a mixed equilibrium where the warden mixes over 0 and 1, the riot probability p is determined by the warden's indifference condition $1 = BNp(1-p)^{N-1}$. As pointed out in the proof of lemma 2, this p is below $1/N$. The first and main step in establishing existence of the mixed equilibrium with $\gamma_1 = 0$ (for large N) is to show that $p < 1/N^2$. By $B = \alpha(N+1)$ with $\alpha \geq 1$, the indifference condition can be written as $p(1-p)^{N-1} - 1/(\alpha(N^2+N) = 0$. Note that the left hand side of this equation is increasing in p by $p < 1/N$. To show $p < 1/N^2$, it is therefore sufficient to show that the left hand side is greater than 0 for $p = 1/N^2$. This is (after multiplying through by N^2) equivalent to showing that

$$\left(1 - \frac{1}{N^2}\right)^{N-1} > \frac{1}{\alpha\left(1 + \frac{1}{N}\right)}$$

which can be rewritten as

$$\left(1 - \frac{1}{N^2}\right)^{N} > \frac{1 - 1/N^2}{\alpha\left(1 + \frac{1}{N}\right)} = \frac{N^2 - 1}{\alpha N(N+1)} = \frac{1 - 1/N}{\alpha}.$$

This inequality holds true as $(1-1/N^2)^N = 1 - 1/N + \sum_{i=2}^{N} \binom{N}{i}(-1/N^2)^i$ and $\sum_{i=2}^{N} \binom{N}{i}(-1/N^2)^i > 0$ because each positive term in the sum is higher than the immediately following negative term (recall that $\binom{N}{i+1} \leq \binom{N}{i} N$). Given $\alpha \geq 1$, the inequality above therefore holds for all N which implies $p < 1/N^2$ (where p is the revolt probability making the warden indifferent between the optimal guard levels 0 and 1).

To show that the mixed equilibrium with mixing over 0 and 1 exists, we have to establish that $\Delta(1) < 0$. Given $p < 1/N^2$, $G_{N-1}(0) = (1-p)^{N-1} > (1 - 1/N^2)^{N-1}$. As $\lim_{N\to\infty}(1 - 1/N^2)^{N-1} = 1$, this implies that $G_{N-1}(0) \to 1$ as $N \to \infty$.[14] Consequently, $\Delta(1) < 0$ for N sufficiently high; i.e. the 0-1 mixed equilibrium exists.

The warden's payoff in the 0-1 mixed equilibrium is $-B(1-(1-p)^N) = -\alpha(N+1)(1 - (1-p)^N) > -\alpha(N+1)(1-(1-1/N^2)^N)$. We now show that the latter term converges to $-\alpha$ as N gets large: This is equivalent to showing that $\lim_{N\to\infty} N - (N+1)\left(\frac{N^2-1}{N^2}\right)^N = 0$. The term in the limit can be written as

$$\frac{N^{2N+1} - (N+1)(N^2-1)^N}{N^{2N}}.$$

[14] Just to be precise, the limit is 1 as $(1-1/N^2)^{N-1} = 1 - N/N^2 + \binom{N}{2}1/N^4 - \ldots$ where all terms but the first approach 0 as N grows large.

Using the binomial expansion and making use of the fact that $\binom{N}{1} = N$, we can see that this is

$$\frac{N^{2N+1} - N^{2N+1} - N^{2N} + N^{2N} + N^{2N-1} - \dots}{N^{2N}}$$

where the first four terms cancel each other out and the remaining expression only contains powers of N smaller than $2N$ in the numerator, so that the expression goes to zero as N gets large. Therefore, $\lim_{N\to\infty}(N+1)(1-(1-1/N^2)^N) = 1$ and the warden's payoff is bounded below by $-\alpha$ in the warden 0-1 mixed equilibrium for N sufficiently large.

Finally, we show uniqueness of the mixed equilibrium with $\gamma_1 = 0$ and N large. To do so, we need two intermediate results that are stated as lemmas below (lemma 3 and 4). To start with, define an *equilibrium candidate* as a (p, γ) such that the warden's indifference condition holds, that is $g(\gamma+1) = \frac{1}{\alpha(N+1)}$, and $p < (\gamma+1)/N$. An equilibrium candidate leads to an equilibrium if $\Delta(\gamma) \geq 0$ and $\Delta(\gamma+1) < 0$, that is if $G_{N-1}(\gamma-1) \leq b/(q+b) \leq G_{N-1}(\gamma)$. We will show that for large N, there are no equilibrium candidates with $\gamma \geq 1$ that satisfy the equilibrium condition $G_{N-1}(\gamma-1) \leq b/(q+b)$.

In the following, we make use of known results on the shape and the tail bounds of the binomial distribution. Recall that $g_N(\gamma) = \binom{N}{\gamma} p^\gamma (1-p)^{N-\gamma}$, i.e. the probability mass of the binomial distribution $\mathcal{B}(N,p)$ at γ. G_N is the corresponding cumulative distribution function; the definitions of g_{N-1} and G_{N-1} are analogous.

Lemma 3. ***(Breakout probability approaches zero for large N)*** *For every $\varepsilon > 0$, there exists an N_ε such that for all models with more than N_ε prisoners $1 - G_N(\gamma) < \varepsilon$ holds in every equilibrium candidate.*

Proof: Using the Chernoff-Hoeffding Theorem (Hoeffding, 1963), we get

$$1 - G_N(\gamma) \leq \left(\frac{N}{\gamma+1}\right)^{\gamma+1} \left(\frac{N}{N-\gamma-1}\right)^{N-\gamma-1} p^{\gamma+1}(1-p)^{N-\gamma-1}. \tag{7}$$

For any equilibrium candidate in which the warden mixes over γ and $\gamma+1$, we therefore obtain

$$1 - G_N(\gamma) \;\leq\; \left(\frac{N}{\gamma+1}\right)^{\gamma+1} \left(\frac{N}{N-\gamma-1}\right)^{N-\gamma-1} \frac{1}{\alpha(N+1)\binom{N}{\gamma+1}}$$

where we plug the warden's indifference condition into (7). It is convenient to define $m = \gamma+1$ as this allows to write the previous expression as

$$1 - G_N(\gamma) \;\leq\; \frac{N^N}{\binom{N}{m} m^m (N-m)^{N-m} \alpha(N+1)}. \tag{8}$$

We are going to show that the RHS term converges to zero as N grows large. We have to show this for any $m \in \{1, \ldots, N\}$ and in particular m might depend on N. That is, we want to show that the expression above converges to zero for any $m(N)$. To do so, let $m^*(N)$ be the m maximizing the expression above. We show that the expression converges to zero even if we plug in $m = m^*(N)$.

Note that the term in (8) is maximal (for a given N) if m minimizes $\binom{N}{m}(m/N)^m(1 - m/N)^{N-m}$. Note that $\binom{N}{m}(m/N)^m(1 - m/N)^{N-m}$ is the probability mass of a binomial distribution with probability $p = m/N$ evaluated at its mode m. Hence, to minimize $\binom{N}{m}(m/N)^m(1 - m/N)^{N-m}$ we have to find the probability $p = m/N$ for which the modal density of a binomial distribution is minimized. This is the case for $p = 1/2$, i.e. $m = N/2$.[15] Consequently, $\forall m(N) : \binom{N}{m} m^m (N - m)^{N-m} \leq \binom{N}{\frac{N}{2}} \left(\frac{N}{2}\right)^N$ and (8) becomes

$$\begin{aligned} 1 - G_N(\gamma) &\leq \frac{N^N}{\binom{N}{N/2}(N/2)^N \alpha(N+1)} \\ &= \frac{2^N}{\binom{N}{N/2}\alpha(N+1)}. \end{aligned} \tag{9}$$

Since the central binomial coefficient $\binom{N}{N/2}$ is bounded from below by $2^N/\sqrt{2N}$ (see the supplementary material for an elementary proof of this), we obtain that the RHS term converges to zero as $N \to \infty$ which implies the lemma. □

We will now use this result to show that not only the probability of a breakout tends to zero if N is large, but also the probability for each prisoner that a revolt will be successful if he decides to revolt. This is given by $1 - G_{N-1}(\gamma - 1)$, i.e. the probability that at least γ other prisoners revolt (so that the remaining prisoner can push the number to $\gamma + 1$ or higher by revolting himself).

Lemma 4. ***(Chance of breakout tends to 0 if*** $\gamma \geq 1$ ***and N large)*** *For every $\varepsilon > 0$, there exists an N_ε such that in all models with more than N_ε prisoners $1 - G_{N-1}(\gamma - 1) < \varepsilon$ in every equilibrium candidate with $\gamma \geq 1$.*

Proof: Note that $1 - G_{N-1}(\gamma - 1) = 1 - G_{N-1}(\gamma) + g_{N-1}(\gamma) \leq 1 - G(\gamma) + g_{N-1}(\gamma)$. From lemma 3 we know that $1 - G(\gamma)$ is arbitrarily close to zero in every equilibrium candidate (for N sufficiently large). If $g_{N-1}(\gamma)$ becomes arbitrarily small as N grows large, we are therefore already done. For the remainder of the proof let us therefore assume that $g_{N-1}(\gamma)$ does not become arbitrarily small. We will show directly that $1 - G_{N-1}(\gamma - 1)$ converges to zero for large enough N in this case.

[15] If N is odd, both $m = \lfloor N/2 \rfloor$ and $m = \lceil N/2 \rceil$ will lead to minimal modal density. We concentrate on the case where N is even for notational convenience. Obviously, our results also hold for odd N.

By the warden's indifference condition, $g_N(\gamma+1) = \frac{1}{\alpha(N+1)}$, and we can write

$$g_{N-1}(\gamma) = g_N(\gamma+1)\frac{\gamma+1}{pN} = \frac{\gamma+1}{\alpha p(N^2+N)} \leq \frac{\gamma+1}{\alpha p N^2}.$$

If $g_{N-1}(\gamma)$ does not become arbitrarily small, neither does $(\gamma+1)/(\alpha p N^2)$ and therefore there is a sequence of tuples $(N, p(N), \gamma(N))$ which are strictly increasing in N such that (i) $(p(N), \gamma(N))$ is an equilibrium candidate (with the respective N) for each tuple $(N, p(N), \gamma(N))$ and (ii) $\gamma(N)+1 \geq \mu p(N) N^2$ for each tuple in the sequence and some $\mu > 0$.

Rearranging the latter condition gives

$$\gamma(N) - p(N)N + p(N) \geq \mu p(N)N^2 - p(N)N + p(N) - 1 = p(N)N^{5/4} * \left(\mu N^{3/4} - \frac{1}{N^{1/4}}\right) + p(N) - 1. \tag{10}$$

We will look at two cases. First, $p(N)N^{5/4}$ does not converge to zero. Then the right hand side of (10) is weakly larger than $\tilde{\mu}N^{3/4}$ for some $\tilde{\mu} > 0$ and N sufficiently large. Therefore, $\frac{(\gamma(N)-p(N)N+p(N))^2}{N-1} \geq \frac{(\tilde{\mu}N^{3/4})^2}{N-1} > \tilde{\mu}^2\sqrt{N}$ for large N which implies that $\frac{(\gamma(N)-p(N)N+p(N))^2}{N-1}$ will grow without bound as N gets large. Hoeffding's inequality (Hoeffding, 1963, Thm. 1) gives the following upper bound for $1 - G_{N-1}(\gamma-1)$:

$$1 - G_{N-1}(\gamma-1) \leq e^{-\frac{2(\gamma - p(N-1))^2}{N-1}}.$$

As we have just shown, this upper bound tends to zero as N grows large. Consequently, we have shown directly that $1 - G_{N-1}(\gamma-1)$ converges to zero. It remains to check the second case in which $p(N)N^{5/4}$ converges to zero. If $p(N)N^{5/4}$ converges to zero, then $p(N) \leq 1/N^{5/4}$ for sufficiently high N. Consequently, $G_{N-1}(0) = (1-p(N))^N \geq (1-1/N^{5/4})^N$ and the latter converges to 1. As $G_{N-1}(0) \leq G_{N-1}(\gamma-1)$ for $\gamma \geq 1$, this implies that $1 - G_{N-1}(\gamma-1)$ converges to zero which completes the proof. □

Lemma 4 implies that $G_{N-1}(\gamma-1)$ is arbitrarily close to one in every equilibrium candidate with $\gamma \geq 1$ as N is sufficiently large. Put differently, for any $\varepsilon > 0$, we can find an N_ε such that $G_{N-1}(\gamma_1) > 1-\varepsilon$ for all $N \geq N_\varepsilon$ and all equilibrium candidates with $\gamma \geq 1$. For given b and q, we can find such an N_{ε^*} for $\varepsilon^* = 1 - b/(q+b)$. For $N \geq N_{\varepsilon^*}$, we have $G_{N-1}(\gamma-1) > b/(q+b)$ for all equilibrium candidates with $\gamma \geq 1$. Hence, no equilibrium candidate with $\gamma \geq 1$ satisfies the equilibrium condition $G_{N-1}(\gamma-1) \leq b/(q+b)$ for N sufficiently high. This means that the equilibrium in which the warden mixes over zero and one is the unique equilibrium for N sufficiently high. □

B. Comparison to Other Information Structures

In this appendix, we compare our main result to the results under different information structures – both in how the warden's behavior is observable to the prisoners (which in turn allows them to condition their behavior on the warden's choice), and in how the prisoners can correlate their behavior with each other. Table B shows how the information structures are related. The panopticon, in which guards are secret, is the model in which we derive our main result; we compare the outcome of the panopticon with two other information structures.

First, we consider a situation ("benchmark") in which the prisoners have no coordination problem, i.e. they can correlate their choices. In this case, it does not matter for the outcome whether the warden's choice is observable or not: All equilibria are (in expectation) payoff-equivalent to the outcome where the warden hires so many guards that a revolt by all prisoners would still be unsuccessful, and all prisoners choose not to revolt.

		Guards are observable:	
		Yes	No
Coordination btw prisoners:	Yes	Benchmark	(=Benchmark)
	No	Transparency	Panopticon

Table 2: The information structures we consider.

Second, we consider what happens if the prisoners cannot coordinate, but the guard level is observable ("transparency"). This in effect turns the situation into a two-stage game and removes the warden from the strategic considerations of the prisoners. Observable guards can deter prisoners from revolting, but if there are visibly fewer guards than prisoners, this also provides the prisoners information which might help them coordinate. Our model is then equivalent to the basic structure of regime-change games that have been studied in the literature on coordination problems and equilibrium selection (e.g. Morris and Shin, 1998). Adding minimal uncertainty, in a way that is close to the literature on global games, can select one of these equilibria for each level of guards. This implies that there is a minimal level of guards that deters the prisoners from revolting. This level is quite high and depends linearly on the number of prisoners, so that deterrence is quite costly.

Figure B shows a comparison of welfare in all three information structures. The comparison confirms that it is essential for our result that the prisoners cannot correlate their choices as well as that the warden's choice remains secret – features that were also the main ingredients in Bentham's "panopticon". The following paragraphs explore the different information structures in more detail.

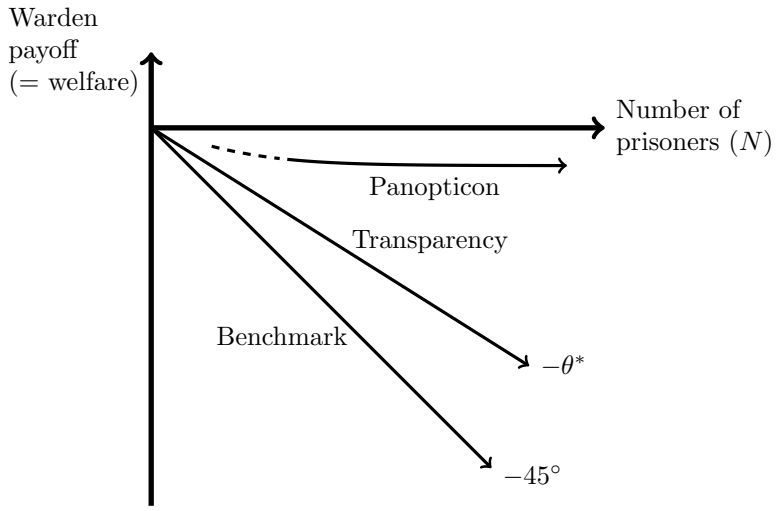


Figure 3: A comparison of welfare (which is equivalent to warden payoff) in the three information structures. The benchmark case is most expensive, as the warden needs as many guards as there are prisoners. In the transparency case, the warden can prevent breakouts with a lower number of prisoners; but the required number of guards still grows linearly in N. In the panopticon, the warden payoff is bounded from below by a constant.

B.1. Benchmark model: Perfect coordination

We first consider a benchmark where we assume that the prisoners do not have trouble coordinating and conditioning their behavior on each other's choices. We distinguish two possibilities: First, the prisoners observe the guard level set by the warden before they have to choose their actions. Assuming the coordination problem away means here that – given the guard level – the prisoners can coordinate on the prisoner optimal Nash equilibrium of any resulting subgame.[16] Hence, all prisoners play r if $\gamma < N$ and all play n otherwise. Given assumption 1, it is then optimal for the warden to choose $\gamma = N$. The payoff of the warden is $-N$ while the payoff of each prisoner is zero.

Second, we consider the possibility that the prisoners do not observe the guard level. As we allow perfect coordination between the prisoners, prisoners will either all revolt or all not revolt. This is due to the strategic complementarity between prisoners: Revolting is relatively better for a given prisoner if other prisoners revolt too. Given that either all or no prisoners revolt, the only two guard levels that can be best responses by the warden are zero and N. Furthermore, the game has no pure strategy equilibrium because of the non-observability of the guard level: If the warden chose a guard level of zero (N), the prisoners would best respond by revolting (not revolting). But then the guard level of zero (N) is not a best response. Therefore, we only have a mixed strategy equilibrium in which the warden mixes

[16]This is equivalent to the prisoner optimal correlated equilibrium of the subgame because of the strategic complementarity in the game among the prisoners.

between the two guard levels of zero and N and the prisoners mix between "all revolt" or "no one revolts". The mixing probabilities are such to keep the other side indifferent. Note that the expected warden payoff is $-N$ since the warden is indifferent between the equilibrium strategy and choosing a guard level of N for sure (which guarantees a payoff of $-N$). The prisoners have an expected payoff of zero as they are indifferent between their equilibrium strategy and not revolting for sure which gives every prisoner a payoff of zero.

Both possibilities of our benchmark lead therefore to the same equilibrium payoffs for all players. In this benchmark model, the warden has to use a large amount of resources to prevent a revolt.

B.2. Transparency model

We will now modify our model slightly so that prisoners first observe the guard level and then choose simultaneously and independently whether to revolt or not. If the guard level is weakly above N, it is a dominant action for each prisoner to play n. If the guard level is strictly below 1, it is a dominant action for each prisoner to play r. For guard levels between 1 and N, the optimal choice of a prisoner depends on what the other prisoners choose: If strictly more than $\gamma - 1$ other prisoners revolt, a given prisoner best chooses r himself. It is, however, optimal to choose n if less than $\gamma - 1$ other prisoners revolt. There are two equilibria in the subgames in which $\gamma \in [1, N)$: All prisoners revolt or no prisoner revolts. Consequently, the prisoners face a coordination problem.

Following the approach in the global games literature, we select one of the two equilibria by relaxing the assumption that γ is common knowledge among the prisoners. More precisely, we show that introducing an arbitrarily small amount of noise into how prisoners observe the guard level leads to a unique equilibrium prediction. Figure 4 shows the intuition behind this equilibrium selection through infection.

The perturbation works in the following way: The warden chooses an intended guard level $\tilde{\gamma}$. The true guard level is then drawn from a normal distribution with mean $\tilde{\gamma}$ and variance $\varepsilon' > 0$.[17] That is, the warden has a "trembling hand". Each prisoner receives a noisy signal of γ: This signal is drawn from a uniform distribution on $[\gamma - \varepsilon, \gamma + \varepsilon]$ with $\varepsilon > 0$. We are interested in the Bayesian Nash equilibrium of this game as $\varepsilon \to 0$. In fact, we show that this Bayesian game generically has a unique Bayesian Nash equilibrium as $\varepsilon \to 0$. Furthermore, this equilibrium does not depend on ε'. We select this equilibrium in the original game.[18]

[17] If γ has to be non-negative, one might think of a normal distribution truncated at zero. The truncation affects neither results nor derivation.

[18] The reader familiar with the global games literature might wonder why we introduce a "tremble" in the warden's action. The reason is that the parameter which is observed with noise (the guard level γ) is an endogenous choice in our model while the usual global game approach would assume noisy observation of an exogenous parameter chosen randomly by nature. Since γ is a strategic choice (made before the prisoners

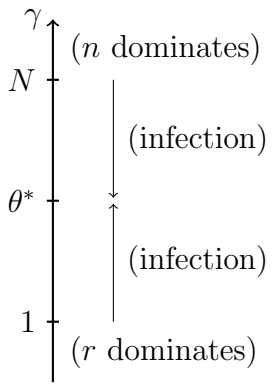


Figure 4: Infection of beliefs among prisoners: If $\gamma \geq N$, not revolting is a strictly dominant strategy for all prisoners. If $\gamma < 1$, revolting is strictly dominant. If $\gamma \in [1, N)$ and γ is common knowledge, there are two pure equilibria: Everybody revolts or no one revolts. When common knowledge is destroyed by the perturbation, beliefs get infected so that for $\gamma < \theta^*$, n is the unique equilibrium action, and r is the unique equilibrium action for $\gamma > \theta^*$.

Note that this setup eliminates common knowledge of the guard level. A prisoner observing signal θ knows that the true guard level is in $[\theta - \varepsilon, \theta + \varepsilon]$; he knows that each other prisoner knows that $\gamma \in [\theta - 3\varepsilon, \theta + 3\varepsilon]$; he knows that each other prisoner knows that he knows that $\gamma \in [\theta - 5\varepsilon, \theta + 5\varepsilon]$ and so on. Higher order beliefs will therefore play a role in determining the equilibrium. This appears to be a natural feature in a coordination game where the driving force of one's choice are exactly the expectations over what others do.

The following lemma contains the main technical result for the Bayesian game.

Lemma 5. ***(Equilibrium in the Bayesian game)*** *Let $\varepsilon' > 0$. Assume that $bN/(q+b) \notin \mathbb{N}$ and define*[19]

$$\theta^* = \left\lceil \frac{bN}{q+b} \right\rceil.$$

Then for any $\delta > 0$, there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \leq \bar{\varepsilon}$, a player receiving a signal below $\theta^* - \delta$ will play r and a player receiving a signal above $\theta^* + \delta$ will play n.

The lemma states that for generic parameter values – whenever $bN/(q+b)$ is not an integer – prisoners in the Bayesian game will revolt when they observe a signal below $\theta^* - \delta$ and will not revolt if they observe a signal above $\theta^* + \delta$. In the limit – as the prisoners' observation noise ε approaches zero – δ approaches zero as well. Put differently, prisoners play a cutoff strategy with cutoff value θ^* in the limit: Whenever they receive a signal below the cutoff, they play r and whenever they receive a signal above the cutoff they play n.

act), prisoners could infer γ correctly in equilibrium despite the noisy observation if the warden did not "tremble". Consequently, prisoners would have common knowledge of γ despite the noise.

[19] The ceiling $\lceil x \rceil$ is the lowest integer above x, i.e. $\lceil x \rceil = \min\{n : n \in \mathbb{N} \text{ and } n > x\}$.

Now consider the warden's decision problem (in the limit as $\varepsilon \to 0$). If the guard level is strictly above θ^*, then all prisoners will receive signals above θ^* and will therefore not revolt. If the guard level is strictly below θ^*, then all prisoners will receive a signal below θ^* and will revolt. Consequently, the optimal guard level for the warden is θ^* (or "slightly above and arbitrarily close" to θ^*). In the limit as $\varepsilon' \to 0$, the warden can ensure this guard level by simply choosing $\tilde{\gamma} = \theta^*$. This gives us the following outcome for our second model.

Result 2. ***(Transparency model)*** *The equilibrium outcome selected by the global game approach is the following: The warden chooses a guard level equal to θ^* and every prisoner plays n.*

Clearly, the warden does better in this equilibrium than in the benchmark model: He prevents a revolt for sure while using guard level θ^* instead of the guard level N. The reason is that he can utilize the coordination problem among prisoners in his favor.

If we compare with the panopticon, however, we can see that the unique equilibrium of the panopticon model is much more advantageous for the warden than the unique equilibrium of the transparency model. Why is that?

The information about γ that the prisoners receive under transparency has two functions: It deters them from attacking if γ is high enough – but it also correlates their beliefs to some degree (even without creating common knowledge) which allows them to eventually correlate their behavior. The main insight of the panopticon model is that the warden gains from the inability of prisoners to be unpredictable as a group; this advantage comes precisely from the impossibility of correlating their behavior. Thus, while the transparency model can deliver payoffs to the warden that are superior to that of our two benchmarks, the panopticon is vastly superior for the warden.

B.3. Comparison of the models

The prisoners are indifferent between all models: In the transparency model and the first benchmark, they did not revolt and therefore had a payoff of zero. In the panopticon and the second benchmark, prisoners were indifferent between revolting and not revolting as they played a mixed strategy. Hence, their expected utility was again zero as this is the payoff from playing n. The warden optimal model will therefore also be the welfare optimal model. Clearly, the two benchmark models are worst for the warden: His payoff is $-N$ which is the cost of preventing a breakout for sure by employing an abundance of guards. If he prevents communication, he can achieve the same outcome at cost $\theta^* \leq N$. In the panopticon model, he is also weakly better off than in the benchmark, since he always has the option of setting a guard level of N and ensuring a payoff of $-N$. He is indeed indifferent to doing so if

the equilibrium in which the warden mixes over $N-1$ and N is the only existing mixed equilibrium. If other equilibria exist, the warden will be strictly better off in those than in the benchmark model.

The interesting comparison is between the transparency model and the panopticon. Which of these two models is warden optimal depends on the parameter values of the model. In general, however, we have shown in section 3.2, the panopticon model has a unique equilibrium for large N in which the warden's payoff is bounded from below by a constant.

In the transparency model, the warden payoff is given by $-\theta^* = -\left\lceil \frac{bN}{q+b} \right\rceil$, which falls linearly in N and therefore becomes very negative for large N. We can therefore always find an $\overline{N}$ such that the panopticon is optimal for all $N > \overline{N}$. Figure B shows a comparison of warden payoff (i.e. welfare) of the different information structures.

Besides this central result for large groups, we present two other comparison results for small N. In this case, either the warden's or the prisoners' payoffs sometimes allow us to say which information structure is optimal.

Proposition 1. ***(Varying** B**)*** *Take q, b, N as given. The transparency model is warden optimal if $\theta^* = 1$. If $\theta^* > 1$, then there exists a $\bar{B}$ such that for all $B \geq \bar{B}$ the warden's payoff in the unique equilibrium of the panopticon model is higher than in the transparency model. The warden mixes over the guard levels zero and one in this unique equilibrium.*

Put differently, if the disutility of a breakout is relatively high compared to the cost of the guards, the panopticon is warden optimal unless a guard level of 1 can completely deter revolts in the transparency model. Given that revolting is dominant for any guard level strictly below one, $\theta^* = 1$ has to be viewed as a special case. Indeed $\theta^* = \lceil bN/(q+b) \rceil$ equals 1 only if the disutility of an unsuccessful revolt is $N-1$ times as high as the utility of a successful breakout which seems somewhat implausible in the applications we have in mind. Hence, the panopticon is – with a small caveat – warden optimal if warden incentives dominate. This might be somewhat surprising as the breakout probability in the panopticon is strictly greater than zero while the breakout probability in the transparency model is zero. There are two reasons explaining why cost savings compared to the transparency model are sizable if $\theta^* > 1$. First, the warden mixes between guard levels of zero and one in the panopticon if B is high. Consequently, a substantial number of guards can be saved compared to the transparency model. Second, the breakout probability in the panopticon – though not zero – is very small. The second follows readily from the first: Given that the warden really dislikes breakouts (high B), he will only be willing to mix between zero and one if the probability of revolt is very small. The reason why no other equilibrium exists is the following. Given that B is very high, the warden is only willing to use $\gamma_1 < N$ guards if the probability of a revolt

is very small. But this implies that for each prisoner it is unlikely that other prisoners revolt. Consequently, each prisoner strictly prefers not to revolt unless $\gamma_1 = 0$.

Next, consider the prisoners' incentives.

Proposition 2. ***(Varying** b/q**)*** *Take N and B as given. For b/q high enough, the warden payoff equals $-N$ in all models. Furthermore,*

- *Suppose $B^{\frac{N-1}{N}} > N$: Then, for $b/q \in (N-1, B^{\frac{N-1}{N}} - 1)$, the warden's payoff in every equilibrium of the panopticon model is higher than in the equilibrium of the transparency model.*
- *Suppose $N > B^{\frac{N-1}{N}}$: Then, for $b/q \in (B^{\frac{N-1}{N}} - 1, N-1)$, there exists an equilibrium in the panopticon model in which the warden's equilibrium payoff is lower than in the transparency model.*

If the prisoners have very strong incentives to break out, the payoff of all models coincides: The warden chooses N guards in the benchmark 1a and transparency model, mixes between N and $N-1$ guards in the panopticon and between N and 0 in benchmark 1b. Hence, the warden payoff is $-N$. For high (but not excessively high) incentives to break out, the comparison between panopticon and transparency model is hampered by the multiplicity of equilibria in the panopticon model. Depending on parameter values, either all (!) equilibria in the panopticon yield a higher warden payoff than the transparency model or the transparency model does better than some equilibria in the panopticon.

References

Angeletos, G. and A. Pavan (2013). Selection-free predictions in global games with endogenous information and multiple equilibria. *Theoretical Economics 8 (3)*, 883–938.

Bentham, J. (1787). *Panopticon; Or, The Inspection-House*. The Works of Jeremy Bentham, published under the superintendence of his executor John Bowring (Edinburgh: William Tait, 1838-1843). 11 vols. Vol. 4.

Bergemann, D. and S. Morris (2016). Information design, Bayesian persuasion, and Bayes correlated equilibrium. *American Economic Review 106*(5), 586–91.

Bergemann, D. and S. Morris (2017). Information design: A unified perspective. Cowles Foundation Discussion Paper No. 2075R2.

Carlsson, H. (1989). Global games and the risk dominance criterion. University of Lund, mimeo.

Carlsson, H. and E. van Damme (1993). Global games and equilibrium selection. *Econometrica 61(5)*, 989–1018.

Chenoweth, E. and M. J. Stephan (2011). *Why civil resistance works: The strategic logic of nonviolent conflict*. Columbia University Press.

Chwe, M. S.-Y. (2003). *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton: Princeton University Press.

Corsetti, G., A. Dasgupta, S. Morris, and H. S. Shin (2004). Does one Soros make a difference? A theory of currency crises with large and small traders. *Review of Economic Studies 71*(1), 87–113.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic Studies 80*, 1422–1458.

Flood, R. P. and P. M. Garber (1984). Collapsing exchange rate regimes: Some linear examples. *Journal of International Economics 17*, 1–13.

Foucault, M. (1975). *Discipline and Punish: The Birth of the Prison (trans. Alan Sheridan)*. New York: Vintage Books.

Frankel, D. M., S. Morris, and A. Pauzner (2003). Equilibrium selection in global games with strategic complementarities. *Journal of Economic Theory 108*(1), 1–44.

Goldstein, I. and A. Pauzner (2005). Demand-deposit contracts and the probability of bank runs. *Journal of Finance 60*(3), 1293–1327.

Harsanyi, J. C. (1973). Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory 2*(1), 1–23.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*(301), 13–30.

Huang, C. (2017). Defending against speculative attacks: The policy maker's reputation. *Journal of Economic Theory 171*, 1–34.

Kurlat, P. (2015). Optimal stopping in a model of speculative attacks. *Review of Economic Dynamics 18 (2)*, 212–226.

Morris, S. and H. Shin (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review 88(3)*, 587–597.

Morris, S. and H. Shin (2003). Global games: Theory and applications. In M. Dewatripont, L. Hansen, and S. Turnovsky (Eds.), *Advances in Economics and Econometrics (Proceedings of the Eighth World Congress of the Econometric Society)*. Cambridge: Cambridge University Press.

Obstfeld, M. (1986). Rational and self-fulfilling balance-of-payments crises. *American Economic Review 76*(1), pp. 72–81.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under almost common knowledge. *American Economic Review 79(3)*, 385–391.

Weinstein, J. and M. Yildiz (2007). A structure theorem for rationalizability with application to robust predictions of refinements. *Econometrica 75*(2), 365–400.

Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. New York: Basic Books.

Supplementary Material

for interested readers – not intended for publication

Proofs for results from the appendix

Proof of lemma 5: The proof is in three steps.

Strategic complementarity: A player finds revolting more attractive if other players are more likely to play revolt. A prisoner's strategy maps from signals into actions. If there are strategy profiles s and s' such that for every signal for which a player $j \neq i$ plays revolt under s he will also play revolt in s', then playing revolt is relatively more attractive for player i given s'_{-i} compared to s_{-i}: Let $G_{N-1}(\gamma - 1)$ be the probability that $\gamma - 1$ or less of the other $N-1$ prisoners revolt (given their strategies and i's signal). Define $\Delta(\gamma) = -qG_{N-1}(\gamma - 1) + b(1 - G_{N-1}(\gamma - 1))$ as the utility of revolting minus the utility of not revolting for a given guard level γ. $G_{N-1}(\gamma - 1)$ is weakly lower under s'_{-i} than under s_{-i} and therefore $\Delta(\gamma)$ is higher. That is, for a given γ revolting is more attractive. Since this is true for any given γ, it is also true in expectation.

Suppose everyone follows a cutoff strategy with cutoff θ. For a given $\delta > 0$, there exists an $\bar{\varepsilon} > 0$ such that the utility of revolting for a prisoner with signal θ is higher (lower) than the utility from not revolting if $\theta \leq \theta^* - \delta$ ($\theta \geq \theta^* + \delta$). The probability that a player observing himself the cutoff signal θ assigns to the event "exactly k other players receive a signal below θ" is

$$g_{N-1}(k) = \int_{\theta-\varepsilon}^{\theta+\varepsilon} \binom{N-1}{k} \left(\frac{\gamma - \theta + \varepsilon}{2\varepsilon}\right)^k \left(1 - \frac{\gamma - \theta + \varepsilon}{2\varepsilon}\right)^{N-1-k} \frac{\phi(\gamma)}{\Phi(\theta+\varepsilon) - \Phi(\theta-\varepsilon)}\, d\gamma.$$

We will now derive a convenient approximation for $g_{N-1}(k)$. Note that for ε small the term $\phi(\gamma)/(\Phi(\theta+\varepsilon) - \Phi(\theta-\varepsilon))$ is approximately constant (and equal to $1/(2\varepsilon)$) as ϕ is continuous and has a bounded first derivative. More precisely, fix θ and define $\phi^{max}(\varepsilon) = \max_{\gamma \in [\theta-\varepsilon, \theta+\varepsilon]} \phi(\gamma)$ and $\phi^{min}(\varepsilon) = \min_{\gamma \in [\theta-\varepsilon, \theta+\varepsilon]} \phi(\gamma)$. Then $g_{N-1}(k)$ and its approximation (where the average $1/(2\varepsilon)$ is used instead of $\phi(\gamma)/(\Phi(\theta+\varepsilon) - \Phi(\theta-\varepsilon))$) are necessarily between the two values

$$\begin{aligned}
\bar{g}(\varepsilon) &= \int_{\theta-\varepsilon}^{\theta+\varepsilon} \binom{N-1}{k} \left(\frac{\gamma - \theta + \varepsilon}{2\varepsilon}\right)^k \left(1 - \frac{\gamma - \theta + \varepsilon}{2\varepsilon}\right)^{N-1-k} \frac{\phi^{max}(\varepsilon)}{\Phi(\theta+\varepsilon) - \Phi(\theta-\varepsilon)}\, d\gamma, \\
\underline{g}(\varepsilon) &= \int_{\theta-\varepsilon}^{\theta+\varepsilon} \binom{N-1}{k} \left(\frac{\gamma - \theta + \varepsilon}{2\varepsilon}\right)^k \left(1 - \frac{\gamma - \theta + \varepsilon}{2\varepsilon}\right)^{N-1-k} \frac{\phi^{min}(\varepsilon)}{\Phi(\theta+\varepsilon) - \Phi(\theta-\varepsilon)}\, d\gamma
\end{aligned}$$

as the integrand is non-negative for all γ in the integration range. By showing that $\lim_{\varepsilon\to 0}\bar{g}(\varepsilon)-\underline{g}(\varepsilon)=0$, we show that the approximation of g becomes arbitrarily close to g for ε small enough:

$$\begin{aligned}\bar{g}(\varepsilon)-\underline{g}(\varepsilon) &= \int_{\theta-\varepsilon}^{\theta+\varepsilon}\binom{N-1}{k}\left(\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{k}\left(1-\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{N-1-k}\frac{\phi^{max}(\varepsilon)-\phi^{min}(\varepsilon)}{\Phi(\theta+\varepsilon)-\Phi(\theta-\varepsilon)}\,d\gamma \\ &\leq \binom{N-1}{k}\int_{\theta-\varepsilon}^{\theta+\varepsilon}\frac{\phi^{max}(\varepsilon)-\phi^{min}(\varepsilon)}{\Phi(\theta+\varepsilon)-\Phi(\theta-\varepsilon)}\,d\gamma=\binom{N-1}{k}\frac{2\varepsilon(\phi^{max}(\varepsilon)-\phi^{min}(\varepsilon))}{\Phi(\theta+\varepsilon)-\Phi(\theta-\varepsilon)}.\end{aligned}$$

From L'Hopital's rule and the fact that $\lim_{\varepsilon\to 0}\phi^{max}(\varepsilon)=\lim_{\varepsilon\to 0}\phi^{min}(\varepsilon)=\phi(\theta)$, it follows that the last term converges to zero as $\varepsilon\to 0$. Therefore, the approximation of $g_{N-1}(k)$ converges to $g_{N-1}(k)$ as $\varepsilon\to 0$. Hence, the approximation is arbitrarily exact for ε sufficiently small (and is totally exact for $\varepsilon=0$). We will use this result later.

Using the approximation we get

$$\begin{aligned}g_{N-1}(k) &\approx \binom{N-1}{k}\int_{\theta-\varepsilon}^{\theta+\varepsilon}\frac{1}{2\varepsilon}\left(\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{k}\left(1-\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{N-1-k}d\gamma \\ &= \binom{N-1}{k}\int_{\theta-\varepsilon}^{\theta+\varepsilon}\frac{N-1-k}{k+1}\left(\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{k+1}\frac{1}{2\varepsilon}\left(1-\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{N-2-k}d\gamma \\ &= \binom{N-1}{k+1}\int_{\theta-\varepsilon}^{\theta+\varepsilon}\left(\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{k+1}\frac{1}{2\varepsilon}\left(1-\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{N-2-k}d\gamma\end{aligned}$$

where the step from the first to the second line uses integration by parts (with $[(\gamma-\theta+\varepsilon)/(2\varepsilon)]^{k}/(2\varepsilon)$ as "first part" and $[1-(\gamma-\theta+\varepsilon)/(2\varepsilon)]^{N-1-k)}$ as "second part"). Using integration by parts for $N-1-k$ times gives

$$g_{N-1}(k)\approx\int_{\theta-\varepsilon}^{\theta+\varepsilon}\left(\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{N-1}\frac{1}{2\varepsilon}\,d\gamma=\left[\frac{1}{N}\left(\frac{\gamma-\theta+\varepsilon}{2\varepsilon}\right)^{N}\right]_{\theta-\varepsilon}^{\theta+\varepsilon}=\frac{1}{N}.$$

Hence, we have obtained that a player receiving the cutoff signal has (approximately) uniform beliefs over the number of players that have received a signal lower than him.

Now we want to consider the expected utility difference between revolting and not revolting of a player receiving cutoff signal θ. If there is no integer $m\in\mathbb{N}$ such that $\theta-\varepsilon\leq m\leq\theta+\varepsilon$, then this utility difference equals $b-(q+b)\lfloor\theta\rfloor/N$ because a breakout cannot succeed if less than $\lfloor\theta\rfloor$ other prisoners play revolt.[20] Given the uniform beliefs derived above, the probability that less than $\lfloor\theta\rfloor$ players play revolt is $\lfloor\theta\rfloor/N$.

[20]Recall that $\lfloor x\rfloor=\max\{n:n\in\mathbb{N}\text{ and }n\leq x\}$, i.e. $\lfloor x\rfloor$ is the highest integer below x.

If there is an integer $m \in [\theta - \varepsilon, \theta + \varepsilon]$, then the expected utility difference is

$$b - (q+b)\left[\frac{(\theta+\varepsilon-m)}{2\varepsilon}\frac{(m+1)}{N} + \left(1 - \frac{\theta+\varepsilon-m}{2\varepsilon}\right)\frac{m}{N}\right].$$

Viewed as a function of θ, the expected utility difference is, therefore, flat on intervals (θ_1, θ_2) such that $\lfloor \theta_1 - \varepsilon \rfloor = \lfloor \theta_2 + \varepsilon \rfloor$ and strictly decreasing in an ε-ball around each integer. As the utility difference is continuous in θ and as it is strictly positive (negative) for $\theta < 1 - \varepsilon$ (for $\theta > N$), there is a unique θ at which the expected utility difference is zero unless the equation $b - (q+b)x/N = 0$ is solved by an integer x, i.e. unless $bN/(q+b) \in \mathbb{N}$, which we ruled out by assumption.[21] As $bN/(q+b) \in \mathbb{N}$ is clearly not true for generic parameter values (q, b, N), there exists a unique θ at which the expected utility difference is zero for generic parameter values. In the limit as $\varepsilon = 0$, we then have – for generic parameter values – that (i) the expected utility difference is strictly positive for $\theta < \theta^*$ and (ii) the expected utility difference is strictly negative for $\theta > \theta^*$. Note that (in the limit $\varepsilon \to 0$) the expected utility difference viewed as a function of θ is discontinuous at θ^*.

The results of the previous paragraph were derived using the approximation of $g_{N-1}(k)$. Now we relax the use of the approximation to obtain the statement we want to show. Take any $\theta < \theta^*$. As the approximation of $g_{N-1}(k)$ converges to $g_{N-1}(k)$, one can find an $\bar{\varepsilon}(\theta) > 0$ such that the expected utility difference is strictly positive for θ for all $\varepsilon \leq \bar{\varepsilon}(\theta)$ (let $\bar{\varepsilon}(\theta)$ be the supremum of all such noise level). Similarly, for each $\theta > \theta^*$ an $\bar{\varepsilon}(\theta)$ can be found such that the expected utility difference at θ is strictly negative for each $\varepsilon \leq \bar{\varepsilon}(\theta)$. Note that $\bar{\varepsilon}(\theta)$ is continuous in θ on $[0, \theta^* - \delta]$ for any given $\delta > 0$: Take $\varepsilon < \bar{\varepsilon}(\theta')$ as given. Since beliefs – i.e. $g_{N-1}(k)$ – change continuously in θ, the expected utility difference is positive not only for θ' but for all θ in some open neighborhood around θ' (given ε). Consequently, $\varepsilon < \bar{\varepsilon}(\theta)$ for every θ in this open neighborhood. A similar argument shows that $\bar{\varepsilon}(\theta)$ is continuous on $[\theta^* + \delta, N]$.

For a given $\delta > 0$, let $\bar{\varepsilon} = \min\{1/2, \min_{\theta \in [0, \theta^* - \delta] \cup [\theta^* + \delta, N]} \bar{\varepsilon}(\theta)\}$. Note that $\min_{\theta \in [0, \theta^* - \delta] \cup [\theta^* + \delta, N]} \bar{\varepsilon}(\theta)$ exists and is strictly greater than zero as it is the minimum over a compact set of an everywhere positive and continuous function. Since revolting is a dominant strategy for signals below $1/2$ (given that $\varepsilon < 1/2$) and not revolting is dominant for signals above $N - 1/2$ (given that $\varepsilon < 1/2$), the expected utility difference is automatically positive (negative) for signals below zero (above N). This concludes the proof of the second step.

For any given $\delta > 0$, there is an $\bar{\varepsilon} > 0$ such that a player with signal below $\theta^* - \delta$ (above $\theta^* + \delta$) plays revolt (not revolt) for all $\varepsilon \leq \bar{\varepsilon}$ in any equilibrium. Hence,

[21] In this case, the expected utility would be zero on one of the flat parts.

each prisoner follows a cutoff strategy with cutoff θ^* in the limit as $\varepsilon \to 0$. We use the $\bar{\varepsilon}$ determined in step 2. Take an arbitrary equilibrium. Denote by θ_1 the infimum of all signals for which some prisoner does not play revolt for sure in this equilibrium. Such a θ_1 exists because of the dominance regions, i.e. revolting (not revolting) is a dominant action for a signal below $1-\bar{\varepsilon}$ (above $N-1+\bar{\varepsilon}$). Then a prisoner receiving any signal below θ_1 should prefer revolting (expected utility difference weakly positive) while there are signals above θ_1 but arbitrarily close to θ_1 where the prisoner prefers not revolting (expected utility difference weakly negative). We will now show that $\theta_1 \geq \theta^* - \delta$: Change all other players strategies such that every player does not revolt if and only if he receives a signal above θ_1. By the first step (supermodularity) and the definition of θ_1, this will make revolting less attractive (decrease the expected utility difference). Hence, a player receiving signal θ_1 will (given that all players use a cutoff strategy with cutoff θ_1) prefer not revolting to revolting. Therefore, by the second step, $\theta_1 \geq \theta^* - \delta$.

Similarly, let θ_2 be the supremum of all signals such that some player plays revolt (with non-zero probability), i.e. for all signals above θ_2 all players prefer not revolting but for some signals below and arbitrary close to θ_2 player i prefers revolting and change the strategies of all other players to cutoff strategies with cutoff θ_2. Player i will then prefer revolting when receiving signal θ_2 (first step). The second step then implies that $\theta_2 \leq \theta^* + \delta$.

In the limit as $\delta, \varepsilon \to 0$, we clearly get $\theta_1 = \theta_2 = \theta^*$. □

Lemma S1. *For sufficiently high b or low q, only the equilibrium in which the warden mixes over N and $N-1$ exists. For sufficiently high B, the equilibrium in which the warden mixes between 0 and 1 is the only mixed equilibrium.*

Proof: As pointed out in the main text, equilibrium p and γ_1 are determined simultaneously by (2) and (1) as the warden's own mixing probability does not play a role in these conditions. Given these two values, (3) will determine the optimal mixing probability of the warden. This insight shows that b and q will not affect the optimal γ_1 or the equilibrium revolt probability p because these parameters do not play a role in (2) and (1). Note that Δ is linearly increasing in b and linearly decreasing in q. Both variables are not part of the warden's maximization problem. Hence, changes in b and q do not affect the equilibrium mixing probability p for a given support of the warden. This implies that for b high enough (q low enough) $\Delta(\gamma)$ is positive for all $\gamma \in \{0, \ldots, N-1\}$. Hence, only the equilibrium where the warden mixes between $N-1$ and N exists if b is sufficiently high (or q sufficiently low).

The payoff of the warden when using N guards is $-N$ while his payoff when using $\gamma < N$ guards is $-B(1-G(\gamma)) - \gamma$. In any mixed equilibrium, the warden has to play an action $\gamma < N$ with positive probability and therefore he must prefer this action (weakly) to the

action $\gamma = N$. For $B \to \infty$, this can only be true if $lim_{B\to\infty} p = 0$. Put differently, the equilibrium mixing probability of the prisoner p in a mixed equilibrium becomes arbitrarily small as B increases. Note that very small p imply high $G_{N-1}(\gamma-1)$ for $\gamma \geq 1$. Consequently, $\Delta(\gamma)$ is negative for sufficiently low p for all $\gamma \geq 1$. As a mixed equilibrium in which the warden mixes over γ_1 and $\gamma_1 + 1$ can only exist if $\Delta(\gamma_1) > 0 > \Delta(\gamma_1 + 1)$, it follows that for sufficiently high B the mixed equilibrium in which the warden mixes over 0 and 1 is the only mixed equilibrium that exists. □

Proof of proposition 1: Lemma S1 establishes that for B high enough the only mixed equilibrium is the one where the warden mixes over 0 and 1. The proof of the lemma also establishes that $\Delta(\gamma) < 0$ for $\gamma \geq 1$ if B is sufficiently high. Consequently, also no semi-mixed equilibrium exists for B high enough. Let $\hat{B}$ be such that only the mixed equilibrium in which the warden mixes over 0 and 1 exists for any $B \geq \hat{B}$. For the rest of the proof, consider only $B \geq \hat{B}$.

In this mixed equilibrium the warden is indifferent between 0 and 1 which means $Bg(1) = 1$ or equivalently $N(1-p)^{N-1}p = 1/B$. Therefore, $\lim_{B\to\infty} p(B) = 0$ where $p(B)$ is the prisoners' equilibrium probability of playing r when the warden's utility is B. Since the warden is indifferent between playing 0 and 1 in equilibrium, his equilibrium payoff equals $\pi(0) = -(1-(1-p)^N)B$. Plugging in the indifference condition $N(1-p)^{N-1}p = 1/B$ derived above yields the warden's equilibrium payoff

$$\pi^* = \frac{(1-p)^N - 1}{N(1-p)^{N-1}p}.$$

Applying L'Hôpital's rule, gives $\lim_{p\to 0} \pi^* = -1$. As we established above, p approaches 0 when $B \to \infty$. Consequently, the warden's payoff in the mixed equilibrium approaches -1 as $B \to \infty$. Furthermore,

$$\begin{aligned}\frac{\partial \pi^*}{\partial p} &= \frac{-N^2(1-p)^{2N-2}p - ((1-p)^N - 1)(-N(N-1)(1-p)^{N-2}p + N(1-p)^{N-1})}{N^2(1-p)^{2N-2}p^2} \\ &= \frac{1 - Np - (1-p)^N}{N(1-p)^N p^2}.\end{aligned}$$

Using L'Hôpital's rule, gives $\partial\pi^*/\partial p|_{p=0} = -(N-1)/2 < 0$. Hence, the warden's payoff approaches -1 from below as $B \to \infty$ and the warden's payoff in the equilibrium where he mixes over 0 and 1 is bounded from above by -1. This proves the proposition because in the transparency model the warden's equilibrium payoff is $-\theta^*$ for any value of B. □

Proof of proposition 2: It was shown in lemma S1 that for b/q high enough, the unique equilibrium in the panopticon model is a mixed equilibrium in which the warden mixes over

$N-1$ and N and his payoff is $-N$. A similar result holds for the transparency model: $\theta^* = N$ if and only if $b/(q+b) > (N-1)/N$ or equivalently if $(b/q) > N-1$. Clearly, $\theta^* = N$ implies that the warden's equilibrium payoff is $-N$. This establishes the result that for b/q high enough all models lead to a warden payoff of $-N$.

Now consider the panopticon. In an equilibrium in which the warden mixes over $N-1$ and N, he has to be indifferent between these two options which implies $1 = Bp^N$, i.e. the mixing probability of the prisoner has to be $p = (1/B)^{1/N}$ in such an equilibrium. To have such an equilibrium, the condition $\Delta(N-1) > 0$ has to be satisfied. Given $p = (1/B)^{1/N}$, this condition becomes $-q\left(1-(1/B)^{(N-1)/N}\right) + b(1/B)^{(N-1)/N} > 0$. This can be rewritten as $b/q > B^{(N-1)/N} - 1$.

If $B^{(N-1)/N} - 1 > b/q > N-1$, then the warden's payoff in the transparency model is $-N$. In the panopticon, however, the equilibrium in which the warden mixes between N and $N-1$ does not exist which means the warden plays N with zero probability in any equilibrium of this game. As the equilibrium guard levels are then strictly preferred to a guard level of N (which would guarantee payoff $-N$), it follows that the warden's payoff in the no information game is strictly larger than $-N$.

If $B^{(N-1)/N} - 1 < b/q < N-1$, the no information game has an equilibrium in which the warden mixes between $N-1$ and N and therefore his expected payoff in this equilibrium is $-N$. In the transparency model, $\theta^* < N$ and therefore the warden's equilibrium payoff is strictly above $-N$. □

Extension: Uncertain punishment

Here we consider a variation of the model in which a prisoner's payoff when revolting unsuccessfully is $-q - \rho\gamma/N < 0$ where $q \geq 0$ is an effort cost and $\rho \geq 0$ is a punishment that happens with probability γ/N. It will become apparent that the the specific linear form chosen here is irrelevant for the analysis, i.e. we could just as well use $-q - h(\gamma, N)$ where $h \geq 0$ increases in its first and decreases in its second argument. Apart from this change in payoff, the model is the same as in the main text.

Note that the arguments in the **benchmark model** go through without change.

In the **transparency model**, lemma 5 holds with a slightly redefined threshold θ^*. Let θ^* be the unique θ such that

- either $\theta \notin \mathbb{N}$ and

$$b - \left(q + b + \frac{\theta}{N}\rho\right)\frac{\lfloor\theta\rfloor}{N}$$

- or $\theta \in \mathbb{N}$ and

$$\begin{aligned} 0 &\geq b - \left(q + b + \frac{\theta}{N}\rho\right)\frac{\theta}{N} \\ 0 &\leq b - \left(q + b + \frac{\theta}{N}\rho\right)\frac{\theta - 1}{N}. \end{aligned}$$

The proof of lemma 5 has to be adjusted only at very few instances: In the first step,

$$\Delta(\gamma) = b - \left(q + b + \frac{\theta}{N}\rho\right) G_{N-1}(\gamma - 1)$$

and everything goes through accordingly.

In the second step, the derivation of the approximation and the resulting Laplacian beliefs remains unaffected. The expected utility difference between rioting and not rioting if there does not exist an $m \in \mathbb{N}$ such that $\theta - \varepsilon \leq m \leq \theta + \varepsilon$ will now be

$$b - \left(q + b + \frac{\theta}{N}\rho\right)\frac{\lfloor\theta\rfloor}{N}.$$

If such an m exists, the expected utility difference is

$$b - \left(q + b + \left(\frac{m}{2} + \frac{\theta + \varepsilon}{2}\right)\frac{\rho}{N}\right)\frac{\theta + \varepsilon - m}{2\varepsilon}\frac{m+1}{N} - \left(q + b + \left(\frac{m}{2} + \frac{\theta - \varepsilon}{2}\right)\frac{\rho}{N}\right)\left(1 - \frac{\theta + \varepsilon - m}{2\varepsilon}\right)\frac{m}{N}.$$

Note that this expected utility difference is strictly decreasing in θ if $\rho > 0$. As rioting is dominant for $\theta < 1 - \varepsilon$ and not rioting is dominant for $\theta > N + \varepsilon$, there is a unique θ at which the expected utility difference is zero. In the limit $\varepsilon \to 0$, we obtain that the expected utility difference is strictly positive for every $\theta < \theta^*$ and strictly negative for every $\theta > \theta^*$. Given this, the remaining parts of the proof of lemma 5 apply without change.

In the **panopticon model**, the indifference condition of the prisoner (3) has to be rewritten as

$$\mathbb{E}\left[b - G_{N-1}(\gamma - 1)\left(b + q + \rho\frac{\gamma}{N}\right)\right] = 0.$$

Lemma 2 remains valid because it uses only the warden's problem which was not changed. The proof of lemma S1 uses the prisoners' indifference condition without using the specific form of the prisoner payoff. Consequently, the proof goes through without change and the lemma remains valid.

The most interesting **comparison** of the models is the result for large N (theorem 1). The proof of this result does again not use the specific form of the prisoners' indifference condition and consequently goes through without change. Hence, all the results for large N

mentioned in the main text remain valid.

Extension: Stochastic breakout

The probability of a breakout was 1 in the main text whenever the number of revolting prisoners exceeded γ and zero otherwise. It is straightforward to extend the model to a framework in which the probability of a breakout is stochastic. In this section, we change the setup in the following way: If m of the N prisoners revolt and the guard level is γ, then the probability of a breakout is

$$\beta \mathbb{1}_{m>\gamma} + (1-\beta)\frac{m}{N}$$

where $\beta \in (0,1)$ and $\mathbb{1}$ is the indicator function.[22] The model of the main text emerges for $\beta = 1$. In this setup, it is necessary to adjust assumption 1 which implies that the warden would prevent a breakout if he knew that all prisoners revolt with probability one. In the setup with stochastic breakouts, the assumption is $\beta B \geq N+1$. We will need additional parameter assumptions in order to ensure that prisoners have dominant strategies if the warden chose zero or N guards. That is, we make the assumption

$$\beta > \frac{b}{q+b} > (1-\beta)\frac{N-1}{N}$$

which (after rearrangement) states that it is dominant to revolt for a given prisoner if $\gamma = 0$ and it is dominant not to revolt if $\gamma = N$.

In the transparency model, θ^* changes to

$$\theta^* = \left\lceil \frac{N}{\beta}\left(\frac{b}{q+b} - \frac{1-\beta}{2}\right)\right\rceil.$$

With this θ^*, lemma 5 applies to the new setup. To see this, note that the first part of the proof (strategic complementarity) still goes through. In the second part, the utility difference between revolting and not revolting if there is no integer $k \in \mathbb{N}$ such that $\theta_\varepsilon \leq k \leq \theta + \varepsilon$ is now $b - (q+b)(\beta\lfloor\theta\rfloor/N + (1-\beta)(N-1)/(2N))$. If there is an integer $k \in \mathbb{N}$ such that $\theta_\varepsilon \leq k \leq \theta + \varepsilon$, then the expected utility difference becomes

$$b - (q+b)\beta\left[\frac{(\theta+\varepsilon-k)}{2\varepsilon}\frac{(k+1)}{N} + \left(1 - \frac{\theta+\varepsilon-k}{2\varepsilon}\right)\frac{k}{N}\right] - (q+b)(1-\beta)\frac{N-1}{2N}.$$

[22] In our prison example, one could think of this story: Fleeing prisoners run into the guards with probability β. In this case, they succeed only if they outnumber the guards. If prisoners find a way out where there are no guards (probability $1-\beta$), they have to overcome obstacles like walls/locks/fences etc. and the more prisoners participate, the more likely it is that they will manage.

Everything else in the proof of lemma 5 goes through without change. Note that by the parameter assumption made above θ^* is still linearly increasing in N.

In the panopticon, the warden's payoff maximization (1) becomes

$$\max_{\gamma\in\{0,1,\dots,N\}} -(1-G(\gamma))\beta B - \gamma - \beta\frac{\sum_{k=0}^{N-1} kg(k)}{N}B.$$

Note that this maximization problem differs from the one in the main text only by a term which is constant in γ. Hence, the warden's maximization problem does essentially not change. The prisoners' indifference condition (3) has to be rewritten as

$$\mathbb{E}_\gamma\left[b - \left(\beta G_{N-1}(\gamma-1) + (1-\beta) * \left(1 - \frac{1+\sum_{k=0}^{N-1} kg_{N-1}(k)}{N}\right)\right)(b+q)\right] = 0.$$

Note that the term in brackets is still decreasing in γ and increasing in p. Lemma 2 remains valid because it only uses the warden's problem which is essentially unchanged (adding a constant does not affect the proofs). The proof of lemma S1 uses the prisoners' indifference condition without using the specific form of the prisoner payoff. Consequently, the proofs go through without change and the lemmas remain valid. It is still true that the mixed equilibrium in which the warden mixes between zero and one is the unique Nash equilibrium if N is large. The proof of this result was only based on the warden's indifference condition which implies that the probability that at least one other prisoner revolts converges to zero as N gets large. By the dominance assumptions (if all other prisoners do not revolt and the warden uses one or more guards, then not revolting is a best response), this implied that only the equilibrium with mixing over zero and one guard can exist. As the warden's indifference condition is unchanged, the whole proof still goes through.

The payoff comparison between transparency model and panopticon is also unaffected: The payoff of the transparency model is linearly decreasing in N while the panopticon payoff is still bounded from below. Hence, the panopticon leads to a higher payoff than the transparency model for large N.

Extension: Heterogeneous attackers

In the model of the paper, all "prisoners" are alike in the sense that they share the same payoff function. A generalization to arbitrarily heterogeneous prisoners leads to an intractable model for two reasons: First, the global game refinement used in the transparency model is no longer able to deliver a clear cut (and noise independent) prediction, see Carlsson (1989), Frankel et al. (2003) or Corsetti et al. (2004). Second, the support of the warden strategy in

the panopticon might contain more than two elements (and his payoff function might have several local optima). While a full generalization is impossible for these reasons the simple extension below proves to be tractable.

Think of the model's interpretation in terms of speculators who can attack a currency peg. Suppose there are K types of attackers who differ in the size of their budget. In particular, type $k \in 1, \ldots, K$ has k units of money to speculate with. For simplicity, assume that a speculator will always either use his complete budget to attack or he will not attack at all. The benefit of a successful attack is then $b * k$. The payoff of not attacking is normalized to zero as in the paper. The payoff from an unsuccessful attack is interpreted as a transaction cost. We assume that there are scale economies in speculating. That is, the transaction cost per unit is strictly decreasing in the budget size. More technically, $q_k \in [q_{k-1}, \frac{k}{k-1} q_{k-1})$ for $k > 1$. The proportion of each type in the population is common knowledge. When we check our result in theorem 1 we will interpret large N as multiplying the number of type k attackers by a large natural number. That is, we increase the number of attackers but keep the proportion of each type in the population fixed.

The main purpose of the extension is to show that the defender prefers the panopticon to the transparency model if N is large. For this, it is unnecessary to derive an equilibrium in the transparency model. It is sufficient to provide an upper bound on the warden's expected payoff in any equilibrium of the transparency model and show that – for large N – this upper bound is below the panopticon payoff. This is exactly what we will do. For the transparency model we can derive a weaker version of lemma 5 where N_K is the number of attackers of type K:

Lemma S2. *Let $\varepsilon' > 0$ and $N_K > 1$. Assume that $bN_K/(q+b) \notin \mathbb{N}$ and define*

$$\theta_K^* = \left\lceil \frac{bN_K}{q+b} \right\rceil.$$

Then for any $\delta > 0$, there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \leq \bar{\varepsilon}$, a player of type K receiving a signal below $\theta_K^* - \delta$ will play r.

The lemma states that type K attackers will attack whenever receiving a signal below $\theta^* - \delta$ where δ can be chosen arbitrarily small. That is, in the limit as $\varepsilon \to 0$ type K players will attack whenever receiving a signal below θ_K^*.

The proof of the lemma is equivalent to the proof of lemma 5 with some small modifications sketched below: Suppose that all types but type K will play n for any signal they get. If we can show that even under this absurd supposition a type K attacker will play attack whenever he receives a signal below $\theta_K^* - \delta$, then – by strategic complementarity – he will

also attack if the other types play any other strategy (and he receives a signal below $\theta_K^* - \delta$). If, however, we focus on the case where all types apart from type K play n for sure, then we basically have the model of the paper where all relevant attackers are homogeneous of type K. The second step of the proof of lemma 5 gives us the following result: *Suppose all type Ks follow a cutoff strategy with cutoff θ while all other types play n for sure for any signal. For a given $\delta > 0$, there exists an $\bar{\varepsilon}$ such that the utility of revolting for an attacker of type K with signal θ is higher than the utility from not attacking if $\theta \leq \theta_K^* - \delta$.* The proof of this statement is equivalent to the proof in the main paper. The third part of the proof is analogous and shows that a type K will attack whenever his signal is below $\theta_K^* - \delta$. By strategic complementarity this is also true if the other types choose to attack as well after some signals. But this implies that the defender has to use currency reserves of at least θ_K^* to prevent an attack. As the defender wants to prevent an attack by assumption 1, the currency reserves will be above θ_K^* in every equilibrium. Note that θ_K^* is linearly increasing in N_K which implies that the defenders equilibrium payoff is arbitrarily low for N (and therefore N_K) sufficiently high.

Now turn to the panopticon. Consider first the game where there are only N_K attackers of type K and no attackers of other types. In this case, the analysis of the paper applies but has to be rescaled by K. For example, the defender will mix only over multiples of K instead of mixing over integers. If N_K is sufficiently large, there will be a unique equilibrium in which the defender mixes over 0 and K; see theorem 1. Following the proof of theorem 1, the expected payoff of the defender is bounded from below in this equilibrium (by $-\alpha K$). Now add one attacker of type $k < K$. We claim that for N_K high enough the best response for this type k is to not attack. To see this note that type K attackers are indifferent between attacking and not attacking in the equilibrium with only type Ks. All we have to show is that a type $k < K$ has a lower expected payoff of attacking than a type K (given the strategies of the type K attackers). This expected payoff equals $(1 - G_{N_K}(0))kb - q_k G_{N_K}(0)$ while the indifference condition for the type K attackers is $(1 - G_{N_K-1}(0))Kb - q_K G_{N_K-1}(0) = 0$. As $q_K < q_k K/k$ by assumption, the indifference conditions implies $(1 - G_{N_K-1}(0))kb - q_k G_{N_K-1}(0) < 0$. The proof of theorem 1 shows that both $G_{N_K}(0)$ and $G_{N_K-1}(0)$ converge to 1 as N_K grows large. Therefore, $(1 - G_{N_K}(0))kb - q_k G_{N_K}(0) < 0$ for N_K sufficiently large which means that indeed type k finds it optimal to not attack. But this implies that in the game with N_K type K and one type $k < K$ there is an equilibrium in which the defender and the type K attackers behave as in the unique equilibrium in which only type K attackers are present and the type k attacker does not attack with probability 1 (for N_K large enough). Adding more type $k < K$ attackers (also with different $k' < K$) does not change this result and we therefore get that the panopticon game has the following equilibrium for N large: defender and type

K attackers use the same strategies as in the game in which only type K attackers were present; all other attackers do not attack with probability 1. The defender's expected payoff is the same as in the equilibrium with only N_K type K attackers and is therefore bounded from below. This establishes that defender payoff is higher in the panopticon than in the transparency model for N sufficiently large.

Note that the central bank will use currency reserves of size K with positive probability in the equilibrium of the panopticon model. If some investors have a lot of money, i.e. K is big, then this implies that the central bank might have substantial reserves in equilibrium (with positive probability). While this differs somewhat from the model in the paper the main point that the panopticon leads to a higher payoff than the transparency model remains valid.

Lower bound of the central binomial coefficient – Proof

We will show the equivalent $\binom{2n}{n} \geq 2^{2n}/(2\sqrt{n})$ as it is notationally more convenient. The first step is to see that

$$\begin{aligned}
\binom{2n}{n}\frac{1}{2^{2n}} &= \frac{1}{2^{2n}}\frac{(2n)!}{n!\,n!} \\
&= \frac{1}{2^n}\frac{(2n)!}{n!\,2^n n!} \\
&= \frac{1}{2^n}\frac{(2n-1)(2n-3)(2n-5)\dots 1}{n!} \\
&= \frac{1}{2^{n-1}}\frac{1}{2n}\frac{(2n-1)(2n-3)(2n-5)*\dots*3}{(n-1)(n-2)*\dots*1} \\
&= \frac{1}{2^{n-1}}\frac{1}{2n}\prod_{j=1}^{n-1}\frac{2j+1}{j} \\
&= \frac{1}{2n}\prod_{j=1}^{n-1}\left(1+\frac{1}{2j}\right).
\end{aligned}$$

The second step is to get a lower bound on the square of the product:

$$\begin{aligned}
\prod_{j=1}^{n-1}\left(1+\frac{1}{2j}\right)^2 &= \prod_{j=1}^{n-1}\left(1+\frac{1}{j}+\frac{1}{4j^2}\right) \\
&\geq \prod_{j=1}^{n-1}\left(1+\frac{1}{j}\right) = n.
\end{aligned}$$

Where the last equality can be easily shown by induction.[23] Taking the first two steps together shows that

$$\left(\binom{2n}{n}\frac{1}{2^{2n}}\right)^2 = \frac{1}{(2n)^2}\prod_{j=1}^{n-1}\left(1+\frac{1}{2j}\right)^2 \geq \frac{1}{4n^2}n = \frac{1}{4n}.$$

Taking square roots on both sides gives

$$\binom{2n}{n}\frac{1}{2^{2n}} \geq \frac{1}{2\sqrt{n}}$$

which is the desired result.

[23] Clearly, it holds for $n = 2$. For higher n, we get $\prod_{j=1}^{n-1}\left(1+\frac{1}{j}\right) = \left(1+\frac{1}{n-1}\right)\prod_{j=1}^{n-2}\left(1+\frac{1}{j}\right) = \left(1+\frac{1}{n-1}\right)(n-1) = n$ where the second equality uses the induction hypothesis.