How Jeremy Bentham would defend against self-fulfilling attacks *

Ole Jann
Nuffield College, University of Oxford

Christoph Schottmüller
University of Copenhagen and TILEC

July 17, 2017

Abstract

We analyze situations like revolutions and speculative attacks: A defender faces an attack by a group, which can only succeed if enough people participate. The defender can increase his strength but that is costly. We show that if the attackers do not observe the defender's strength, there is a unique Nash equilibrium in which the defender has almost no strength, and attacks almost never occur. This result is identical to Bentham's (1787) idea of the "panopticon". We show how it emerges from the interaction between the attackers' and the defender's beliefs, and does not require game-theoretic refinements.

Keywords: panopticon, coordination games, global games, transparency
JEL classification: D23 (Organizational Behavior), D74 (Conflict, Revolutions), D82 (Asymmetric and Private Information), E58 (Central Banks and Their Policies), F31 (Foreign Exchange), Z13 (Economic Sociology)

*Jann: Nuffield College and Department of Economics, University of Oxford, Oxford OX1 1NF; ole.jann@economics.ox.ac.uk. Schottmüller: Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K, Denmark; christoph.schottmueller@econ.ku.dk.
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, Tomas Sjöström, Joel Sobel, Peter Norman Sørensen and Adrien Vigier as well as from audiences at Bocconi University, the Universities of Bonn, Copenhagen, Lund and Oxford, Tilec, SING 2016 (Odense) and GAMES 2016 (Maastricht).

> Morals reformed – health preserved – industry invigorated – instruction diffused – public burthens lightened – Economy seated, as it were, upon a rock – the gordian knot of the Poor-Laws not cut, but untied – all by a simple idea in Architecture! *(Jeremy Bentham, 1787)*

1. Introduction

The general problem which we analyze in this article has the following structure: A group can act together to overpower an opponent; that opponent can defend himself by investing in defenses. Consider, for example, the warden of a prison who faces a possible prison riot. A riot is more likely to lead to a breakout if more prisoners participate, but less likely to do so if the warden has many well-equipped guards at his disposal. A single prisoner is therefore more willing to riot if many others do so as well – and no one wants to be the only one to show up against a massive number of guards. The warden can protect himself against riots by investing in additional guards, but only wants to do so if a riot is sufficiently likely since guards are costly. Other examples of this problem include governments facing a revolution, or central banks defending a currency peg against speculators.

We construct a simple theoretical model that fits all these situations. For ease of exposition, we stay with the graphic example of the prison. A prison warden chooses how many guards he wants to hire; guards are costly. Then a number of prisoners decide whether they want to revolt or not. If the number of prisoners who revolt is higher than the number of guards, these prisoners win and break out from the prison. Otherwise, the prisoners who revolted get punished. Prisoners who don't revolt will neither break out nor be punished.

Our model thus follows the same lines as canonical models of speculative attacks or regime change games.[1] The crucial difference is that we consider the defender an active player, whose strength is a strategic (and costly) choice and not simply a variable drawn by nature. This also allows us to analyze the influence of different information structures (which the defender might be able to design) on the outcome of the game.

Our main result (Theorem 1) is that if the number of guards is kept secret from the prisoners, there is a unique Nash equilibrium in which the warden hires almost no guards and prisoners almost never attack. 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 guard level is secret, and it only occurs if the number of prisoners is sufficiently high.[2]

[1] See Diamond and Dybvig (1983) for a model of bank runs, Krugman (1991), Obstfeld (1986) for models of speculative attacks. For global games approaches to these problems, see Goldstein and Pauzner (2005) (bank runs), Morris and Shin (1998) (speculative attacks), as well as the survey in Morris and Shin (2003).

[2] We discuss robustness of the result and consider several extensions in section 5.2.

This 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. His prediction was that this construction would make revolts impossible at a low cost.[3]

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 5.1. The following paragraphs provide a game-theoretic intuition into how these assumptions together lead to the existence and uniqueness of a Nash equilibrium which is extremely favorable to the warden.

The situation between the warden and the group of prisoners is one of pure 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.[4] This is a feature which our model shares with simple games such as "matching pennies" or "rock, paper, scissors".

If it is revealed to the prisoners how many guards the warden has hired, this makes their problem easier in the sense that they only need to solve a coordination problem among themselves. Their decision has not quite become as simple as that of a player of "rock, paper, scissors" who knows that her opponent is playing rock, but the warden has essentially been eliminated from the game. In the remaining coordination problem, the prisoners' beliefs are "self-fulfilling", i.e. if the prisoners believe that a successful breakout is likely, they act in a way that makes it more likely.

If the number of guards is kept secret, however, the warden and the prisoners effectively make a simultaneous choice. In any Nash equilibrium, the beliefs of both the warden and the prisoners must be consistent with the strategies of the other players. But where the beliefs of the prisoners are self-fulfilling, those of the warden add what one could call a "self-defeating" element: If he believes that a successful breakout is likely, he acts in a way (i.e. he hires additional guards) that makes a breakout less likely. The interplay between these countervailing belief effects – self-fulfilling and self-defeating – determines the Nash

[3]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? ... But who would think of beginning a work of hours and days, without any tolerable prospect of making so much as the first motion towards it unobserved?" 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."

[4]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.

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

Table 1: The information structures we consider.

equilibrium.

How does a sufficiently high number of prisoners guarantee the uniqueness of Nash equilibrium? Why is it such an extreme equilibrium in which breakouts almost never occur even though there are almost no guards? Consider again our analogy to "rock, paper, scissors". There, as in our game, players gain from being unpredictable. A player who can never play "rock" is an easier opponent, just as it makes things easier for the warden if he can more accurately predict the number of prisoners who will revolt.

But that is precisely what happens if there are many prisoners: 20 prisoners, who each have to decide on their own, can simply not be as unpredictable as a single prisoner in relative terms. If the prison contains a single prisoner who flips a fair coin, with probability one-half he revolts (i.e. everybody revolts) and with probability one-half he does not (i.e. nobody revolts). But if 20 prisoners each flip a coin to make their decision, the number of those who revolt will be quite tightly distributed around 10. More generally, it follows from the law of large numbers that a larger group, whose members have no way of correlating their behavior, can simply not be as unpredictable as a smaller group. It is a direct consequence of this "lack of unpredictability" that there can be no Nash equilibria in which breakouts happen with a large probability.

We compare our main result with the outcomes under different information structures. Table 1 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.

Firstly, 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.

Secondly, 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

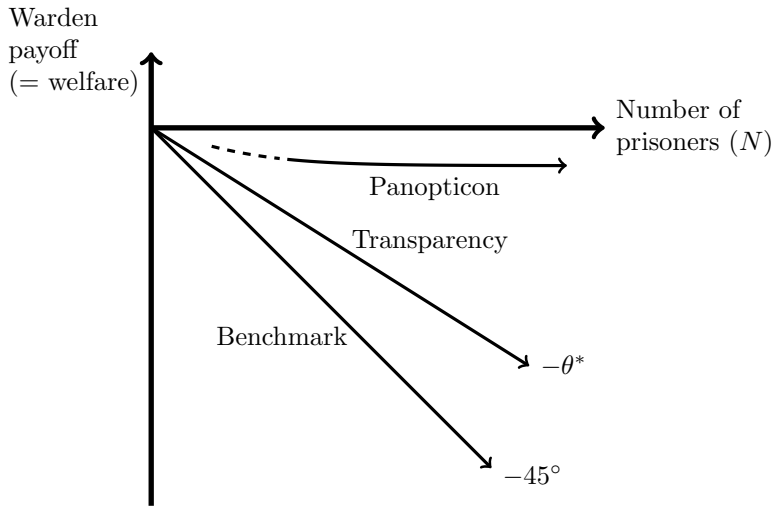


Figure 1: A comparison of welfare (which is equivalent to warden payoff) in the three information structures. The benchmark case is most expensive, as the warden needs as many guards as there are prisoners. In the transparency case, the warden can prevent breakouts with a lower number of prisoners; but the required number of guards still grows linearly in N. In the panopticon, the warden payoff is bounded from below by a constant.

then equivalent to the basic structure of regime-change games that have been studied in the literature (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 1 shows a comparison of welfare in all three information structures. This 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. These features were also the main ingredients in Bentham's "panopticon".

We do not claim that our model "shows" that central banks 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 take our model to have two main predictions. Firstly: The defender is best off by keeping his own strength secret. Secondly: There will be a unique equilibrium in which few resources are used on defense, yet successful attacks almost never occur.

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 extremely 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 second point about equilibrium existence and uniqueness. 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 on policymakers to save money by downsizing the police forces). We live, therefore, in an intermediate equilibrium in which there are relatively few police officers per population, but large-scale riots are rare.

Our model is directly related to the literature on 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 models predict that the self-fulfilling nature of attacks leads to multiple equilibria. This insight in itself 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. We replicate this result in our model (section 4.2). Our main result, however, shows that a unique equilibrium naturally emerges from the interaction of the attackers' and the defender's beliefs if – in contrast to the previous literature – the guard level is chosen endogenously and secretly. Uniqueness does not require any ad-hoc assumptions about higher-order beliefs or a signaling structure that creates a particular structure of higher-order beliefs.

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, 2014) 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.

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 this way, we have a limit result like Bolton and Farrell (1990), who consider a coordination game with strategic substitutes. In their model, however, the law of large numbers makes coordination easier (because the realized distribution gets closer to the expected distribution), while in our model the predictability of large groups means that all equilibria in which prisoners coordinate on attacking disappear.

Chwe (2003) provides a discussion of the panopticon and higher-order knowledge. The panopticon, he argues, creates common knowledge among prisoners of being in the same situation – an idea that is connected to Bentham's plan of having a chapel above the watchtower in his panopticon. Indeed we find that there are no asymmetric equilibria in our panopticon model, i.e. all the prisoners use the same strategy in equilibrium.

2. Model

This section describes the general setup of our model. The specific assumptions of the three information structures that we consider are described in the following sections.

First, the warden chooses a guard level $\gamma \in \mathbb{R}_+$. Second, N 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. The payoffs are as follows: Each prisoner values breaking out by $b > 0$. If the prisoner revolts but cannot break out, 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 2 for a summary of these payoffs.

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

Table 2: Payoff prisoner 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. The assumption implies that the warden would prevent a revolt (by setting $\gamma = N$) if he knew that all prisoners play r for sure.

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

The reasoning behind this assumption is as follows. If $B < N$, there is – independent of the specific 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 technical reasons, we assume $B \geq N + 1$ (instead of $B > N$) as it significantly simplifies the analysis.

We want to point out two other modeling choices we made: First, the warden's utility depends only on whether there is a breakout and not on how many prisoners break out (or by how much the number of revolting prisoners exceeds the guard level). In this sense, the disutility B corresponds to an image or reputation concern, or a regime preference. Also in the other applications mentioned in the introduction this assumption appears reasonable: A

central bank will mainly care about whether it was able to hold the announced peg (and less about how many speculators attacked the peg in case of a successful attack), a government about whether it can stay in power or not.

Second, prisoners that do not revolt will not break out (or have at least no benefit from doing so). Think of a prisoner sitting calmly in his cell who will not escape even if others do. Again this fits also the example of speculating against a currency peg: If one does not speculate against the peg, one cannot benefit from a successful attack. It should be noted, however, that our model is robust to deviations from this assumption as long as they do not destroy the strategic complementarity which is at the core of our model – see section 5.2 for details.

3. Analysis of the Panopticon

3.1. Preliminary Analysis

In this section, we analyze the model in which the warden's choice of γ cannot be observed by the prisoners, and prisoners cannot coordinate themselves. This setting closely resembles Bentham's idea of the "panopticon", which is how we call this model.[5] In section 4, we consider two alternative information structures by allowing the prisoners to observe γ before they make their choice, and allowing them to coordinate perfectly.[6]

We begin by showing that there exist only equilibria in which all prisoners play r with the same probability p in equilibrium:[7]

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.

[5]Bentham (1787) emphasized the lack of communication possibilities (leading directly to a coordination problem): "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, and extending as many feet as shall be thought necessary to form the largest dimension of the cell."

[6]We could allow prisoners to communicate by cheap talk, but such communication is usually not considered in stag hunt type coordination problems as every prisoner weakly benefits if the other prisoner plays revolt; messages are therefore not credible.

[7]The reason for lemma 1 lies in 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 $p_i < p_j$.

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 given that 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.*

We illustrate the lemma using figure 2. 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

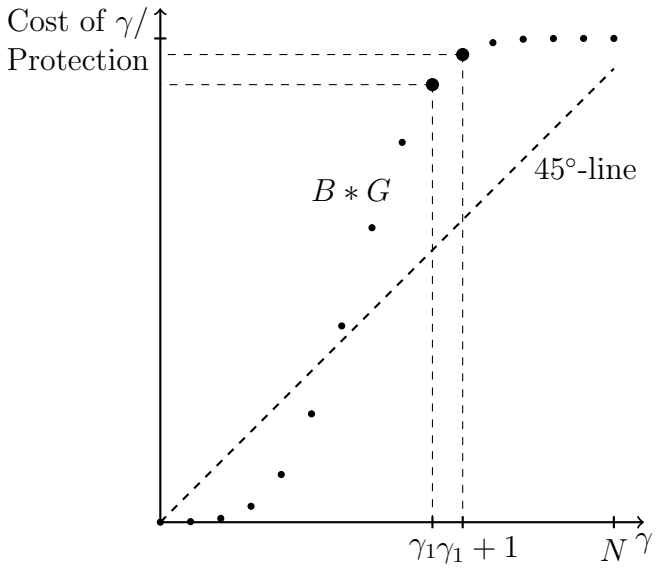


Figure 2: Equilibrium in the panopticon-model.

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 2.

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 $\gamma_1 + 1$ for

$\gamma_1 \in \{0, \ldots, N-1\}$. Furthermore, the warden's incentives do not pose an obstacle for the existence of such an equilibrium for any $\gamma_1 \in \{0, \ldots, N-1\}$ as there is always a p for which γ_1 and $\gamma_1 + 1$ are optimal. Whether an equilibrium exists for $\gamma_1 \in \{0, \ldots, N-1\}$ is determined by the prisoner's indifference condition. More precisely, a mixed strategy equilibrium where the warden mixes over γ_1 and $\gamma_1 + 1$ 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 $\gamma_1 + 1$ 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 $\gamma_1 + 1$ 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.

Given that potentially several equilibria exist, we are especially interested in the warden optimal equilibrium. (Note that prisoners have payoff zero in all equilibria as they are indifferent to playing n which yields zero.) The following lemma shows that the warden optimal equilibrium is the one with the lowest guard level. This equilibrium will also have the lowest revolt probability p.

Lemma 3. ***(Payoff ordering of equilibria)*** *Suppose there are two mixed equilibria: In equilibrium 1, the warden mixes over γ_1 and $\gamma_1 + 1$ and in equilibrium 2 the warden mixes over γ_2 and $\gamma_2 + 1$. Then the warden's equilibrium payoff is higher in equilibrium 1 if and only if $\gamma_1 < \gamma_2$. Furthermore, the prisoners' equilibrium probability of playing r is lower in equilibrium 1 if and only if $\gamma_1 < \gamma_2$.*

So far, we focused on completely mixed equilibria. However, there can be semi-mixed equilibria as well: the warden plays a pure strategy while the prisoners mix. Take a guard

level $\gamma \in \{1, \ldots, N-1\}$. There is a range of values for p such that γ is the warden's optimal choice. The prisoner is willing to mix if he is indifferent between revolting and not revolting, that is, if $\Delta(\gamma) = 0$. This indifference condition holds for exactly one p. If the p solving the indifference condition is accidentally within the range of p values for which γ is the maximizer of the warden's utility we have an equilibrium. The following lemma, however, states that semi-mixed equilibria are not warden optimal.

Lemma 4. ***(Semi-mixed equilibria are payoff dominated)*** *For every semi-mixed equilibrium, there is a completely mixed equilibrium in which the expected warden payoff is higher.*

We have therefore established the following for the panopticon model:

Result 1. ***(Panopticon)*** *In every equilibrium, the prisoners mix over r and n with identical probabilities. The warden mixes between some γ_1 and γ_1+1 in the warden optimal equilibrium. However, other equilibria (in which the warden mixes over γ_2 and γ_2+1 with $\gamma_2 > \gamma_1$ or the warden does not mix) can exist.*

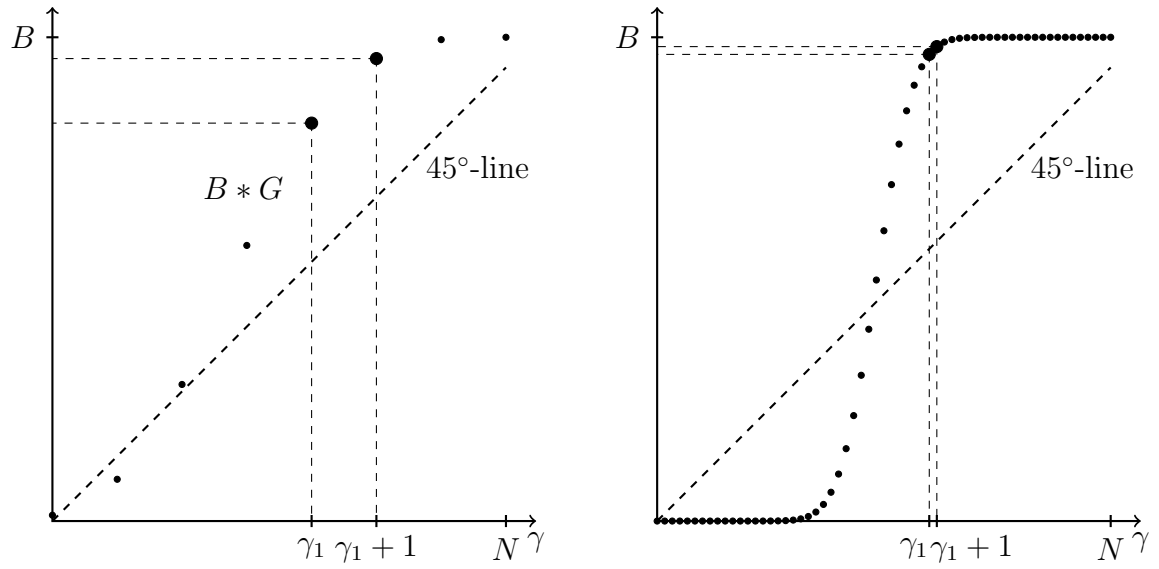
3.2. Unique Equilibrium for large N

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.*[8] *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.*

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 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 3 shows, similar to figure 2, 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.

[8] 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 $\bar{N}$ sucht that for all $N \geq \bar{N}$ and all B satisfying assumption 1, the results hold.



(A) small N (B) large N

Figure 3: An illustration of theorem 1.

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 to one without any prospects of support by his fellow inmates. The panopticon exploits, in this sense, the prisoners' coordination problem maximally.

Our preferred interpretation of the mixed equilibrium in the panopticon is in terms of

Harsanyi's (1973) purification. According to this interpretation, we can view the mixed strategy equilibrium as a 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 extreme prisoners who fear punishment the very least will revolt and every prisoner knows that it is extremely unlikely that such a prisoner is around. For every other prisoner, not revolting is the unique best response in the Bayesian game (and a non-unique best response in the limit). This interpretation is perfectly in line with Bentham's idea that in a panopticon (almost all) prisoners would not even consider revolting as they are only focused on the possibility of being punished.

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. This equilibrium need not necessarily be mixed, as we can see with a small change in the strategy sets. Assume that while the warden is free to vary the guard level, there is a minimum guard level γ_{min} which he cannot go below: Guards might have administrative duties, police forces have other jobs than to prevent a revolution and currency reserves facilitate the daily business of central banking.

To capture this assumption we could easily change the model in the following way (assume N sufficiently high so that only the 0-1 equilibrium would exist in the panopticon if we did not impose restrictions): Let z_{eq} be the probability of playing $\gamma = 1$ in the equilibrium without minimum guard requirement. Now say we require $\gamma \geq \gamma_{min}$ with probability $z_{min} > z_{eq}$ for some $\gamma_{min} \geq 1$. It follows straightforwardly from our results that there is a unique equilibrium in the panopticon in which the warden uses the minimum guard level and each prisoner plays n with probability 1.

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 seem unrealistic in some applications. Instead the main results are that (i) the probability of a breakout approaches zero for large N, and that (ii) this is hugely advantageous for the warden.

4. Comparison to Other Information Structures

4.1. Benchmark model: Perfect coordination

We will first compare the result for the panopticon model to a benchmark where we assume the coordination problem of the prisoners away. 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.[9] 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 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.

4.2. 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

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

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.

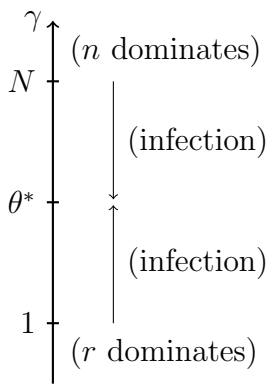


Figure 4: Infection of beliefs among prisoners: If $\gamma \geq N$, not revolting is a strictly dominant strategy for all prisoners. If $\gamma < 1$, revolting is strictly dominant. If $\gamma \in [1, N)$ and γ is common knowledge, there are two pure equilibria: Everybody revolts or no one revolts. When 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 $\gamma > \theta^*$.

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 $\varepsilon' > 0$.[10] 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.[11]

[10] In the context of a prison, one might think here of a normal distribution truncated at zero. The truncation affects neither results nor derivation.

[11] 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 act), prisoners could infer γ correctly in equilibrium despite the noisy observation if the warden did not

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 (which itself is driven by what others believe I do and therefore beliefs over beliefs and beliefs over beliefs over beliefs etc.).

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*[12]

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

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.

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.

"tremble". Consequently, prisoners would have common knowledge of γ despite the noise.

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

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.

4.3. 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 1 in the introduction 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.

5. Discussion

5.1. 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, and would simply concentrate their thinking on the possibility of being caught and disciplined. If the number of prisoners is large, our model exhibits the same property: For any prisoner, the probability that any of the other prisoners will revolt is close to zero, and the prisoner de facto finds himself in a game only between himself and the warden – where the warden chooses a mixing between having one guard and having no guards at all that just assures the prisoner's docility. By putting each prisoner in a situation where he is almost sure that no other prisoner will revolt, the panopticon thus makes optimal use of the prisoners' coordination problem.

In the 230 years since Bentham first proposed the panopticon, many scholars have interpreted it as a metaphor for 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 by overwhelming force or a "contest of violence" between those opposing and those defending it, as in our benchmark model. Instead, the docility of individuals allows for cost-saving minimal enforcement: There is neither wasteful use of resources through unused guard capacity nor fruitless attempts at revolting.[13] 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. It was this "accumulation of men" (p. 220)

[13]"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)

that, besides the accumulation of capital, made the industrial take-off of the early 18th century possible. Our result captures the intuition on how and why panopticism would work in a formal, game-theoretical model.

Moreover, modern society has at its center the individual, not the family or tribe or any other unit. This is crucial for maintaining the self-disciplining aspect of the panopticon, which relies on every prisoner reasoning on his own and choosing what is optimal for him, and facing strategic uncertainty about the choices of others. 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 – an idea that has gained credence by recent revelations of mass surveillance by intelligence agencies. Our results, especially the comparison of transparency and the panopticon, suggest that if the true level of surveillance is revealed (or there is a danger of revelation), efficacious enforcement becomes much more expensive in equilibrium – a reason why whistleblowers might indeed pose a threat to enforcement by panopticon.

These results show that "order" as used by Foucault, or the central prison metaphor of our theory, are neutral concepts: The free, democratic society might defend itself against an uprising for the sake of social welfare, while a repressive dictatorship might deploy secret surveillance methods to suppress dissent and rebellion. In our model, we are interested in the mechanisms by which this is done, and our results are positive, not normative.

5.2. Extensions and Robustness

Our main result has two parts: Firstly, the warden can almost always deter attacks in the panopticon by mixing between minimal guard levels if N is large. Secondly, this means that the panopticon is the optimal information structure for the warden if N is large. In this section, we consider several extensions and generalizations of our model and show that our main results are robust to such changes. In particular, we show how the fundamental property of large populations upon which our proof relies is still present in models with stochastic payoff functions, richer payoff functions, stochastic breakouts or heterogenous attackers.

So far, we assumed that revolting leads to a payoff of $-q$ for the prisoner if there was no successful breakout. In particular, this payoff did not depend on the guard level. This is in line with the interpretation of an effort cost in the prison or a transaction cost in the speculation application. One could, however, imagine that revolting prisoners are punished. In the application of a revolution, it is not unreasonable to assume that those that participated in a failed coup d'état might face severe consequences. Punishment, however, requires that the subversive activities are detected and the revolutionaries are identified. One could argue that the probability of being detected and identified depends on the guard level; e.g. the guards might not detect/identify all unsuccessful revolutionaries if there are few guards monitoring

a lot of "prisoners". One way to capture this is to say that the payoff of a revolting prisoner that does not break out is $-q - \rho\gamma/N < 0$ where $\rho \geq 0$ denotes a punishment and the probability of a punishment is proportional to the guard/prisoner ratio.

As we show in the supplementary material, our analysis covers this more general case. While the specific threshold level θ^* in the transparency model and the precise equilibrium mixing probabilities in the panopticon are different, the analysis remains qualitatively the same. In particular, the result that the panopticon is much better than the transparency and benchmark model for large N remains true. Also the result that the equilibrium probability of revolting in the panopticon is arbitrarily close to zero for large N holds. This captures an idea which has been central in understanding the effect of the panopticon: The prisoners behave as if they are watched because there is a slight chance that they are watched.[14] One could interpret γ/N as the fraction of prisoners that are watched or the chance of being discovered. With $q \to 0$ and $\rho > 0$, the only reason not to riot is the possibility of being watched (and punished if caught). Since prisoners almost always do not riot in equilibrium, they arguably behave as if they were watched because they are afraid that they might be watched.

Another possible extension of our model allows the payoff of a non-revolting prisoner to depend on whether a breakout occurs or not. Assume that the payoff of a non-revolting prisoner is $w \neq 0$ if a breakout occurs and zero if no breakout occurs. In the revolution example, w could be negative: If there is a successful coup, the new rulers might punish those that did not participate in the revolt. While the equilibria change quantitatively, all our qualitative results still hold in this setting. The crucial part is that $w < 0$ preserves the supermodular structure of the coordination game: A prisoner is more willing to revolt if other prisoners are more likely to revolt. If, on the other hand, $w > 0$, i.e. if there is a free riding problem, then our results only hold if w is not too big. More precisely, our derivations go through unless the free riding possibility destroys the supermodularity: A prisoner would then be less willing to revolt if others are more likely to revolt because he is more likely to get a high free rider benefit w when not revolting.

In our model, the probability of a breakout is 1 if the number of guards is less than the number of revolting prisoners and 0 otherwise. It is possible to generalize the model by introducing some randomness in the probability of a breakout. In the supplementary material, we show that all our results still hold if the the probability of a breakout is $\beta \mathbb{1}_{m>\gamma} + (1-\beta)m/N$ where m is the number of revolting prisoners, $\mathbb{1}$ is the indicator function and $\beta \in (0, 1]$ is a parameter (note that the model in the main text corresponds to $\beta = 1$). In terms of the

[14] This dates back to Bentham (1787) who writes "You will please to observe, that though perhaps it is the most important point, that the persons to be inspected should always feel themselves as if under inspection, at least as standing a great chance of being so, yet it is not by any means the only one."

revolution example, this setup could be interpreted as a probability β that the current regime fights an uprising using force and a probability $1-\beta$ that it is forced by international pressure to respond peacefully – for example by holding an election. The probability that protesters win the election increases in the number of initial protesters.

Finally, we consider an extension where the attackers differ in their size. Think, for example, of speculators who have different budgets. The central bank will then mix not between 0 and 1 but between 0 and the highest speculator budget in the panopticon model for large N. Intuitively, this is clear: If the central bank used (with probability 1) currency reserves less than the budget of the biggest speculator, this speculator would have a dominant strategy to speculate which would then always break the peg. We show in the supplementary material that the central bank is better off in the mixed equilibrium of the panopticon than in the transparency model if N is large.

6. Conclusion

This paper analyzes how a single player can defend against a group of opponents by making use of their coordination problem. Our model formalizes and replicates earlier results showing that "infection" in the absence of common knowledge can be used for this purpose, but our main result is to show that absolute secrecy is often optimal.

In the general debate between secrecy and transparency, this reminds us that we have to think clearly about the purpose and effect of information revelation. Revealing information to a single actor has the effect of informing and influencing that actor, but if that actor is part of a group it will also make him consider what kind of information the others have received, how they reason about his information and so on. These higher-order effects have to be considered and can be substantial. Showing one's power in the hope of deterring attackers might just give them the higher-order knowledge they need to coordinate on an attack, while not giving them any information may make them so predictable that attacks almost never happen in any equilibrium.

We can conceive of other situations for which our model offers only limited guidance. For example, the idea of transparency and forward guidance by central banks is not necessarily at odds with our result that secrecy is optimal: While our result is based on a conflict between the central bank and speculators, one could imagine other situations in which the interests of central bank and market participants are not opposed. In such a situation with aligned interests, transparency might indeed be an optimal policy. Our results show that the optimal information policy depends crucially on the degree of (mis-)alignment of interests between central bank and market participants.

We have seen that for a large number of prisoners, minimal enforcement with secrecy is optimal. This is in line with Bentham's original concept. But while prisons indeed rely more on cameras and prisoner separation than on massive numbers of guards, one might wonder why in many other situations massive presence of enforcement is publicly observable. For example, large numbers of police officers are deployed to uphold public order during (potentially violent) demonstrations and sport events. This does not contradict our theory. Demonstrators (or football hooligans) do not face a large coordination problem. By being in the same place, being able to observe each other and possibly even having some hierarchy among them, they can condition their choices upon each other's behavior and thereby achieve coordination. And, as we have shown in our benchmark model: when coordination problems do not matter, the warden chooses maximum enforcement in equilibrium.

Appendix

Proofs and limit results: Panopticon

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).[15] The binomial distribution G has the property that G is convex on $\{0, \dots, \gamma^m\}$ and G is concave on $\{\gamma^m, \dots, N\}$. Therefore, the maximization problem of the warden over the domain $\{0, \dots, \gamma^m\}$ is convex and consequently only the boundary values 0 and γ^m can be local maxima (on this restricted domain). If we take $\{\gamma^m, \dots, N\}$ 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 $\gamma \in \{1, \dots, \gamma_1\}$ 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 $\gamma \in \{1, \dots, \gamma_1\}$ (recall that $g(\gamma_1 + 1) = 1/B$) which contradicts our earlier result that the average $g(\gamma)$ for $\gamma \in \{1, \dots, \gamma_1\}$ is at most $1/B$.[16]

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 $g(\gamma_1 + 1) \leq 1/B$.[17] 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 $\gamma \geq \gamma_1 + 1$. The warden prefers γ_1 to $\gamma_1 - 1$ which implies $g(\gamma_1) \geq 1/B$. Furthermore, the warden has

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

[16] 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 $\gamma_0 = 0$.

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

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 lemma 3. Let $\gamma_1 < \gamma_2$. We first show that the equilibrium revolting probability p is lower in equilibrium 1. Suppose otherwise, i.e. suppose $p_1 > p_2$. As the warden prefers $\gamma_2 + 1$ over $\gamma_1 + 1$ given p_2, we have $G^{p_2}(\gamma_2 + 1) - G^{p_2}(\gamma_1 + 1) \geq (\gamma_2 - \gamma_1)/B$ where G^{p_2} is the

binomial cdf under p_2. This last inequality is equivalent to $\sum_{\gamma=\gamma_1+2}^{\gamma_2+1} g^{p_2}(\gamma) - (\gamma_2 - \gamma_1)/B \geq 0$. Note that $\gamma_1 + 1$ is strictly above the mode of g^{p_2}: We know that $\gamma_1 + 1$ is above the mode of g^{p_1} and as $p_1 > p_2$ the mode of g^{p_2} is lower than the mode of g^{p_1}. Similarly, any $\gamma \geq \gamma_1 + 1$ is strictly above the mode of any binomial distribution g^p with $p \in [p_2, p_1]$. This implies that $\sum_{\gamma=\gamma_1+2}^{\gamma_2+1} g^p(\gamma) - (\gamma_2 - \gamma_1)/B$ is strictly increasing in p for $p \in [p_2, p_1]$ and therefore $p_1 > p_2$ and $\sum_{\gamma=\gamma_1+2}^{\gamma_2+1} g^{p_2}(\gamma) - (\gamma_2 - \gamma_1)/B \geq 0$ imply that $\sum_{\gamma=\gamma_1+2}^{\gamma_2+1} g^{p_1}(\gamma) - (\gamma_2 - \gamma_1)/B > 0$. But this is equivalent to saying that the warden strictly prefers $\gamma_2 + 1$ over $\gamma_1 + 1$ under p_1 contradicting that $\gamma_1 + 1$ is the warden's equilibrium choice. Hence, $p_1 > p_2$ cannot hold and we have $p_2 \geq p_1$ whenever $\gamma_2 > \gamma_1$. In fact, $p_2 > p_1$ as otherwise the warden would have to be indifferent between at least three guard (γ_1, $\gamma_1 + 1, \gamma_2$ and $\gamma_2 + 1$) levels above the mode which is impossible by the concavity of G on $\{\gamma^m, \ldots, N\}$.

Given that $p_2 > p_1$, G^2 first order stochastically dominates G^1. Therefore, the warden's payoff $-(1 - G(\gamma))B - \gamma$ in equilibrium 1 is higher than his payoff in equilibrium 2 (i.e. if he played γ_2 under p_1, he would have a higher payoff than in equilibrium 2 and he can do even better by playing γ_1). □

Proof of lemma 4. Denote by $p(\gamma)$ for $\gamma \in \{0, \ldots, N-1\}$ the value of p for which the warden's payoff is maximized by γ and $\gamma + 1$. The proof of the previous lemma showed that $p(\gamma)$ is strictly increasing in γ. Denote by $\tilde{p}(\gamma)$ the value of p such that $\Delta(\gamma) = 0$. Clearly, $\tilde{p}$ is strictly increasing as well.

Now let there be a semi-mixed equilibrium at γ'. This implies that the $\tilde{p}(\gamma')$ is between $p(\gamma' - 1)$ and $p(\gamma')$. If $\tilde{p}(\gamma' - 1)$ is below $p(\gamma' - 1)$, then there is a completely mixed equilibrium where the warden mixes between $\gamma' - 1$ and γ' which leads to a higher payoff for the warden than the γ' equilibrium as the probability of revolting is $p(\gamma' - 1)$ in the mixed equilibrium which is lower than in the semi-mixed equilibrium. Therefore, let's proceed by supposing that $\tilde{p}(\gamma' - 1)$ is above $p(\gamma' - 1)$. This implies that $\tilde{p}(\gamma' - 1)$ is also above $p(\gamma' - 2)$.[18] If $\tilde{p}(\gamma' - 2)$ is below $p(\gamma' - 2)$, then there is a completely mixed equilibrium where the warden mixes between $\gamma' - 1$ and $\gamma' - 2$ which gives him a clearly higher payoff than the γ' semi-mixed equilibrium. Therefore, let us proceed by assuming that $\tilde{p}(\gamma' - 2)$ is above $p(\gamma' - 2)$ which implies that $\tilde{p}(\gamma' - 2)$ is also above $p(\gamma' - 3)$. Iterating further in this way, we finally reach the case where $\tilde{p}(1)$ is above $p(0)$. But this implies that there is an equilibrium where the warden mixes over 0 and 1 and $p = p(0)$: Since $\tilde{p}(1) > p(0)$, $\Delta(1) < 0$ while obviously $\Delta(0) > 0$. □

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 in the panopticon (for this 0-1 mixed equilibrium) and show that

[18] If $\tilde{p}(\gamma' - 2)$ does not exist, then the prisoner prefers not revolting to revolting for all values of p where $\gamma' - 2$ is weakly above the mode (in particular for $p(\gamma' - 2)$ and $p(\gamma' - 3)$) and the same argument as follows still applies.

it is above the warden payoff in the transparency model. Last we will show uniqueness of the equilibrium in the panopticon 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 $N \to \infty$.[19] Consequently, $\Delta(1) < 0$ for N sufficiently high; i.e. the 0-1 mixed equilibrium exists. Lemma 3 establishes that this is the warden optimal equilibrium in the panopticon.

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

[19] 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.

term in the limit can be written as

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

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. As the warden's payoff is $-\theta^* = -\lceil Nb/(q+b)\rceil$ in the transparency model, the warden has a higher payoff in the panopticon for N high enough.[20]

Finally, we show uniqueness of the mixed equilibrium with $\gamma_1 = 0$ in the panopticon (for large N). To do so, we need two intermediate results that are stated as lemmas below (lemma 6 and 7). 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 6. ***(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

[20]Note that the result does not depend on using a fixed α. More precisely, take a sequence of N and $B_N = \alpha_N(N+1)$ with $\alpha_N \geq 1$ for all N. The previous steps above still apply (for each given N) and the warden will prefer the no information 0-1 mixed equilibrium to $-\theta^*$ for N high enough as long as the sequence of α_N is bounded by some $\bar{\alpha}$.

obtain

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

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

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

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. $m = N/2$.[21] 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

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

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).

[21] 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.

Lemma 7. ***(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 6 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.

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 7 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. □

Proofs transparency model

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 revolt-

ing 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.[22] Given the uniform beliefs derived above, the probability that less than $\lfloor\theta\rfloor$ players play revolt is $\lfloor\theta\rfloor/N$.

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.[23] 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 $N - 1/2$

[22] 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.

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

(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, 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 $\theta_1 = \theta_2 = \theta^*$. □

Further comparison proofs

Lemma 8. *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 8 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 8 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 $-N$. □

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.

Bolton, P. and J. Farrell (1990). Decentralization, duplication, and delay. *Journal of Political Economy 98*(4), 803–826.

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.

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.

Diamond, D. and P. Dybvig (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy 91(3)*, 401–419.

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. (2014). Defending against speculative attacks: Reputation, learning, and coordination. Working paper, University of California, Irvine. Available at SSRN: http://ssrn.com/abstract=1960673.

Krugman, P. (1991). History versus expectations. *Quarterly Journal of Economics 106*(2), 651–667.

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.

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

Supplementary Material

not intended for publication

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

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

Lemmas 2 and 3 remain valid because they use only the warden's problem which was not changed. The proofs of lemmas 8 and 4 use 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.

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.[24] 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

[24] 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.

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

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

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

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

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

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

Note that the term in brackets is still decreasing in γ and increasing in p. Lemmas 2 and 3 remain valid because they use only the warden's problem which is essentially unchanged (adding a constant does not affect the proofs). The proofs of lemmas 8 and 4 use 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 S1. *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.

Example: N=2

To illustrate the results of the paper, we give the solved model for the simple case where $N = 2$.

Denoting the expected warden payoff by $\pi(\gamma)$, we get for the $N = 2$ case

$$\begin{aligned}\pi(0) &= -(2p + p^2)B \\ \pi(1) &= -p^2B - 1 \\ \pi(2) &= -2.\end{aligned}$$

This implies that $\pi(0) = \pi(1)$ iff $p = 1/(2B)$. Given the assumption $B \geq N + 1 = 3$, $\pi(0) = \pi(1) > \pi(2)$ holds if $p = 1/(2B)$.

Furthermore, $\pi(1) = \pi(2)$ iff $p = \sqrt{\frac{1}{B}}$ and $B \geq 3$ implies in this case that $\pi(1) = \pi(2) > \pi(0)$. To determine the equilibrium we will have to check the prisoners' indifference condition. Denoting the utility difference from revolting and not revolting given γ guards by $\Delta(\gamma)$ we get

$$\begin{aligned}\Delta(0) &= b \\ \Delta(1) &= -q(1 - p) + bp \\ \Delta(2) &= -q.\end{aligned}$$

If $\Delta(1) < 0$ with $p = 1/(2B)$, then there is an equilibrium in which the warden mixes over 0 and 1 with probability $z_{0,1} = \frac{-\Delta(1)}{-\Delta(1)+\Delta(0)} = \frac{q-b/(2B-1)}{q+b}$. The inequality $\Delta(1) < 0$ is, given $p = 1/(2B)$, equivalent to $b/q < 2B - 1$.

If $\Delta(1) > 0$ with $p = \sqrt{\frac{1}{B}}$, then there exists an equilibrium in which the warden mixes over 1 and 2 with probability $z_{1,2} = \frac{q}{p(b+q)} = \sqrt{B}\frac{q}{q+b}$. Then the inequality $\Delta(1) > 0$, given $p = \sqrt{1/B}$, is $b/q > \sqrt{B} - 1$.

Note that $\sqrt{\frac{1}{B}} > 1/(2B)$ and $2B - 1 > \sqrt{B} - 1$ by $B \geq N + 1 = 3$. This implies the structure in figure 5 for existence of the different equilibria.

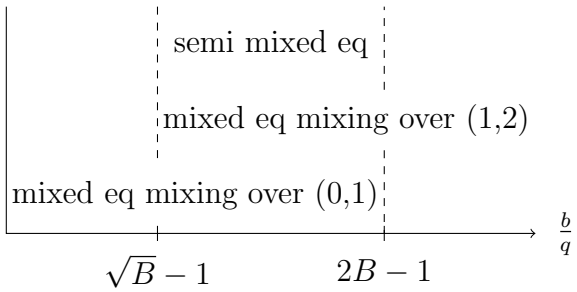


Figure 5: Equilibria for N=2 case

The warden payoff in the 0,1 mixing equilibrium equals $\pi(1) = -p^2B - 1 = -\frac{1}{4B} - 1$. The warden payoff in the 1,2 mixing equilibrium equals $\pi(2) = -2$.

Last, we look at semi-mixed equilibria, i.e. the warden plays a pure strategy while the prisoners play completely mixed strategies. Note that the warden cannot play the pure strategies 0 or 2 in such an equilibrium because the prisoners would then have a dominant action contradicting that they mix. Hence, we can focus on the equilibrium where the warden plays $\gamma = 1$. Playing $\gamma = 1$ is optimal for the warden if $p \in \left[1/(2B), \sqrt{1/B}\right]$. The prisoner is willing to mix only if $\Delta(1) = 0$, i.e. if $b/q = (1-p)/p = 1/p - 1$. Note that $1/p - 1$ equals $2B - 1$ for $p = 1/(2B)$ and $1/p - 1$ equals $\sqrt{B} - 1$ for $p = \sqrt{1/B}$. Consequently, the semi-mixed equilibrium exists if $\frac{b}{q} \in \left[\sqrt{B} - 1, 2B - 1\right]$.

The warden payoff in the panopticon were already established above. In particular, the mixed equilibrium with mixing over zero and one existed if $b/q < 2B - 1$ and the warden payoff in this game was $-1/(4B) - 1$. For $b/q > 2B - 1$, only the mixed equilibrium with mixing over 1 and 2 existed where the warden payoff is -2. In the transparency model, $\theta^* = 1$ if $b/q < 1$ and $\theta^* = 2$ if $b/q > 1$. This implies that the warden payoff is higher in the transparency model than in the panopticon if $b/q < 1$. For $1 < b/q < 2B - 1$, the warden optimal equilibrium of the panopticon gives the warden a higher payoff than the transparency model. The worst equilibrium in the panopticon model gives the warden the same payoff as the transparency model in this case. If $b/q > 2B - 1$, all models give payoff -2 to the warden.

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.[25] 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.

[25]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.