DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

PhD Thesis

Ole Jann

Multiple Equilibria in Markets and Games

Supervisor: Peter Norman Sørensen

Date of submission: August 4, 2016

Contents

Introduction – English

> "All that I have to say has already crossed your mind," said he.
> "Then possibly my answer has crossed yours," I replied.
> "You stand fast?"
> "Absolutely."
> *(Arthur Conan Doyle: The Memoirs of Sherlock Holmes – The Final Problem)*

In economics, most decision problems are interdependent. That can make things quite complicated. If people stay at home because the weather looks gloomy, the weather does not change in response to their decision. Looking at the sky will still give them all the information they need to make a choice. But if people decide to sell a stock in anticipation of a market crash, the stock price responds and hence the very likelihood of a market crash changes. Whether it is best to buy or to sell cannot be learned from just examining the fundamentals, however closely – one would also have to find some way to look into everybody else's mind. And what is everybody else thinking about? Why, they are trying to figure out what everybody else may be thinking – and so on.

How can we fruitfully analyze such situations without ending up in an infinite regress of "he thinks that I think that he thinks that ..."? The most productive and most widely-used thought construct to understand this problem has been the notion of equilibrium, first formalized by Nash (1950) and refined, looked for and found ever since. The idea is simple: If everybody behaves optimally given what he expects the others to do, and has the correct expectation about what the others are doing, then the game is in equilibrium. Just as in the conversation between Sherlock Holmes and Professor Moriarty quoted above: In equilibrium, each player correctly anticipates the action of the other – and still "stands fast" in his own chosen action.

But this concept, as powerful as it is, does not always yield a unique outcome. An example of that is already given by Jean-Jacques Rousseau in his metaphor of the stag hunt:

> If it was a matter of hunting a deer, everyone would well realize that he must remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple. *(Discourse on Inequality, 1754)*

Rousseau did not write about equilibria, but he effectively proposes that there are two equilibria in this game: If all hunters stay "faithful to their post", they can catch the stag and no single hunter would be better off by running after hares instead. If they all hunt hares, however, none of them would be better off by trying to hunt a stag on his own.

Of course, "Discourse on Inequality" is not a hunting manual. Rousseau sees the stag hunt as a metaphor for social cooperation – just as economic models, of which this thesis contains a few, are meant as metaphors for specific economic problems, expressed in the formal language of mathematics. What is so puzzling about the stag hunt is that it appears impossible to say how rational people would behave in it, since several beliefs can be self-fulfilling: If everybody believes that the others participate in the stag hunt, they will, too. But if everybody doubts that the others cooperate, then no one will cooperate and their doubts have been justified. It appears that everything hinges on what everybody believes on how the others will act – and, given that the others are facing the same problem, what they believe about what everybody else believes, and so on.

The first four of the five chapters in this thesis are about such problems of multiple equilibria in the contexts of financial markets, voting in committees, speculative attacks, revolutions, investment decisions and price setting. One particular technique that will appear throughout is the "global game" developed by Carlsson and van Damme (1993), who actually use the stag hunt as their leading example. They propose that if we consider higher-order uncertainty – if, to stay in Rousseau's metaphor, each hunter is sure that there is a stag to be killed but is not entirely sure whether the others know, or whether the others know that others know, and so on – then we can describe thresholds of model parameters above which people will find it uniquely optimal to hunt the stag, and below which they will go for the hare.

In the first essay, I consider mechanisms of information aggregation (like markets or committees) in which individuals have strategic complements – i.e., an action becomes more attractive if more people take it. I show how such complements can naturally arise from short-term constraints in financial markets, or from wanting to be on the winning side in a committee vote. If individuals in such situations lack common knowledge (i.e. if they are unsure about what the other knows), but they do know something about each other's knowledge, I show that as people start worrying about each other's actions, about others' beliefs about actions and so on, it is impossible for people to coordinate on informative behavior. In any equilibrium, people behave according to rumors or ideas, not according to their knowledge. (This chapter is partially based on my master's thesis.)

The second essay, written together with Christoph Schottmüller, is about how to defend against coordinated attacks – such as speculative attacks against currencies, revolutions, or prison riots. Is it better for the defender to show his own strength, or to keep it secret, to deter potential attackers so that no attack occurs? 230 years ago, Jeremy

Bentham proposed the "panopticon" as an ideal solution to this problem. He suggested to construct a prison in which the prisoners are separated from each other and guarded by a central watchtower, while they are unable to see who (if anyone) is in the watchtower. Bentham argued that this setup would make it impossible for prisoners to revolt, and a large philosophical and sociological literature has built on his idea as a metaphor for modern society.

From a game-theoretical perspective, Bentham's argument seems unconvincing at first glance: Why should the fact that prisoners are unable to see the guards make it impossible for them to coordinate? And why should they not, in equilibrium, hold correct beliefs about how many guards (if any) were on duty? We show, however, that Bentham's intuition is remarkably correct if we consider large groups of attackers, which are relatively more predictable than smaller groups by the law of large numbers. We show that this is a fundamental property of defense against large groups – something that Bentham did not argue formally, but probably understood intuitively.

In the third essay, I show how information asymmetries can lead to strategic substitutes in risk-taking: a firm only wants to take risks if other firms take less risks. In the event of an economic crisis, that will guarantee that there are firms left with enough capital to acquire valuable assets, so that they don't need to be sold at fire-sale prices. This leads to subtle ways in which firms try to anti-coordinate and push each other into taking a particular decision with welfare-decreasing actions, and market crashes can simply result from an impossibility to anti-coordinate completely.

In the fourth essay, written with Christoph Schottmüller, we take up one of the most enduring results in microeconomics: the Bertrand paradox, which states that price competition between two firms is equivalent to perfect competition. (It can also be rephrased as an impossibility result: That there are no stable cartels in finitely-repeated interactions.) We ask: Can firms construct any informational mechanism that allows them to circumvent this result so that there exists an equilibrium in which they make positive profits? The answer turns out to be no, even though the argument is somewhat subtle. (This chapter has been published in the *Journal of Mathematical Economics*.)

Information – what people or firms know about each other and about the world around them – has played a central role in these first four chapters. In the fifth essay, also written with Christoph Schottmüller, we consider the role of information in society from a slightly different angle. We develop an informational theory of privacy: What is privacy, why would individuals care about it, (why should economists care about it), and how can privacy increase welfare even though, on the face of it, it introduces an additional information asymmetry?

We argue that privacy allows individuals to express themselves freely without having to worry about being discriminated against because of their choices. This benefits the individuals, and it benefits society because it improves information aggregation.

All of these chapters start with a short abstract; omitted proofs and supplementary material can be found at the end of each chapter. The bibliography for all chapters is at the end of this thesis.

Introduction – Danish

> "All that I have to say has already crossed your mind," said he.
> "Then possibly my answer has crossed yours," I replied.
> "You stand fast?"
> "Absolutely."
> *(Arthur Conan Doyle: The Memoirs of Sherlock Holmes – The Final Problem)*

De fleste beslutningsproblemer i økonomi er indbyrdes afhængige. Det kan gøre tingene ret komplicerede, for at sige det forsigtigt. Hvis folk bliver hjemme fordi vejret ser mørkt ud, ændrer vejret sig ikke på grund af deres beslutning. De kan fortsat bare se op til himlen og få al den information de har brug for at træffe en beslutning. Men hvis folk til gengæld beslutter sig for at sælge en aktie fordi de forventer et markedscrash, reagerer aktieprisen og selve sandsynligheden for et crash forandrer sig. Om det er bedst at købe eller sælge kan man ikke bare lære fra fundamentalinformationen, ligegyldigt hvor nøje man ser på den – man er også nødt til at finde en måde at se ind i folks hoveder på. Og hvad tænker alle andre så? De forsøger selvfølgelig at finde ud af hvad alle andre tænker – og så videre.

Hvordan kan vi alligevel analysere sådanne situationer på en frugtbar måde, uden at vi ender i en uendelig gentagelse af "han tænker at jeg tænker at han tænker ..."? Den mest produktive og mest brugte idé for at forstå adfærd har været ligevægts-ideen, som Nash (1950) først formaliserede og som er blevet forfinet, opsøgt og fundet lige siden. Ideen er simpel: Hvis alle opfører sig individuelt optimalt, givet hvad de forventer at de andre gør, og alle har korrekte forventninger om hvad de andre gør, så er spillet i ligevægt. Ligesom i samtalen mellem Sherlock Holmes og Professor Moriarty ovenfor: I ligevægten forudser hver spiller den andens aktion – og står stadigvæk fast på sin egen aktion.

Men selvom konceptet er meget kraftfuldt, giver det ikke altid et unikt resultat. Et eksempel for dette kan vi allerede finde hos Jean-Jacques Rousseau og hans metafor om hjortejagten:

> Var man i Begreb med at fange en Hjort, saa følte enhver saare vel, at han maatte være aarvaagen paa sin Post, men naar en Hare løb forbi, som han

> kunde naae, saa er der ingen Tvivl om, at han uden Betænkning satte efter den.[1] *(Om Oprindelsen til Uligheden blandt Menneskene, 1754)*

Rousseau skriver ikke om ligevægte, men han foreslår faktisk at der findes to ligevægte her: Hvis alle jægere bliver "årvågen på deres post", kan de fange hjorten og ingen jæger ville få et bedre resultat af at jagte efter harer. Men hvis de allesammen jagter efter harer, vil ingen af dem have det bedre hvis han alene prøvede at fange en hjort.

"Om Oprindelsen til Uligheden" er selvfølgelig ikke en lærebog om at jage. Rousseau ser hjortejagten som en metafor for socialt samarbejde – lige som økonomiske modeller, som der findes nogle af i denne afhandling, er ment som en metafor for specifikke økonomiske problemstillinger, udtrykt i matematikkens formelle sprog. Det der er så gådefuldt i hjortejagten er at det forekommer umuligt at sige hvordan rationelle folk vil opføre sig i denne situation, fordi forskellige forventninger kan være selvopfyldende: Hvis alle tror på at de andre deltager i hjortejagten, skal de også selv gøre det. Men hvis alle tvivler på at de andre samarbejder, vil ingen af dem samarbejde og deres tvivl har været berettiget. Det ser ud som om alt kommer an på hvad alle tror om hvad alle andre vil gøre – og, fordi de andre jo har det samme problem, hvad de tænker om hvad alle andre vil gøre, og så videre.

De første fire af i alt fem kapitler i denne afhandling handler om sådanne problemer med multiple ligevægte i relation til finansielle markeder, afstemninger i komiteer, spekulative angreb, revolutioner, investeringsbeslutninger og prissætning. En særlig teknik der vil blive brugt flere gange er det "globale spil", udviklet af Carlsson og van Damme (1993), som faktisk bruger hjortejagten som deres indledende eksempel. De foreslår at hvis vi tager hensyn til højere-grads-usikkerhed – hvis, for at blive i Rousseaus metafor, hver jæger ved at der findes en hjort, men ikke ved om de andre ved det, eller om de ved at de andre ved, og så videre – så kan vi beskrive tærskelsværdier for modelparametrene over hvilke folk optimalt vil jage hjorten, og under hvilke folk vil jage efter harerne.

I det første essay betragter jeg informationsaggregationsmekanismer (som fx markeder eller komiteer) hvor individer har strategiske komplementer – dvs en aktion bliver mere attraktiv hvis flere andre vælger den. Jeg viser hvordan sådanne komplementer naturligt opstår gennem kortsigtighed i finansielle markeder, eller fra at ville være på den vindende side af en afstemning i en komité. Hvis folk i sådanne situationer ikke har fællesviden (dvs hvis de ikke er sikre på hvad de andre ved), men de ved noget om andre folks viden, kan jeg vise at folk begynder at bekymre sig om andre folks handlinger, andre folks forventninger om andre folks handlinger osv, og det bliver umuligt for dem at koordinere på opførsel der er baseret på deres viden. Der eksisterer kun ligevægte i hvilke folk handler efter rygter eller ideer. (Dette kapitel er delvist baseret på mit speciale.)

Det andet essay, som er skrevet sammen med Christoph Schottmüller, handler om

[1] Citeret efter dansk oversættelse af Salomon Goldin, 1800.

optimalt forsvar mod koordinerede angreb – som for eksempel spekulative angreb mod valuta, revolutioner, eller fængsleoprør. Er det bedre for forsvareren at vise sin egen styrke, eller at holde den hemmelig, for at afskrække potentielle angribere? For 230 år siden foreslog Jeremy Bentham et "panopticon" som en ideal løsning til problemet. Han foreslog at bygge et fængsel hvor fangerne er adskilt fra hinanden og bevogtet fra et centralt vagttårn, mens de ikke kan se hvem (om nogen) der opholder sig i vagttårnet. Bentham argumenterede for at denne opsætning ville gøre det umuligt for fangerne at gøre oprør, og en stor filosofisk og sociologisk litteratur har siden bygget på hans idé som en metafor for det moderne samfund.

Fra et spil-teoretisk perspektiv forekommer Benthams argument ikke rigtigt overbevisende: Hvorfor skulle fangerne ikke kunne koordinere bare fordi de ikke kan se vagterne? Og hvorfor skulle de ikke, i ligevægt, have en korrekt forventning om hvor mange vagter (om nogen) der var på vagt? Vi viser dog at Benthams intuition er bemærkelsesværdigt korrekt hvis vi betragter store grupper, der er mere forudsigelige end mindre grupper på grund af store tals lov. Vi viser at det er en fundamental egenskab ved forsvar mod store grupper – noget som Bentham ikke formelt kunne vise, men som han sandsynligvis forstod intuitivt.

I det tredje essay viser jeg hvordan informationsasymmetrier kan føre til strategiske substitutter i risiko: En virksomhed vil kun påtage sig risiko hvis andre virksomheder påtager sig mindre risiko. Hvis der kommer en økonomisk krise, garanterer dette at der findes andre virksomheder som har kapital nok til at købe aktiver, så de ikke behøver at blive solgt til lave priser. Det fører til at virksomhederne forsøger at anti-koordinere og presse hinanden til at vælge en særlig beslutning, og markedscrashes kan ganske enkelt resultere ved at det er umuligt at anti-koordinere fuldstændigt.

I det fjerde essay, som er skrevet sammen med Christoph Schottmüller, undersøger vi et af de mest kendte resultater i mikroøkonomi: Bertrand paradokset, der siger at priskonkurrence mellem to virksomheder fører til perfekt konkurrence. (Man kan også formulere det som et umulighedsresultat: Der findes ingen stabile karteller i engangsinteraktion.) Vi spørger: Kan virksomheder konstruere en informationsmekanisme som ville gøre det muligt for dem at omgå dette resultat så der findes en ligevægt hvor de tjener positive profitter? Svaret viser sig at være nej, men argumentet er lidt subtilt. (Dette kapitel er blevet udgivet i *Journal of Mathematical Economics*.)

Information – hvad folk eller virksomheder ved om hinanden, og om verden omkring dem – har spillet en central rolle i de første fire kapitler. I det femte essay, også skrevet sammen med Christoph Schottmüller, undersøger vi informationens rolle i samfundet fra en lidt anden vinkel: vi fokuserer på privacy (som kun utilstrækkeligt kan oversættes til "privatliv" på dansk). Vi udvikler en informationel teori om privacy: Hvad er det, hvorfor skulle individer bekymre sig om det, (hvorfor skulle økonomer bekymre sig om det), og kan privacy øge velfærd selvom det introducerer en informationsasymmetri?

Vi argumenterer for at privacy gør det muligt for individer at udtrykke deres præferencer frit og uden at de behøver at være bekymret over at blive diskrimineret på grund af deres valg. Det gavner dem, men det gavner også samfundet fordi folks præferencer bliver bedre aggregeret.

Alle kapitler starter med et kort abstract; de fleste beviser og noget supplerende materiale kan findes i slutningen af hvert kapitel. Litteraturoversigten til alle kapitler findes i slutningen af denne afhandling.

Chapter 1

Is Beauty Contagious? Higher-Order Uncertainty and Information Aggregation[1]

Ole Jann

I investigate the robustness of information aggregation to higher-order uncertainty. Consider an asset market where short-lived speculators have information both about the asset's fundamental value and the amount and direction of noise trading. In equilibrium, each speculator's trading takes account of both pieces of information and the market price adjusts to the fundamental information. But if speculators lack common knowledge about noise trading, they worry about other speculators' beliefs, beliefs about beliefs, and so on. Even with minimal higher-order uncertainty, informative trading is not rationalizable and no informationally efficient equilibrium exists. I discuss how this result relates to historical events and what it says on how markets should be organized to make them informationally efficient. In a second application to expert committees, I show that lack of common knowledge among experts with very precise information makes them unable to communicate their information to a decision maker whose interests are aligned with theirs.

[1]This chapter is partially based on my master's thesis "Speculation and Inefficient Market Equilibria", which I handed in for evaluation two years into my PhD studies, and reuses some text from that thesis. An earlier draft has also been circulated under the title: "When Only the Market Can Vindicate You: Speculation and Inefficient Market Equilibria". I am grateful to Olga Balakina, Amil Dasgupta, Eddie Dekel, Jeppe Druedahl, Nicola Gennaioli, Itay Goldstein, Nenad Kos, Stephen Morris, Michael Møller, Marco Ottaviani, Alessandro Pavan, Sönje Reiche, Jesper Rüdiger, Christoph Schottmüller, Peter Norman Sørensen, Annette Vissing-Jørgensen, Xavier Vives and Asher Wolinsky as well as audiences at Copenhagen University, at the DGPE 2013, EEA 2015 (Mannheim), EDGE 2015 (Marseille) and Oxford University for helpful comments.

1 Introduction

This paper shows that information aggregation in financial markets can be paralyzed by minimal higher-order uncertainty among traders. I demonstrate a novel mechanism by which small doubt about the beliefs of others (and beliefs about beliefs, and so on) leads to a contagion of beliefs which destroys any informationally efficient equilibrium. The result rests mainly on two realistic assumptions: Traders have a short horizon, and they have information about the number and opinion of irrational noise traders.

To understand the mechanism that drives the result, consider the following example: A rational speculator believes that irrationally optimistic traders will drive up the price of a stock. He concludes that he should buy the stock and sell it at a higher price later. A speculator who believes that other speculators believe that optimists will drive up the price will therefore believe that these other speculators will buy the stock and drive up the price, and he will conclude that he should also buy the stock regardless of what he actually believes about the optimists. And so on, to an ever higher degree, until the idea that there are optimists in the market takes on a life of its own without anyone having to believe in it.

If speculators have common knowledge about the actions of noise traders, such contagion of beliefs cannot happen, since everybody knows what everybody knows about the noise traders, and so on. But with the smallest seed of higher-order uncertainty, belief contagion eradicates any connection of price to fundamental value. Note that I am not claiming that the unfounded buying frenzy in the above example is an equilibrium. But under the assumptions mentioned above, this paper argues that a contagion can be unavoidable, and it can destroy all informationally efficient equilibria.

This result offers an explanation for the frequent observation that well-informed speculators seem to trade against their own better knowledge, such as hedge fund managers who bought tech stocks in 1999 or well-connected bankers who did not get out of the market in 1929. A similar contagion can also occur in other institutions for information aggregation, such as expert committees, where it can result in anticipatory obedience to the biases of a decision maker. Understanding the contagion mechanism, and under which conditions and assumptions it emerges, has implications for how financial markets (and other institutions for information aggregation) should be designed to make them informationally efficient.

I develop a model of an asset market in which speculators have information about the fundamental value v of the asset. v is realized in period 3, but speculators only live until period 2 and can either buy in period 1 and sell in period 2, or the other way around. The speculators therefore want to forecast p_2, the price in period 2, which is determined by the actions of other speculators and noise traders. This problem of trying to forecast a price that is determined by the actions of others who are trying to make the same

forecast is what Keynes (1936) famously called the "beauty contest".[2] I naturally extend Keynes' metaphor by considering the contagion of beliefs if speculators lack common knowledge.[3]

Assume that speculators can observe the order flow from noise traders, x_N. If speculators have common knowledge about v and x_N, there exists an equilibrium in which speculators base their trading on both pieces of information (Proposition 1). In this equilibrium, p_2 is a function of both v and x_N. However, p_2's dependence on v is an equilibrium effect (it only happens if speculators trade on their fundamental information), while noise trading is independent of any strategic reasoning. Only x_N predicts p_2 independently of the strategic decisions of other speculators.

If a speculator believes that x_N is very large (i.e. many noise traders are buying), he expects that p_2 will be high, regardless of what other speculators do. (Even if they all sold, the many buy orders from noise traders would still push the price up.) He will therefore condition his trading more on what he knows about noise traders than on what he knows about v. This amplifies the influence that noise traders have on the price, and the expectation of strong noise trading has been self-fulfilling. Even a speculator who knows that the actual order flow from noise traders is small can therefore be swayed to base his behavior on that of noise traders by a self-fulfilling worry that others are doing the same. Even small higher-order uncertainty can "infect" the beliefs of speculators, as they worry that others think that there is lots of noise trading, or they worry that others worry about this, and so on.

I show that this contagion of beliefs among speculators is so powerful that there exists no Nash equilibrium in which speculators base their trading on any fundamental information (Proposition 2). This is true even for the smallest possible amount of higher-order uncertainty. In fact, informative trading is not rationalizable, i.e. any speculator who holds any consistent set of beliefs about the actions of other speculators will conclude that it is never optimal to condition his trading on v. In effect, the contagion of beliefs leads to a "contagion of types," since all informed speculators act like noise traders, and no trades based on fundamental information are made. Prices in periods 1 and 2 are completely independent of v.

Figure 1.1 illustrates the intuition of the first contagion step. The noise order flow x_N from noise traders determines the sign of the price change only in extreme cases, if x_N is very small or very large (shaded areas in the graph). Otherwise (i.e. in the white center

[2]Keynes took the name from a popular newspaper competition where participants had to choose the six most beautiful faces among a hundred photographs, and those who chose the most popular faces could win a prize.

[3]I use the term "common knowledge" in the sense of Aumann (1976), i.e. something is common knowledge if everybody knows it, everybody knows that everybody knows it, and so on. "Higher-order uncertainty" is n-th order uncertainty with an arbitrarily large n, where first-order uncertainty is uncertainty about a variable, second-order uncertainty is uncertainty about someone else's belief about the variable, and so on. Thus higher-order uncertainty implies lack of common knowledge.

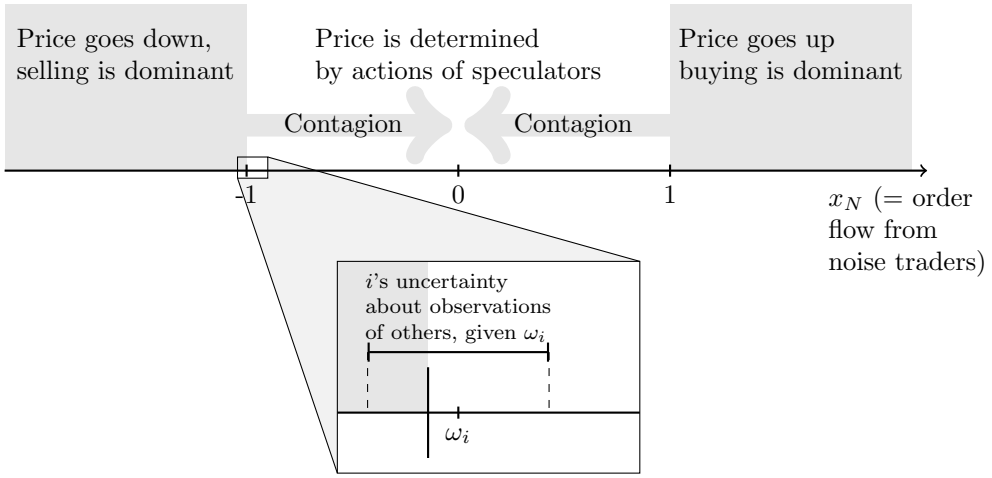


Figure 1.1: Illustration of the intuition of contagion. The prize is in equilibrium determined by the trading of the speculators, and only in extreme cases by the noise traders. With higher-order uncertainty, a speculator who thinks that the noise trader order flow is close to the cutoff believes that some other speculators believe that it is beyond the cutoff (small, magnified rectangle). It is therefore optimal for him to behave as if x_N was beyond the cutoff.

of the graph), the sign of the price change is determined by the trading of speculators, who trade according to their knowledge about fundamental value. But now assume that there is small belief uncertainty, so that speculator i observes signal ω_i, which is a very precise signal of x_N.[4] The small, magnified rectangle shows the problem of a speculator who receives a signal very close to the lower cutoff, below which speculators have no influence on the price change. Because of the small uncertainty, he knows that some other speculators receive a signal that is below the cutoff. These speculators follow their dominant strategy and sell, which effectively turns them into noise traders. This moves the cutoff slightly, since there are fewer speculators "available" that could push the price upwards. It is therefore only rationalizable for i to sell, independently of his knowledge about v.

The main result of this paper is quite stark, and it needs to be qualified to be useful. A general prediction that there is no informative trading in financial markets would quite obviously be at odds with reality. But the results of this paper offer an explanation at what went wrong in times when asset prices were substantially detached from fundamentals, such as just before the Great Crash of 1929 or during the dot-com bubble. In both cases, experts who understood the mispricings refrained from trading on their information.[5] In section 5.2, I discuss several such episodes and relate them to

[4] I model higher-order uncertainty with the standard global games methodology of Carlsson and van Damme (1993).

[5] There was considerable uneasiness in regulatory circles, the Fed, large banks and the media on the eve of the Great Crash (Galbraith, 1954); fund managers were aware that internet stocks were overpriced but felt they couldn't afford to stay out or short-sell (cf. Abreu and Brunnermeier, 2003).

the model.

Even more importantly, understanding under which conditions the result emerges can provide insight into how to design financial markets to make them informationally efficient. The result is robust to the inclusion of some well-informed and long-lived investors. A market that is made-up mostly of long-term oriented traders, however, would not exhibit beauty-contest features and this model does not apply. This suggests that markets in which there is more short-term trading, and times in which people are looking to make profits quicker, are more susceptible to the belief contagion, and more likely to lead to informationally inefficient prices. The information structure that leads to the contagion is also crucial, since contagion does not occur in models with either more information (i.e. common knowledge) or less information (i.e. no knowledge about noise trading). While the information structure in this model is not implausible, it is clearly important whether traders observe (and think about) the current market sentiment when making their decisions. This offers an explanation for the role that rumors play in financial markets: Rumors are not necessarily influential because everybody believes them, but because traders don't have common knowledge about the fact that no one believes them. In section 5.1, I discuss which conditions need to be in place to obtain the result, and which tentative policy implications we can draw.

The contagion mechanism is not exclusive to financial markets; it applies in principle to all beauty-contest type models. In section 5.3, I briefly sketch an application to voting in expert committees. Consider an expert committee that has to give a recommendation between two options. Experts receive a payoff from voting for the winning option, which is larger than the intrinsic payoff from voting for the better option. A contagion of beliefs can make it unrationalizable for the experts to systematically vote for the better option.

The speculators in this paper have two pieces of information on which to base their decision: Fundamental information about the value of the asset, and information about the behavior of noise traders. This has similarities to the studies on private and public information by Morris and Shin (2002) and on beauty contests by Allen, Morris, and Shin (2006), whose central finding is that agents can overweigh information that forecasts the actions of others, and underweigh information about the state of the world. In this model, the "predictive" information is about the behavior of noise traders. Since noise traders sometimes influence the price to an extreme degree (similar to the "noise trader risk" of De Long et al., 1990), either buying or selling can become the dominant speculator action for some realizations of noise trader behavior.

The central result of theoretical models of contagion is that small higher-order uncertainties can be magnified (Rubinstein, 1989) and select between equilibria (Carlsson and van Damme, 1993). These insights have usually been applied to models in which the contagion occurs on beliefs about a fundamental variable. But this is not necessary. As this paper shows, an otherwise insubstantial variable can completely determine behavior.

It is only necessary that people do not find it inconceivable that this variable could take values that would make an action strategically dominant. Whether anyone thinks this is likely is not important. That is how it can become uniquely rationalizable for traders to condition their behavior on the irrational ideas of a small group of noise traders.

Global games analysis is usually taken to describe selection between two outcomes. In Morris and Shin (1998), for example, the result is (broadly speaking): For some fundamental values, a speculative attack occurs, for others it does not. In this model, the important multiple-equilibrium structure is the existence of two intermediate equilibria: "all speculators trade in the right direction" (i.e. buy if value is high and sell if it is low) and "all speculators trade in the wrong direction." It is the ability of speculators to choose one of the equilibria that makes informative trading possible. But through contagion, they are made unable to choose any equilibrium, and informational efficiency is destroyed. Informative trading happens for no realization of noise trading. Thus the result is qualitatively different from global games applications that are outcome-centered and present a selection between the outcomes. The application to voting in committees (section 5.3) presents this insight even more starkly.

Abreu and Brunnermeier (2003) develop a model in which informed arbitrageurs may delay selling for some time despite knowing that an asset is overvalued. Since arbitrageurs are not aware of how many others know of the mispricing, they temporarily cannot coordinate on selling. This argument is based on a similar intuition as the contagion which leads to a durable disconnect of prices from value in this paper.

The following section introduces the model under common knowledge and describes the assumptions in some detail. Section 3 describes the equilibrium without higher-order uncertainty; section 4 derives the main result if speculators lack common knowledge. The discussion (section 5) relates the assumptions and result of the model to policy implications (5.1) and to historical events (5.2). Section 5.3 applies the contagion mechanism to voting in expert committees; section 6 concludes. All proofs are in the appendix.

2 The Model

2.1 General Structure

Consider the market for one asset. The asset has a fundamental value v of either either v_H or v_L, where $v_H > v_L$ and both values are equally probable. v is realized in period 3. There is a group of speculators who know v, but they only live until period 2 and can therefore only trade in periods 1 and 2. In period 1, there are also noise traders who buy or sell randomly and have a net order flow of x_N.

Trading occurs in two periods: In the first period, speculators and noise traders post buy or sell orders, which are executed according to a linear pricing function with

unknown market depth. In period 2, prices are set by the market (which we can think of as being composed of long-term investors, market makers and others). The market does not know v, but can observe p_1 and therefore make inferences about v from observing p_1.

At $t = 1$, speculators are also informed about x_N, i.e. the direction and size of order flow from noise traders. This seeks to capture the phenomenon that speculators not only have some information about the value of the asset, but that they can also observe the current market sentiment – and hence whether their private information is in line with this sentiment or not. The main result about non-robustness to higher-order uncertainty emerges when we remove common knowledge about x_N by introducing small, idiosyncratic observation noise – see section 4.1 below.

There is no discounting and speculators are risk-neutral, and there are no budget or inventory considerations. However, we restrict the amount that any single trader can trade; such a restriction would otherwise arise naturally from assuming risk-aversion or budget restrictions.

2.2 Assumptions in Detail

The Players There is a continuum of speculators and a continuum of noise traders.[6] The speculators, who are perfectly informed about v, are ordered on the unit interval. They get a utility of $p_2 - p_1$ if they buy in period 1 and $p_1 - p_2$ if they sell in period 1, and 0 if they do nothing.

The uninformed noise traders decide randomly (not necessarily without correlation) whether to buy or sell one unit of the asset (because of liquidity needs, or irrational ideas about v). Denote by x_S the net order flow from the speculators, and by x_N the net order flow from noise traders in period 1. The order flow x_S from speculators is the result of their strategic trading decisions, while x_N is simply the result of some random process. Specifically, we assume that x_N is distributed according to a continous distribution F that is symmetric around 0 (so that $F\left(x_N'\right) = 1 - F\left(n - x_N'\right)$) and has density everywhere on an interval $[-n, n]$ where $n \in \mathbb{R}$. We restrict our attention to cases where $n > 1$, i.e. there could potentially be more trades from uninformed than from informed traders.

The Pricing Function in Period 1 Let p_t be the market price of the asset in period t. At the beginning, the asset is trading at price p_0, which is the unconditional expected

[6] The assumption that there are infinitely many traders need not be taken literally - it is simply meant to represent the fact that traders do not take account of the price impact of their trades, or assume that they cannot influence the price. The model works just as well in a discrete setting with finite numbers of speculators and noise traders, see appendix 9.

value:

$$p_0 = \frac{v_H + v_L}{2}. \tag{1.1}$$

Let $x = x_S + x_N$ be the total order flow from speculators and noise traders in period 1. These orders will be cleared by market makers (or a residual market) according to the linear pricing function

$$p_1 = p_0 + \lambda x \tag{1.2}$$

where λ is an unknown reverse market depth parameter. While λ is similar to Kyle's Lambda (Kyle, 1985) in the role it plays in the pricing function, note that it is exogenously given here. We assume that λ is uniformly distributed on the open interval $\left(0, \hat{\lambda}\right)$. To guarantee existence of well-behaved equilibria, we will impose a maximum condition on $\hat{\lambda}$ (i.e. a minimum condition on market depth) below.

The randomness of λ is mainly a technical assumption to make the mapping from order flow x to price p_1 noisy, so that p_1 is not fully revealing about x. If that were the case, the market would be able to observe p_1 and perfectly infer who had been trading (since speculators and noise traders exist in different measure). With a random λ, the size of the price change is still informative about the trading volume and therefore about whether informed or uninformed traders were trading more, but it is not fully revealing. The fact that speculators don't know λ also precludes the existence of spurious equilibria in which the speculators submit information by precisely encoding it into the price.

The randomness of λ can also be understood differently if we recall that the mass of speculators is normalized to one. Without changing the model, we could fix market depth at a constant, and assume that the mass of noise traders is given by λn and that of speculators by λ, which would leave the pricing function unchanged. Now the market depth would be known, but the mass of speculators would be unknown instead.

The Market in Period 2 In period 2, the market is an intelligent player, who has to set a price p_2 at which it is willing to buy or sell any quantity of the asset. I assume that it gets a utility of $-(v - p_2)^2$, so that it will always always maximize utility by setting $p_2 = \mathrm{E}[v \mid p_1]$. We can think of the market in period 2 as a large number of rational long-term investors, market makers and the like, who are in Bertrand-style competition and therefore make zero profit and are willing to buy or sell the asset for its expected value.

Restriction to Trade Size The speculators in period 1 can only buy or sell one unit of the asset each. The main intuition of this assumption is that the market is large compared to any single speculator. In the context of this model it is also a technically desirable assumption, since perfectly informed speculators with no trading or budget restrictions would otherwise have an incentive to trade arbitrarily large quantities and

completely correct the price (as there is no fundamental risk for them). Just like in Glosten and Milgrom (1985), our focus is on the informational content of trades, not on their size.

With the introduction of fundamental risk or agency concerns, a similar size restriction would emerge endogenously. It also does in the real world: Even a trader who is absolutely sure of himself will normally not be allowed to trade very large quantities.[7]

All traders (speculators, noise traders and the market) are free of inventory considerations. Speculators can either buy one unit in period 1 and then sell it in period 2, or sell in period 1 and buy back in period 2, or they can abstain from trading at all. Selling and later buying back can also be thought of as a short sale (which has an inherent short horizon, even if we were to assume that speculators were not short-term interested). The market in period 2 is willing to trade any number of units at a fixed price.

Perfectly informed Speculators The speculators in this model are perfectly informed not only about the true value of the asset, but also about how the noise traders are (overall) trading. We could think of x_N being as the market sentiment, i.e. the current (irrational) movement of prices or the direction of the current mispricing.[8] The speculators learn v and x_N at the beginning of period 1.

The speculators do not know the inverse market depth λ, the other source of noise in the model. The noise in λ mostly serves to reduce the informativeness of p_1 such that p_1 doesn't fully reveal who has been trading in which direction (and thus reveal v). The market only knows the probability distributions of v and x_N, and observes p_1 at the beginning of period 2 before deciding which price to offer.

Timing of the Model The timing is shown in figure 1.2.

While the market behaves rationally in using all information that is contained in p_1, it is conceivable that it could also condition on order flow in period 2 when speculators liquidate their holdings. In particular, it could act similar to the market makers of Glosten and Milgrom (1985) and adjust p_2 conditional on whether it receives buy or sell orders. But the assumption that all speculators liquidate their holdings in period 2 is merely a simplification. In reality, many or even most traders are short-term oriented not because they have to liquidate their holdings every few days or weeks, but because their holdings get evaluated, by themselves or their superiors, *at market prices* in short time intervals. For their motivation and strategic choice, this is equivalent to a world in

[7] Securities trading companies usually institute rules that limit trading by any one trader, similar to the trade size restriction of this model. Several scandals of "rogue trading" in recent years have highlighted the importance of such restrictions by illustrating the damage than can be done if they are circumvented.

[8] Cf. the discussion on insider trading by Leland (1992), who works with a similar assumption, and the "private learning channel" that speculators have in Cespa and Vives (2015).

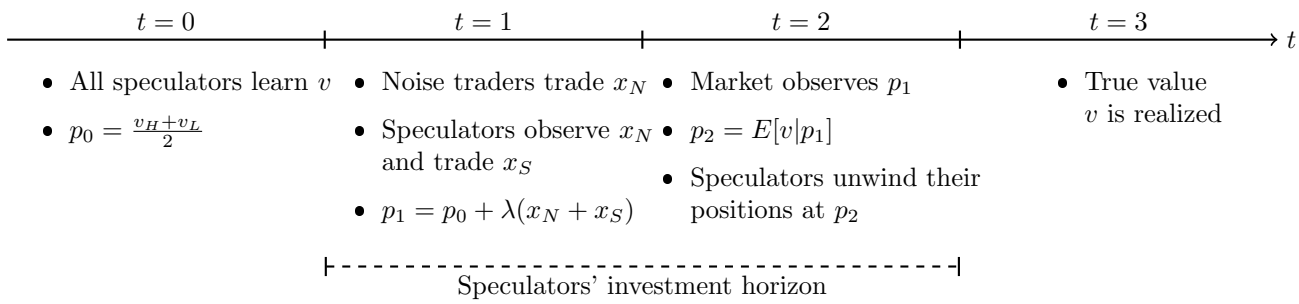


Figure 1.2: The timing of the model. The dashed line shows the speculators' investment horizon, which does not stretch to the realization of fundamental value in period 3 as speculators need to unwind their position in period 2.

which they had to completely sell off and rebuild their portfolio frequently – but it does not per se allow the investors in our model to deduce any information about the order flow from the orders they face in period 2.

3 The Adjustment Equilibrium under Common Knowledge

We start by analyzing the game without higher-order uncertainty, where there exists an equilibrium in which p_1 and p_2 adjust to v on average. In this equilibrium, the market in period 2 assumes that p_1 is informative about v. In particular, it assumes that if $p_1 > p_0$, it is more likely that $v = v_H$ and vice versa. p_2 is set accordingly. If $|x_N| < 1$, the speculators can therefore influence p_2 by their trading, and they buy if v_H and sell otherwise. If $|x_N| \geq 1$, however, whether p_1 is above or below p_0 is determined by the direction of the noise trading, and speculators cannot influence p_1 sufficiently. It is then optimal for them to just trade in the same direction as the noise traders.

The market adjusts its expectation of v according to the function $p_2(p_1)$, which takes the behavior of the speculators and the distribution of x_N into account. If $p_1 > p_0$, for example, they know that overall order flow in the first period was positive, and that therefore either $|x_N| < 1$ and $v = v_H$, or that $|x_N| \geq 1$ and the speculators just followed the herd. The existence of the equilibrium is assured by a maximum condition on inverse market depth, which guarantees that it will always be optimal for the speculators to follow their equilibrium strategy.

Proposition 1. *(Adjustment equilibrium) It is an equilibrium if every speculator follows the strategy given by table 1.1 and the market sets* $p_2 = \pi(p_1)v_H + (1 - \pi(p_1))\, v_L$*, where* $\pi(p_1)$ *is the belief of the market that* $v = v_H$*, given* p_1*. This is under the condition that market depth is sufficient, i.e.*

$$\hat{\lambda} \leq \phi \frac{(v_H - v_L)}{n + 1}. \tag{1.3}$$

	v_L	v_H
$x_N \geq 1$	Buy	Buy
$x_N \in (-1, 1)$	Sell	Buy
$x_N \leq -1$	Sell	Sell

Table 1.1: Equilibrium strategy of the speculators. Only trading for $x_N \in (-1, 1)$ is informative.

The precise expressions of $\pi(p_1)$ and ϕ are given in the proof.

The intuition of the proof is the following: If speculators follow their equilibrium strategies, p_1 will contain some information about v. The function $\pi(p_1)$, which takes account of the distributions of x_N and λ, gives the probability (and hence the equilibrium belief of the market) that $v = v_H$ for every p_1. Since all possible prices occur in equilibrium, we do not need to consider out-of-equilibrium beliefs.

The speculators, on the other hand, will make an expected profit by following their equilibrium strategies, since the price movement in period 1 is always small enough (if market depth is sufficient, which is where the maximum condition on $\hat{\lambda}$ comes from). In particular, it is always either $p_0 < p_1 < p_2$ or $p_0 > p_1 > p_2$. Because a single speculator has only limited influence on p_1, no single speculator has an incentive to deviate. If a speculator would deviate from his equilibrium strategy, he would make a loss equal to the profit of his equilibrium strategy. The maximum condition on $\hat{\lambda}$ in (1.3) guarantees that there is no "overshooting" in expectation, i.e. if all speculators buy in period 1, p_1 still doesn't rise above v.

Figure 1.3 shows an exemplary price path in the adjustment equilibrium, where noise is small (i.e. $|x_N| < 1$). The speculators then face a coordination problem: They can either all buy or all sell, which will place p_1 either above or below p_0. In both cases they make a profit, and both cases constitute an equilibrium of their coordination game. In the Nash equilibrium, however, the market must optimally extract information from p_1, which is only the case if speculators trade towards the fundamental value v.

If $|x_N| \geq 1$ the speculators cannot influence whether p_2 will be above or below p_1, since they cannot neutralize the noise trading and there is no coordination game among them. It is dominant for them to follow the herd of noise traders – regardless of whether it is right or wrong. If the noise traders are wrong, that means that the speculators will drive the market price further away from its correct value even though they know better, and even though the investors would gladly enrich them by buying the asset at a more correct price. Figure 1.4 shows such a price path. The price gets pushed too far away from p_0 (into the grey area), so that the speculators are not able to move it above p_0 again. Once noise trading has pushed the price into the "grey area" of the graphs, p_1 and p_2 will not return to the informative "white area" and are therefore uninformative.

De Long et al. (1990) describe a similar effect when they consider "noise trader risk":

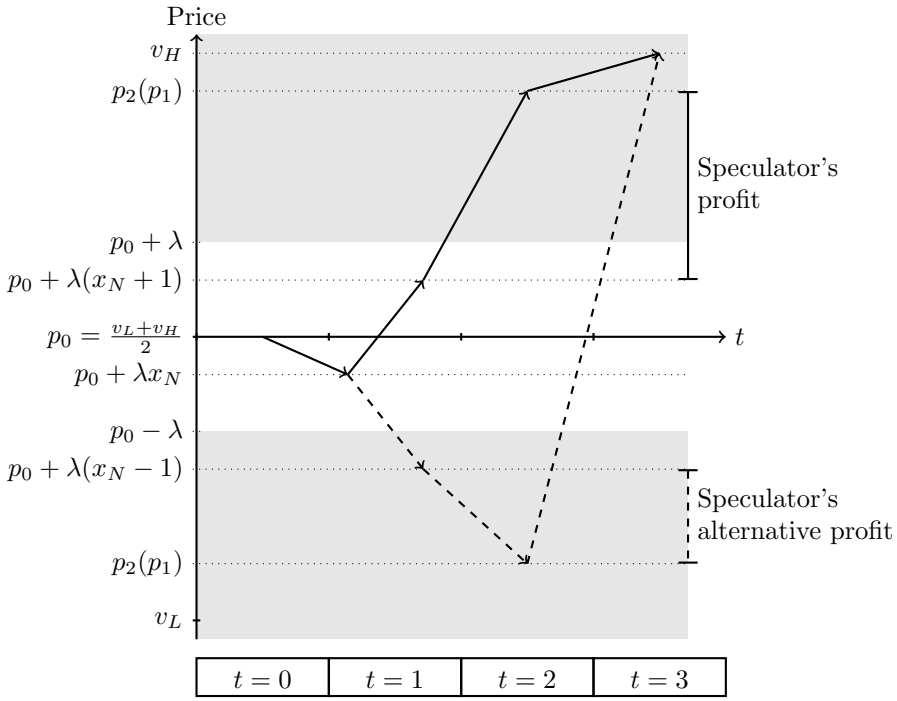


Figure 1.3: The equilibrium price path in the adjustment equilibrium for a given set of parameters, where $v = v_H$ (asset value high) and $0 > x_N > -1$ (noise traders sell the asset). Noise trading is small, i.e. noise traders do not push the price outside the white area in the center. All speculators buy the asset, thus pushing the price to $p_1 = p_0 + \lambda(x_N + 1)$. The market observes $p_1 > 0$ and sets $p_2(p_1) > p_1$, so that speculators make a profit.

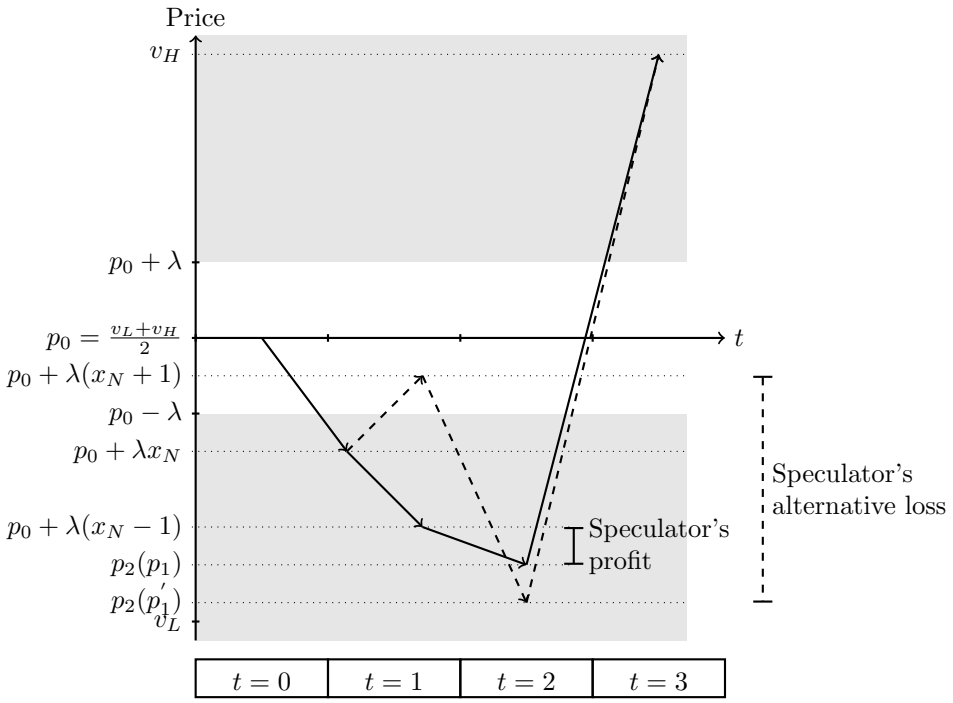


Figure 1.4: Another equilibrium price path in the adjustment equilibrium. Now $x_N < -1$, i.e. noise traders push the price into an area (given by the gray shade) where it is the dominant action for speculators to sell. Unlike in the previous figure, the coordination game between speculators does not have multiple equilibria and thus no information submission is possible.

In their model, rational and informed arbitrageurs in an overlapping-generations model could correct mispricings that arise through noise trading. But since arbitrageurs are short-lived and the market could get even more irrational (noise trade is randomly distributed), they refrain from fully correcting the mispricings. In my model, the direction of the noise order flow (and therefore also the direction of the mispricing in the next period) is known to the speculators, and they can therefore choose to trade against their information and therefore avoid the noise trader risk. They drive prices further away from fundamentals while doing so, as in models of speculative bubbles such as Abreu and Brunnermeier (2003).

4 Higher-Order Uncertainty

4.1 The Adjustment Equilibrium without Common Knowledge

Now consider the problem of the speculators in the adjustment equilibrium described above and take the market's strategy as given. As we have seen above, for $x_N \in (-1, 1)$ the speculators are playing a coordination game with two equilibria: They can either all sell or all buy; in either case they make a positive profit and no speculator would

optimally choose a different action (cf. figure 1.3 on page 24). Only trading in the correct direction (buying for v_H and selling for v_L) can be part of a Nash equilibrium where the market optimally plays its equilibrium strategy.

This coordination among speculators works under the assumption that x_N is common knowledge, i.e. the ratio between informed and noise traders is common knowledge among the informed traders. Given that both v and x_N are information that is not publicly available to everyone (otherwise the market would be fully informed), it is plausible to consider what happens if x_N is known very precisely to the speculators, but not common knowledge. We will see that in this case, no set of values of x_N remains for which the speculators ever trade on their information. Their worries about other speculators' information about x_N (and their worries about other speculators' worries and so on) lead them to completely disregard any fundamental information, and the adjustment equilibrium collapses.

We loosen the common knowledge assumption in a way that is similar to the canonical models of Carlsson and van Damme (1993) and Morris and Shin (1998). Assume that instead of learning x_N, every speculator i observes some ω_i. All ω_i are independently drawn from a uniform distribution on $[x_N - \varepsilon, x_N + \varepsilon]$, i.e. an interval of length 2ε around the true x_N, with $\varepsilon > 0$ but small. Every single speculator will then know after observing ω_i that $x_N \in [\omega_i - \varepsilon, \omega_i + \varepsilon]$. But about the signal of another speculator j he will only know that $\omega_j \in [\omega_i - 2\varepsilon, \omega_i + 2\varepsilon]$, and he only knows that j believes that $x_N \in [\omega_i - 3\varepsilon, \omega_i + 3\varepsilon]$, and so on. Then, even if the observation of every single speculator is extremely precise, it is only common knowledge that $x_N \in [-n, n]$ – which is identical to the prior. We are interested in the case where $\varepsilon \to 0$, i.e. all speculators are arbitrarily well-informed about x_N, but lack common knowledge of it. In this slightly modified game, we can show the following proposition:

Proposition 2. *Assume that the market follows a strategy where $p_2 > p_1$ if $p_1 > p_0$ and $p_2 < p_1$ if $p_1 < p_0$. Then any rationalizable strategy of the speculators' coordination game has the property that all speculators buy if they observe $\omega_i > \varepsilon$ and sell if $\omega_i < -\varepsilon$. For $\varepsilon \to 0$, this gives a uniquely rationalizable equilibrium where all speculators buy if $\omega_i \geq 0$ and sell otherwise.*

Note that this is not a canonical application of the global games refinement, since the speculators' coordination game is not supermodular: Once p_1 is on the right side of p_0, every additional speculator who trades *decreases* the profits of the other speculators.[9] Still, it is possible to show that the above strategy is uniquely rationalizable, as the

[9] As an illustration, consider the case where $v = v_H$ and $x_N = -0.2$. If all other speculators increase their probability of buying from 0.5 to 0.7, buying becomes more attractive. If they increase their probability of buying from 0.8 to 0.9, however, buying becomes *less* attractive. Thus the game is not supermodular.

dominance regions where it is always optimal to buy or sell "infect" the undominated region where there were multiple equilibria in the complete information game.

Intuitively, every single speculator reasons along the following lines:

> I know that x_N is within a small interval around my observation ω_i. If ω_i is in $(-1, 1)$, it is my best guess that all speculators together could overcome the noise so that p_1 correctly reflects our private information. I also know that the other speculators have a very precise idea about x_N—but my knowledge about their knowledge is a little less precise than my own knowledge about x_N. If I consider my knowledge about their knowledge about my knowledge, it gets even less precise.
>
> In particular, if ω_i is very close to 1, I think it is very likely that many other speculators have received a signal above 1 and will therefore play what they believe is the dominant strategy of buying. So I should buy if I observe ω_i very close to but below 1, regardless of what my information about v.
>
> The others will reason the same way, so that if I observe ω_i somewhat less close to 1, I know that many others will observe a ω_j closer to 1, and buy for the reason outlined above. Such contagion carries on, and vice versa from ω_i close to -1. So I will choose to sell if $\omega_i < 0$ and buy if $\omega_i \geq 0$, and disregard my private information about v.

Figure 1.5 depicts the intuition of the contagion argument in a graph similar to the ones above. The proof formalizes this iterative reasoning by defining an elimination process that starts with the set of all possible strategies and then removes, in each step, strategies that are never a best response to any other strategy in the remaining set. In this way, the proof is in the vein of the original work by Carlsson and van Damme (1993) while taking up and modifying some ideas from Frankel et al. (2003).

If the trading of all speculators is only dependent on ω_i and independent of v, p_1 will actually be completely uninformative about v. Hence there is no Nash equilibrium of a game in which x_N is not common knowledge where the market treats p_1 as informative.

4.2 A Non-Adjustment Equilibrium

If informative trading is not rationalizable, which equilibrium can we expect the whole market to be in? It depends on the perspective we take on the role the market in period 2. If we see it as a non-thinking actor who simply follows the decision role laid out in proposition 1, the story ends here: Without common knowledge, speculators never trade informatively, and still the market sets p_2 as if p_1 were informative. All price movements in p_1 and p_2 are pure noise.

Almost the same happens if we treat the lack of common knowledge as an unlikely event, or an event that the market does not expect. Since the market cannot observe

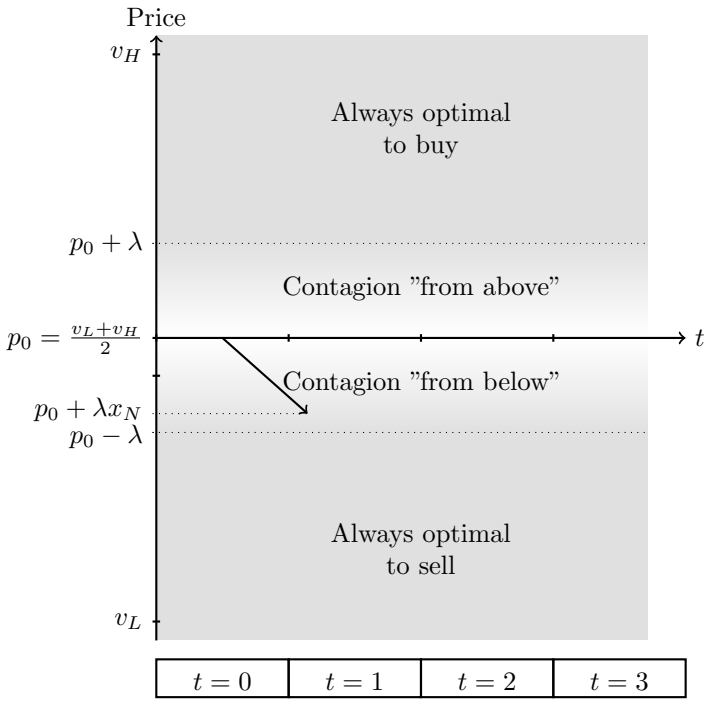


Figure 1.5: Contagion in the absence of common knowledge. Even though $x_N > -1$, the speculators' coordination game no longer has multiple equilibria. A speculator observing $x_N > -1$ but close to -1 is worried that others might believe that $x_N < -1$, or that others might believe that others believe this, etc. This contagion carries on, such that selling becomes optimal for all $x_N < 0$ and buying for all $x_N > 0$.

whether speculators trade informatively or not, it would continue in treating p_1 as informative.

But if we are to take the role of the market as as rational actor seriously, we must assume that in equilibrium it is not "fooled" by p_1 and correctly believes that p_1 is uninformative, in which case it would set $p_2 = p_0$, the prior. Consequently, the trading strategy of the speculators derived in proposition 2 would no longer be optimal. Trading on v, however, does not become optimal. Instead there exists a different equilibrium:

Proposition 3. *(Non-adjustment equilibrium) It is an equilibrium if speculators with probability* $\min\{|\omega_i|, 1\}$ *either buy if* $\omega_i < 0$ *or sell if* $\omega_i \geq 0$ *(and neither buy or sell with the complementary probability), and the market believes that* p_1 *is completely uninformative and therefore sets* $p_2 = p_0$*.*

If the market believes p_1 to be uninformative, the speculators already know that $p_2 = p_0$ and the only gain they can make is by providing liquidity to noise traders. Since this means they do not act on their information about v, the market is correct to believe that p_1 is uninformative.

This equilibrium actually exists independently of whether there is common knowledge or not, as there is no strategic complementarity in the speculators' actions. No player has an incentive to deviate from their equilibrium strategies: The market would not benefit

from assuming that prices contain information, and the speculators cannot gain from unilaterally (or as a group) submitting information (and thereby driving p_1 away from p_0). In this interplay of "not talking" and "not listening", the equilibrium is similar to the "babbling equilibrium" of cheap-talk games (Farrell and Rabin, 1996). There, the sender randomizes between messages such that her message has no correlation to her private information, and the receiver ignores any messages by the sender. This constitutes an equilibrium, albeit (when it comes to everyday communication) perhaps not a plausible one.

The non-adjustment equilibrium is also similar to uninformative equilibrium of Benhabib and Wang (2015), and it is an extreme case of the less informative equilibrium of Cespa and Vives (2015). In both cases, the uninformativeness of the equilibria also emerges through short-term constraints in the models.

4.3 Without Common Knowledge, the Market Cannot Be Informationally Efficient

The rationalization result derived in proposition 2 is clear and general: Consider any equilibrium of the complete-information game in which the market treats p_1 as informative by setting $p_2 > p_1$ if $p_1 > p_0$ and vice versa. Clearly, the adjustment equilibrium and any small perturbation of it belong in this class. Now relax the common knowledge assumption about x_N by introducing the smallest seed of doubt about whether the other speculators are making the same observation as you. For ε very small, which is the case we are interested in, the speculators are still generically 100% sure that x_N is small, they are almost equally sure that everybody else knows that x_N is small, and so on ... but not ad infinitum. They no longer have common knowledge about this fact, and this small seed of doubt means that the strategies that would constitute the adjustment equilibrium are no longer rationalizable – which means that they cannot be part of a Nash equilibrium. Without common knowledge, therefore, no equilibrium can exist in which the market correctly beliefs that p_1 is informative.

Recall also that this result requires no assumptions on the shape of F, the distribution of x_N, except that it is continuous, symmetric and has density everywhere. In particular, this means that F could be shaped such that an arbitrarily large mass of F is inside $[-1, 1]$. Then the noise was almost always small and trading in the adjustment equilibrium would be almost fully revealing. Even in this model, contagion would occur and the adjustment equilibrium would not exist. The frequency of large $|x_N|$ is therefore not important for the relevance of the model – the pure possibility that x_N is outside $[-1, 1]$ is sufficient.

5 Discussion

5.1 When Does Contagion Occur, and What Can We Learn from It?

The main result of this paper is quite stark, as it shows that informative trading only happens if speculators have common knowledge of all model variables. Common knowledge is an exacting requirement that is often unlikely to be met in reality, especially given that the very idea of information aggregation is that the information is not known to everyone. I would therefore like to point out which assumptions of the model are crucial for obtaining the contagion result. That allows us to make predictions about which real-world conditions promote or preclude informational ineffiency through contagion.

In general, it can be said that:

- Contagion does not occur if speculators do not have *short-horizons*, but live until period 3. However, if we endow only a few speculators with long horizons, there is no qualitative change – the remaining speculators are still subject to the same contagion.
- While contagion occurs regardless of any specific assumptions about the likelihood of certain actions by *noise traders*, contagion does not occur if speculators do not consider a large enough set of trading behavior ex ante conceivable.
- If speculators have no knowledge about x_N, contagion does not happen (since there is no information about the beliefs of others that would be reason for worry). Introducing higher-order uncertainty of other variables than x_N does not appear to change anything.

Short horizons A central assumption of the model is that the information about the value v is known only to short-term speculators. This is not to suggest that *all* information arrival at financial markets works in this way, but just that the theory of contagion only applies to situations where this is the case. In general, however, it does not seem a wholly unreasonable assumption that speculators could be better informed than some long-term investors. Just consider that most professional money managers would count as "speculators" in the context of this model if we consider sufficiently long time periods—a few weeks, say, or a quarter. Few of them are allowed and capable of raking up massive losses over such a time frame even if they claim to have superior knowledge that will in the end be vindicated.

Empirical evidence suggests likewise that a large proportion of stock positions are opened for a very limited amount of time, with the expectation of making a profit in less time that it takes to see two quarterly earnings reports. The average holding period

of stocks in the United States is three to four months—not even enough to receive a full dividend payment, let alone profit from long-term business or macroeconomic developments.[10] And even where assets are not bought and sold within days or seconds, those who decide about trading them have their performance evaluated at market prices at very short intervals. If a trader buys an asset at time t for the price p_t, it does not matter to him whether he sells the asset at $t+1$ and it contributes $p_{t+1} - p_t$ to his cash holdings, or whether he still holds it at $t+1$ and it contributes $p_{t+1} - p_t$ to the overall appreciation of his holdings since t.

If all speculators lived until period 3, they would always trade on their information and no contagion would occur. But if we start out with the model in this paper and add a number of long-lived informed investors, the result is robust – up to a point. Consider, for example, a modified model in which there is a measure $\mu < 1$ of informed investors, who always buy if $v = v_H$ and sell otherwise. This would be akin to shifting the distribution of x_N by μ, so that noise trading is given by $\hat{x_N} = x_N \pm \mu$, depending on v. As long as the distribution of $\hat{x_N}$ has density both below -1 and above 1 so that it reaches into the dominance regions, contagion occurs.

In general, the contagion result is remarkably robust to small changes in the payoffs of the speculators. This matters, for example, if we assume that speculators get a small additional payoff from trading in the "right" direction, because there was an exogenous chance that they could live longer. To see why this is the case, note that the payoff structure of a speculator looks like this (+ denotes positive profits, − negative profits):

Speculator's action at $t=1$:		Result: $p_2 > p_1$	Result: $p_2 < p_1$
	Buy	+	-
	Sell	-	+

A speculator that decides whether to buy or to sell will only ever compare two values in the same colum, since there is no uncertainty in the rationalization argument as to which way the price will move. (Assuming that a speculator lives until period three with a certain probability would mean that he plays the game given by the matrix above with a certain probability, and another game otherwise.) The chain of rationalizability arguments that led to the contagion result therefore only relies on the fact that the values in the main diagonal are larger than the other values in the same column. As long as the intrinsic payoff of trading on v, and the probability of being long-lived, are small enough, the contagion result obtains.

[10] The "World Bank Financial Development Indicators" show stock market turnover ratios, which is the inverse of average holding period. In the United States in 2008, for example, trade volume was 4.35 times as high as total market capitalization. Since this is the mean holding period and the distribution is truncated at 0, the median holding period is probably much lower.

Noise trading Some authors (e.g. Dow and Gorton, 1994, p. 825) argue that the presence of noise traders has to be explained. But the absence of noise traders would mean that all traders, at all times, act rationally to maximize their expected payoff from trading. There are two main types of traders for whom that does not apply. Firstly, substantial research on behavioral finance has shown that traders, institutional or private, fall prey to a large number of irrational biases. Secondly, even a rational trader might find it optimal to sell an asset (whose price he expects to rise) for liquidity reasons – for example when he needs to access his savings to retire or pay unforeseen expenses.

Once we accept the assumption that there are noise traders in the market, the question naturally arises whether additional assumptions about the actions of noise traders are necessary. Note, however, that the only two assumptions about the distribution F of x_N that are used in the proof of proposition 2 are (a) that the probability density function of F is continuous and (b) that F has density everywhere on $[-n, n]$. It is therefore only required that speculators consider any order flow from noise traders conceivable – they don't have to think it likely. In fact, if we assume that x_N was normally distributed around 0, we could make the standard deviation of this distribution arbitrarily small without in any way containing the contagion. The distribution of the order flow from noise traders could be so concentrated that speculators were almost sure that x_N was in $(-1, 1)$. In that case, trading in the common-knowledge equilibrium (proposition 1) would almost always be informative and $\mathbb{E}[(v - p_2)^2]$ would get arbitrarily small in this equilibrium. Yet as soon as we introduce the smallest higher-order uncertainty about x_N, contagion carries through all the way and informative trading is not rationalizable.

The information available to speculators When I have considered higher-order uncertainty in this paper, I have limited this uncertainty to the realization of x_N and continued to assume common knowledge of v. This begs the question of what would happen if there was also higher-order uncertainty of v, so that every speculator would worry also about other speculators' belief about v. Could there be a similar contagion of beliefs that might even restore dependence of the speculators' actions on v?

The answer is no, at least in a setup like in this paper where there are no possible values of v for which any action by the speculators would be dominant. There is simply no possible belief about v to "start" a contagion of beliefs. In the case of uncertainty about x_N, this is the belief that another speculator might think that another speculator might think etc. that x_N is so large or so small that buying or selling is the dominant action. Without such "dominance regions", there can be no contagion. In the teminology of Weinstein and Yildiz (2007), the "richness assumption" fails on v, since the parameter space of v is not rich enough to contain dominance regions.

It is possible to think of situations where there are conceivable fundamental values that make buying or selling dominant. If, for example, speculators know that they live

until period 3 with some probability, and v is extremely large with some probability, they might find it dominant to buy the asset. This reinforces the point (made above) that a sufficient long-term orientation of the speculators can break the chain of contagion.

Finally, a crucial requirement of contagion is that speculators actually have an observation of x_N, since it is the worry about other speculators' beliefs of x_N that keeps the contagion alive. If speculators are completely unaware of x_N and only observe v, their only consideration is whether x_N is outside the dominance regions with enough probability to make informative trading profitable. We therefore have the seemingly curious result that contagion fails both if speculators have less and more information (i.e. no information or common knowledge of x_N). If speculators fall prey to contagion, the market would function much better if they did not have access to information about the market sentiment. The sort of coverage that is most beloved by newspapers and tv stations the world over – "Panic at NYSE! Euphoria as Asian Markets Open!" – can thus have a hugely detrimental effect by giving informed speculators information about the noise in the market without generating common knowledge about it. Common knowledge would only be generated if all speculators followed the same news sources, had common knowledge about this fact, and also had common knowledge about the fact that they all understand the news in the same way – a tall order. Ultimately, the contagion argument rationalizes a folk argument among economists: The hype and sensationalist coverage surrounding financial markets can magnify the "psychological moods" of the market and eradicate cool-headed, rational trading – and everybody would be better off without it.

5.2 Examples of the Mechanism at Work

As an example of the paralysis of informative trading described in this paper, consider the so-called Dot-com bubble in the late 1990s and early 2000s. In the context of this model, we could think of internet stocks as being worth either v_L ("most of these companies will never make a profit") or v_H ("they will change the economy forever and be hugely profitable"). Many market participants did not know which was the case, but because v_H was extremely large their unconditional prior $\frac{v_L+v_H}{2}$ was also large. The uncertainty was large enough to make it plausible that it would only be resolved quite far into the future (what if internet companies needed to grow for a decade before turning huge profits?), far beyond the investment horizon of most investors.

Many sophisticated fund managers knew that internet stocks were overvalued, i.e. that $v = v_L$.[11] But to coordinate on an informative sell-off of internet stocks, they would need common knowledge about the fact that there were enough informed traders. As we have seen, it does not matter how large the number of informed speculators is in relation

[11] See for example the discussion in Abreu and Brunnermeier (2003, p. 175). Brunnermeier and Nagel (2004) document that hedge funds were heavily invested in tech stocks, and argue that this was not because they believed prices to be reasonable.

to the number of noise traders. Without common knowledge, the sheer possibility that there could be many noise traders infects everyone's beliefs, despite the fact that all speculators know this not to be the case. So even a well-informed and sophisticated fund manager who knew that stocks were overvalued, and who knew that there were enough others to support a sufficiently large sell-off, feared that others would not sell because they feared that still others would not sell, and he would therefore not sell himself.

A similar pattern emerges when we consider what is perhaps the most notorious market movement in history, the "Great Crash" of 1929. The crash was by no means unexpected, as many experts had come to realize throughout 1929 that stock prices were unsustainably high. Galbraith (1954, ch. 2) describes the uneasiness in regulatory circles and the various attempts to deflate the bubble, and also documents prescient warnings by well-known bankers, financial services and the *New York Times*. But without common knowledge about the fact that informed traders could outnumber noise traders, there was no informative sell-off.

1929 also offers a glimpse into how an equilibrium shift from the non-adjustment to the adjustment equilibrium can occur when common knowledge is generated. On October 24 ("black thursday"), prices fell suddenly and violently by nearly 13%. They swiftly recovered (the closing was only 2.1% down that day), but the event had shown that there were many traders in the market willing to sell. What was even more important was that, since everybody could reasonably assume that everybody else would follow the market closely enough to notice such an event, the preponderance of informed traders was now also common knowledge. In the following days, despite no substantial economic news (cf. Shiller, 2000, p. 94), informed market participants could now coordinate on selling, and the Dow fell over 23% in two days.

As an example of (unprofitable) out-of-equilibrium behavior, consider the spread between Royal Dutch and Shell stocks in the late 1990s. The stocks were trading at different exchanges, but prices should have been at a fixed proportion, because cash flows were paid in a fixed proportion. Instead, there was a spread that was quite stable around 8% (cf. Froot and Dabora, 1999). It appears that the market was in a non-adjustment equilibrium where the many traders who were aware of the unreasonable spread could not coordinate on trading to narrow it, since they didn't have common knowledge about their on combined strength in the market. When the hedge fund Long-Term Capital Management (LTCM) began to trade against the spread in 1997, there was sufficient trading in the opposite direction to maintain the spread – as we would expect from the model, as informed speculators had settled on trading against changes in the spread instead of betting on it to close.[12]

[12]The managers at LTCM, however, were "mystified" – cf. Lowenstein (2000, p. 148).

	A is chosen	B is chosen
Vote for A	1000	0
Vote for B	0	1

Table 1.2: Payoffs of committee members, assuming that A is the better option. (If B is the better option, the payoffs 1 and 1000 change places.)

5.3 An Application to Voting in Committees

The main theoretical insight of this paper can be applied to other settings besides financial markets. This section sketches an application to voting in committees.

A decision maker has to decide between two options, A and B. One of them is better than the other; the decision maker gets a payoff of 1 for choosing the better option and 0 otherwise. The decision maker does not know which option is better, but he gets help from a committee of experts, who all know which option is better. As an example, consider a department head who has to choose between two applicants for an academic position, or an authority that has to decide whether to approve a new drug.

Committee members (experts) are on the unit interval. They strongly prefer that the better option gets chosen, but they also get a small payoff if they back the worse option and it gets chosen. They get nothing if they vote for a losing option. Their payoffs, assuming that A is the better option, are shown in table 1.2.

The timing is as follows:

1. All experts observe whether A or B is better.
2. All experts decide simultaneously whether to vote for A or B.
3. The decision maker observes the proportion a of experts that voted for A and makes a choice.

There is a simple and robust Nash equilibrium where all experts vote for the better option, and the decision maker implements the option that gets the majority of the votes. Now assume that there are some outside forces that influence the vote count or its transmission, so that the decision maker observes $\hat{a} = a + \theta$ instead of a, where $\theta \sim \mathrm{N}(0, \sigma)$. Let σ be small. This could also be thought of as influencing the decision maker himself, for example a bias in his perception or another source of information that he has besides the expert committee. One could think of a pharmaceutical company lobbying for or against the approval of a new medication, or an unknown bias on the side of the department head for one candidate or another. For σ sufficiently small, the equilibrium strategies remain the same, and the worse option is only chosen in very few cases (only if $\theta \notin (-0.5, 0.5)$).

Now assume that all experts observe θ at the same time that they observe which option is better. The equilibrium of this game is almost the same: Experts vote for

the better option if $\theta \in (-0.5, 0.5)$, vote for A if $\theta \geq 0.5$ and for B if $\theta \leq -0.5$. The decision maker chooses A if $\hat{a} \geq 0.5$ and B otherwise, and the better option is chosen if $\theta \in (-0.5, 0.5)$, which is usually the case since σ is small by assumption.

Now, however, assume that instead of observing θ perfectly, each expert i observes signal ω_i which is i.i.d. uniformly distributed on $[\theta - \epsilon, \theta + \epsilon]$, and we are interested in the case where $\epsilon \to 0$. Experts are still very precisely informed, but lack common knowledge about θ. They still have common knowledge about which option is better.

We can now use the same technique as in the proof of proposition 2 to show that the dominance regions infect the multiple-equilibria region. Taking the decision maker's strategy as given, there is no rationalizable strategy for any expert that conditions voting on which option is best. The informationally efficient equilibrium gets destroyed completely, and instead we get an equilibrium in which each expert votes A if $\omega_i > 0$ and B otherwise, and the decision maker follows their recommendation. The options are chosen randomly depending on the realization of θ, and the better option gets chosen only half of the time. This is despite the fact that it is common knowledge among the experts which option is better, the "transmission noise" $|\theta|$ is small with very high probability, and experts as well as the decision maker prefer the better option. It is enough that the experts consider it remotely conceivable that the department head would ignore their recommendation (i.e. that the distribution of θ has density outside of $(-0.5, 0.5)$) to make them use their very precise observation of her bias to always choose the candidate to which she is leaning. This anticipatory obedience even occurs if her bias is almost always negligibly small.

The implications of this model are similar to recommendations for financial markets above. If experts derive sufficient intrinsic motivation from voting for the correct option, the contagion collapses. No contagion occurs if we simply place 2 in the upper right, and -2 in the lower left field of table 1.2. Furthermore, experts should, if possible, be kept in ignorance of the biases that influence the decision maker: As we have seen above, the informationally efficient equilibrium continues to exists if we introduce θ but keep it hidden from the experts.

6 Conclusion

Perhaps the main reason for the triumph of market-based economic systems is that no other mechanism can transmit information about scarcity, efficiency and ability as realiably, fast and cheaply as the price mechanism (cf. Hayek, 1945). We live in a system of financial capitalism because financial markets are the ultimate way of transmitting information: Financial assets are standardized and fungible, all information other than prices is stripped away, information flow is immediate and transaction costs minimal.

But the well-functioning of financial markets requires that they actually incorporate the information that is held by market participants.

This paper describes a mechanism that can destroy informational efficiency if traders only care about the short run and have knowledge about the irrational moods and passions of the market. Both assumptions are compatible with empirical observations. The effect of the latter assumption also supports the conclusion that the spread of rumors and ideas can be highly destructive even in a market that is mainly populated by rational traders. Rumors need a medium to spread, and accordingly Shiller (2000) hast pointed out that "the history of speculative bubbles begins roughly with the advent of newspapers."

The concrete uses of the model lie in providing advice on how to prevent belief contagion in financial markets (section 5.1) and explaining observed behavior (section 5.2). But the theoretical contribution goes beyond. As we have seen in section 5.3, contagion can destroy information aggregation in other settings if actions are strategic complements. Ultimately, the role of contagion in magnifying noise trading and detaching market prices from fundamentals is only one application, if perhaps the most important, of the general theoretical insight. Contagion only requires that actions are strategic complements, and that people find it conceivable that the world would be in a state where each of them had a uniquely optimal action. Then, with even minimal higher-order uncertainty, contagion guarantees that for any state of the world, there is a uniquely rationalizable action. And crucially, as this paper argues, the signal that they condition their actions on need not be fundamental. It can be irrational ideas about the prospects of dot-com companies or the biases of a decision maker, but it might just as well be any other idea that is not ruled out by prior beliefs. A general pattern emerges by which higher-order uncertainty can detach outcomes from the fundamental variables that actually matter. Instead, behavior is determined by the spurious realizations of meaningless signals, purely out of the self-fulfilling belief that others are following these signals. There are connections to the theories of groupthink (Janis, 1972) and preference falsification (Kuran, 1997), which suggest other applications to political behavior, decision making in groups and the collection of knowledge in organizations.

7 Appendix: Proofs

Proof of Proposition 1. **Part 1: The market has no incentive to deviate (and $\pi(p_1)$ is obtained by Bayes' rule).**

Assume that the speculators follow their equilibrium strategies and consider the case where $p_1 > p_0$. The market can then, from observing p_1, draw conclusions about v. Let $\pi(p_1)$ be the conditional probability that $v = v_H$ after observing a certain p_1, $\Pr(v_H|p_1)$.

It is

$$\begin{aligned}
\pi(p_1) &= \Pr(v_H|p_1) = \frac{\Pr(p_1 \cap v_H)}{\Pr(p_1)} \\
&= \frac{\Pr(p_1 \cap v_H \cap |x_N| < 1) + \Pr(p_1 \cap v_H \cap x_N \geq 1)}{\Pr(p_1 \cap |x_N| < 1) + \Pr(p_1 \cap x_N \geq 1)} \\
&= \frac{\int_{-1}^{n} g\left(\frac{p_1 - p_0}{x_N + 1}\right) dF(x_N)}{\int_{-1}^{n} (1 + \mathbf{1}_{x_N > 1})\, g\left(\frac{p_1 - p_0}{x_N + 1}\right) dF(x_N)}
\end{aligned}$$

where g is the density of λ. Since $g\left(\frac{p_1 - p_0}{x_N + 1}\right) = \frac{1}{\hat{\lambda}}$ if if $0 < \frac{p_1 - p_0}{x_N + 1} < \hat{\lambda}$ and 0 otherwise, we can rewrite this as

$$\begin{aligned}
\pi(p_1) &= \frac{\int_{\frac{p_1 - p_0}{\hat{\lambda}} - 1}^{n} dF(x_N)}{\int_{\frac{p_1 - p_0}{\hat{\lambda}} - 1}^{n} (1 + \mathbf{1}_{x_N > 1})\, dF(x_N)} \\
&= \frac{1 - F\left(\frac{p_1 - p_0}{\hat{\lambda}} - 1\right)}{2 - F\left(\frac{p_1 - p_0}{\hat{\lambda}} - 1\right) - F\left(\max\left\{\frac{p_1 - p_0}{\hat{\lambda}} - 1,\, 1\right\}\right)}.
\end{aligned}$$

$\Pr(v_L|p_1)$ is the complementary probability $1 - \pi(p_1)$, so that the expected value of v given p_1 is $E[v|p_1] = \pi(p_1)v_H + (1 - \pi(p_1))\, v_L$. A similar argument applies to the case where $p_1 < p_0$. If $p_1 = p_0$, the price contains no information and p_2 should be set equal to the prior.

p_1 is between $p_0 - \hat{\lambda}(n+1)$ and $p_0 + \hat{\lambda}(n+1)$. For $x_N \in \{-n, n\}$, all possible p_1 occur with positive density, so that in equilibrium all possible p_1 occur with positive probability and there can be no out-of-equilibrium beliefs.

Part 2: Speculators make a positive profit in equilibrium.

Now assume that the market follows its equilibrium strategy. Consider the case where $p_1 > p_0$, meaning that either $|x_N| < 1$ and $v = v_H$ or simply $x_N \geq 1$.[13] If they follow their equilibrium strategies, the speculators' buy orders will drive the price to

[13] An analogous argument applies where $p_1 < p_0$.

$p_0 + \lambda(x_N + 1) > p_0$, and in period 2 all speculators will be able to sell their holdings at $\pi(p_1)v_H + (1 - \pi(p_1))\, v_L$. Their profit is then $\pi(p_1)v_H + (1 - \pi(p_1))\, v_L - p_0 - \lambda(x_N + 1)$, which can also be written as

$$\frac{F\left(\max\left\{\frac{p_1-p_0}{\hat{\lambda}} - 1,\, 1\right\}\right) - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)}{4 - 2F\left(\max\left\{\frac{p_1-p_0}{\hat{\lambda}} - 1,\, 1\right\}\right) - 2F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)} (v_H - v_L) - \lambda\,(x_N + 1)\,.$$

Since every speculator knows x_N, the expected profit is

$$\mathbb{E}\left[\left.\frac{F\left(\max\left\{\frac{p_1-p_0}{\hat{\lambda}} - 1,\, 1\right\}\right) - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)}{4 - 2F\left(\max\left\{\frac{p_1-p_0}{\hat{\lambda}} - 1,\, 1\right\}\right) - 2F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)}\right| x_N\right] (v_H - v_L) - \frac{\hat{\lambda}}{2}\,(x_N + 1)\,.$$

Since p_1 is increasing in x_N, $F\left(\max\left\{\frac{p_1-p_0}{\hat{\lambda}} - 1,\, 1\right\}\right)$ and $F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)$ are also (weakly) increasing in x_N. The expression therefore becomes minimal for $x_N = n$. If at this minimal point it is still non-negative, speculators make a positive expected profit in equilibrium; this is the case if

$$\begin{aligned}\hat{\lambda} &\leq \mathbb{E}\left[\left.\frac{F\left(\max\left\{\frac{p_1-p_0}{\hat{\lambda}} - 1,\, 1\right\}\right) - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)}{4 - 2F\left(\max\left\{\frac{p_1-p_0}{\hat{\lambda}} - 1,\, 1\right\}\right) - 2F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)}\right| x_N = n\right] \frac{(v_H - v_L)}{n+1}. \\ &\leq \phi\frac{(v_H - v_L)}{n+1}\end{aligned}$$

This gives a minimum condition for market depth, which is simply given by the spread between high and low value, adjusted for the number of market participants and some adjustment factor ϕ that depends on the precise shape of F. If this minimum condition is fulfilled, speculators make a non-negative expected profit in equilibrium. Note that $\phi \in (0, 1/2)$ since the expression in the expectation is at least 0 (if $\frac{p_1-p_0}{\hat{\lambda}} > 2$) and at most $1/2$ (if $\frac{p_1-p_0}{\hat{\lambda}} \leq 2$), and both cases occur.

Part 3: No single speculator has an incentive to deviate from his equilibrium strategy.

Part 2 shows that every speculator has, after having observed v and x_N, a non-negative expected profit from following his equilibrium strategy. If his equilibrium action is to buy, then $p_1 - p_0 \geq 0$, and $p_0 - p_1 \geq 0$ if his equilibrium action is to sell. If he were to do nothing instead, his profit would be 0, which is not better. If he were to do the opposite, his profit would be non-positive, which is also not an improvement. All speculators hence optimally follow their equilibrium strategies. □

Proof of Proposition 2. A strategy is a function $s(\omega)$, where $s : [-n - \epsilon,\, n + \epsilon] \to [0, 1]$

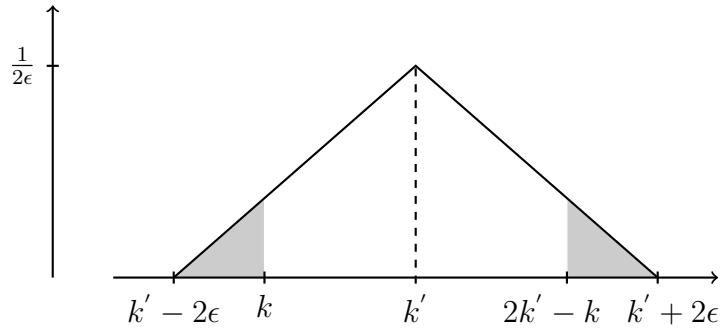


Figure 1.6: Illustration of the proof for the lemma. For small ϵ, the distribution of the signals of the other speculators is a symmetric triangular distribution around the own signal. Given that all speculators that receive a signal lower than k sell, a mass of speculators that is given by the shaded area on the left will always sell. Their sell orders will at least cancel out the buy orders by a mass of speculators given by the shaded area on the left, so that the maximum number of buy orders is given by the white area between the two shaded areas. If it should be undominated to buy with positive probability after receiving signal k', the white area would have to be larger than $-k' - \epsilon$. If $k' - k$ is below the upper bound given by the definition of $B(k, \epsilon)$, that is not possible.

gives the probability of buying, given any observation ω. Let Σ be the set of all strategies. Define the iterative-dominance function $\rho : \mathcal{P}(\Sigma) \to \mathcal{P}(\Sigma)$ where $\mathcal{P}$ is the power set. Given a set of strategies Σ', ρ returns a set of strategies $\rho(\Sigma')$ that is identical to the first one except that all strategies in Σ' that are never a best-reponse to any strategy in Σ' have been removed. Let $\rho^2(\Sigma) = \rho(\rho(\Sigma))$ and so on; a strategy s is rationalizable if $\forall n \in \mathbb{N} : s \in \rho^n(\Sigma)$.

What does $\rho(\Sigma)$ look like, where Σ is the set of all strategies? Clearly, no strategy that puts probability higher than 0 on buying for any $\omega_i \in [-n - \epsilon, -1 - \epsilon]$ is in $\rho(\Sigma)$, since otherwise the speculator would be buying with positive probability even though he knows for sure that $p_1 > p_2$.

Let $B(k, \epsilon) = \left\{ k' \in \mathbb{R} | \left| k' - k \right| < 2\epsilon \left(1 + \epsilon - \sqrt{(1+\epsilon)^2 + k + \epsilon} \right) \right\}$ be an open ball around k with a size that depends on k and ϵ. Note that the size of $B(k, \epsilon)$ is always below 4ϵ if $k \geq -1 - \epsilon$. The following lemma establishes that we can use this ball $B(k, \epsilon)$ to exclude elements from $\rho(\Sigma')$ if no strategy that buys for k is in Σ'. The proof is illustrated in figure 1.6.

Lemma 1. *If Σ' contains no strategy that puts positive probability on buying for any ω_i with $-1 - \epsilon \leq \omega_i < k < -\epsilon$, then $\rho(\Sigma')$ contains no strategy that puts positive probability on buying for any $\omega_i \in B(k, \epsilon)$.*

Proof. Consider the reverse, i.e. there exists a $k' \in B(k, \epsilon)$ such that there is a strategy $s \in \rho(\Sigma')$ with $s(k') > 0$. (Only $k' > k$ is possible by assumption.) If speculator i gets the signal $\omega_i = k'$, he knows that $\int_{k'-2\epsilon}^{k} dH$ speculators will sell, with H being the distribution of the signals of other speculators conditional on receiving signal $\omega_i = k'$. For ϵ very

small, this conditional distribution is approximately a symmetric triangular distribution on $[k^{'} - 2\epsilon, k^{'} + 2\epsilon]$, and therefore $\int_{k^{'}-2\epsilon}^{k} dH \approx \frac{(2\epsilon-(k^{'}-k))^2}{8\epsilon^2}$. If a mass $\frac{(2\epsilon-(k^{'}-k))^2}{8\epsilon^2}$ of speculators is always selling, the maximum mass of net buy orders is (since every sell order cancels one buy order)

$$1 - \frac{(2\epsilon - (k^{'} - k))^2}{4\epsilon^2} = \frac{4\epsilon(k^{'} - k) - (k^{'} - k)^2}{4\epsilon^2}.$$

Since the signal is $\omega_i = k^{'}$, the minimum number of buy orders to make $x_N + x_S$ positive and therefore make buying profitable is $-k^{'} - \epsilon$ (remember that $k^{'} < 0$). Buying can therefore only make sense after receiving $\omega_i = k^{'}$ if the maximum number of buy orders is larger than the minimum number of buy orders required to make buying profitable, i.e.

$$\begin{aligned} -k^{'} - \epsilon &< \frac{4\epsilon(k^{'} - k) - (k^{'} - k)^2}{4\epsilon^2} \\ (k^{'} - k)^2 - 4\epsilon(k^{'} - k) - 4\epsilon^2(k^{'} + \epsilon - k + k) &< 0 \\ (k^{'} - k)^2 - (4\epsilon + 4\epsilon^2)(k^{'} - k) - 4\epsilon^2(k + \epsilon) &< 0 \end{aligned}$$

The last inequality is not true for $k^{'}$ very large or very small, so that it must be true between the two solutions for the corresponding equality (since these solutions exist) and we get that buying can only be profitable for $k^{'}$ if

$$2\epsilon\left(1 + \epsilon + \sqrt{(1+\epsilon)^2 + k + \epsilon}\right) > k^{'} - k > 2\epsilon\left(1 + \epsilon - \sqrt{(1+\epsilon)^2 + k + \epsilon}\right).$$

But that is incompatible with $k^{'} \in B(k, \epsilon)$. □

Using this lemma, we can show that there is no $k < -\epsilon$ such that there exists a strategy s with $s(k) > 0$ and $s \in \rho^m(\Sigma)$ for all $m \in \mathbb{N}$. Assume to the contrary that there exists a non-empty set of such ks and let $\hat{k}$ be the infimum of this set. Then $\forall k < \hat{k} : s(k) = 0$, and we can pick a $\bar{k}$ that is arbitrarily close to but below $\hat{k}$. Since $\hat{k} < -\epsilon$, there exists a $\bar{k}$ such that $\hat{k} \in B(\bar{k}, \epsilon)$. Then it follows from the lemma that there cannot exist a strategy that has positive probability of buying anywhere in an open ball around $\bar{k}$.

We can show analogously that there is no $k > \epsilon$ such that there is a strategy s with $s(k) < 1$ and and $s \in \rho^m(\Sigma)$ for all $m \in \mathbb{N}$. Hence the only rationalizable strategies are those that sell with probability 1 for all $\omega_i < -\epsilon$ and buy for all $\omega_i > \epsilon$. □

Proof of proposition 3. First, I show existence if there is common knowledge of x_N. Assume that the market follows its equilibrium strategy, so that $p_2 = p_0$. It is then profitable

for any speculator to buy at $p_1 < p_0$ or sell at $p_1 > p_0$. Speculators therefore trade against the noise traders until either all of them have posted an order or $x = 0$ and $p_1 = p_0$. No speculator has any incentive to deviate: Those who post orders either make a positive profit (if $|x_N| > 1$) or no profit (otherwise), and those who do not post orders (since other informed speculators have already driven the price back to p_0) would lose money by trading (since they would move the price above p_0 if they bought or below p_0 if they sold).

Now assume that the speculators follow this strategy. Then p_1 contains absolutely no information about v, since the speculators only either do nothing or counteract the noise traders (whose actions are independent of v), and none of their behavior is conditional on v. The market can therefore only follow its prior and set $p_2 = p_0$.

In the game without common knowledge, consider the following argument: If each speculator can only observe his signal ω_i, it is still optimal to buy if $\omega_i \leq -1$, because in expectation $p_1 < p_0$ regardless of the behavior of other speculators. Now consider the case where $\omega_i \in (-1, 0)$. If all other speculators buy with probability $-\omega_j$ upon observing $\omega_j \in (-1, 0)$, they will on average buy with probability $-x_N$, which means that p_1 will be 0 in expected terms. Every single speculator is then indifferent between buying or selling or doing nothing. Therefore, it is an equilibrium if all speculators buy for $\omega_i \leq -1$, buy with probability $-\omega_i$ for $\omega_i \in (-1, 0)$, sell with probability ω_i if $\omega_i \in (0, 1)$ and always sell if $\omega_i \geq 1$. The non-adjustment equilibrium remains completely undisturbed if x_N is no longer common knowledge.

A brief remark on out-of-equilibrium beliefs: In this equilibrium, total order flow x will be between $1-n$ and $n-1$, meaning that $p_1 \in \Big[p_0 + \hat{\lambda}(1-n),\, p_0 + \hat{\lambda}(n-1)\Big]$. Out-of-equilibrium beliefs are what the market thinks if p_1 should lie outside that interval. But it is clearly not optimal for the market to assume that prices outside this interval are informative. If it did, and accordingly set some $p_2 > p_0 + \hat{\lambda}(n-1)$ after observing $p_1 > p_0 + \hat{\lambda}(n-1)$, the speculators would have an incentive to try to push p_1 above $p_0 + \hat{\lambda}(n-1)$ regardless of whether $v = v_H$ or $v = v_L$, so that p_1 would not be any more informative than it was before. If, on the other hand, they were to set p_2 with $p_0 < p_2 < p_0 + \hat{\lambda}(n-1)$ after observing $p_1 > p_0 + \hat{\lambda}(n-1)$, the speculators would have no incentive to drive prices out of equilibrium range at all, even if they could submit information in this way. □

8 Appendix: Which Equilibrium is Pareto-Preferred?

Proposition 4. *If f (the density of x_N) is single-peaked, speculators prefer the adjustment to the non-adjustment equilibrium.*

To simplify notation, let $p_2^H(p_1)$ be the expected value of v given p_1 if $p_1 > p_0$, and

$p_2^L(p_1)$ the expected value of v given p_1 if $p_1 < p_0$. We make use of the following lemma:

Lemma 2. *If* $2 > \frac{p_1-p_0}{\hat{\lambda}}$ *it is* $\frac{\partial p_2^H(p_1)}{\partial p_1} < 0$ *(and hence also* $\frac{\partial p_2^L(p_1)}{\partial p_1} > 0$*). If* $2 \leq \frac{p_1-p_0}{\hat{\lambda}}$*, then* $p_2^H = p_2^L = p_0$ *and consequentially* $\frac{\partial p_2^H(p_1)}{\partial p_1} = \frac{\partial p_2^L(p_1)}{\partial p_1} = 0$.

Proof. It is $p_2^H(p_1) = \pi(p_1)v_H + (1 - \pi(p_1))\, v_L$, or

$$p_2^H(p_1) = \frac{1 - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)}{2 - F(k_H) - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)} v_H + \frac{1 - F(k_H)}{2 - F(k_H) - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)} v_L.$$

Since $k_H = \max\left\{\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right), 1\right\}$, there are two possible cases:

1. $2 > \frac{p_1-p_0}{\hat{\lambda}}$. Then $k_H = 1$ and

$$p_2^H(p_1) = \frac{1 - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)}{2 - F(1) - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)} v_H + \frac{1 - F(1)}{2 - F(1) - F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)} v_L.$$

   As $F\left(\frac{p_1-p_0}{\hat{\lambda}} - 1\right)$ is monotonously growing in p_1, and since $v_H > v_L$, it is then $\frac{\partial p_2^H(p_1)}{\partial p_1} < 0$.

2. $2 \leq \frac{p_1-p_0}{\hat{\lambda}}$. Then $k_H = \frac{p_1-p_0}{\hat{\lambda}} - 1$ and

$$p_2^H(p_1) = \frac{v_H + v_L}{2} = p_0.$$

□

Proof of Proposition 4. Speculators' expected profit from the efficient equilibrium is the sum of expected profits if $|x_N| < 1$ and $|x_N| \geq 1$. More precisely, it is

$$\begin{aligned} &\Pr(|x_N| < 1)\left(\mathbb{E}\left[p_2^H(p_1)||x_N| < 1\right] - p_0 - \frac{\hat{\lambda}}{2}\left(\mathbb{E}\left[x_N||x_N| < 1\right] + 1\right)\right) \\ &+ \Pr(|x_N| = 1)\left(\mathbb{E}\left[p_2^H(2\lambda)\right] - p_0 - \frac{\hat{\lambda}}{2}(2)\right) \\ &+ \Pr(|x_N| > 1)\left(\mathbb{E}\left[p_2^H(p_1)|x_N > 1\right] - p_0 - \frac{\hat{\lambda}}{2}\left(\mathbb{E}\left[x_N|x_N > 1\right] + 1\right)\right) \end{aligned} \tag{1.4}$$

(Note that, because of symmetry, we can restrict ourselves to the expected prices if $p_1 > p_0$.) All three summands are clearly positive, as we can see from lemma 2 and the proof of proposition 1.

In the inefficient equilibrium, expected profit for any speculator is positive only if $|x_N| > 1$, so that overall expected profit from the inefficient equilibrium is

$$\Pr(|x_N| > 1) \frac{\hat{\lambda}}{2} (\mathbb{E}[x_N | x_N > 1] - 1).$$

If the expression "(Expected profit from efficient equilibrium)−(Expected profit from inefficient equilibrium)" is positive, speculators prefer the efficient equilibrium. We can write this expression as the sum of some positive terms and the term

$$\Pr(|x_N| > 1) \left(\mathbb{E}\left[p_2^H(p_1) | x_N > 1\right] - p_0 - \frac{\hat{\lambda}}{2} (\mathbb{E}[x_N | x_N > 1] + 1) - \frac{\hat{\lambda}}{2} (\mathbb{E}[x_N | x_N > 1] - 1) \right). \tag{1.5}$$

From the proof of proposition 1 we know that $\mathbb{E}\left[p_2^H(\lambda(n+1))\right] - p_0 - \frac{\hat{\lambda}}{2}(n+1) > 0$. From lemma 2, it follows that then also $\mathbb{E}\left[p_2^H(\lambda(x_N+1)) | x_N > 1\right] > p_0 + \frac{\hat{\lambda}}{2}(n+1)$. That means that if

$$\frac{\hat{\lambda}}{2}(n+1) - \frac{\hat{\lambda}}{2} (\mathbb{E}[x_N | x_N > 1] + 1) - \frac{\hat{\lambda}}{2} (\mathbb{E}[x_N | x_N > 1] - 1) \tag{1.6}$$

is positive, then expression (1.5) is also positive. (1.6) simplifies to $n+1-2\mathbb{E}[x_N | x_N > 1]$, which is positive if $\frac{n+1}{2} > \mathbb{E}[x_N | x_N > 1]$. If $f(x)$ is falling in $|x|$, that is the case. □

It should be noted that this is a sufficient, but not a necessary condition: The difference between expected payoffs from the efficient and the inefficient equilibrium can well be positive even if $\frac{n+1}{2} < \mathbb{E}[x_N | x_N > 1]$. But it can be shown that the efficient equilibrium is not always preferred: If f is not falling in its argument, it is possible that speculators actually prefer the inefficient equilibrium. Intuitively, that is the case if f has a lot of mass towards n and $-n$, so that large bubbles (which are profitable for rational speculators in the inefficient equilibrium) become very likely. In the efficient equilibrium, the market adjusts $\pi(p_1)$ accordingly, and speculators' expected profit margins in the efficient equilibrium (which is now not very efficient) become very low. In the inefficient equilibrium, on the other hand, speculators could now make large expected gains, since their profit is higher the further noise traders drive p_1 away from p_0.

Corollary. *There are distributions of x_N so that the efficient equilibrium exists, but speculators ex ante prefer the inefficient equilibrium.*

Proof. Consider the the case where $\Pr(x_N = 1) = \Pr(x_N = -1) = \varepsilon$ and $\Pr(x_N = n) = \Pr(x_N = -n) = \frac{1}{2} - \varepsilon$. Then the expected payoff in the inefficient equilibrium is $(1-2\varepsilon)\frac{\hat{\lambda}}{2}(n-1)$, while the expected payoff from the efficient equilibrium is

$$\mathbb{E}\left[\varepsilon p_2^H(p_0 + \lambda(1+x_N)) \middle| x_N = 1\right] + \mathbb{E}\left[\varepsilon p_2^H(p_0 + \lambda(1+x_N)) \middle| x_N = -1\right]$$

$$+\mathbb{E}\left[(1-2\varepsilon)\,p_2^H\left(p_0+\lambda(1+x_N)\right)\middle|\, x_N=n\right]-\frac{\hat{\lambda}}{2}-(1-2\varepsilon)\,\frac{\hat{\lambda}}{2}n-p_0.$$

Let $D_i=\mathbb{E}\left[p_2^H\left(p_0+\lambda(1+x_N)\right)\middle|\, x_N=i\right]-p_0$. Then the difference between profits from the efficient and inefficient equilibrium is

$$\varepsilon D_1+\varepsilon D_{-1}+(1-2\varepsilon)D_n-\hat{\lambda}\left(\varepsilon+(1-2\varepsilon)n\right). \tag{1.7}$$

If we take the maximal $\hat{\lambda}$ such that the efficient equilibrium still exists,[14] we have $\hat{\lambda}=2\frac{D_n}{1+n}$, and (1.7) becomes

$$\varepsilon D_1+\varepsilon D_{-1}+\frac{(1-4\varepsilon)-(1-2\varepsilon)n}{1+n}D_n.$$

For this always to be positive, it would have to be

$$\frac{D_1+D_{-1}}{2}/D_n>\frac{4\varepsilon-2\varepsilon n-1+n}{2\varepsilon(1+n)}.$$

Intuitively, this means that as ε gets arbitrarily small, the prices that result in period 1 from $x_N=1$ and $x_N=-1$ would have to become infinitely more informative than the prices that result from $x_N=n$ and $x_N=-n$. But a price p_1 that results from $x_N=n$ lies within the price range $\left[p_0+\hat{\lambda}(-2),\, p_0+2\hat{\lambda}\right]$ with constant probability $\frac{2}{1+n}$ because of the price formation process through noisy λ. Therefore, the prices resulting from $x_N=1$ and $x_N=-1$ can never be infinitely more informative than the prices resulting from $x_N=n$. Therefore, there exists a distribution for x_N so that for a large enough $\hat{\lambda}$ speculators prefer the inefficient to the efficient equilibrium. □

These conditions on the shape of f might seem rather abstract, but they have an intuitive interpretation. f is falling in distance from 0 if the correlation between noise traders' decisions is sufficiently small (they might make their decisions independently, or their actions might even be negatively correlated). In these cases, speculators will always prefer the efficient equilibrium. But high correlation between the decisions of the noise traders means nothing else than strong herding. If noise traders are sufficiently prone to strong herding, all rational market participants weakly prefer an equilibrium in which no information is transmitted to a partially revealing equilibrium.

[14] For very small $\hat{\lambda}$, speculators always prefer the efficient equilibrium.

9 Appendix: A Discrete Model where Speculators have Market Power

The model can also be written with a finite number of speculators and noise traders, such that signle speculators actually have market power and can influence the price. While this makes some of the expressions less tractable and slightly changes the proofs, the main theorems remain intact and the two equilibria still exist. Assume in the following that there is a finite number n of noise traders and m of speculators.

Proposition 5. *(Efficient equilibrium) It is an equilibrium if every speculator follows the strategy "If $x_N \leq -1$, sell and if $x_N \geq 1$ buy. If $|x_N| < 1$, buy if $v = v_H$ and sell if $v = v_L$." and the market sets*

$$p_2 = p_2^H(p_1) = \pi(p_1)v_H + (1 - \pi(p_1))\, v_L \quad \text{if} \quad p_1 > p_0 \tag{1.8}$$

$$p_2 = p_2^L(p_1) = (1 - \pi(p_1))\, v_H + \pi(p_1)v_L \quad \text{if} \quad p_1 < p_0 \tag{1.9}$$

$$p_2 = p_0 \quad \text{if} \quad p_1 = p_0, \tag{1.10}$$

where

$$\pi(p_1) = \begin{cases} \frac{1-F\left(\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor\right)}{2-F(k_H)-F\left(\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor\right)} & \text{if } p_1 > p_0 \\ \\ \frac{1-F\left(\left\lceil \frac{p_1-p_0}{\hat{\lambda}} + m \right\rceil\right)}{2-F(k_L)-F\left(\left\lceil \frac{p_1-p_0}{\hat{\lambda}} + m \right\rceil\right)} & \text{if } p_1 < p_0 \end{cases}$$

where $\pi(p_1)$ is the market's belief that $v = v_H$ if $p_1 > p_0$ or that $v = v_L$ if $p_1 < p_0$, respectively, with $k_H = \max\left\{\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor, m-1\right\}$ and $k_L = \min\left\{\left\lceil \frac{p_1-p_0}{\hat{\lambda}} + m \right\rceil, -m+1\right\}$, if and only if

$$\hat{\lambda} \leq \mathbb{E}\left[\frac{F(k_H) - F\left(\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor\right)}{2 - F(k_H) - F\left(\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor\right)} \middle| x_N = n \right] \frac{v_H - v_L}{m+n}. \tag{1.11}$$

Proof (similar to the continuous case). **Part 1: The market has no incentive to deviate (and $\pi(p_1)$ is the correct belief).**

Assume that the speculators follow their equilibrium strategies and consider the case where $p_1 > p_0$. The market can then, from observing p_1, draw conclusions about v. Let $\pi(p_1)$ be the conditional probability that $v = v_H$ after observing a certain p_1, $\Pr(v_H|p_1)$.

It is

$$\pi(p_1) = \Pr(v_H|p_1) = \frac{\Pr(p_1 \cap v_H)}{\Pr(p_1)} = \frac{\Pr(p_1 \cap v_H \cap |x_N| < m) + \Pr(p_1 \cap v_H \cap x_N \geq m)}{\Pr(p_1 \cap |x_N| < m) + \Pr(p_1 \cap x_N \geq m)}$$

since $\Pr(p_1 \cap x_N \leq -m) = 0$.

If g is the probability density function of λ, we can express this as

$$\pi(p_1) = \frac{\frac{1}{2}\sum_{y=-m+1}^{m-1} f(y)g\left(\frac{p_1-p_0}{y+m}\right) + \frac{1}{2}\sum_{y=m}^{n} f(y)g\left(\frac{p_1-p_0}{y+m}\right)}{\frac{1}{2}\sum_{y=-m+1}^{m-1} f(y)g\left(\frac{p_1-p_0}{y+m}\right) + \sum_{y=m}^{n} f(y)g\left(\frac{p_1-p_0}{y+m}\right)}.$$

The product in all the sums, $f(y)g\left(\frac{p_1-p_0}{y+m}\right)$, gives the probability that $x_N = y$ and $\lambda = \frac{p_1-p_0}{y+m}$, in which case the parameters would lead to the given p_1 if speculators always bought in period 1. The first sum in the numerator is hence the overall probability that p_1 would be observed as a result of some $x_N \in [-m+1, m-1]$ if speculators always bought the asset. Since, if $x_N \in [-m+1, m-1]$, speculators buy the asset only if $v = v_H$, this probability has to be multiplied by $\frac{1}{2}$ to give the probability $\Pr(p_1 \cap v_H \cap |x_N| < m)$. The second sum in the numerator gives the probability that p_1 would be observed as the result of some $x_N \geq m$. Since $v = v_H$ in only half of these cases, we again need to multiply with $\frac{1}{2}$ (albeit for different reasons) to get the unconditional probability that p_1 would happen as the result of some $x_N > m$ *and* that also $v = v_H$. In the numerator, therefore, we have the overall probability that a given p_1 is observed and is informative.

In the denominator, we then have the overall probability that a given p_1 is observed. This is given by the expression from the numerator, only that now *all* cases in which $x_N > m$ are considered (since they all lead to $p_1 > p_0$), whereas only half of them are informative. The fraction therefore gives the ratio between the number of cases in which p_1 is observed and it is $v = v_H$ and the overall number of cases in which p_1 is observed. This is the conditional probability $\Pr(v_H|p_1)$.

We can simplify the expression: Since λ is uniformly distributed on the interval $\left(0, \hat{\lambda}\right)$, $g\left(\frac{p_1-p_0}{y+m}\right) = \frac{1}{\hat{\lambda}}$ if $0 < \frac{p_1-p_0}{y+m} < \hat{\lambda}$ and 0 otherwise. For any $p_1 > 0$, it is $0 < \frac{p_1-p_0}{y+m}$, but $g\left(\frac{p_1-p_0}{y+m}\right)$ is nonzero only for $y > \frac{p_1-p_0}{\hat{\lambda}} - m$. We can write

$$\begin{aligned}
\pi(p_1) &= \frac{\sum_{y=\left\lceil \frac{p_1-p_0}{\hat{\lambda}}-m \right\rceil}^{k_H} f(y) + \sum_{y=k_H+1}^{n} f(y)}{\sum_{y=\left\lceil \frac{p_1-p_0}{\hat{\lambda}}-m \right\rceil}^{k_H} f(y) + 2\sum_{y=k_H+1}^{n} f(y)} \\
&= \frac{1 - F\left(\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor\right)}{2 - F\left(\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor\right) - F\left(k_H\right)}
\end{aligned}$$

where $k_H = \max\left\{\left\lfloor \frac{p_1-p_0}{\hat{\lambda}} - m \right\rfloor, m-1\right\}$. Therefore, given the speculators' strategies, $\pi(p_1)$ gives the correct beliefs in equilibrium.

$\Pr(v_L|p_1)$ is the complementary probability $1-\pi(p_1)$, so that the expected value of v given p_1 is $E[v|p_1] = \pi(p_1)v_H + (1-\pi(p_1))\,v_L$. A similar argument applies to the case where $p_1 < p_0$. If $p_1 = p_0$, the price contains no information and p_2 should be set equal to the prior.

p_1 is between $p_0 - \hat{\lambda}(m+n)$ and $p_0 + \hat{\lambda}(m+n)$. For $x_N \in \{-n, n\}$, all possible p_1 occur with positive probability, so that in equilibrium (where $x_N \in [-n, n]$) all possible p_1 occur with positive probability and there can be no out-of-equilibrium beliefs.

Part 2: Speculators make a positive profit in equilibrium.

Now assume that the market follows its equilibrium strategy. Consider the case where $p_1 > p_0$, meaning that either $|x_N| < m$ and $v = v_H$ or simply $x_N \geq m$. If they follow their equilibrium strategies, the speculators' buy orders will drive the price to $p_0 + \lambda(m + x_N) > p_0$, and in period 2 all speculators will be able to sell their holdings at $p_2^H = \pi(p_1)v_H + (1-\pi(p_1))\,v_L$. Their profit is then $p_2^H - p_1$, or $\pi(p_1)v_H + (1-\pi(p_1))\,v_L - p_0 - \lambda(m+x_N)$, which can also be written as

$$\left[\frac{1-F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}{2-F(k_H)-F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}-\frac{1}{2}\right]v_H+\left[\frac{1-F(k_H)}{2-F(k_H)-F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}-\frac{1}{2}\right]v_L-\lambda(m+x_N) \tag{1.12}$$

$$=\frac{F(k_H)-F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}{4-2F(k_H)-2F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}(v_H-v_L)-\lambda(m+x_N)$$

x_N is known to the speculators. Then we can write expression 1.12 in expected terms (given x_N):

$$\mathbb{E}\left[\frac{F(k_H)-F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}{4-2F(k_H)-2F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}\middle| x_N\right](v_H-v_L)-\frac{\hat{\lambda}}{2}(m+x_N).$$

Since p_1 is monotonically increasing in x_N, and therefore $F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)$ and $F(k_H-1)$ are weakly increasing in x_N, the whole expression becomes minimal for $x_N = n$, where it is

$$\mathbb{E}\left[\frac{F(k_H)-F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}{4-2F(k_H)-2F\left(\left\lfloor\frac{p_1-p_0}{\hat{\lambda}}-m\right\rfloor\right)}\middle| x_N=n\right](v_H-v_L)-\frac{\hat{\lambda}}{2}(m+n).$$

If this is positive, then speculators will make an expected profit by following their equilibrium strategies for all x_N (the case where x_N is negative is analogous and leads

to the same result). We can reformulate the condition as

$$\hat{\lambda} \leq \mathbb{E}\left[\frac{F(k_H) - F\left(\left\lfloor \frac{p_1 - p_0}{\hat{\lambda}} - m \right\rfloor\right)}{2 - F\left(k_H\right) - F\left(\left\lfloor \frac{p_1 - p_0}{\hat{\lambda}} - m \right\rfloor\right)} \middle| x_N = n \right] \frac{v_H - v_L}{m + n}$$

which is simply the spread between high and low value, adjusted for the number of market participants and some adjustment factor that depends on the precise shape of f.

Part 3: No single speculator has an incentive to deviate from his equilibrium strategy.

As speculators always make a profit in equilibrium, it would not be profitable for any speculator to deviate by not trading at all. But what if a speculator decided to sell if his equilibrium action would be to buy? We have to distinguish three cases (note that "buy" would never be an equilibrium action if $x_N \leq -m$):

1. $x_N = -(m-1)$. In this case it is $x = 1$ in equilibrium, and if a single speculator decided to sell instead of buying, x would be -1. Since $p_2(p_1)$ is point-symmetric around (p_0, p_0) (i.e. $p_2(p_1) - p_0 = p_0 - p_2(p_0 - (p_1 - p_0))$ because of the symmetry assumption on f), the speculator who sold would gain just as much in expectation as he would have by buying. Since he is thus indifferent, there is no incentive to deviate from equilibrium strategies.

2. $x_N = -(m-2)$. Then $x = 2$ in equilibrium, but if a single speculator sold instead of buying, the resulting net order flow would be 0, so that $p_1 = p_0$. Then it would also be $p_2 = p_0$, so that the speculator would make no gain at all by selling, whereas he could have made a positive profit by buying.

3. $x_N > -(m-2)$. Then $x > 2$ in equilibrium, and a single speculator can only lower x to some slightly lower, but still positive number. Then $p_2 = p_2^H(p_1) > p_1$, so that the speculator would actually make a loss by selling in period 1.

We can therefore conclude that no speculator has an incentive to deviate from his equilibrium strategy if $p_1 > p_0$. A similar argument applies where $p_1 < p_0$ (i.e. if speculators bought instead of selling). □

Chapter 2

How Jeremy Bentham would defend against coordinated attacks[1]

Ole Jann and Christoph Schottmüller

We consider the use of information in deterring coordinated attacks, for example a central bank defending a currency peg or a government facing a revolution. Bentham (1787) proposed the "panopticon" as an ideal solution: Potential attackers are deterred by secrecy about the defender's strength. We compare different information structures in a model of coordinated attacks. We uncover a fundamental property of defending against a large but finite group and show that Bentham's intuition is correct. Our results provide insights into the applications of Bentham's ideas across the social sciences, and recommendations for the general problem of defense against coordinated attacks.

[1]A previous version of this paper has been published as TILEC Discussion Paper 2015-018 and as University of Copenhagen Discussion Paper 15-11. We are grateful for helpful comments by Alberto Alesina, Eric van Damme, Eddie Dekel, Jeff Ely, Nicola Gennaioli, Heidi Kaila, Nenad Kos, Pablo Kurlat, Marco Ottaviani, Alessandro Pavan, Jens Prüfer, Tomas Sjöström, Joel Sobel and Peter Norman Sørensen as well as from audiences at Bocconi University, the University of Copenhagen, the University of Lund, 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! *(Bentham, 1787)*

1 Introduction

We analyze situations in which a single player is in conflict with a group of others, and the group members' actions are strategic complements. Consider, for example, a government threatened by a revolution: Each potential revolutionary has to decide whether to show up for a demonstration, and larger demonstrations are more likely to succeed – but no one wants to be the only one to show up. A speculative attack on a currency peg requires the participation of many speculators – but if the attack fails because not enough speculators participate, those who participated will lose money.

In both cases, the single player would like to prevail with a minimum use of resources (security forces, currency reserves) by discouraging the group from acting in the first place. In this paper, we consider how he can accomplish this goal by choosing the right information structure. The information structure determines which information about his own strength will be revealed when he chooses a costly strength level at a later stage. Our surprising result is that in many situations, complete secrecy is optimal. That is, the single player foregoes the option to publicly commit himself to a strength level. Secrecy mirrors the idea of the "panopticon" proposed by Bentham (1787) – an innovative prison concept in which prisoners were to be kept unable to see the guards as well as separated from each other.

The general problem that we consider has many applications, some of which we discuss later in the paper. Our main analysis concentrates on a succinct and graphic example close to Bentham's idea: The question of how to construct a prison.[2] The prison warden faces a trade-off, as guards are costly but more guards offer more protection. The prison design allows a choice over how much information about the guard strength is available to the prisoners. This can make coordination among individual prisoners, in the absence of institutions that allow for explicit coordination, easier or harder. Ideally, the prison warden would prefer to maintain order in the prison and prevent revolts and breakouts while using a minimum of guards. The optimal prison design will exploit the prisoners' coordination problem in order to prevent them from revolting.

Bentham proposed that the isolation of the prisoners, together with their lack of

[2]Bentham tried to construct the actual panopticon according to his plans, using considerable time on the purpose while trying to convince successive governments of the idea. Unlike him, we mostly see the prison as a metaphor for the mechanisms we want to analyze. Taking our formal model as a practical guide to prison construction is done at the reader's own risk.

knowledge of how many guards (if any) were on duty, would make coordination and thus a successful revolt impossible.[3] Through the lense of game theory, this argument appears unconvincing. Rational prisoners should be able to implicitly coordinate, and in equilibrium they should be able to infer the choice of the prison warden about guard strength. We find, however, that Bentham's intuition plays out: In a large prison, where prisoners have no information about guard strength before independently choosing whether to revolt or not, there is only one equilibrium in which the warden randomizes between minimal guard levels and prisoners almost never revolt.

The result arises from a fundamental property of large populations. If there are many prisoners who have no way of coordinating, their aggregate behavior is relatively more predictable than if there were only a handful of them.[4] This means that there can be no equilibria in which the warden chooses a positive guard strength which he neither wants to adjust up- or downwards, and in which a successful revolt is sufficiently likely to induce prisoners to revolt (since otherwise the warden would like to use fewer guards). We show that this main insight is robust to several modifications and extensions, which we discuss in section 4.2.

We construct a simple model in which a warden chooses a costly guard level. Afterwards, each prisoner decides for himself whether to revolt or not. A revolt is successful if the number of revolting prisoners is larger than the guard level; otherwise the revolt fails and revolting prisoners get punished.

We compare four different information structures, also shown in table 2.1: (1a) Prisoners can observe the guard level and coordinate ("benchmark model"). (1b) Prisoners cannot observe the guard level but can coordinate ("benchmark model"). (2) Prisoners can observe the guard level but face a coordination problem ("transparency model"). (3) Prisoners cannot observe the guard level and face a coordination problem ("panopticon").

In cases (1a) and (1b), preventing a revolt is only possible when choosing the guard level such that a revolt by all prisoners would not be successful. In (2), "a union of hands" is required for a successful revolt for any intermediate guard level, i.e. any guard level that doesn't offer protection against a revolt by all prisoners. As the actions of the prisoners are strategic complements, there are two equilibria in the prisoners' subgame (after the warden has chosen an intermediate guard level): All prisoners revolt, or none does. One of these, the successful revolt, is preferred by the prisoners, but in this

[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 also 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]This follows from the law of large numbers, i.e. by the same logic that the average of a number of dice rolls tends to be closer to 3.5 if the dice is rolled more often.

		Guard level observable	
		Yes	No
Coordination problem between prisoners	No	(1a) Benchmark	(1b) Benchmark
	Yes	(2) Transparency	(3) Panopticon

Table 2.1: The four information structures we consider.

equilibrium each of them puts himself at the mercy of the others – he does not want to be caught as the only one revolting. Following the global games literature, we select an equilibrium by assuming that the prisoners, being isolated from each other, do not achieve *common* knowledge of the guard level. Without common knowledge, they need to reason about each others' beliefs to make an optimal choice as in Rubinstein (1989) or Carlsson and van Damme (1993): 'I believe that a revolt can be successful, but what if the others think that it cannot? Then they would not revolt, and neither should I.' This "infection of beliefs" makes it possible to reliably prevent a revolt with a much lower guard level than in the benchmark model. While the number of guards needed to deter revolts still rises linearly in the number of prisoners, the slope is usually much lower than one.

Finally, in the last model, the panopticon, it is not immediately obvious what kind of equilibria there are. Knowing that the guard level will not be observed, the warden has an incentive to choose a low guard level, but that will make him very vulnerable to revolts by even a few prisoners. Especially if there are many prisoners, it might seem sensible to always set a sufficiently high guard level to prevent substantial revolts.

Instead, we find that – if the number of prisoners is large – there is a unique equilibrium in mixed strategies in which the warden randomizes between the lowest possible guard levels, and each prisoner randomly chooses whether to revolt or not. The individual probability of revolting and the probability of a successful breakout are very small if the number of prisoners is large. This guarantees that revolts can be prevented almost surely with just one guard, as Bentham predicted. No other equilibria exist – neither pure nor mixed, symmetric or asymmetric.

To get a sense of the intuition, consider the following arguments.[5] There can be no equilibria in pure strategies, since that would either mean that there is a successful attack for sure, or never. In the former case, the warden would like to increase the number of guards; in the latter case he would like to decrease it (or the prisoners would like to switch from revolting to not revolting). In any mixed equilibrium, each prisoner revolts with a certain probability. For large numbers of prisoners, the overall behavior of the prison population becomes relatively more predictable by the law of large numbers. This implies that the probability of a breakout is very low in equilibrium because the

[5]We deepen this intuition in section 3.4 and provide a formal proof in the appendix.

warden would otherwise have an incentive to increase the guard level due to the high predictability of a breakout. If a successful breakout is unlikely, however, the prisoners would never want to revolt in the first place if the warden chooses relatively high guard levels with positive probability. But then the warden would want to lower his guard level, so that there cannot be any equilibria in which the warden chooses a relatively high guard level. This argument leads to the conclusion that there is only one equilibrium, in which the warden mixes between the two lowest guard levels. This fact makes the panopticon the optimal information structure for large groups of prisoners, where it performs far better than the other structures. Graph 2.1 compares the expected warden payoff in the three information structures for large numbers of prisoners.

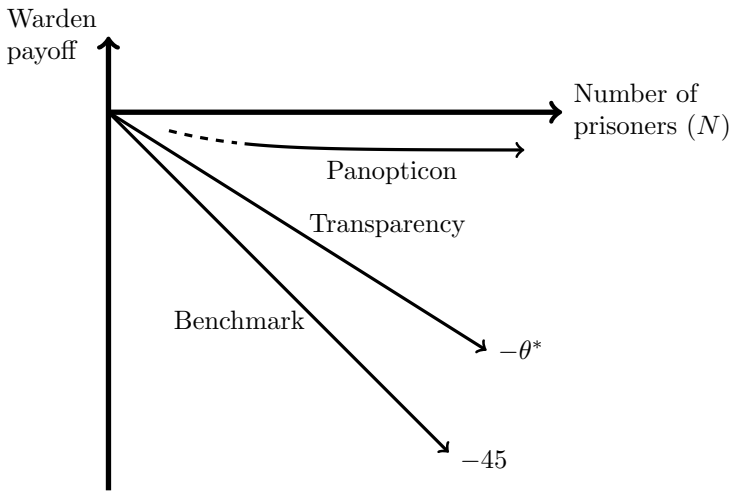


Figure 2.1: A comparison of the warden payoff in the three information structures. The benchmark case is most expensive, as the warden needs as many guards as there are prisoners. In the transparency case, the warden can prevent breakouts with a lower number of prisoners; but the required number of guards still grows linearly in N. In the panopticon, the warden payoff is bounded from below by a constant as N grows larger. For small N, the panopticon is not necessarily optimal.

This result resembles Bentham's ideas.[6] 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,

[6]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;"

the panopticon thus makes optimal use of the prisoners' coordination problem.

The contribution of this paper is two-fold. Firstly, our model allows us to theoretically underpin the social science literature that followed Bentham. The idea of the panopticon has been influential in philosophy, sociology and political science; our model is the first that formally examines and explains how and why the panopticon can work. This allows us, for example, to identify the law of large numbers as the driving force behind the mechanics of the panopticon. Secondly, our model is descriptive of some situations in which a central player has to defend against coordinated attacks and our results have policy implications for these situations, for example, the optimality of secrecy.

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 first 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.[7] 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) that, besides the accumulation of capital, made the industrial take-off of the early 18th century possible. Our result captures some of 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 information structures 2 and 3, 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

[7]"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)

against an uprising for the sake of social welfare, while a repressive dictatorship might deploy secret surveillance methods to suppress dissent and rebellion. We are interested in the mechanisms by which this is done, and our results are positive, not normative.

Our results also have much more direct applications to situations where one actor can use the coordination problem of his opponents against them. Especially the problem of a central bank defending a currency peg against speculators has received much attention in the economic literature (e.g. Flood and Garber, 1984; Obstfeld, 1986; Morris and Shin, 1998). The problem of a player (usually called "policymaker") who is attacked by a group has 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). In contrast to these studies, we consider strength to be a costly choice of the defending player instead of a randomly drawn type and we consider a basic, one-shot game in which the defending player cannot lie about his strength.

The main contrast to these studies is that we consider the information structure – how much information the central player releases about his own strength – as an instrument, under the realistic assumption that he also has to choose how many resources he should use to defend himself. This leads us to different conclusions from any of the other studies – namely, that the single player can optimally exploit the coordination problem of the attackers by maintaining absolute secrecy about his own strength. In section 4.1, we discuss how our results apply to the problem of a central bank defending against speculators. Under assumptions that are very similar to those made by seminal papers in the field, we can show that the central bank optimally keeps the level of reserves a secret.

Our paper is also related to the game-theoretic literature on global games and common knowledge. In the model where the guard level is known and prisoners face a coordination problem, we make use of the seminal results on global games; see Carlsson and van Damme (1993), Morris and Shin (1998), and Morris and Shin (2003) for a survey. The "infection of beliefs" that occurs among prisoners was already described by Rubinstein (1989). We build on this literature but endogenize the "state of nature"as an active choice of the central player, by adding an extra perturbation to the model.

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 no asymmetric equilibria exist in our panopticon model, i.e. all the prisoners behave exactly the same in equilibrium.

2 Model

This section describes the general setup common to all three models. Details concerning the information structure that differ across the three models are described in the following section.

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 2.2 for a summary of these payoffs.

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

Table 2.2: 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 an 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 4.2 for details.

3 Analysis

3.1 Benchmark model: Perfect coordination

The first model is 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 the resulting subgame.[8] 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

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

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. The reason is that we assumed that the prisoners had no coordination problem. In the following model, we introduce the coordination problem and show how the warden can exploit this problem to his advantage. In terms of prison design, one might view the benchmark model as a prison in which all prisoners are kept in the same room and find it easy to resolve their coordination problem by communicating with each other. In this interpretation, prisoners are – as Bentham suggested – kept separately in the following models and will therefore face a coordination problem.

3.2 The transparency model

In the second model, prisoners first observe the guard level and then choose simultaneously and independently whether to revolt or not. If the guard level is weakly above N, it is a dominant action for each prisoner to play n. If the guard level is strictly below 1, it is a dominant action for each prisoner to play r. For guard levels between 1 and N, the optimal choice of a prisoner depends on what the other prisoners choose: If strictly more than $\gamma - 1$ other prisoners revolt, a given prisoner best chooses r himself. It is, however, optimal to choose n if less than $\gamma - 1$ other prisoners revolt. There are two equilibria in the subgames in which $\gamma \in [1, N)$: All prisoners revolt or no prisoner revolts. Consequently, the prisoners face a coordination problem. Following the approach in the global games literature, we select one of the two equilibria by relaxing the assumption that γ is common knowledge among the prisoners. More precisely, we show that introducing an arbitrarily small amount of noise into how prisoners observe the guard level leads to a unique equilibrium prediction. Figure 2.2 shows the intuition behind this equilibrium selection through infection.

The perturbation works in the following way: The warden chooses an intended guard level $\tilde{\gamma}$. The true guard level is then drawn from a normal distribution with mean $\tilde{\gamma}$ and variance $\varepsilon' > 0$.[9] 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.[10]

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

[10] 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

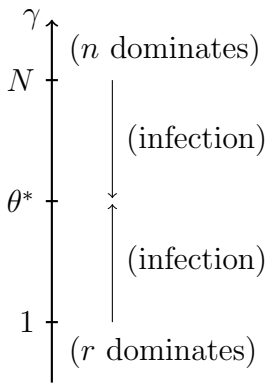


Figure 2.2: Infection of beliefs among prisoners: If $\gamma \geq N$, not revolting is a strictly dominant strategy for all prisoners. If $\gamma < 1$, revolting is strictly dominant. If $\gamma \in [1, N)$ and γ is common knowledge, there are two pure equilibria: Everybody revolts or no one revolts. When common knowledge is destroyed by the perturbation, beliefs get infected so that for $\gamma < \theta^*$, n is the unique equilibrium action, and r is the unique equilibrium action for $\gamma > \theta^*$.

Note that this setup eliminates common knowledge of the guard level. A prisoner observing signal θ knows that the true guard level is in $[\theta - \varepsilon, \theta + \varepsilon]$; he knows that each other prisoner knows that $\gamma \in [\theta - 3\varepsilon, \theta + 3\varepsilon]$; he knows that each other prisoner knows that he knows that $\gamma \in [\theta - 5\varepsilon, \theta + 5\varepsilon]$ and so on. Higher order beliefs will therefore play a role in determining the equilibrium. This appears to be a natural feature in a coordination game where the driving force of one's choice are exactly the expectations over what others do (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 1. *Let $\varepsilon' > 0$. Assume that $bN/(q+b) \notin \mathbb{N}$ and define*[11]

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

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,

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

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

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 1. *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. More technically, the so-called "infection argument" is at work: Consider a prisoner receiving a noisy signal above N. It is then quite likely that the guard level is above N and also quite likely that one other prisoner receives a signal above $N+\varepsilon$ (where it is a dominant action to play n). Consequently, a prisoner receiving a signal above N finds it optimal to not revolt. Now consider a prisoner receiving a signal just below N: This prisoner will consider it quite likely that at least one other prisoner receives a signal above N in which case this prisoner will play n (as we just established). So, even if the guard level is below N, it is unlikely that all other prisoners revolt and therefore a prisoner receiving a signal just below N will still play n. In this way, the dominance region (signals above $N+\varepsilon$) "infects" lower and lower signals in the sense that players with these lower signals also find it optimal to play n. A similar infection starts from signals below 1 where it is optimal to play r. Eventually (in the limit), this infection from both sides leads to the unique equilibrium.

3.3 The Panopticon

The third model is the one that closely mirrors Bentham's original idea. Now the warden chooses γ, but it cannot be observed by the prisoners, who also face a coordination problem.[12][13] We concentrate on equilibria in which all prisoners play r with the same

[12]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."

[13]If we allowed prisoners to communicate in a cheap talk way and selected the prisoner optimal equilibrium in this communication game, we would be back in the benchmark model. Such communication

probability p in equilibrium. In the supplementary material, we show that this is without loss of generality, i.e. no prisoner asymmetric equilibria exist in this game.

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

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 for γ. Therefore, the warden's maximization problem is

$$\max_{\gamma \in \{0,1,\dots,N\}} -(1 - G(\gamma))B - \gamma. \tag{2.1}$$

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

$$B\left(G(\gamma_0) - G(\gamma_1)\right) = \gamma_0 - \gamma_1 \tag{2.2}$$

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. *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 $\gamma_1 \in \{0, \dots, N-1\}$. For any $\gamma_1 \in \{0, \dots, N-1\}$, 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 2.3. 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

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 very credible.

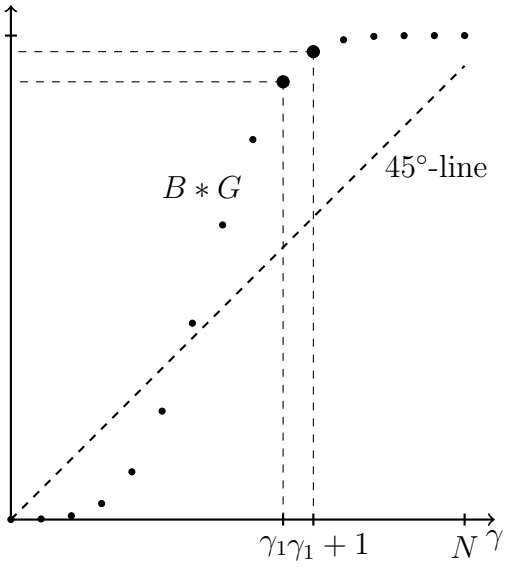


Figure 2.3: Equilibrium in the panopticon-model.

γ, which is simply γ. The warden optimally mixes between guard levels that maximize the difference between $B * G(\gamma)$ and γ. Intuitively, he trades off the additional cost of increasing the guard strength with reducing the probability of a breakout. Choosing a higher γ than $\gamma_1 + 1$, for example, would increase the cost by much more than the probability of preventing breakouts (weighted by the disutility of a breakout), and is therefore not optimal. If there are several guard levels where the difference is equivalent, the warden is indifferent between them. The example illustrates our two intermediate results: (a) The warden will never mix between more than two guard levels, since the concavity of G (above the mode) means that the difference between 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 2.3.

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

$$\mathbb{E}_\gamma \left[-qG_{N-1}(\gamma - 1) + b(1 - G_{N-1}(\gamma - 1)) \right] = 0 \tag{2.3}$$

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 (2.1) and (2.2) as the warden's own mixing probability does not play a role in these conditions. Given these two values, (2.3) will

determine the equilibrium mixing probability of the warden.

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

$$\Delta(\gamma) = -qG_{N-1}(\gamma-1) + b(1-G_{N-1}(\gamma-1)) \tag{2.4}$$

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

$$z = \frac{-\Delta(\gamma_1+1)}{\Delta(\gamma_1)-\Delta(\gamma_1+1)}. \tag{2.5}$$

Note that several equilibria can exist because Δ is not necessarily monotone: While both terms in (2.4) 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. 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. *Suppose there are two mixed equilibria: In equilibrium 1, the warden mixes over γ_1 and γ_1+1 and in equilibrium 2 the warden mixes over γ_2 and γ_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. *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:

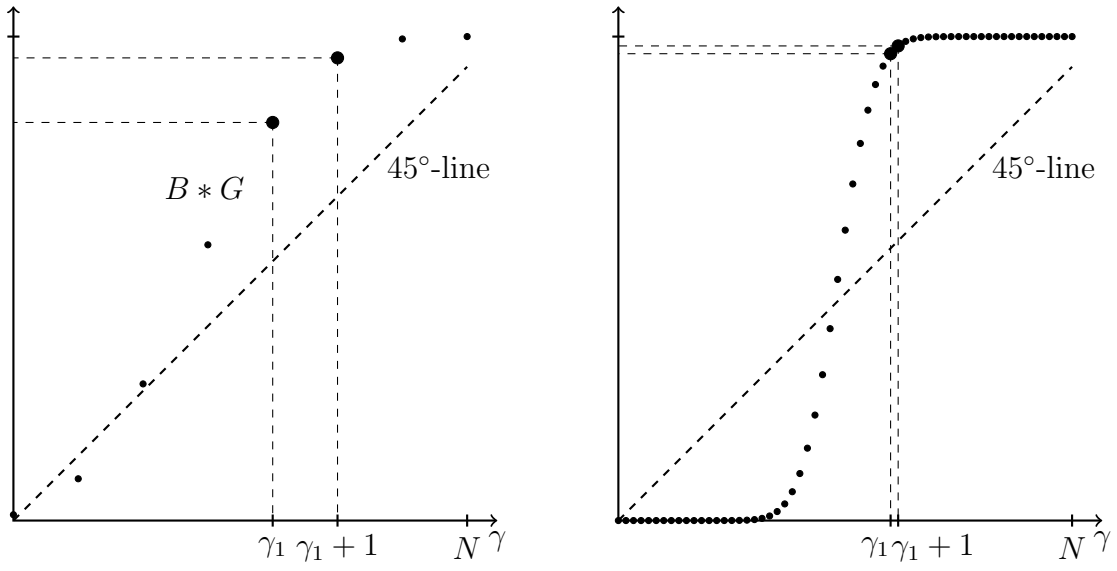
Result 2. *In every equilibrium, the prisoners mix over r and n. The warden mixes between some γ_1 and $\gamma_1 + 1$ in the warden optimal equilibrium. However, other equilibria (in which the warden mixes over γ_2 and $\gamma_2 + 1$ with $\gamma_2 > \gamma_1$ or the warden does not mix) can exist.*

3.4 Comparison of the models

The prisoners are indifferent between all models: In the transparency model and benchmark 1a, they did not revolt and therefore had a payoff of zero. In the panopticon and benchmark 1b, 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 can show that for large values of N, the panopticon model has a unique equilibrium 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 $N > \overline{N}$. Also, since the proof of the following theorem establishes that $G(0) \to 1$ in the unique equilibrium for $N \to \infty$, the probability of successful breakouts in the panopticon converges to zero.

Theorem 1. *Take b and q as given. Let N be sufficiently large and B such that assumption 1 is satisfied. Then, the warden mixes between 0 and 1 in the unique equilibrium of the panopticon model. The warden's payoff is – for N sufficiently high – higher in this equilibrium than in the transparency model.*
In the panopticon, the probability of a breakout is arbitrarily close to zero and $G_{N-1}(0)$ is arbitrarily close to one for sufficiently high N.



(A) small N (B) large N

Figure 2.4: An illustration of theorem 1.

After having derived the intermediate results about the panopticon model in section 3.3, 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 γ_1+1: (i) The warden must be indifferent between the guard levels, (ii) both guard levels must be better than all other guard levels, and (iii) the prisoners must be indifferent between revolting and not revolting. Figure 2.4 shows, similar to figure 2.3, a distribution G of attacking prisoners so that the first two requirements are fulfilled. In particular, by (2.2), a line through the points $(\gamma_1, BG(\gamma_1))$ and $(\gamma_1 + 1, BG(\gamma_1 + 1))$ would be parallel to the 45° line.

The third requirement can only be fulfilled if the probability of a successful revolt is sufficiently high, since it is otherwise optimal for the prisoners to never revolt. In panel (A), where N is relatively small, this is possible: There is a positive probability that the number of revolting prisoners is larger than γ_1. 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 around pN

(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) can not 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.

To get some more intuition for the uniqueness result in the panopticon, consider also – as an example – the equilibrium where the warden mixes over $N-1$ and N (which is used here because it is particularly tractable). The warden is only indifferent between the two guard levels if the marginal cost of adding the Nth guard, which is 1, equals the marginal benefit of reducing the probability of a breakout by increasing the guard level by one. This marginal benefit is $Bg(N) \geq (N+1)p^N$ where the inequality holds by assumption 1. Hence, $p^N \leq 1/(N+1)$ and $p \leq \sqrt[N]{1/(N+1)}$ in this equilibrium. Now consider the problem of a prisoner. In this equilibrium, he prefers to revolt only if all other prisoners revolt. The probability that all other prisoners revolt is $g_{N-1}(N-1) = p^{N-1}$. Since $p \leq \sqrt[N]{1/(N+1)}$ by the warden's indifference condition, we get $g_{N-1}(N-1) \leq (N+1)^{-\frac{N-1}{N}}$. This term converges to 0 for large N, so that it becomes extremely unlikely that there is a successful revolt. The prisoner therefore strictly prefers not revolting to revolting, i.e. $\Delta(N-1) < 0$. Consequently, there is no equilibrium where the warden mixes between $N-1$ and N for N sufficiently large. A similar logic applies to all other equilibria in which the warden mixes between $\gamma_1 \geq 1$ and $\gamma_1 + 1$: The warden's indifference condition requires a revolt probability p that is – for sufficiently large N – incompatible with the prisoner's indifference condition.

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 while the transparency model exploits this coordination problem only to a certain degree.

In the unique equilibrium for large N, the prisoners correctly believe that there is at most one guard on duty. Yet they still find it impossible to coordinate on attacking, even though an attack by just two prisoners would be successful for sure. Early readers of this paper have pointed out to us that this can seem "unrealistic": Should the prisoners not be able to implicitly coordinate on revolting, given that they know that the decision of the warden not to have more than one guard has already been made? We would like to point out that this argument amounts to a broad critique of the concept of equilibrium itself. We know of no epistemological argument that would distinguish between the situation where the warden and the prisoners reason simultaneously, and the one where they do so sequentially without knowing about the other's choice. Any argument that viewed

the prisoners' coordination problem as a subgame, however, would make precisely such a distinction.

Besides this central result for large groups, we present two 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. *Take q, b, N as given. If $\theta^* = 1$, then the warden is best off in the transparency model. 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 a bit 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. *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.

4 Discussion

4.1 Central Bank Defending Against Speculators

Our results can be applied to many situations of conflict where a central player can use the coordination problem of his opponents against them. An example that has received much attention in economics is the problem of defending a currency peg against speculators. The coordination aspect of this problem, which often leads to multiple equilibria, was pointed out by Flood and Garber (1984) and Obstfeld (1986). The equilibrium multiplicity resulting from the speculators' coordination problem was contentious until Morris and Shin (1998) established equilibrium uniqueness for each parameter value if speculators lack *common* knowledge about the strength of the currency and their beliefs get infected as in Carlsson and van Damme (1993). This insight has since been applied to other coordination problems like bank runs (Goldstein and Pauzner, 2005) or civil war (Chassang and Miquel, 2009). These models, however, concentrate on the coordination problem of the opponents. In fact, they would be equivalent to our transparency model if we did not allow the warden to choose the guard level but had this variable drawn from an exogenous distribution. The first difference to our model is therefore that the underlying "strength" (of the currency, the bank etc.) is exogenous in this literature while it is endogenous in our setup. The second difference is that we introduce the panoptical information structure and give the warden a choice between information structures. Our model allows us therefore to ask how the central player should use these instruments – strength level and information policy – to defend himself against the coordinated threat.

In the remainder of this section, we reinterpret our results in terms of the classical example of defending a currency peg against speculators as modeled in Morris and Shin

(1998). This should help to illustrate our results and to facillitate comparisons with the literature.

Consider the situation of a central bank that has to defend a currency peg against speculation. For this purpose, it can build up foreign exchange holdings that it can then use to counteract speculation. Doing so is costly, since it requires holding liquid bonds with low yields, so that the central bank would prefer to prevent a breaking of the peg with a minimum of reserves.

The transfer from our prison model is relatively straightforward. Assume that there are N speculators who can each decide to do nothing or to take a costly speculative position against the currency. Before the speculators make their choice, the central bank builds up foreign exchange reserves of size γ at cost γ. If there is a speculative attack against the currency and the central bank cannot defend the peg, its payoff is $-B-\gamma$ with some $B \geq N+1$, otherwise it receives $-\gamma$.

In this context, the assumption that $B \geq N+1$ means that our model only applies to cases where, if the central bank knew exactly the strength of the speculative attack that was coming, it would always prefer to build a large enough reserve to fight it off. We would argue that this is usually the case in the real world, and that in most cases where a central bank was overwhelmed by speculators it was because of the unexpected extent of the speculative attack.

A speculative attack is successful if more than γ out of the N speculators speculate against the currency. In that case, those who attacked the currency get a payoff of $b > 0$. If they speculate against the currency but the central bank can defend the peg, the speculators lose $q > 0$ on their positions. This loss q denotes the transaction costs of taking the speculating position and also includes the opportunity costs of forgoing an alternative investment. This alternative payoff, which speculators get if they do not speculate against the peg, is normalized to zero.

Should the central bank make its foreign exchange reserves public?[14] If the reserves are public, speculators are in the same situation as in Morris and Shin (1998) and – just as in their paper – we use the global game approach to select an equilibrium. If the reserves are kept secret, speculators and central bank find themselves in the panopticon model.

From our results in the previous sections, we can make several observations about which information policy the central bank should choose in revealing the size γ of the foreign exchange reserve. The optimal choice depends on the interplay of all parameter values, so that the following observations are *ceteris paribus*:

[14]We will interpret the publicity of information concerning the reserves as *transparency*. This term has occasionally been used in the literature on speculative attacks in a different way, see Heinemann and Illing (2002); Huang (2014). Huang analyzes a model where the central bank can have a behavioral type who always defends the peg. Speculators can learn the central bank's type over time and "transparency" refers to the precision of speculators' private signals about the central bank's type.

- If there are many speculators, the central bank should always choose to keep the reserve level secret.

- If speculators have a lot more to gain from breaking the peg than they can lose by speculating against the peg (in relation to the next-best investment), it may be optimal to keep the reserve level secret. This is especially the case if the cost (economic or reputational) of giving up the peg is high.[15] If the proportion between the speculators' possible earnings and their potential losses grows without bounds, however, the choice of information structure does not matter much since speculators are likely to speculate in any case and the reserve level always has to be maximal.

Especially, the first point, which follows directly from theorem 1, adds a new perspective to the literature on this topic. The uniqueness result of Morris and Shin (1998) has usually been understood to mean that a currency peg can be defended even in cases where coordination among all speculators could bring it down.

Our result shows that the central bank can make even better use of the speculators' coordination problem by keeping its own strength secret. Especially if there are many speculators (i.e. the coordination problem is worse), this will guarantee an extremely low probability of losing the peg with a minimal exertion of resources.[16] It should be noted, however, that the massive savings in costly reserves come at the cost of a strictly positive chance of the peg being broken. Observing a central bank that kept its reserves secret being overwhelmed by speculators would, therefore, not necessarily be a sign of a bad policy. While we know of no instance where a central bank actually maintained complete secrecy about the size of its reserves, secrecy about the existence and size of foreign exchange interventions is not uncommon. The reasons for this have been debated in the literature; see Vitale (2007) for a discussion.

4.2 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

[15] While it may seem like speculators usually have little to lose by speculating against a peg (because they can exchange their money back at the peg if they "lose"), this also includes the cost of transaction and any interest rate differential. Also, re-converting might not be costless if all speculators want to get out at the same time: When the pressure on the Danish krone/Euro peg let off in spring 2015, the Danish central bank suddenly had to stabilize the market *on the other side* of the peg since so many traders reversed or unwound their positions simultaneously.

[16] Theorem 1 might seem to imply that the currency reserves will be unrealistically low ("0" or "1") in equilibrium. However, "1" has to be interpreted as the highest budget of any speculator; see the extension in the supplementary material where we derive the panopticon equilibrium in a model where speculators can have varying budgets. This could be quite substantial – especially if some speculators are big institutional investors. Also, central banks might in reality hold some amount of currency reserves for other purposes than deterring speculators (providing liquidity etc.).

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.[17] 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

[17]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."

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 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 in the mixed equilibrium of the panopticon described above better off than in the transparency model if N is large.

5 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 results go further in arguing that absolute secrecy is often optimal. While secrecy is optimal for all larger groups, the transparency model may be optimal for smaller groups of opponents.

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.[18]

Our model suggests which is the optimal information structure in a conflict between one central player and a group. However, other situations are conceivable 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 the 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 without any problem of cheap talk. And, as we have shown in our benchmark model: when coordination problems do not matter, the warden chooses maximum enforcement in equilibrium.

[18] Practitioners of "fedspeak" have clearly understood this.

Appendix

Proofs transparency model

Proof of lemma 1. The proof is in three steps.

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

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

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

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

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

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

to g for ε small enough:

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

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

Using the approximation we get

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

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

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

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

Now we want to consider the expected utility difference between revolting and not revolting of a player receiving cutoff signal θ. If there is no integer $m \in \mathbb{N}$ such that $\theta-\varepsilon \leq m \leq \theta+\varepsilon$, then this utility difference equals $b-(q+b)\lfloor\theta\rfloor/N$ because a breakout cannot succeed if less than $\lfloor\theta\rfloor$ other prisoners play revolt.[19] 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

[19] 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.

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.[20] As $bN/(q + b) \in \mathbb{N}$ is clearly not true for generic parameter values (q, b, N), there exists a unique θ at which the expected utility difference is zero for generic parameter values. In the limit as $\varepsilon = 0$, we then have – for generic parameter values – that (i) the expected utility difference is strictly positive for $\theta < \theta^*$ and (ii) the expected utility difference is strictly negative for $\theta > \theta^*$. Note that (in the limit $\varepsilon \to 0$) the expected utility difference viewed as a function of θ is discontinuous at θ^*.

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

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

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

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

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 $\theta_1 = \theta_2 = \theta^*$. □

Proofs and limit results: Panopticon

After the proofs of the results in the main text, we derive another limit result (lemma 5) that we will use when comparing the different models.

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).[21] The binomial distribution G has the property that G is convex on $\{0, \ldots, \gamma^m\}$ and G is concave on $\{\gamma^m, \ldots, N\}$. Therefore, the maximization problem of the warden over the domain $\{0, \ldots, \gamma^m\}$ is convex and consequently only the boundary values 0 and γ^m can be local maxima (on this restricted domain). If we take $\{\gamma^m, \ldots, N\}$ as domain of the warden's maximization problem, the problem is concave and therefore (because γ takes integer values) this problem can have at most two local maxima γ_1 and γ_2 such that $\gamma_2 = \gamma_1 + 1$ (clearly, it could have only one local maximizer as well in which case we are already done). This implies that (2.1) has (at most) three local maxima: one at $\gamma_0 = 0$, γ_1 weakly above γ^m and possibly $\gamma_2 = \gamma_1 + 1$. Therefore, f's support will contain at most three elements.

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

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

g implies that $g(\gamma) > 1/B$ for all $\gamma \in \{1, \ldots, \gamma_1\}$ (recall that $g(\gamma_1 + 1) = 1/B$) which contradicts our earlier result that the average $g(\gamma)$ for $\gamma \in \{1, \ldots, \gamma_1\}$ is at most $1/B$.[22]

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 $g(\gamma_1 + 1) \leq 1/B$.[23] 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 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 $\gamma_1 + 1$ 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

$$\frac{G(\gamma_1 + 1) - G(0)}{\gamma_1 + 1} \geq \frac{1}{B} \tag{2.6}$$

[22] 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 $\gamma_0 = 0$.

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

(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 (2.6) 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 γ_2+1 over γ_1+1 under p_1 contradicting that γ_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, $\gamma_1 + 1$, γ_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 $p(\gamma' - 2)$.[24] If $\tilde{p}(\gamma' - 2)$ is below $p(\gamma' - 2)$, then there is a completely mixed

[24]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

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 $\Delta(0) > 0$. □

Lemma 5. *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 (2.2) and (2.1) as the warden's own mixing probability does not play a role in these conditions. Given these two values, (2.3) 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 (2.2) and (2.1). 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. □

Proofs model comparison

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 it is above the warden payoff in the transparency model. Last we will

as follows still applies.

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 $N \to \infty$.[25] 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 term in the limit can be written as

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

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

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

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

where the first four terms cancel each other out and the remaining expression only contains powers of N smaller than $2N$ in the numerator, so that the expression goes to zero as N gets large. Therefore, $\lim_{N\to\infty}(N+1)(1-(1-1/N^2)^N) = 1$ and the warden's payoff is bounded below by $-\alpha$ in the warden 0-1 mixed equilibrium for N sufficiently large. 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.[26]

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 $B(N,p)$ at γ. G_N is the corresponding cumulative distribution function; the definitions of g_{N-1} and G_{N-1} are analogous.

Lemma 6. *The probability $1 - G_N(\gamma)$ that $\gamma+1$ or more prisoners revolt in any equilibrium candidate (and therefore the probability of a breakout) converges to zero as N grows large.*

Proof. Using the Chernoff bound (Chernoff, 1952), we get

$$1 - G_N(\gamma) \leq \left(\frac{N}{\gamma+1}\right)^{\gamma+1}\left(\frac{N}{N-\gamma-1}\right)^{N-\gamma-1}p^{\gamma+1}(1-p)^{N-\gamma-1}. \tag{2.7}$$

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

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

where we plug the warden's indifference condition into (2.7). It is convenient to define

[26] 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 $\bar{\alpha}$.

$m = \gamma + 1$ as this allows to write the previous expression as

$$1 - G_N(\gamma) \leq \frac{N^N}{\binom{N}{m} m^m (N-m)^{N-m} \alpha (N+1)}. \tag{2.8}$$

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 (2.8) 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. $m = N/2$.[27] 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 (2.8) becomes

$$\begin{aligned} 1 - G_N(\gamma) &\leq \frac{N^N}{\binom{N}{N/2}(N/2)^N \alpha (N+1)} \\ &= \frac{2^N}{\binom{N}{N/2} \alpha (N+1)}. \end{aligned} \tag{2.9}$$

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 $1 - G(\gamma)$ converges to zero in any equilibrium candidate. □

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

Lemma 7. *In any equilibrium candidate with $\gamma \geq 1$, $1 - G_{N-1}(\gamma - 1)$ converges to zero as N grows large.*

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 $1 - G(\gamma)$ converges to zero in any equilibrium candidate. If $g_{N-1}(\gamma)$ converges to zero 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 converge to zero.

[27] 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.

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 converge to zero, 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

$$\gamma(N) - p(N)N + p(N) \geq \mu p(N)N^2 - p(N)N + p(N) - 1 = p(N)N^{5/4} * \left(\mu N^{3/4} - \frac{1}{N^{1/4}}\right) + p(N) - 1. \tag{2.10}$$

We will look at two cases. First, $p(N)N^{5/4}$ does not converge to zero. Then the right hand side of (2.10) 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. □

Lemma 7 implies that $G_{N-1}(\gamma - 1)$ converges to one for any equilibrium candidate with $\gamma \geq 1$ as N gets large. Put differently, for any $\varepsilon > 0$, we can find an $\bar{N}(\varepsilon)$ such that $G_{N-1}(\gamma_1) > 1 - \varepsilon$ for all $N \geq \bar{N}(\varepsilon)$ and all equilibrium candidates with $\gamma \geq 1$. In particular, we can find such an $\bar{N}(\varepsilon^*)$ for $\varepsilon^* = 1 - b/(q + b)$. For $N \geq \bar{N}(\varepsilon^*)$, 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. □

Proof of proposition 1. Lemma 5 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 5 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 $-N$. □

6 Appendix: No asymmetric equilibria in the panopticon

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.

7 Appendix: 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 1 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 1 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 1 apply without change.

In the **panopticon model**, the indifference condition of the prisoner (2.3) 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 5 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.

8 Appendix: 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.[28] The model of the main text emerges for $\beta = 1$. In this setup, it is necessary to adjust assumption 1 which implies that the warden would prevent a breakout if he knew that all prisoners revolt with probability one. In the setup with stochastic breakouts, the assumption is $\beta B \geq N + 1$. We will need additional parameter assumptions in order to ensure that prisoners have dominant

[28] 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.

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 1 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 1 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 (2.1) 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 (2.3) 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 5 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.

9 Appendix: Heterogenous 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 heterogenous 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 1 where N_K is the number of attackers of type K:

Lemma 8. *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 1 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 homogenous of type K. The second step of the proof of lemma 1 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 whenver 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.

10 Appendix: 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

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

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 2.5 for existence of the different equilibria.

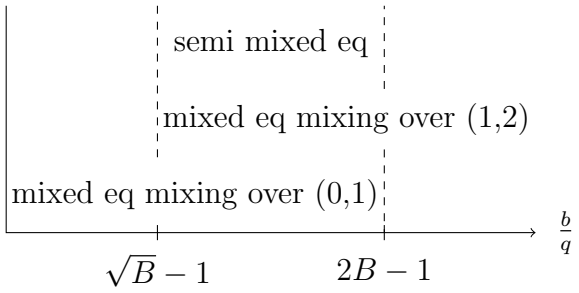


Figure 2.5: Equilibria for N=2 case

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

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

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.[29] 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.

[29] 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.

Chapter 3

Risk Capacity and the Chicken Game[1]

Ole Jann

Information asymmetries between firms and lenders mean that a firm would often rather turn for financing to another firm in the same industry or region, who can more easily appraise the value of the firm's assets. This leads to strategic substitutes in risk-taking: the more risk other firms take, the more attractive it becomes to hoard cash and then acquire assets in the event of a crisis. I examine the consequences of these strategic substitutes: There can be an inefficient race to risk between firms, and even if coordination does not fail fire sales happen in equilibrium. Regulations on the interest rate or asset prices can make equilibrium unattainable.

[1]Part of this paper is based on a note that I wrote in Martin Gonzalez Eiras' course "Financial Frictions, Liquidity and the Business Cycle." I am grateful for comments by Martin Gonzalez Eiras, Christoph Schottmüller and Peter Norman Sørensen.

1 Introduction

Economic projects often rely on acquiring outside funding, and outside funding coupled with limited liability comes with problems of asymmetric information. Financiers have limited control over what their money is used on, and it could be wasted by the borrowing firm on projects that benefit the firm's owners but increase the chance of default.

A possible solution to this principal-agent-problem is to channel financing through informational insiders who have equity to lose in the process – i.e., instead of lending to a firm with unknown quality, lending to an insider who then uses his knowledge to lend to a firm at his own risk, thus making sure that the money eventually goes to a productive use. This channeling can either occur through indirect lending or through allowing insiders to actually acquire the assets of failing firms.

But this method comes with a problem at the linkage between insiders and outsiders, too: The advent of a crisis which requires firms to seek outside funding might also impose borrowing constraints on those insiders through which funds are channeled. Consider the following story, in the style of Shleifer and Vishny (1992): Two farmers are in identical economic positions. Each of them can decide to modernize the farm by using his money to buy new machinery (with or without taking a loan for the expense), which promises superior returns in the future. If a crisis hits, however, a modernizing farmer cannot meet his expenses and needs to sell his land. If the other farmer has also invested in modernization, they both fail and have to sell at a low price to outside investors. If, on the other hand, only one of the farmers invested in modernization, the other can use his own money (and possibly leverage his equity by borrowing more) to buy the liquidated assets at a high price.

The risk-taking decisions of the farmers are strategic substitutes, as each farmer prefers to do the opposite of the other. The economy (which consists of the two farms) has a risk capacity: If both farmers engage in risky modernization, their whole sector stands to fail painfully in the event of a crisis. But it is also optimal for the economy to operate at the risk capacity limit: If none of the farmers modernizes, their combined output will continue to stay low.

There are three theoretical contributions of this paper: First, to establish how the adverse selection problems of outside financing leads to a risk capacity limit. It is both individually rational and pareto-optimal for the firms to keep the economy below the risk capacity limit, but it is worse than the informational first-best. Second, I examine how the multiple equilibria at the risk capacity limit give rise to a socially inefficient "race to risk" among firms. Finally, in an economy with many firms, fire sales of assets occur randomly and despite all firms behaving optimally.

Consider an economy where two firms can each decide to engage in a risky, but promising project or just to store their money. With some probability, the economy is

hit by a crisis in which case the project turns either good or bad. Good projects are still profitable, while bad projects only produce non-transferable returns for their owners. At the same time, a crisis means that all projects need extra cash to continue.

In a crisis, outside financiers are reluctant to lend such extra cash, since they do not know whether the firm they are lending to has a good or a bad project. Other firms in the same economy, however, can verify the quality of a project, so that a firm that has stored money can buy good projects. Since this possibility exists, firms with good projects will take it and firms who turn directly to outside investors will automatically be assumed to have bad projects and will receive no funding. The only funding channel available is therefore through firms that hoarded cash.

Depending on the liquidation price of good projects, it can be much more attractive to be the firm taking the risk (and being rewarded if there is no crisis) than being the firm that hoards cash for the remote possibility of a crisis, even though both firms find it optimal to stick to their choice. The firms will therefore both want to convince the other that they will take the risk, and that the other should hoard cash – leading to a "race to risk" in which the firms dispose of cash and which can leave all firms worse off.

Outside investors, who may lend to inside firms so that the firms can buy projects, are worried that the firm might acquire lots of bad projects and reap short-term, non-transferable benefits from those, and subsequently default. They therefore require that a firm uses some amount of its own cash for every acquisition that it makes – that it has "skin in the game". That means that a firm that has hoarded cash can, even when it goes to the limit of its borrowing capability, only buy a limited number of good projects. Since the overall number of projects that turn out bad or good in a crisis is not ex ante predictable, this can produce a mismatch between good projects that are for sale and firms that have the ability to buy them. If few good projects are for sale, they are sold at high prices. If the number of good projects for sale is high, a fire sale ensues and prices collapse. Such a collapse is unpredictable and occurs with positive probability if all participants behave optimally.

This paper takes up ideas from several strands of literature. Strategic substitutes in the decisions of firms also appear in the theory of Shleifer and Vishny (1992), who describe a "debt capacity" that also arises from a multiple-equilibrium structure, albeit under the assumption that firms need to take up debt to undertake their projects. The firms in their model are ex ante different, so that it is clear who will be buyer and seller in a crisis and they abstract from the resulting coordination problem. The effects in this paper, which arise because of the identical positions of the firms, are therefore absent. Perotti and Suarez (2002) discuss failed banks being taken over by their competitors, which allows the surviving bank to acquire cheap capital and profit from higher market concentration. They also note that this "last bank standing effect" creates a desirable strategic substitutability as "temporary consolidation in the aftermath of a crisis has the

ex ante desirable effect of promoting stability by rewarding those banks that remained solvent during the crisis."

Shleifer and Vishny (1997) also describe a "hold back effect" where arbitrageurs hoard cash in the hope that mispricings could deepen, for example through fire sales. A more general discussion and an overview of the empirical evidence on banks "keeping their powder dry" in the anticipation of fire sales can be found in Shleifer and Vishny (2011).

In analyzing the multiple-equilibria structure of this model and the effects of incomplete information, I make use of the results of Carlsson and van Damme (1993) on global games, where incomplete information leads to the selection of the risk-dominant equilibrium. A related application of these methods on "endogenous leadership" is by Hurkens and van Damme (2004).

2 The Model with Two Firms

Consider a three-period economy. There are two firms who each have 1 unit of cash. At $t = 0$, they make an investment choice. At $t = 1$, they might have financing needs, for which they can apply for funding from banks.

The firms can either store their money, or they can engage in a risky project. The risk-free storage makes all their money available in periods 1 and 2 with certainty, while the risky project requires an initial investment of 1 but can provide an attractive return of $R > 1$ after two periods. The firms aim to maximize their cash holdings at $t = 2$.

In $t = 1$, a crisis occurs with probability θ. The crisis has two effects: Firstly, every risky project will turn "bad" with some probability λ and will stay "good" otherwise. At $t = 2$, both projects provide a benefit. But a bad project only provides a non-transferable (i.e. non-monetary) private benefit of b to the firm that owns the project, while a good project returns R money units to whoever owns the project at $t = 2$. Only firms can observe which projects have turned bad – banks cannot observe this and it can not be credibly communicated to them. Secondly, the crisis also means that each project (regardless of being good or bad) requires an additional cash payment of c to be made at $t = 1$. If this payment is made, the project continues and returns R or b at $t = 2$. If the payment is not made, the project is closed down immediately and without any return.

At $t = 1$, it is possible to lend from banks at some per-period interest rate r (so that when borrowing 1 unit in $t = 1$, $1 + r$ units would have to be repaid in $t = 2$). We assume that r is determined by (international) capital markets and is therefore taken as given. Furthermore, r is relatively small and the banks have the alternative to just store their money, so that they are not willing to accept large risks by lending out (otherwise there would beo no problem with asymmetric information). In particular, we assume that $r < \frac{\lambda}{1-\lambda}$, i.e. the interest rate is small compared to the proportion of projects that

turn bad.[2] This means that if a firm that has a risky project comes to a bank in $t = 1$ and asks to borrow c, and the bank has the prior belief that the project is bad with probability λ, the bank will not lend. This introduces the adverse-selection problem: Even firms who have good projects will find themselves in a situation where they have a project that provides a good return, but they cannot raise the money they would have to spend to keep the project going.

However, we assume that other firms can costlessly observe whether projects have turned bad or good, since they are informational insiders – as opposed to the financiers or banks, who are outsiders. A firm that has chosen to store its cash, and therefore has it available in $t = 1$, can now use its money to buy a project from another firm – and it can avoid buying bad projects, since it can observe project quality. Moreover, it can even raise money from outside banks for this undertaking, thus potentially leveraging its own capital of 1 for the purpose of buying several projects if there were several other firms.

There are two possible ways financing of c could be arranged: Either the firm that previously engaged in storage can use its own cash to lend c to firms with good projects. Or the firm with a good project, which cannot meet its financing need of c, can liquidate and sell all their assets to the firm that previously engaged in storage, who also borrows from a bank to finance the purchase. Either way, the financing problem that is created by the combination of illiquidity and asymmetric information is solved. Actual loss only occurs through the exogenous event of a project turning bad, not through the financing problem.

This solution, however, requires ex ante coordination between the firms. If both firms engage in storage, they will both only have 1 unit of cash in $t = 2$, as much as they started with. If both firms start a risky project, they have no way of selling their assets to another firm in period 1 in case of a crisis, and they both have to shut down their project and get 0 regardless of whether their projects turned out good or bad. The socially optimal case (and also the case preferred by both firms to any symmetrical outcome) is the one where exactly one firm engages in the risky project, while the other engages in storage and stands ready to acquire the failing firms' assets if a crisis occurs.

The payoffs are shown in of figure 3.1. $A(p)$ and $B(p)$ are simple placeholders for how the two firms distribute the surplus generated by their trade by deciding on a price p for the assets. The individual rationality restrictions are $A(p) \geq 0$ and $B(p) \geq 1$, and $A(p) + B(p) \leq R - \max\{p + c - 1, 0\} - c$. Depending on the parameters of the model, this game has several possible equilibrium structures.

1. The project could be so unpromising (R low) or crises so likely (θ high) or severe

[2] Yet not terribly small – if half of all projects turn bad in a crisis, it simply means that the per-period interest rate is less than 100%.

		Firm 2	
		Storage	Project
Firm 1	Storage	1,1	$1-\theta+\theta\lambda+\theta(1-\lambda)B(p), (1-\theta)R+\theta(1-\lambda)A(p)$
	Project	$(1-\theta)R+\theta(1-\lambda)A(p), 1-\theta+\theta\lambda+\theta(1-\lambda)B(p)$	$(1-\theta)R, (1-\theta)R$

Figure 3.1: The game played by two firms.

(λ or c high) that both firms always prefer Storage and (Storage, Storage) is the unique Nash equilibrium.

2. The project could be so promising (R high) or crises so unlikely (θ low) that both firms always prefer Project, despite the certainty of going bankrupt in a crisis, and (Project, Project) is the unique Nash equilibrium.

3. There is an intermediate parameter space where the firms prefer to anti-coordinate so that one firm chooses Project and the other chooses Storage. Then one firm takes the risk, while the other stands ready to buy the assets in the case of a crises. This gives the game a chicken-like structure, and there are three Nash equilibria: (i) (Project, Storage), (ii) (Storage, Project) and (iii) a mixed-strategy equilibrium where both firms place positive probability on both strategies. This mixed strategy equilibrium (iii), however, is strictly pareto-inferior to any of the two pure-strategy equilibria, since the firms land in (Project, Project) and (Storage, Storage) with positive probability.

4. (There is also a fourth possible equilibrium structure, in which crises are so likely and the terms of liquidation are so bad that it is only attractive to undertake a project if the other firm is also doing so.)

The first and the second case are not very interesting: Each firm has an optimal choice regardless of what the other firm does, and the risky project is either so unattractive or attractive that either no firm or both firms engage in it. In the fourth case, the firms' actions are strategic complements – we will ignore this case here as we want to focus on the interplay of symmetry and strategic substitutes.

In the third case the firms' optimal choices are interdependent and it is socially optimal for only one firm to engage in the project. This gives rise to the risk capacity constraint in its simplest form. Figure 3.2 shows the R-θ-parameter space, and where the different equilibrium structures lie. We are interested in the shaded area, where firms' choices are strategic substitutes; note that there is a dominance region for each action.

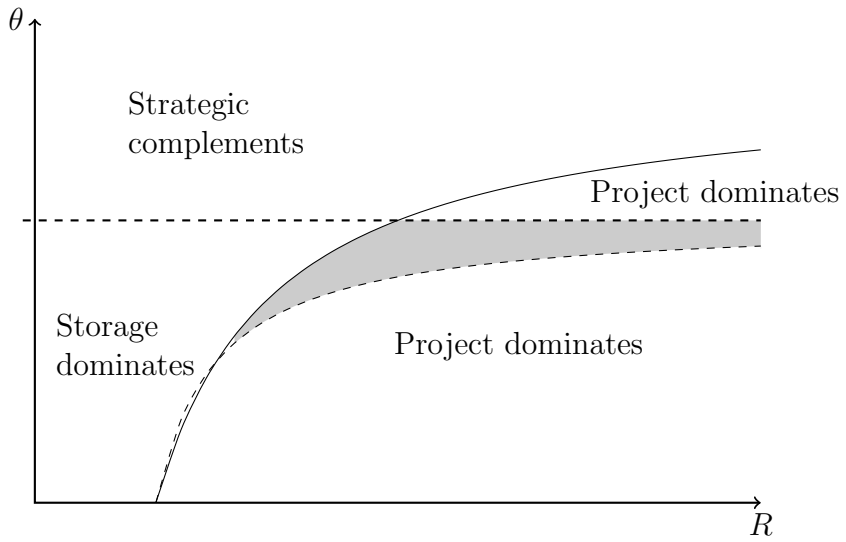


Figure 3.2: Storage is best reponse to Storage to the left of the solid curve, and best response to Project inside the dashed lines. The shaded area indicates the parameter space in which the game has a chicken structure.

3 Symmetry and the Race to Risk

3.1 Symmetric equilibria but opposed preferences

The risk capacity constraint is symmetric; it doesn't tell us which of the asymmetric equilibria the firms will choose. However, the firms have strictly opposing preferences over the two pareto-optimal (pure-strategy) equilibria. Given the choice of the other firm, they prefer to do the opposite – but if they could choose first, they have a preferred choice. In this way, their situation is comparable to that of a Stackelberg competition, where the leader can make a larger profit by forcing the follower to optimally make a more defensive choice.

Let us assume in the following that the parameters and the price-setting are such that it is more attractive to be the firm undertaking the risky project than the one storing money. Intuitively, the project has a large upside, while the inherent risk is mitigated by the fact that the assets from a good project can still be sold to an insider. Especially if crises are sufficiently unlikely, this should make the (Project, Storage) pairing more attractive to firm 1 than (Storage, Project).

Under this assumption, each firm wants to convince the other that it is going to choose Project, so that the other chooses Storage. We have so far assumed that decisions are made simultaneously. If one firm were able to go first, the implication would be clear: That firm would choose Project, leaving the other to settle for Storage. Since each firm would want to go first, this would lead to a race to be the first to decide.

3.2 Equilibrium selection through incomplete information?

A large literature (starting with Carlsson and van Damme, 1993) has shown that in games with several strict equilibria, lack of common knowledge among the players can select one of the equilibria.

Consider the following incomplete-information game: Both firms believe that the parameter tuple $x = (R, \theta, \lambda, c)$ is distributed on the space $(0, \infty) \times (0, 1) \times (0, 1) \times (0, \infty)$ with strictly positive density everywhere that is continuously differentiable. Before making their (simultaneous) choices, firm i observes a signal x_i^ε. These signals are independently uniformly drawn from the interval $[x - \varepsilon\eta, x + \varepsilon\eta]$, where $\eta \in \mathbb{R}^4$ lies in a ball with radius 1 around 0 and ε is a scale parameter; we are interested in the cases where ε is very small.

These assumptions map to the assumptions of Carlsson and van Damme (1993), we would expect their main result to apply. We can see, however, that in the limit of this incomplete-information game for ε small, the space of rationalizable strategies is the same as in the complete-information game. (Recall that in the classical uses of equilibrium selection by global games, the set of rationalizable strategy profiles is shrunk down to a single point, which is the unique equilibrium.)

To understand why global games has no bite here, consider the concept of risk-dominance, with which the global games criterion is inseparably intertwined. There are several (equivalent) definitions of risk-dominance, but given the symmetry of the game it is easy to see that there will neither be a difference in the product of deviation losses (the definition used by Harsanyi and Selten 1988) or the sum of probabilities that players would put on pure strategies in the mixed equilibrium (the definition used by Carlsson and van Damme 1993). The two pure equilibria weakly risk-dominate each other, and risk-dominance therefore doesn't allow us to select an equilibrium here.[3]

We can see directly how this makes the main proof of Carlsson and van Damme (1993) unapplicable: Since the belief about j playing Storage that makes i indifferent between his two options is exactly the same as the belief about i playing Storage that would make j indifferent, we get (in the notation of their proof) $\bar{s}_2(x^*) + \bar{s}_1(x^*) = 1$, which means that the proof no longer leads to a contradiction on p. 1003.

Intuitively, we can think of every parametrized 2×2 game as having a "knife-edge" parameter configuration where no pure equilibrium is risk-dominant. (In Carlsson and van Damme's leading example, this is the case at $x = 2$.) The "infection" of beliefs that occur without common knowledge (where each player worries about what the other player will do, given that he worries about what the other player will do and so on) makes one action unrationalizable for each point in the parameter space that is not on

[3] We disregard the axiomatic postulate by Harsanyi and Selten, p. 88, who say that in this case the mixed equilibrium should be considered risk-dominant.

		Firm 2	
		Storage	Project
Firm 1	Storage	$1-\delta, 1$	$1-\theta+\theta\lambda+\theta(1-\lambda)B(p)-\delta, (1-\theta)R+\theta(1-\lambda)A(p)$
	Project	$(1-\theta)R+\theta(1-\lambda)A(p), 1-\theta+\theta\lambda+\theta(1-\lambda)B(p)$	$(1-\theta)R, (1-\theta)R$

Figure 3.3: The payoffs with a small asymmetry $\delta > 0$.

the knife-edge, so that the knife-edge becomes the threshold at which players change their behavior. In any stag-hunt game, for example, the knife-edge is non-generic.

In the symmetric game that we face here, however, the whole parameter space is on the knife-edge, so to speak, since the pure equilibria weakly risk-dominate each other for any parameter value for which they exist.

3.3 Small asymmetries allow for equilibrium selection

Even though lack of common knowledge does not allow us to select a unique equilibrium, it can magnify small asymmetries which otherwise do not change the equilibrium structure of the complete-information game.

Consider the same game as above, but now assume that storage is slightly less attractive for firm 1 (see figure 3.3). If the parameters are common knowledge, there are still two pure and one mixed equilibria. But everything has changed in the incomplete-information game. Since the symmetry is broken, (Project, Storage) now has a slightly higher product of deviation losses. This makes it the risk-dominant equilibrium of the complete-information and therefore the unique equilibrium of the incomplete-information game.

Note that, given our assumption that firm 1 prefers to (Project, Storage) to (Storage, Project), this equilibrium selection insures that the firms will play firm 1's preferred equilibrium. Firm 1 is hence better off by giving up some potential payoff – and, what is more, it can actually use this fact to influence equilibrium selection by pre-emptively disposing of some of the utility it can receive from storage. This could take the form of imparing its own storage technology, contractually signing away some of the stored money or staking their reputation on choosing project, or any number of other things.[4] It is does not even matter how large δ is, since it is payoff-irrelevant in the (Project,Storage) equilibrium.

Of course, our analysis would not be complete if we did not consider the possibility that *both* firms can choose to make such a disposal before actually having to decide between Storage and Project. Consider a dynamic game where in period 1, the firms simultaneously choose to lower their own payoff from Storage by δ_i, and in the second period they observe the payoff matrix given in panel A of figure 3.4 and choose simultane-

[4]See Chassang (2008) for an analysis of the global games approach in settings where the structure of the game itself is endogenous.

		Firm 2: Storage	Firm 2: Project
Firm 1	Storage	$1-\delta_1, 1-\delta_2$	$1-\theta+\theta\lambda+\theta(1-\lambda)B(p)-\delta_1, (1-\theta)R+\theta(1-\lambda)A(p)$
	Project	$(1-\theta)R+\theta(1-\lambda)A(p), 1-\theta+\theta\lambda+\theta(1-\lambda)B(p)-\delta_2$	$(1-\theta)R, (1-\theta)R$

(A)

		Firm 2: Storage	Firm 2: Project
Firm 1	Storage	0,0	$-\theta+\theta\lambda+\theta(1-\lambda)B(p), (1-\theta)R+\theta(1-\lambda)A(p)$
	Project	$(1-\theta)R+\theta(1-\lambda)A(p), -\theta+\theta\lambda+\theta(1-\lambda)B(p)$	$(1-\theta)R, (1-\theta)R$

(B)

Figure 3.4: If both firms can decide to dispose of δ_i in case they choose Storage.

ously between Storage and Project. This throws the firms into a game that is somewhat like a "loser pays" auction (or an all-pay auction), where the "winner" (the one with higher δ_i) gets to pick his preferred option (Project). This is a "race to risk": Each wants to be the one that gets to take the risk in the second-period subgame, and both are willing to dispose of ressources to do so.

The precise equilibrium of such a dynamic game depends on several factors. One of them is the difference between the payoffs from (Storage,Project) and (Project,Project) for player 1 (and the symmetric difference for player 2). This determines whether the firms can dispose of so much that Project becomes a dominant strategy for them, or whether the equilibrium structure of the second-period subgame persists independently of δ_1 and δ_2. In the latter case we would have to specify which subgame equilibrium is played if $\delta_1 = \delta_2$. If we assume that firms play the mixed equilibrium if the period 2 subgame is symmetric and that Project does not become dominant for $\delta_i = 1$, for example, firms will mix between all $\delta_i \in [0, 1]$ and put positive probability on $\delta_i = 1$.

Regardless of the precise equilibrium of the dynamic game, we can easily see that there are no equilibria in which both firms always choose $\delta_i = 0$. Given our assumptions, any $\delta_j > 0$ would be a costless best-response to this that guarantees the preferred equilibrium in period 2. Regardless of the exact parameter values, the race to risk among firms therefore makes the economy worse off, as firms dispose of possible profits in order to put themselves in a better position by jumping along the risk capacity constraint line. No firm is better off, and both firms might very well be worse off (for example in the case where they both choose play $\delta_i = \delta_j = 1$ and Project becomes dominant, or they both play Project with positive probability in the second period – both cases are followed by fire sales with positive probability). The "small" disposal that would give firms an edge in the chicken-like risk capacity game has therefore had a similar effect as any other way to get an advantage in the classical chicken game: If both drivers disable their steering wheels, for example, they will simply crash.

4 Many Firms and Endogenous Price Setting

4.1 The Model with n Firms

Now assume that there are not two firms, but n firms which simultaneously have to decide on which project to choose. The parameters of the model are the same, and again we allow a firm engaged in storage to use their cash to buy the assets of a firm that engaged in the project. Extrapolating from the case of two firms, one could expect that the information asymmetry problems that the firms face can be solved by one firm choosing Storage, and everybody else choosing Project. The single firm could then potentially use its own stored money, and money it has borrowed, to buy assets of several failed firms.

If insider firms only buy good projects, outside banks should be willing to lend to insider firms without restrictions. But because of the non-monetary utility b of owning a bad project, an insider company could simply borrow a lot of money, buy an amount m of bad projects, and afterwards default and get a payoff of mb which is larger than 1 if m is large enough. Banks therefore need to make sure that for every project that is bought with borrowed money, the firm has enough "skin in the game" to induce it to buy only good projects. In this model, that means that for every project acquired, the firm has to use at least b of its own money to convince banks that it is not buying bad projects.

This restriction limits the number of good projects any single firm can buy to $\lfloor \frac{1}{b} \rfloor$. In equilibrium, bad projects will never be traded since it doesn't pay off for a firm to use at least b of its own money to buy a project worth b. The equilibrium price p of good projects in a crisis is determined by demand and supply, since now the number of buyers and sellers can be mismatched in several ways. Let B and S be the number of buyers and sellers, i.e. the number of firms who chose storage and the number of fims who chose Project and ended up with a good project. If we simply assume that all projects are traded at one price, Bertrand competition gives the following pricing structure:

- If $\frac{S}{B} < \lfloor \frac{1}{b} \rfloor$, sellers can capture the whole surplus since some buyers would be able to borrow and buy more, and $p = \frac{R+(1+r)(1-c)}{2+r}$.
- If $\frac{S}{B} > \lfloor \frac{1}{b} \rfloor$, buyers can capture the whole surplus since some firms with good projects are not able to find a buyer, and $p = 0$.
- If $\frac{S}{B} = \lfloor \frac{1}{b} \rfloor$, any price between $\frac{R+(1+r)(1-c)}{2+r}$ and 0 that the buyers and sellers can agree on is stable.

Since B is simply the number of firms that chose Storage and $E[S] = (1-\lambda)(n-B)$, $E[p]$ is decreasing in the number of firms who choose Project. Let $\Delta(B, n)$ be the difference in expected utility between choosing Project and Storage if B out of n fims choose Storage.

From the assumptions of the model it follows that $\Delta(n, n) > 0$, since a single firm could choose Project and be sure to sell its asset at a maximum price if there is a crisis and the project is good. But $\Delta(0, n) < 0$, since the firms choosing Project would lose everything in a crisis. Thus there exists some B^* such that $\Delta(B^*, n) < 0$ and $\Delta(B^* + 1, n) > 0$.[5] There can be no equilibrium with $B \neq B^*, B^* + 1$, since some firms could profitably change their choice. In equilibrium, it is therefore B^* or $B^* + 1$ firms that store their money while the remaining firms engage in the project. As in the two-firm game, this means that there is a set of $\binom{B^*}{n}$ or $\binom{B^*+1}{n}$ equilibria in pure strategies.

Now, however, it matters that the true realization of S in a crisis is stochastic. When there were only two firms, there would either be a firm with a good project or a firm with a bad project. Now the number of firms with good projects can be smaller or larger than the number of firms that can be bought, so that Bertrand competition pushes the price up and down.[6] Intermediate prices can only exist in the case where $S = B$, which is exceedingly unlikely as n grows lagers.

If the risk capacity constraint is at work, we will therefore see that some firms take risks while others hoard cash, so that the economy as a whole nears the risk capacity limit in expectation. When a crisis hits and the true states of the projects are realized, there are either so many good projects that they need to be sold off at fire-sale prices, or the firms have been hoarding so much cash that failing firms can comfortably sell of their projects at high prices. It is not possible to tell in advance which will occur, and in equilibrium there can either be consolidations and fire sales, where both occur with positive probability. Fire sales will therefore occur in equilbrium.

4.2 An Exogenous Price and a Continuum of Firms

Given that the coordination problem between firms stems from the fact that there is a finite number of firms, and the number of failed projects is therefore stochastic, we could assume that in the limit as n grows larger, the coordination problem disappears.[7] In this section, I show that while this is the case, an infinite number of firms can lead to a different problems, as there does not necessarily exist an equilibrium if the liquidation price of assets cannot be chosen freely.

The considerations of the preceding section were under the assumption that the single firm considers its influence on the proportion between B and S (and therefore the price of the project in $t = 1$) when making a its choice. The equilibria were stable because in every equilibrium, those firms that chose storage knew that choosing the project instead would raise the number of projects to a point where the expected price of the project would be sufficiently low to make the payoff from project worse than from storage.

[5] There could be equality on only one of these expressions.

[6] To be precise, note that S is binomially distributed according to $\mathbb{B}(B - n, 1 - \lambda)$.

[7] Cf Bolton and Farrell (1990), where all coordination problems appear in the limit.

Similarly, firms undertaking the project in any equilibrium received a superior or equal payoff to those engaging in storage and were hence also not interested in changing their strategy.

This is not the case in very a large economy with small firms, where every firm has no influence on the overall proportions and hence takes the price of the project in $t = 1$ as given. Assume that the model is as before, except that there is now a continuum of firms with measure 1. If a proportion β of firms chooses storage, a proportion of $(1-\lambda)(1-\beta)$ of firms will end up with good projects. Every firm that has stored money can still buy $\left\lfloor \frac{1}{b} \right\rfloor$ in a crisis, so that in equilibrium it should be $\beta \left\lfloor \frac{1}{b} \right\rfloor = (1-\lambda)(1-\beta)$ or

$$\beta^* = \frac{1-\lambda}{\left\lfloor \frac{1}{b} \right\rfloor + 1 - \lambda}.$$

At this proportion, there are β^* buying firms and $(1-\lambda)(1-\beta^*)$ selling firms, so that each buyer goes to the maximum amount that he can borrow and each firm with a good project is able to sell it. Since buyers and sellers are so exactly matched, they can negotiate any price, and the equilibrium price at which no firm wants to change its strategy is

$$p^* = \frac{R - 1 + \theta - \lambda\theta - \theta r + \lambda\theta r - \theta R + \left\lfloor \frac{1}{b} \right\rfloor (\theta c - \theta\lambda c + \theta cr - \lambda\theta cr - \theta R + \lambda\theta R)}{(\lambda - 1)\theta(1 + (1+r)\left\lfloor \frac{1}{b} \right\rfloor)}.$$

Unlike in the model with n firms, the price now matters very much. If some exogenous factor means that good projects have to be traded at any price that is different from p^*, there is no equilibrium.

Proposition 1. *If $p \neq p^*$, there is no Nash equilibrium.*

Proof. Assume that there was a Nash equilibrium in which a proportion β of firms chose storage and the remaining firms started a project. Then the expected profit of choosing storage or project must be the same. This cannot be the case if $\beta \neq \beta^*$, since then all the surplus from trading the project goes either to the firms engaged in storage or in the project, and all firms in the other group would optimally like to switch strategies. But for β^*, where buyers and sellers are matched, the price $p \neq p^*$ also means that one group makes a higher expected profit and that all firms in the other group could therefore gain by changing their strategy. □

If we consider an economy where, for example, the regulatory framework is such that in selling (or liquidating) a company, buyers and seller cannot freely choose the price at which they trade, this economy will not find a stable equilibrium. There will always be firms that choose to take a risk, or firms that choose not take a risk and "stay behind"

to organize the financing and the consolidation in case the risky firms fail, who would prefer if they had chosen differently.

To be sure, the economy is always driven towards what would be the equilibrium: If $\beta < \beta^*$, i.e. there are too few storage firms, firms that engage in the project will expect to do badly and therefore have an incentive to change their strategy. The converse occurs when $\beta > \beta^*$. But there is no consistent alignment of beliefs at which all firms will be happy with their choice.

This non-existence of equilibrium is a direct result of the fact that, in a continuum of players, no single player has to consider the influence of his choice on the system as a whole, or on the probability distribution of prices. The argument is thus in a way similar to that of Lorenzoni (2008), where ex-ante inefficient credit booms can occur in equilibrium because continuum agents do not consider the consequences of their actions.

5 Discussion and Conclusion

A central assumption of this paper is that firms in the "economy" can observe the quality of projects, while those outside cannot. What is the "economy" that the model applies to? Its most important characteristic is that while insiders can observe the value of each other's assets, this is difficult to do for outside financiers, who can only rely on some prior distribution. Outsiders who want to lend or otherwise get engaged can only do so through an inside partner, who can help them pick worthy projects and assets. Crucially, this also leads to an adverse selection: An outsider who is approached by an insider with a business proposition will not apply his prior belief, but must assume that the offer is worthless or disingenuous, since the insider would have turned to another insider for indirect financing if his offer was any good.

We could therefore think of the economy of this model being a region, such as Eastern Europe or Sub-Saharan Africa, that can be hard for outside investors to understand. Or we could think of an industry in which it can be hard for outsiders to tell profitable assets from non-profitable ones, such as farming or internet companies. Both a region or an industry can be in dire need of outside financing, and investment can be ex ante worthwhile (R large enough), but the risk capacity constraint inefficiently limits outside investments and leads to a race to risk and random fire sales.

6 Appendix: If Every Project Requires Debt

In this extension of the model, I consider the implication of risk capacity if all risky investments require external financing to get started. It turns out that this leads to a debt capacity very similar to the model by Shleifer and Vishny (1992), and that this debt capacity is enforced by the banks.

Consider the case of two firms who each have 1 unit of capital. They can decide to store the capital with perfect liquidity and no return, or invest in a risky project that requires an investment of 2 at $t = 0$ but returns $R > 2$ at $t = 2$. That is, a risky project requires outside financing already at $t = 0$, which can be provided by banks at per-period interest rate r.

A risky investment will turn bad (i.e. providing idiosyncratic and non-transferable payoff b) at $t = 1$ with probability λ and remain good otherwise. Only insiders (i.e. firms) can observe whether a project is good or bad.

At $t = 1$, a crisis hits and every risky project needs a further cash injection of c. This can either be provided by another firm (if it has engaged in storage) or by outside banks (who can't tell whether a project is good or bad and therefore don't know whether they'll get their money back). Assume that the banks make the decision about lending c independently of credit given at $t = 0$ – so that the loans given at $t = 0$ are either seen as "sunk" or that firms are each dealing with a different bank.

Consider the simple case with 2 firms. We have the same situation as in the main part: If both firms engage in the project, none of them is able to get additional financing in period 2 under the assumption of relatively low interest rates, $r < \frac{\lambda}{1-\lambda}$. Now, however, it is no longer the binding constraint that not both firms want to engage in the project, but that in equilibrium the banks do not want to lend to both firms so they can engage in a project (since the banks know that if both firms engage in the project, they will both fail in period 1 and no money will ever be recovered). Thus, instead of a precarious equilibrium keeping firms below the *risk capacity*, the threat of the risk capacity forces the banks to impose a *debt capacity* on the economy.

Chapter 4

Correlated equilibria in homogenous good Bertrand competition[1]

Ole Jann and Christoph Schottmüller

We show that there is a unique correlated equilibrium, identical to the unique Nash equilibrium, in the classic Bertrand oligopoly model with homogenous goods and identical marginal costs. This provides a theoretical underpinning for the so-called "Bertrand paradox" as well as its most general formulation to date. Our proof generalizes to asymmetric marginal costs and arbitrarily many players in the following way: The market price cannot be higher than the second lowest marginal cost in any correlated equilibrium.

[1]This paper has been published in the *Journal of Mathematical Economics (57), March 2015, 31-37.* We would like to thank one anonymous referee and the editor (Atsushi Kajii) for helpful suggestions. We have also benefitted from comments by Jan Boone, Gregory Pavlov and Peter Norman Sørensen.

1 Introduction

A substantial body of theory in industrial organization and other fields of economics is built on the idea that there are no equilibria with positive expected profits in a simple Bertrand competition model with homogenous goods and symmetric firms—in other words, that there are no profitable cartels and that price competition between $n > 1$ firms will drive prices down to marginal cost in one-shot price competition. The fact that price competition between two firms is equivalent to perfect competition is often referred to as the "Bertrand paradox".

Yet the theoretical foundation for this idea is not fully clear, especially where correlated equilibria are concerned. In a correlated equilibrium, players can construct a correlation device which gives each player a private recommendation before the players choose their actions. In correlated equilibrium, the device is such that it is an equilibrium for the players to follow the recommendation. Every (mixed strategy) Nash equilibrium is a correlated equilibrium where the recommendations are independent. Players can in many games achieve higher payoffs in correlated equilibrium than in Nash equilibrium because the device is able to correlate recommendations; see Aumann (1974). In Bertrand competition, it is conceivable that players could correlate their prices in such a way as to achieve high prices while still (through the shape of the joint price distribution) making sure that none of them wants to deviate. We show that this is not possible, although the argument is somewhat subtle.

More precisely, we show that no correlated equilibrium (and hence also no mixed Nash equilibrium) with positive expected profits can exist in a symmetric Bertrand game with homogenous products and bounded monopoly profits.[2] This is the most general formulation of the Bertrand paradox yet. Our result is certainly desirable because a statement like the Bertrand paradox – implying that zero profits are inevitable in a price competition setting – should naturally be shown using an equilibrium concept that is "permissive", i.e. a solution concept that allows the players to coordinate as much as possible within the paradigm of a one-shot, non-cooperative game. This is exactly what correlated equilibrium does.[3] Our result is not obvious given that the set of rationalizable actions is large: In symmetric, homogenous good Bertrand competition all non-negative prices are rationalizable.[4] This is, for example, in stark contrast to Bertrand games

[2]Wu (2008) claims to prove a similar theorem for symmetric linear costs and linear demand. Note, however, that he does not provide a proof for the central second case in his case distinction and implicitly limits his analysis to a finite action space which is incompatible with the standard version of the Bertrand game.

[3]Correlated equilibrium has been shown to have many other attractive properties as well: For example, several simple learning procedures converge to correlated equilibria, see for example Foster and Vohra (1997), Fudenberg and Levine (1999), Hart and Mas-Colell (2000), and unique correlated equilibria are robust to introducing incomplete information, see Kajii and Morris (1997). It should, however, be noted that these papers limit themselves to finite games for technical reasons.

[4]Every $p_i \in \Re_+$ is in our model rationalizable because p_i is – assuming zero marginal costs – a best

with differentiated products: Milgrom and Roberts (1990) show for a large class of demand functions that there is a unique rationalizable action in a differentiated goods Bertrand game. This clearly implies that there is a unique Nash equilibrium and also a unique correlated equilibrium in these games. Their reasoning, however, applies only to supermodular games. A Bertrand game with homogenous goods is not supermodular since the profit functions (i) do not have increasing differences and (ii) are not order upper semi-continuous in the firm's price.

Our proof is by contradiction: We show that if there was a correlated equilibrium in which prices higher than marginal cost were played with positive probability, then there would be an interval of recommendations in which each player prefers to deviate downwardly from his recommendation. This interval consists of the highest recommendations that a player might get (with positive probability) in the assumed equilibrium.

The contribution of this paper lies in the proof that in Bertrand games with arbitrary demand functions (in which the set of rationalizable actions is infinite), the Bertrand Nash equilibrium is the unique correlated equilibrium.

Apart from that, it is also a generalization (by different methods) of results of Baye and Morgan (1999) and Kaplan and Wettstein (2000) on mixed-strategy equilibria in Bertrand games. Baye and Morgan (1999) show that if monopoly profits are unbounded, any positive finite payoff vector can be achieved in a symmetric mixed-strategy Nash equilibrium, and Kaplan and Wettstein (2000) prove that unboundedness of monopoly profits is both necessary and sufficient for the existence of such mixed-strategy Nash equilibria. These insights have led Klemperer (2003, section 5.1) to conclude that "there are other equilibria with large profits, for some standard demand curves." We show that expected profits in any correlated equilibrium (and therefore in any mixed Nash equilibrium) are zero if demand is such that monopoly profits are bounded. Finally, unlike the cited results, our proof is generalizable to games with asymmetric costs and arbitrarily many players: We show that the highest market price in any correlated equilibrium equals the second lowest marginal cost. This establishes an (outcome) equivalence of Nash and correlated equilibria also in this more general setup.

A related result is derived in Liu (1996). Liu shows that the unique Nash equilibrium in Cournot competition with linear demand and constant marginal costs is also the unique correlated equilibrium.

This paper is organized as follows. The next section introduces the Bertrand model with two symmetric firms as well as the concept of correlated equilibrium. Section 3 derives our result. This result is generalized for the case of n non-symmetric firms in section 4. Section 5 concludes.

response to $p_j = 0$ which is the Bertrand equilibrium price and therefore itself rationalizable.

2 Model

There are two firms with constant marginal costs which are normalized to zero. Firms set prices simultaneously. The price of firm i is denoted by p_i. If $p_i < p_j$, consumers buy quantity $D(p_i)$ of the good from firm i (and 0 units from firm j). If both firms quote the same price p', consumers buy $D(p')/2$ from each firm. $D(p)$ denotes market demand where $D : \Re_+ \to \Re_+$ is a (weakly) decreasing, measurable function and $\Re_+$ denotes the non-negative real numbers. We assume that the demand function is such that a strictly positive monopoly price $arg\max_p pD(p)$ exists. We define p^{mon} as the supremum of all prices maximizing $pD(p)$ and assume that p^{mon} is finite. Firms maximize expected profits.

A correlated equilibrium in this game is a probability distribution F on $\Re_+ \times \Re_+$. This probability distribution is interpreted as a correlation device. The correlation device sends recommended prices (r_1, r_2) to the two firms. Each firm i observes r_i but does not observe the other firm's recommendation r_j. $F(p_1, p_2)$ is the probability that $(r_1, r_2) \leq (p_1, p_2)$. Roughly speaking, a distribution F is called a *correlated equilibrium* if both firms find it optimal to follow the recommendation.

To be more precise denote the profits of firm i given prices p_i and p_j with $i, j \in \{1, 2\}$ and $i \neq j$ as

$$\pi_i(p_i, p_j) = \begin{cases} p_i D(p_i) & \text{if } p_i < p_j \\ p_i D(p_i)/2 & \text{if } p_i = p_j \\ 0 & \text{else.} \end{cases} \tag{4.1}$$

Note that we define the profit function such that the *own* price is the first argument, i.e. the first argument of π_2 is p_2.

A strategy for firm i is a mapping from "recommendations" to prices. Both recommendations and prices are in $\Re_+$. Hence, a strategy is a measurable function $\zeta_i : \Re_+ \to \Re_+$. The identity function represents the strategy of following the recommendation. F is a correlated equilibrium if no firm can gain by unilaterally deviating from a situation where both firms use ζ_i = identity function. More formally, we follow the definition of correlated equilibrium for infinite games given in Hart and Schmeidler (1989) and also used in Liu (1996): A correlated equilibrium is a distribution F on $\Re_+ \times \Re_+$ such that for all measurable functions $\zeta_i : \Re_+ \to \Re_+$ and all $i \in \{1, 2\}$ and $i \neq j \in \{1, 2\}$ the following inequality holds:

$$\int_{\Re_+ \times \Re_+} \pi_i(p_i, p_j) - \pi_i(\zeta_i(p_i), p_j)\, dF(p_1, p_2) \geq 0. \tag{4.2}$$

In words, a distribution F is a correlated equilibrium if no player can achieve a higher expected payoff by unilaterally deviating to a strategy ζ_i instead of simply following the

recommendation. Last, we define a *symmetric* correlated equilibrium as a correlated equilibrium F in which $F(p_1, p_2) = F(p_2, p_1)$ for all $(p_1, p_2) \in \Re_+ \times \Re_+$.

It is well known that both firms set prices equal to zero in the unique Nash equilibrium of this game (usually this is called "Bertrand equilibrium"); see, for example, Kaplan and Wettstein (2000).

3 Analysis and Result

We start the analysis by noting that whenever there is a correlated equilibrium F then there is a symmetric correlated equilibrium G in which the aggregated expected profits are the same as in F. This result is, of course, due to the symmetry of our setup. It will allow us later on to focus on symmetric correlated equilibria.[5]

Lemma 1. *Let F be a correlated equilibrium. Then there exists a symmetric correlated equilibrium G such that*

$$\int_{\Re_+ \times \Re_+} \pi_1(p_1, p_2) + \pi_2(p_2, p_1)\, dF(p_1, p_2) = \int_{\Re_+ \times \Re_+} \pi_1(p_1, p_2) + \pi_2(p_2, p_1)\, dG(p_1, p_2).$$

Proof. Let F be a correlated equilibrium. Define $\tilde{F}(p_1, p_2) = F(p_2, p_1)$. Then, $\tilde{F}$ is also a correlated equilibrium as for any measurable function $\zeta : \Re_+ \to \Re_+$

$$\begin{aligned} & \int_{\Re_+ \times \Re_+} \pi_i(p_i, p_j) - \pi_i(\zeta(p_i), p_j)\, d\tilde{F}(p_1, p_2) \\ = & \int_{\Re_+ \times \Re_+} \pi_i(p_j, p_i) - \pi_i(\zeta(p_j), p_i)\, dF(p_1, p_2) \\ = & \int_{\Re_+ \times \Re_+} \pi_j(p_j, p_i) - \pi_j(\zeta(p_j), p_i)\, dF(p_1, p_2) \geq 0 \end{aligned}$$

where the first equality holds by the definition of $\tilde{F}$, the second holds by the symmetry of the setup, i.e. $\pi_1(x, y) = \pi_2(x, y)$, and the inequality holds as F is a correlated equilibrium.

Define $G(p_1, p_2) = \frac{1}{2}F(p_1, p_2) + \frac{1}{2}\tilde{F}(p_1, p_2)$. Then G is a correlated equilibrium as for

[5] Intuitively, we make use of the fact that the set of correlated equilibria in this game is convex—as could be shown by generalizing the following lemma with arbitrary weights instead of $\frac{1}{2}$ and $\frac{1}{2}$.

any measurable function $\zeta : \Re_+ \to \Re_+$

$$\begin{aligned} & \int_{\Re_+ \times \Re_+} \pi_i(p_i, p_j) - \pi_i(\zeta(p_i), p_j)\, dG(p_1, p_2) \\ = \; & \frac{1}{2} \int_{\Re_+ \times \Re_+} \pi_i(p_i, p_j) - \pi_i(\zeta(p_i), p_j)\, dF(p_1, p_2) \\ & + \frac{1}{2} \int_{\Re_+ \times \Re_+} \pi_i(p_i, p_j) - \pi_i(\zeta(p_i), p_j)\, d\tilde{F}(p_1, p_2) \geq \frac{1}{2}0 + \frac{1}{2}0 = 0 \end{aligned}$$

where the equality follows from the definition of G and the inequality follows from the fact that F and $\tilde{F}$ are correlated equilibria. Clearly, G is symmetric as $G(p_1, p_2) = \frac{1}{2}F(p_1, p_2) + \frac{1}{2}\tilde{F}(p_1, p_2) = \frac{1}{2}F(p_1, p_2) + \frac{1}{2}F(p_2, p_1) = \frac{1}{2}\tilde{F}(p_2, p_1) + \frac{1}{2}F(p_2, p_1) = G(p_2, p_1)$ by the definition of G and $\tilde{F}$. Finally, expected profits under F and G are the same as

$$\begin{aligned} & \int_{\Re_+ \times \Re_+} \pi_1(p_1, p_2) + \pi_2(p_2, p_1)\, dG(p_1, p_2) \\ = \; & \frac{1}{2} \int_{\Re_+ \times \Re_+} \pi_1(p_1, p_2) + \pi_2(p_2, p_1)\, dF(p_1, p_2) + \frac{1}{2} \int_{\Re_+ \times \Re_+} \pi_1(p_2, p_1) + \pi_2(p_1, p_2)\, dF(p_1, p_2) \\ = \; & \frac{1}{2} \int_{\Re_+ \times \Re_+} \pi_1(p_1, p_2) + \pi_2(p_2, p_1)\, dF(p_1, p_2) + \frac{1}{2} \int_{\Re_+ \times \Re_+} \pi_2(p_2, p_1) + \pi_1(p_1, p_2)\, dF(p_1, p_2) \\ = \; & \int_{\Re_+ \times \Re_+} \pi_1(p_1, p_2) + \pi_2(p_2, p_1)\, dF(p_1, p_2) \end{aligned}$$

where the first equality follows from the definition of G and $\tilde{F}$ and the second equality follows from the symmetry of setup, i.e. $\pi_1(x, y) = \pi_2(x, y)$. □

Let F be a symmetric correlated equilibrium. Define $\bar{p} := \inf\{p' : \int_{(p', \infty)^2} dF(p_1, p_2) = 0\}$. Intuitively, $\bar{p}$ is the price such that (i) the probability that the market price is greater than $\hat{p}$ is strictly positive for any $\hat{p} < \bar{p}$ and (ii) the probability that the market price is greater than $\hat{p}$ is zero for any $\hat{p} > \bar{p}$. That is, if we consider the distribution of prices that consumers pay in the correlated equilibrium F, $\bar{p}$ is the essential supremum of this "market price distribution". The following lemma establishes that $\bar{p}$ exists by showing that $\int_{(p^{mon}, \infty)^2} dF(p_1, p_2) = 0$ in any correlated equilibrium F. This implies $\bar{p} \leq p^{mon}$ and consequently a finite $\bar{p}$ exists. The intuitive reason for lemma 2 is that setting prices above p^{mon} is a weakly dominated strategy.

Lemma 2. *In a correlated equilibrium F, $\int_{(p^{mon}, \infty)^2} dF(p_1, p_2) = 0$.*

Proof. Consider the strategy

$$\zeta_1(r_1) = \begin{cases} r_1 & \text{if } r_1 \leq p^{mon} \\ p^* & \text{if } r_1 > p^{mon}, \end{cases}$$

where $p^* \in \arg\max_{p \in \Re_+} pD(p)$, i.e. $p^* \leq p^{mon}$ is an arbitrary monopoly price. Firm 1's

payoff difference between following the recommendation and using the deviation strategy ζ_1 is

$$\int_{(p^{mon},\infty)\times(p^*,\infty)} [\pi_1(p_1,p_2) - p^*D(p^*)]\; dF(p_1,p_2) + \int_{(p^{mon},\infty)\times\{p^*\}} -p^*D(p^*)/2\, dF(p_1,p_2).$$

The integrand of the first integral is strictly negative as $p^*D(p^*) = \max_{p\in\Re_+} pD(p)$ and larger than the profit at any price above p^{mon}. The second integral is non-positive. Consequently, F can only be a correlated equilibrium, i.e. satisfy (4.2), if $\int_{(p^{mon},\infty)\times(p^*,\infty)} dF(p_1,p_2) = 0$ which implies $\int_{(p^{mon},\infty)^2} dF(p_1,p_2) = 0$ by $p^* \leq p^{mon}$. □

Before we proceed, it is useful to define the following sets which will serve as the domain of integration multiple times in the following proofs. For some $\hat{p} \in (0,\bar{p})$ and $\varepsilon \in (0,1)$, define the sets

$$\begin{aligned}
A(\hat{p}) &= \{(p_1,p_2):\; p_1 \in (\hat{p},\bar{p}] \text{ and } p_2 \in [p_1,\bar{p}]\} \\
B(\hat{p}) &= \{(p',p'):\; \hat{p} < p' \leq \bar{p}\} \\
C(\hat{p},\varepsilon) &= \{(p_1,p_2):\; p_1 \in (\hat{p},\bar{p}] \text{ and } p_2 \in [\varepsilon p_1,\bar{p}]\} \\
E(\hat{p}) &= \{(p_1,p_2):\; p_1 \in (\hat{p},\bar{p}] \text{ and } p_2 \in [\hat{p},\bar{p}]\} \\
E'(\hat{p}) &= \{(p_1,p_2):\; p_1 \in (\hat{p},\bar{p}] \text{ and } p_2 \in (\hat{p},\bar{p}]\}.
\end{aligned}$$

Figure 4.1 depicts the sets.

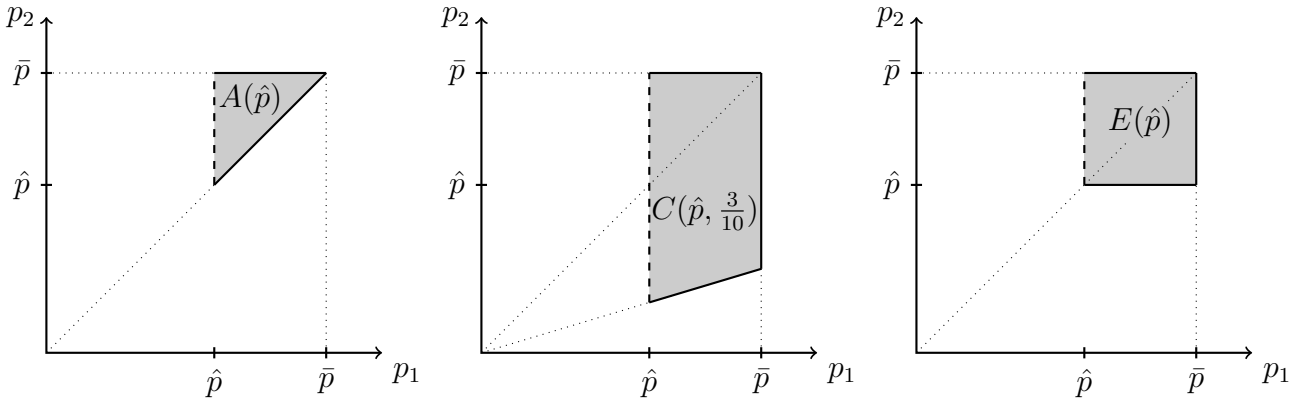


Figure 4.1: $A(\hat{p})$ is shown in panel 1, while $B(\hat{p})$ is simply the diagonal between $(\hat{p},\hat{p})$ and $(\bar{p},\bar{p})$, including the latter but not the former point. Panel 2 shows $C(\hat{p},0.3)$. Panel 3 shows $E(\hat{p})$; $E'(\hat{p})$ is identical to $E(\hat{p})$ except that the border where $p_2 = \hat{p}$ is not part of the set.

It follows immediately from the definition of $\bar{p}$ and the symmetry of the setup that $\int_{A(\hat{p})} dF(p_1,p_2) > 0$ for any $\hat{p} \in (0,\bar{p})$.[6] That is, a firm deviating by charging $\hat{p} < \bar{p}$

[6]To be precise, note that $\int_{p_1\in(\hat{p},\bar{p}),\, p_2>\bar{p}} dF(p_1,p_2) = 0$ in any correlated equilibrium F: Otherwise, firm 2 could profitably deviate by setting $p_2 = \hat{p}$ whenever receiving a recommendation above $\bar{p}$ (while following recommendations below $\bar{p}$).

given any recommendation will sell with positive probability. This observation will be important later on.

The following lemma shows that there is no probability mass on the diagonal of the distribution F above $(\hat{p}, \hat{p})$ if F is a symmetric correlated equilibrium. This is quite intuitive: The diagonal represents situations in which both firms get the same recommendation. Hence, each firm could discontinuously increase its profits by lowering the price only slightly (thereby capturing all instead of half of the demand) in this situation. If the event that both firms get the same recommendation above $\hat{p}$ had positive probability mass, each firm could therefore gain by deviating to a price slightly below its recommendation whenever it receives a recommendation above $\hat{p}$.

Lemma 3. *Let F be a symmetric correlated equilibrium. Then, $\int_{B(\hat{p})} dF(p_1, p_2) = 0$ for any $\hat{p} \in (0, \bar{p})$.*

Proof. The proof is by contradiction. Suppose to the contrary that $\int_{B(\hat{p})} dF(p_1, p_2) > 0$. Recall that π_1 is discontinuous at points on the diagonal of the (p_1, p_2) plane. Therefore, (4.2) is violated for

$$\zeta^{\varepsilon}(r_1) = \begin{cases} r_1 & \text{if } r_1 \notin (\hat{p}, \bar{p}] \\ \varepsilon r_1 & \text{if } r_1 \in (\hat{p}, \bar{p}] \end{cases}$$

for $\varepsilon \in (0, 1)$ sufficiently close to 1: Firm 1's payoff difference between following the recommendation and playing ζ^{ε} can be written as

$$\begin{aligned} \Delta &= \int_{A(\hat{p})} \pi_1(p_1, p_2)\, dF(p_1, p_2) - \int_{C(\hat{p},\varepsilon)} \pi_1(\varepsilon p_1, p_2)\, dF(p_1, p_2) \\ &= \int_{A(\hat{p})\setminus B(\hat{p})} \pi_1(p_1, p_2) - \pi_1(\varepsilon p_1, p_2)\, dF(p_1, p_2) + \int_{C(\hat{p},\varepsilon)\setminus A(\hat{p})} -\pi_1(\varepsilon p_1, p_2)\, dF(p_1, p_2) \\ &\quad + \int_{B(\hat{p})} \frac{p_1 D(p_1)}{2} - \varepsilon p_1 D(\varepsilon p_1)\, dF(p_1, p_2). \end{aligned}$$

The first term continuously approaches 0 as $\varepsilon \nearrow 1$. To see this, note that the first term is (weakly) less than $(1-\varepsilon)\int_{A(\hat{p})\setminus B(\hat{p})} \pi_1(p_1, p_2) dF(p_1, p_2)$ because $p_1 < p_2$ in $A(\hat{p}) \setminus B(\hat{p})$. The second term is non-positive and the third term is strictly negative and bounded away from 0 as $\varepsilon \nearrow 1$ because $\int_{B(\hat{p})} dF(p_1, p_2) > 0$. Consequently, $\Delta < 0$ for sufficiently high $\varepsilon < 1$. This contradicts that F is a correlated equilibrium and therefore $\int_{B(\hat{p})} dF(p_1, p_2) = 0$ has to hold. □

After this auxiliary result, we come to the main result: In any correlated equilibrium, both firms set prices equal to zero with probability 1 and therefore make zero profits. That is, every correlated equilibrium is essentially equivalent to the Bertrand Nash equilibrium.[7]

[7]The qualifier "essentially" stems from the definition of correlated equilibrium in infinite games: A

The intuition behind this result is the following: Take a symmetric correlated equilibrium and suppose that $\bar{p} > 0$. Take some $\hat{p} \in (0, \bar{p})$. If a firm, say firm 1, deviates by charging $\hat{p}$ instead of its recommendation whenever firm 1 receives a recommendation above $\hat{p}$, there are two effects of the deviation: If $r_1 > r_2 > \hat{p}$, firm 1 gains $\hat{p}$ because it sells while it would not have sold by following the recommendation. If $r_2 > r_1 > \hat{p}$, firm 1 loses $r_1 - \hat{p}$ by deviating because it would have sold at the higher price r_1 if it followed the recommendation. In a symmetric equilibrium both events are equally likely.[8] If one chooses $\hat{p}$ sufficiently high, the deviation is therefore profitable as then $\hat{p} > \bar{p} - \hat{p} > r_1 - \hat{p}$.

Theorem 1. *In every correlated equilibrium F, $\bar{p} = 0$. That is, $p_1 = p_2 = 0$ with probability 1 in every correlated equilibrium.*

Proof. By lemma 1, it is sufficient to show that in any *symmetric* correlated equilibrium F, we have $\bar{p} = 0$. Therefore, we concentrate on symmetric F in the remainder of the proof.

The proof is by contradiction. Suppose to the contrary that $\bar{p} > 0$. Define $\hat{p} = \frac{3}{4}\bar{p}$. As F is a correlated equilibrium, player 1 must get a higher expected payoff from following the recommendation r_1 than from following the deviation strategy

$$\zeta(r_1) = \begin{cases} r_1 & \text{if } r_1 \notin (\hat{p}, \bar{p}] \\ \hat{p} & \text{if } r_1 \in (\hat{p}, \bar{p}]. \end{cases}$$

Making use of the sets $E(\hat{p})$ and $E'(\hat{p})$ as defined above, the difference between the expected payoff when following the recommendation and the expected payoff under ζ is

$$\begin{aligned} \Delta &= \int_{A(\hat{p})} \pi_1(p_1, p_2)\, dF(p_1, p_2) - \int_{E(\hat{p})} \pi_1(\hat{p}, p_2)\, dF(p_1, p_2) \\ &\leq \int_{A(\hat{p})} D(\hat{p}) p_1\, dF(p_1, p_2) - \int_{E'(\hat{p})} \pi_1(\hat{p}, p_2)\, dF(p_1, p_2) \\ &= D(\hat{p}) \int_{A(\hat{p})} (p_1 - \hat{p})\, dF(p_1, p_2) - D(\hat{p})\hat{p} \int_{E'(\hat{p}) \setminus A(\hat{p})} dF(p_1, p_2) \\ &= D(\hat{p})\hat{p} \left(\int_{A(\hat{p})} \frac{p_1 - \hat{p}}{\hat{p}}\, dF(p_1, p_2) - \int_{A(\hat{p})} dF(p_1, p_2) \right) \end{aligned}$$

where the last equality follows from the symmetry of F and lemma 3 (which states that $\int_{B(\hat{p})} dF(p_1, p_2) = 0$). By the definition of $\hat{p} = \frac{3}{4}\bar{p}$, $\frac{p_1 - \hat{p}}{\hat{p}} < 1$ for all $p_1 \in (\hat{p}, \bar{p}]$. Therefore,

$$\int_{A(\hat{p})} \frac{p_1 - \hat{p}}{\hat{p}}\, dF(p_1, p_2) < \int_{A(\hat{p})} dF(p_1, p_2) \tag{4.3}$$

strategy ζ_i that differs from the identity function on a set of points that has zero probability under F is also an equilibrium strategy.

[8] Note that we do not have to consider the case $r_1 = r_2$ because of the previous lemma.

as $\int_{A(\hat{p})} dF(p_1, p_2) \neq 0$ by the definition of $\bar{p}$ and $\hat{p} < \bar{p}$. Note that (4.3) implies $\Delta < 0$ which contradicts that F is a correlated equilibrium. □

4 The General Case

Consider the general case of asymmetric marginal costs c_i and n firms. The main idea of our proof also applies in this case, and we can show that the market price paid by consumers is less or equal to the second lowest marginal costs with probability one in every correlated equilibrium. Hence, correlated equilibrium is essentially equivalent to the Bertrand Nash equilibrium also in this more general framework.

The setup of this more general model is as follows: Market demand is, as before, $D(p)$ where $D : \Re_+ \to \Re_+$ is a weakly decreasing, measurable function. There are n firms. All firms have constant marginal costs c_i, $i \in \{1, \ldots, n\}$, where – without loss of generality – we assume $c_1 \leq c_2 \leq \ldots \leq c_n$. Firms set prices simultaneously. If $p_i < p_j$ for all $j \neq i$, consumers buy quantity $D(p_i)$ units of the good from firm i (and 0 units from the other firms).

If $k \geq 2$ firms post the same lowest price $p' = \min\{p_1, \ldots, p_n\}$, we assume that consumers do the following: The firms with the lowest marginal costs among those k firms quoting p' share the demand $D(p')$ equally. More formally, denote the k firms quoting p' as $\{m_1, \ldots, m_k\}$ and let – without loss of generality – the ordering be such that $c_{m_1} \leq c_{m_2} \leq \cdots \leq c_{m_k}$. Define $\tilde{k}$ as $max_{j \in \{1,\ldots,k\}}\{j : \ c_{m_1} = c_{m_j}\}$. Then firms m_1 to $m_{\tilde{k}}$ sell $D(p')/\tilde{k}$ units and all other firms sell zero units. We assume that the demand is such that $p^{mon} = \max\{p_1^{mon}, \ldots, p_n^{mon}\}$, where p_i^{mon} is the supremum of $\arg\max_p (p - c_i)D(p)$, is finite and strictly positive.

The assumption that all consumers buy from the low cost firms in case several firms charge the same price deserves some comment. We make this assumption to ensure the existence of the standard Bertrand Nash equilibrium. If c_2 is lower than the (lowest) monopoly price of firm 1, this well known equilibrium postulates that $p_1 = p_2 = c_2$ (and arbitrary $p_i \geq c_i$ for $i \in \{3, \ldots, n\}$). This is indeed a Nash equilibrium with our tie-breaking rule above but can fail to be an equilibrium with other tie-breaking rules. If, for example, $c_1 < c_2$ and a mass of consumers does not buy from firm 1 whenever $p_1 = p_2$, then $p_1 = p_2 = c_2$ is not an equilibrium as firm 1 could increase its profits by decreasing its price by a sufficiently small amount. Assuming a tie-breaking rule such that a Nash equilibrium exists has two advantages: First, it gives us a benchmark to which we can compare correlated equilibria. Second, as every Nash equilibrium can be interpreted as a correlated equilibrium, we know that a correlated equilibrium exists.[9] Note also that the behavior of the consumers that corresponds to this assumption is optimal, and that

[9]It should be noted that the equal sharing assumption (in case $\tilde{k} > 1$) is not important for our analysis and any other rule would work as well.

the Nash equilibrium would therefore also be a Nash equilibrium of the wider game in which a group of consumers acts as players.

An alternative to our "buy from the low cost firm" assumption would be to assume that all k firms that charge the same lowest price p' sell the same quantity $D(p')/k$ ("equal splitting"). Blume (2003) shows that a Nash equilibrium in mixed strategies exists with equal splitting if D is continuously differentiable. Since we did not assume differentiability of D (and as it is unclear whether Blume's result holds without differentiability), we do not go this path. However, it should be noted that – with some minor modifications – all proofs would go through with the alternative assumptions of equal splitting and continuously differentiable demand.

Our setup gives therefore the following profits for firm i at a price vector $p = (p_1, \ldots, p_n)$:

$$\pi_i(p) = \begin{cases} (p_i - c_i)D(p_i) & \text{if } p_i < p_j \text{ for all } j \neq i \\ (p_i - c_i)D(p_i) & \text{if } p_i = p_{m_1} = \cdots = p_{m_k} < p_j \text{ for all } j \notin \{i, m_1, \ldots, m_k\} \\ & \quad \text{and } c_i < c_l \text{ for all } l \in \{m_1, \ldots, m_k\} \\ (p_i - c_i)D(p_i)/\tilde{k} & \text{if } p_i = p_{m_1} = \cdots = p_{m_k} < p_j \text{ for all } j \notin \{i, m_1, \ldots, m_k\} \\ & \quad \text{and } c_i = c_{m_1} = \cdots = c_{m_{\tilde{k}}} < c_{m_{\tilde{k}+1}} \leq \cdots \leq c_{m_k} \\ 0 & \text{else.} \end{cases}$$

As before, a strategy for firm i is a measurable function $p_i : \Re_+ \to \Re_+$ and a distribution F on $\Re_+^n$ is a correlated equilibrium if it satisfies (4.2) for all firms and all deviation strategies. We obtain our main theorem:

Theorem 2. *Let F be a correlated equilibrium. Then, $\bar{p} = \inf\{p' : \int_{(p',\infty)^n} dF(p) = 0\} \leq c_2$.*

The proof, which is similar to the proof of theorem 1 though without using the shortcut of symmetry, is relegated to the appendix.

5 Conclusion

We have shown that marginal cost pricing is the unique correlated equilibrium in a symmetric Bertrand game with homogenous products and constant marginal costs. This establishes the well-known Bertrand paradox in its most general form. The idea of the Bertrand paradox is that the perfectly competitive outcome is unavoidable if two firms compete in prices (in a market for homogenous products). The set of correlated equilibria establishes – in an equilibrium sense – the set of payoffs that players can achieve non-cooperatively. It might allow for forms of (self-enforcing) coordination and cooperation

that are unattainable in other equilibrium concepts, e.g. Nash equilibrium. Therefore it is the natural solution concept to state an unavoidability-result like the Bertrand paradox.

We have also shown that the result generalizes to Bertrand settings with n firms and non-identical marginal costs. In this case, the market price paid by consumers is less or equal to the second lowest marginal costs with probability one in every correlated equilibrium, making correlated equilibrium equivalent to Bertrand Nash equilibrium also in this more general framework.

The key ingredients in our proof are (i) the discontinuities in the payoff functions that are typical for the Bertrand model and (ii) boundedness of possible equilibrium prices (induced by a bounded monopoly price). Loosely speaking, firm 1 sets prices above its costs only if firm 2 is sufficiently likely to set even higher prices. In a symmetric correlated equilibrium, ingredient (ii) gives us an essential supremum on the price distribution $\bar{p}$. Whenever firm 2 receives a recommendation very close to $\bar{p}$, it must therefore be sure that — when following the recommendation — there is a sufficiently high chance of firm 1 getting a recommendation that is even closer to $\bar{p}$. Both firms, however, know that there is a significant chance that the other firm will just undercut them — otherwise the other firm would not follow such high recommendations! Since the marginal loss in lowering the price to some $\bar{p}-\varepsilon$ when getting a recommendation in $(\bar{p}-\varepsilon, \bar{p})$ is minimal while the upside of possibly undercutting the other firm is immense (the all-or-nothing nature of Bertrand competition, ingredient i), such deviations increase profits and contradict the existence of correlated equilibria with prices above costs.

The discontinuities of the profit functions explain also why our proof is unrelated to proofs of (essential) uniqueness of correlated equilibrium in other industrial organization models, e.g. Cournot competition (Liu, 1996) or price competition with differentiated goods (Milgrom and Roberts, 1990). Our proof does not work in these models because they have no payoff discontinuity. Vice versa, their proofs will not work for homogenous good Bertrand competition since it is not a supermodular game.[10] We conjecture that similar arguments as in our paper would also hold in other games that share the two ingredients mentioned above: (i) a threshold (depending on the other players' actions) such that player i's payoff discontinuously decreases from a positive value to zero when passing the threshold and (ii) an upper bound on actions possible in equilibrium.

[10]Liu (1996) considers an n-firm game that is not supermodular. However, the first step of his proof shows that the non profitability of not deviating to the Nash equilibrium output already implies that the market quantity equals the Nash equilibrium quantity in any correlated equilibrium. No such result obtains in our model as deviating to marginal cost pricing is never strictly profitable and therefore cannot restrict potential correlated equilibria (in a symmetric Bertrand model).

6 Appendix: Proof of Theorem 2

Given a correlated equilibrium F, we define $\bar{p} \in \Re_+$ in the following way: $\bar{p} = \inf\{p' : \int_{(p',\infty)^n} dF(p) = 0\}$ where $p = (p_1, \ldots, p_n)$. Intuitively, $\bar{p}$ is the price such that (i) the probability that the market price is greater than $\hat{p}$ is strictly positive for any $\hat{p} < \bar{p}$ and (ii) the probability that the market price is greater than $\hat{p}$ is zero for any $\hat{p} > \bar{p}$. That is, if we consider the distribution of prices that consumers pay in the correlated equilibrium F, $\bar{p}$ is the essential supremum of this "market price distribution".

$\bar{p}$ is weakly below p^{mon} where $p^{mon} = \max\{p_1^{mon}, \ldots, p_n^{mon}\}$ and p_i^{mon} is the supremum of $\arg\max_p (p - c_i)D(p)$: If $\bar{p} > p^{mon}$, the event that all firms charge a price above p^{mon} would have positive probability. Hence, at least one firm i would – with positive probability – sell goods at a price higher than p_i^{mon}. For this firm, it would be a profitable deviation to charge $p^* \in \arg\max_p (p - c_i)D(p)$ whenever receiving a recommendation r_i above p_i^{mon}. This can be shown more formally as in lemma 2 in the main text. The main point is that $\bar{p} \leq p^{mon}$ exists because $\int_{(p^{mon},\infty)^n} dF(p) = 0$.

Define the following sets analogously to the main text (again p denotes a vector of prices): $A(\hat{p}) = \{p : p_1 \in (\hat{p}, \bar{p}] \text{ and } p_1 \leq p_i \text{ for all } i = 1, 2, \ldots, n\}$ is the set of price vectors for which firm 1 sells with a price between $\hat{p}$ and $\bar{p}$; $K(\hat{p}) = \{p : p_2 \in (\hat{p}, \bar{p}] \text{ and } p_2 \leq p_i \text{ for all } i = 2, \ldots, n \text{ and } p_2 < p_1\}$ is the set of price vectors where firm 2 sells at a price between $\hat{p}$ and $\bar{p}$ (and firm 1 does not sell). Furthermore, define $B(\hat{p}) = \{p : p_1 \in (\hat{p}, \bar{p}] \text{ and } p_1 = p_2 \leq p_i \text{ for all } i = 1, \ldots, n\}$, i.e. B is the set of price vectors where firm 1 and 2 charge both the same price above $\hat{p}$ and all other firms set weakly higher prices.

Lemma 4. *Let F be a correlated equilibrium and suppose $\bar{p} = \inf\{p' : \int_{(p',\infty)^n} dF(p) = 0\} > c_2$. Then, $\int_{B(\hat{p})} dF(p) = 0$ for any $\hat{p} \in (c_2, \bar{p})$.*

Proof. Suppose to the contrary that there exists a $\hat{p} < \bar{p}$ such that $\int_{B(\hat{p})} dF(p) > 0$. We will show that it is then profitable for firm 2 to use the following deviation strategy for $\varepsilon > 0$ sufficiently small

$$\zeta_2^{\varepsilon}(r_2) = \begin{cases} r_2 & \text{if } r_2 \notin (\hat{p}, \bar{p}] \\ (1-\varepsilon)r_2 & \text{if } r_2 \in (\hat{p}, \bar{p}]. \end{cases}$$

The profit difference of firm 2 between sticking to the recommendation and using ζ_2^ε is

$$\begin{aligned}
\Delta_2^\varepsilon &= \int_{K(\hat{p})\cup B(\hat{p})} \pi_2(p)\, dF(p) - \int_{[(1-\varepsilon)\hat{p},\infty)\times(\hat{p},\bar{p}]\times[(1-\varepsilon)\hat{p},\infty)^{n-2}} \pi_2(p_1,(1-\varepsilon)p_2,p_3,\ldots,p_n)\, dF(p) \\
&\leq \int_{K(\hat{p})} \pi_2(p) - \pi_2(p_1,(1-\varepsilon)p_2,p_3,\ldots,p_n)\, dF(p) \\
&\quad + \int_{B(\hat{p})} \pi_2(p) - \pi_2(p_1,(1-\varepsilon)p_2,p_3,\ldots,p_n)\, dF(p) \\
&\leq \int_{K(\hat{p})} D((1-\varepsilon)p_2)\varepsilon p_2\, dF(p) + \int_{B(\hat{p})} \pi_2(p) - \pi_2(p_1,(1-\varepsilon)p_2,p_3,\ldots,p_n)\, dF(p) \\
&\leq \varepsilon \int_{K(\hat{p})} D((1-\varepsilon)p_2)p_2\, dF(p) \\
&\quad + \int_{B(\hat{p})} \frac{D(p_2)(p_2-c_2)}{2} - D((1-\varepsilon)p_2)((1-\varepsilon)p_2-c_2)\, dF(p).
\end{aligned}$$

Note that the first integral in the last line continuously converges to 0 as $\varepsilon \to 0$. The second integral in the last line is, however, negative and bounded away from 0: First, we show that the integrand is strictly negative and bounded away from zero. $\frac{D(p_2)(p_2-c_2)}{2} - D((1-\varepsilon)p_2)((1-\varepsilon)p_2-c_2) < D((1-\varepsilon)p_2)\left(\frac{-(p_2-c_2)}{2}+\varepsilon p_2\right)$ which for $\varepsilon < \frac{p_2-c_2}{4\bar{p}}$ is less than $D((1-\varepsilon)p_2)\frac{-(p_2-c_2)}{4} < D(\bar{p})\frac{-(\hat{p}-c_2)}{4}$. Hence, the integrand is bounded from above by $-D(\bar{p})\frac{\hat{p}-c_2}{4} < 0$ if $\varepsilon \in (0, \frac{\hat{p}-c_2}{4\bar{p}})$ because $\frac{\hat{p}-c_2}{4\bar{p}} < \frac{p_2-c_2}{4\bar{p}}$ for all elements of $B(\hat{p})$. By assumption, $\int_{B(\hat{p})} dF(p) > 0$ which implies that the second integral is bounded from above by $-D(\bar{p})\frac{\hat{p}-c_2}{4}\int_{B(\hat{p})} dF(p) < 0$ for $\varepsilon \in (0, \frac{\hat{p}-c_2}{4\bar{p}})$. Consequently, $\Delta_2^\varepsilon < 0$ for $\varepsilon > 0$ small enough which contradicts that F is a correlated equilibrium. □

We need one further auxiliary result. Roughly speaking, the result says that in a correlated equilibrium firm 1 will sell at a price in $(\hat{p},\bar{p}]$ with positive probability for any $\hat{p} < \bar{p}$. Given the definition of $\bar{p}$, this should be hardly surprising.

Lemma 5. *Let F be a correlated equilibrium such that $\bar{p} = \inf\{p' : \int_{(p',\infty)^n} dF(p) = 0\} > c_2$. Then, $\int_{A(\hat{p})} dF(p) > 0$ for any $\hat{p} \in (c_2,\bar{p})$.*

Proof. Take an arbitrary $\hat{p} \in (c_2,\bar{p})$. First, we show that $\int_{p_1\in(\hat{p},\bar{p}],\, p_i>\bar{p}\ \forall i\neq 1} dF(p) = 0$. Suppose otherwise. Then firm 2 receiving recommendation r_2 can profitably deviate by playing

$$\zeta_2(r_2,\hat{p}) = \begin{cases} r_2 & \text{if } r_2 \leq \bar{p} \\ \hat{p} & \text{if } r_2 > \bar{p}. \end{cases}$$

This is a profitable deviation as it increases firm 2's profits by at least $\int_{p_1\in(\hat{p},\bar{p}],\, p_i>\bar{p}\ \forall i\neq 1}(\hat{p}-c_2)D(\hat{p})dF(p) > 0$. Hence, $\int_{p_1\in(\hat{p},\bar{p}],\, p_i>\bar{p}\ \forall i\neq 1} dF(p) = 0$. This means that firm 1 would never sell at a price in $(\hat{p},\bar{p}]$ if $\int_{A(\hat{p})} dF(p)$ was zero.

Second, consider the following deviation strategy for firm 1:

$$\zeta_1(r_1, \hat{p}) = \begin{cases} r_1 & \text{if } r_1 \notin (\hat{p}, \bar{p}] \\ \hat{p} & \text{if } r_1 \in (\hat{p}, \bar{p}]. \end{cases}$$

The payoff difference between sticking to the recommendation and using ζ_1 is[11]

$$\Delta = \int_{A(\hat{p})} \pi_1(p) - \hat{p}D(\hat{p})\, dF(p) + \int_{(\hat{p},\bar{p}]\times[\hat{p},\infty)^{n-1}\setminus A(\hat{p})} -\pi_1(\hat{p}, p_{-1})\, dF(p).$$

By the definition of $\bar{p}$, $\int_{(\hat{p},\bar{p}]\times[\hat{p},\infty)^{n-1}} dF(p) > 0$ (recall that firm 1 never wants to set a price above $\bar{p}$ as the probability of selling at such a price is zero). If $\int_{A(\hat{p})} dF(p) = 0$, this would imply that the second integral in Δ is strictly negative while the first integral in Δ would be zero. Hence, ζ_1 is a profitable deviation if $\int_{A(\hat{p})} dF(p) = 0$ contradicting that F is a correlated equilibrium. □

The following observation is related to lemma 5: For any $\hat{p} < \bar{p}$, a firm i using the strategy

$$\zeta_i(r_i, \hat{p}) = \begin{cases} r_i & \text{if } r_i \notin (\hat{p}, \bar{p}] \\ \hat{p} & \text{if } r_i \in (\hat{p}, \bar{p}] \end{cases}$$

will sell $D(\hat{p})$ units at price $\hat{p}$ with positive probability: By the definition of $\bar{p}$, the event that all firms get a recommendation above $\hat{p}$ has positive probability. Hence, firm i sells with positive probability at price $\hat{p}$ when using the strategy ζ_i.

Using lemma 4, we can now show the main result: In any correlated equilibrium, $\bar{p} \leq c_2$. This means that the price that consumers pay will be weakly less than c_2 with probability 1. Consequently, the expected profits for firms $2, \ldots, n$ are zero and the expected profits of firm 1 are bounded from above by $D(c_2)(c_2 - c_1)$ in any correlated equilibrium (assuming that c_2 is lower than the lowest monopoly price of firm 1; otherwise, firm 1's monopoly profits are, of course, the upper bound of firm 1's equilibrium profits).

Suppose to the contrary $\bar{p} > c_2$ in a correlated equilibrium F. Let $\hat{p} = \frac{1}{4}c_2 + \frac{3}{4}\bar{p}$ and distinguish the two cases

1. $\int_{K(\hat{p})} dF(p) \geq \int_{A(\hat{p})} dF(p)$
2. $\int_{K(\hat{p})} dF(p) < \int_{A(\hat{p})} dF(p)$.

In the first case, the profit difference of firm 1 from using $\zeta_1(r_1, \hat{p})$ (see above) and from

[11] We use $p_{-1} = p_2, \ldots, p_n$ to denote the prices of all firms but firm 1.

following the recommendation is

$$\begin{aligned}
\Delta_1 &= \int_{A(\hat{p})} \pi_1(p_1, p_{-1})\, dF(p) - \int_{(\hat{p},\bar{p}]\times[\hat{p},\infty)^{n-1}} \pi_1(\hat{p}, p_{-1})\, dF(p) \\
&\leq \int_{A(\hat{p})} D(\hat{p})(p_1 - c_1)\, dF(p) - \int_{(\hat{p},\bar{p}]^n} D(\hat{p})(\hat{p} - c_1)\, dF(p) \\
&= D(\hat{p})(\hat{p} - c_1)\left(\int_{A(\hat{p})} \frac{p_1 - \hat{p}}{\hat{p} - c_1}\, dF(p) - \int_{(\hat{p},\bar{p}]^n \setminus A(\hat{p})} dF(p)\right) \\
&\leq D(\hat{p})(\hat{p} - c_1)\left(\int_{A(\hat{p})} \frac{p_1 - \hat{p}}{\hat{p} - c_1}\, dF(p) - \int_{K(\hat{p})} dF(p)\right).
\end{aligned}$$

By $\hat{p} = \frac{1}{4}c_2 + \frac{3}{4}\bar{p}$, $\frac{p_1 - \hat{p}}{\hat{p} - c_1} \in (0,1)$ for all $p_1 \in (\hat{p}, \bar{p}]$. Therefore,

$$\int_{A(\hat{p})} \frac{p_1 - \hat{p}}{\hat{p} - c_1}\, dF(p) < \int_{K(\hat{p})} dF(p) \tag{4.4}$$

because $0 < \int_{A(\hat{p})} dF(p) \leq \int_{K(\hat{p})} dF(p)$ by the definition of case 1 and lemma 5. Note that (4.4) implies $\Delta_1 < 0$ which contradicts that F is a correlated equilibrium.

In the second case, the profit difference of firm 2 from using $\zeta_2(r_2, \hat{p})$ and from following the recommendation is

$$\begin{aligned}
\Delta_2 &= \int_{K(\hat{p})\cup B(\hat{p})} \pi_2(p_2, p_{-2})\, dF(p) - \int_{[\hat{p},\infty)\times(\hat{p},\bar{p}]\times[\hat{p},\infty)^{n-2}} \pi_2(\hat{p}, p_{-2})\, dF(p) \\
&\leq \int_{K(\hat{p})} \pi_2(p_2, p_{-2})\, dF(p) - \int_{[\hat{p},\bar{p}]\times(\hat{p},\bar{p}]\times[\hat{p},\bar{p}]^{n-2}} \pi_2(\hat{p}, p_{-2})\, dF(p) \\
&\leq \int_{K(\hat{p})} D(\hat{p})(p_2 - c_2)\, dF(p) - \int_{A(\hat{p})\cup K(\hat{p})} \pi_2(\hat{p}, p_{-2})\, dF(p) \\
&= \int_{K(\hat{p})} D(\hat{p})(p_2 - \hat{p})\, dF(p) - \int_{A(\hat{p})} D(\hat{p})(\hat{p} - c_2)\, dF(p) \\
&= D(\hat{p})(\hat{p} - c_2)\left(\int_{K(\hat{p})} \frac{p_2 - \hat{p}}{\hat{p} - c_2}\, dF(p) - \int_{A(\hat{p})} dF(p)\right).
\end{aligned}$$

Note that the step from the first to the second line uses lemma 4. The step from the second to the third line uses the fact that the intersection of $A(\hat{p})$ and the set of price vectors with $p_2 > \bar{p}$ has zero probability in a correlated equilibrium F: Otherwise, firm 2 could profitably deviate to $\hat{p}$ whenever receiving a recommendation above $\bar{p}$.

Now, $\hat{p} = \frac{1}{4}c_2 + \frac{3}{4}\bar{p}$ implies that $\frac{p_2 - \hat{p}}{\hat{p} - c_2} \in (0,1)$ for all $p_2 \in (\hat{p}, \bar{p}]$. The definition of case 2 therefore implies $0 \leq \int_{K(\hat{p})} \frac{p_2 - \hat{p}}{\hat{p} - c_2}\, dF(p) \leq \int_{K(\hat{p})} dF(p) < \int_{A(\hat{p})} dF(p)$. Hence, $\Delta_2 < 0$ which contradicts that F is a correlated equilibrium. □

Chapter 5

An Informational Theory of Privacy[1]

Ole Jann and Christoph Schottmüller

We develop a theory that explains how and when privacy can increase welfare. Without privacy, some individuals misrepresent their preferences, because they will otherwise be statistically discriminated against. This "chilling effect" hurts them individually, and impairs information aggregation. The information gain from infringing privacy (e.g. by electronic surveillance) can be much smaller than expected ceteris paribus. Overall, privacy is essential for any mechanism of information aggregation, such as markets or a democratic society. It is also redistributive: Like free spech, privacy benefits some and hurts others.

[1]We are grateful for helpful comments by Sebastian Barfort and Peter Norman Sørensen, as well as a seminar audience at the University of Copenhagen.

1 Introduction

Privacy is one of the most pressing issues of the information age. It is at the center of debates about government response to extremism and terrorism, especially after revelations that many western governments systematically infringe the privacy of their own citizens by engaging in indiscriminate electronic surveillance (cf. Schneier (2015), Greenwald (2014) or Economist (2013)). The fact that people are willing to give up some privacy in exchange for lower prices or better services is part of the business model of many companies. Some of the most successful and fastest-growing businesses are even built on the fact that people are willing to trade their data in exchange for a free service, thus turning their data into a product which can then be profitably sold.

As these examples show, issues of privacy often involve trade-offs: Between privacy and security, privacy and thriftiness, or even privacy and participation in public life. Understanding the value and the effects of privacy is crucial for how voters, consumers and regulators approach these trade-offs. Classical economic theory suggests that privacy is usually welfare-reducing because it creates asymmetric information – an idea that is echoed by probably the best-known economic treatise on privacy, Posner (1981).

In this paper, we develop an informational theory of privacy.[2] By considering the role of privacy in allowing individuals to express their preferences and in information aggregation, we show that privacy can enhance, not lessen, welfare. One part of this insight is that as privacy allows people to express their preferences and opinions more freely, it can actually improve overall information aggregation of a society. Moreover, the welfare gains from infringing privacy, such as better information about individuals, are often not as large as the losses of the individuals, because individuals will react to a loss of privacy by changing their behavior and thus providing less information. Privacy can also protect minorities: Those with opinions or preferences which are different from the median of the population.

To illustrate our results, consider the following example. Alice would prefer if marijuana was legalized. She considers publishing an overview of her arguments on a social network to try to convince her friends. However, we assume that in Alice's world there is very little privacy: If she does something online, everyone can see it – not just her friends, but also future employers, her parents, the police, and so on.

There is some correlation between preferences on legalization and actual drug use, in that people who actually use drugs are more likely to support legalization. The correlation is of course far from perfect – many people might support legalization for philosophical or practical reasons without being users, and some drug users might even

[2]By "privacy", we mean the ability to take actions without being observed, and having interactions with others confined to the intended recipients. This is only one of many possible definitions and understandings of the term "privacy"; see for example Solove (2010) for an overview.

oppose it.

Employers do not want to hire drug users, but drug use is not observable. An employer will therefore use the observable characteristic (whether Alice did or did not publicly support legalization) to make a hiring decision: People who have supported legalization will not be hired. We can show that this happens in any equilibrium where the correlation between types (i.e. drug use and preference for legalization) is high enough. Being unable to observe the attribute that he is really interested in, the employer will statistically discriminate (as in Arrow (1973) and Phelps (1972)) based on observed preference.

Then, however, Alice has to make a choice: Voice her preference, and risk not being hired for a job – or stay quiet, and face no such consequence. If she doesn't feel strongly about the subject (i.e. if she only has a weak preference for legalization), she will choose not to express her opinion. That is our first result: Lack of privacy causes a "chilling effect". Despite Alice's preference being not only legal and legitimate, but also insubstantial for the job (recall that even the employer does not take issue with her preference itself), she decides not to express it for fear of the consequence.[3]

If Alice had been an opponent of legalization, there would have been no chilling effect. The spectrum of expressed preferences that are present in the public debate will therefore be skewed: Those who oppose legislation speak out freely, while those who support it tend to stay quiet. Since the optimal policy should be an unbiased aggregation of individual preferences, the policy that is implemented will systematically deviate from this optimum.

If her views on the matter are strong, however, she decides to post her arguments anyway since the expected gain from doing so outweighs the disadvantage of not being hired. In this case the lack of privacy hurts Alice in two ways. She suffers from not being hired and – due to the chilling of others with more moderate opinions – her preferred policy of legalization is less likely to be implemented. This is our second result: Lack of privacy hurts those with non-mainstream preferences the most – those that care too strongly to protect themselves by adapting their behavior (i.e. giving in to the chilling effect). Again, note that this effect is asymmetric, too, and affects only those who strongly support legalization.

The change in behavior that results from the chilling effect also has another consequence: It makes the statistical discrimination that the employer uses less effective. Since many people (both drug users and non-users) misrepresent their preferences now, observing what someone posts about drug legalization becomes less informative. This is in particular true if actual drug users are more afraid of not being hired, say because of worse outside options, than non-users as in this case actual drug users are chilled more than non-users. This is our third result, which is especially important in debates about

[3]The term "chilling effect" has been used by legal scholars since at least 1952, when U.S. Supreme Court Justice Felix Frankfurter used it in a concurring opinion in Wieman v. Updegraff, 344 U.S. 183.

government surveillance. Instead of "employer", think for a moment that it was the police who decides to watch those closer who voice a certain opinion, or use certain means of communication, or visit certain places. Surveillance programs and other privacy intrusions should therefore never be naively evaluated *ex ante* – they will always be made less effective by the resulting shift in behavior. Herein lies the crux of our analysis of the trade-offs involved in privacy – that we don't just need to weigh the loss of those who lose privacy with the gain of those who gain information, but also consider how large that gain actually is, given that people rationally change their behavior if they know they are being observed. Such changes in behavior happen: Martews and Tucker (2015) show a significant shift in search engine search terms after the Snowden revelations in 2013; a survey of American writers has found that 1 in 6 has avoided writing or speaking on a particular topic for fear of surveillance (PEN America, 2013).

Consider, for comparison, a world in which Alice has more privacy, so that only her friends (i.e. the intended recipients) can see her message. Now the employer cannot discriminate, since there is no position on drug legalization for him to observe. Alice will therefore be uninhibited in expressing herself, and there is no systematic bias in who speaks up about their opinion. Neither do people with strong opinions get statistically discriminated against. One could argue that the employer loses out, since he has less information on which to base his decision. We can show, however, that the privacy case provides higher welfare compared to the no-privacy case in large populations (and the employer will not even lose from privacy) – this is our fourth result. Note that in this particular case, some people (those who support legalization) would prefer privacy, while opponents of legalization are more likely to get their preferred policy if there is no privacy and the chilling effect silences people with different opinions. Some people therefore rationally oppose privacy. If there were a large number of issues at stake, however, where every individual can sometimes find themselves on either side of the debate, the whole population gains from privacy.

In an extension, we ask: What if Alice herself can choose to keep her message private? Then the employer has to treat all applicants about whom he can find no information equally. Would he hire someone about whom he can find no information? That depends on how people who oppose legalization behave: Would they also choose to keep their messages private? Intuitively, they have nothing to fear from not doing so. We argue that a possible equilibrium where everybody, supporters and opponents of legalization, individually choose privacy for themselves is not stable. Another equilibrium, in which only supporters of legalization choose privacy and in which privacy therefore is meaningless, is more stable. To work well, privacy can therefore not always be left to the individual – sometimes it needs to be mandated.[4]

[4]There are parallels to the obligatory secret ballot, made for example by Schelling (1960): If ballot secrecy was optional, voters could be intimidated into making their ballot public. Forbidding them from

Finally, in our second extension, we look more closely at Alice's rational response to having no privacy. We have established above that Alice might misrepresent her opinion if she prefers legalization (the "chilling effect"), but that she will otherwise express herself truthfully if she strongly supports legalization. If she does so, however, she is aware that this will come with negative consequences, such as worse employment prospects. Now, assume that Alice can take a costly action to mitigate the damage to herself – however, this is costly to the employer. In our example, imagine that Alice could bring a lawyer to every job interview – which slightly decreases, ceteris paribus, the probability of not being hired because of her expressed opinions (thus making it her rational response to the employer's statistical discrimination), while making the interaction much more cumbersome for the employer. In this way, the statistical discrimination that results from lack of privacy can erode trust among the players, and can mean that by a chain of rational responses to each other's behavior, they end up in a Pareto-inferior equilibrium.

Our general model, which we introduce in section 2, considers a problem of information aggregation, in which a group of individuals have cardinal preferences over two options and express their preference by supporting one of the two options. We do not restrict our arguments to any specific information aggregation problem – in fact, our only assumptions about the information aggregation mechanism are that the probability of an option being implemented increases in the number of supporters that it has, and that the process is not systematically biased towards one option. Our model is therefore applicable to a large variety of situations. The example above already points to political information aggregation through public debate or voting. However, the mechanism might just as well be a market in which two providers of goods or services compete for customers. Efficiency demands that the provider who is preferred by most customers also does more business. But if using one of the providers is in some way disreputable, or can bring adverse consequences, lack of privacy and the chilling effect will systematically bias the result. We discuss examples of the mechanism in section 7.

What kind of privacy problem do we have in mind when we assume, as in our example, that some observable behavior is predictive of an unobservable type? Here, too, we keep our assumptions quite general, as we only assume that the two unobservable types (in our example: policy preference and drug use) are positively correlated. It is crucial to note that this does not require any sort of causal relationship – only correlation. We think that in the real world, almost any variable can be "predictive", in the sense of our model, of almost any other variable. Meehl (1990) calls this the "crud factor" and notes that "in social science, everything is somewhat correlated with everything." Even minor choices are correlated with political preferences – a fact which is being used by many political parties and candidates to identify their potential voters. Hamburger and

doing so protects them from any such intimidation.

Wallsten (2005), for example, report about microtargeting efforts by the Republican party in the United States that "... bourbon drinkers were more likely to be Republican, while gin was a Democrat's drink. ... Democrats preferred Volvos; Ford and Chevy owners were most likely Republicans. People with call-waiting service on their telephones were predominantly Republican." As the availability of data and of cheap processing power has grown substantially in recent years, this has become much more acute. The consulting firm Deloitte advertises that it can reliably predict people's life expectancy from observing their buying decisions (Robinson et al., 2014, p. 6), and the "big data underwriting firm" Zest Finance simply claims: "All Data is Credit Data."

In understanding privacy as the creation and maintenance of asymmetric information, our study takes a similar point of departure as the "Chicago school", exemplified by Stigler (1980) and Posner (1981). However, they go on to argue that since asymmetric information creates economic inefficiencies and reduces welfare, privacy must be welfare-reducing. This line of thought echoes the ubiquitous "nothing to hide"-argument, which Schneier (2006) has called "the most common retort against privacy advocates." As Solove (2010) points out, this argument usually takes the form: "If you aren't doing anything wrong, what do you have to hide? ... If you have nothing to hide, what do you have to fear?"

Our model allows us to argue that this argument, and hence the claim that privacy necessarily reduces welfare, is based on two faulty assumptions. Firstly, it assumes that all information is precise and unambiguous. But decisions that are made under uncertainty are routinely based on statistical discrimination. Not everybody who travels to Yemen is doing so to attend a terrorist training camp; yet it might be rational for Western governments to watch people who undertake such travels more closely – to the detriment of someone who is planning to visit his family in Yemen.

The Chicago argument therefore offers only limited guidance when it comes to actual problems of privacy. It is plausible that the first-best could be achieved in the total absence of asymmetric information. But in the real world, asymmetric information is a fact of life, and questions of privacy are therefore about *how much* asymmetric information there should be, and how it should be structured. The Chicago argument addresses an imaginary ideal case and has little to say about intermediate cases (and whether, for example, welfare is monotone in the amount of asymmetric information), which limits its use substantially.[5]

Secondly, it takes a naive ex-ante view of rational behavior. People who know that their actions are being observed often optimally change their behavior. Since traveling to Yemen can make one the target of extensive surveillance, both terrorists and non-terrorists might choose to abstain from such travels. This is clearly welfare-reducing for

[5] A similar argument against the Chicago school is made by Hermalin and Katz (2006).

the non-terrorist whose plans are disrupted, and not necessarily counterbalanced by a welfare gain from deterring the terrorists' travels.

Accepting our argument that privacy can be welfare-enhancing, and that sometimes privacy even needs to be mandated to work, also means refuting the Chicago argument that any regulation of privacy can at best be ineffective and at worst damaging.

Two recent papers have proposed rationales for privacy in public good settings where agents have an intrinsic motivation to contribute and also care about their image. That is, each agent would like others to believe that he has a high intrinsic motivation. Daughety and Reinganum (2010) show that privacy can be optimal in this setting if a lack of privacy would lead to excessive contributions due to image concerns. Ali and Bénabou (2016) add a principal who has to decide on his own contribution in a setting where agents and principal have only noisy information about the usefulness of the public good. More privacy implies that the aggregate contribution by the agents is – as a signal of the usefulness of the public good – more informative and therefore allows the principal to better choose his contribution. The mechanism in our model differs in two important ways: First, we do not rely on image concerns but micro model the downside of taking a certain action (e.g. supporting drug legalization) through an interaction with another player (e.g. a future employer). Note that image concerns are not a reduced form for this because the utility of the interacting player will be an integral and indispensable part of our welfare analysis. Second, the inference is somewhat more subtle in our model as the interacting player is not interested in the preference for action (e.g. the preference for drug legalization) but only in unobservables that are correlated with this preference (e.g. drug use). In this sense, we link the literature on statistical discrimination (Arrow, 1973; Phelps, 1972) and the literature on privacy.

Apart from such general economic studies of privacy, there is a large literature in industrial organization and related fields that deals with demand for privacy and the meaning of privacy for issues like pricing. Acquisti (2010) and Acquisti et al. (2015) provide excellent overviews; here we want to point to some studies that are loosely related to ours.

Hirshleifer (1971) argues that information revelation before trading can impair risk-sharing and therefore reduce welfare. This "Hirshleifer effect" means, for example, that providing health data about buyers of life insurance transfers risk from the seller to the more risk-averse buyers. This can be understood as an argument for privacy. Hermalin and Katz (2006) follow in a similar vein and show that privacy can be efficient in a model of price discrimination by a monopolist and a model of a competitive labor market. They also show that allocating property rights to control information does not affect equilibrium outcomes (and therefore their results) in their setup.

Similar to the second extension of our model, Acquisti and Varian (2005) consider rational reactions by people who lack privacy – for example, that internet users em-

ploy anonymization tools. They argue that this can make it unprofitable (and hence inefficient) for the seller of goods to collect information.

2 Model

The model has two stages. First an information aggregation stage in which each of n citizens has to decide whether he supports a given policy or not. Second an interaction stage in which each citizen interacts with an (outside) player.

In the information aggregation stage, each of n citizens has to voice his opinion on whether a given policy p should be implemented ($p = 1$) or not ($p = 0$). Citizen i's voiced opinion is denoted by $p_i \in \{0, 1\}$. If m citizens support implementation of p, the probability that p is carried out is $q(m/n)$. We assume that q is continuously differentiable and strictly increasing in m/n, that is, the policy is more likely to be implemented the higher the proportion of citizens who support it.[6] We also assume that q is not systematically biased towards one of the two policies, i.e. we assume that it is point-symmetric around 0.5. The payoff of the policy $p \in \{0, 1\}$ for citizen i is then $\theta_i p$. Citizen i's valuation θ_i is privately known by i. We assume that the θ_is are iid draws from a standard normal distribution Φ.

Before describing the interaction stage let us give an example for the information aggregation stage.

Example 1. *There is a petition to liberalize drug laws to a certain degree. The more citizens sign the petition, the more likely it is that its demands will be implemented. Every citizen has to decide whether to sign the petition ($p_i = 1$) or not ($p_i = 0$). Every citizen has an expected payoff consequence of liberalization of θ_i.*

We now turn to the interaction stage. Each citizen interacts in this stage with one opposing player (OP). We will describe this player as one central outside player with which each citizen interacts although nothing in the model rules out the alternative case where each citizen interacts with a different player (possibly even another citizen). OP has to choose how he interacts with citizen i and he can choose from the actions A (aggressive) or M (mild). We normalize OP's payoff from playing M to 0 and assume that the payoff of playing A against a type τ_i is simply τ_i which is a private characteristic of citizen i. The characteristics τ_i are drawn independently from a distribution Γ_{θ_i} with support in $[\underline{\tau}, \bar{\tau}]$. We assume that $\Gamma_{\theta_i'}$ first order stochastically dominates $\Gamma_{\theta_i''}$ if and only if $\theta_i' \geq \theta_i''$. This implies that θ_i and τ_i are positively correlated as higher θ_i make higher τ_i more likely.

[6] Differentiability will allow us to later analyze the effect of large n in proposition 3 – it has no effect on our other results.

To make the problem interesting, we assume that A is OP's best response if $\tau_i = \bar{\tau}$ and M is the best response if $\tau_i = \underline{\tau}$. That is, $\bar{\tau} > 0$ and $\underline{\tau} < 0$. However, OP does not observe τ_i when choosing his action and will only be able to form expectations about the citizen's τ_i. We will distinguish two cases: In the *privacy* case, we consider OP's problem when he has no information on τ_i apart from the priors Γ_{θ_i} and Φ; in particular OP does not know p_i in this case. The more complicated part of the analysis, however, will deal with the case *without privacy* in which OP observes which opinion i voiced in the information aggregation stage, i.e. OP can observe p_i and can condition his expectation of τ_i on this information. The citizen's payoff is normalized to 0 when OP plays M. If OP plays A against citizen i, then i will have a payoff of $-\delta(\tau_i)$ where $\delta > 0$ and δ is strictly increasing in τ_i. We assume that citizen payoffs from the two stages are additive. All players are assumed to maximize their expected payoff.

Figure 5.1 shows a graph of the model which we will use and modify in the following sections to illustrate our main points.

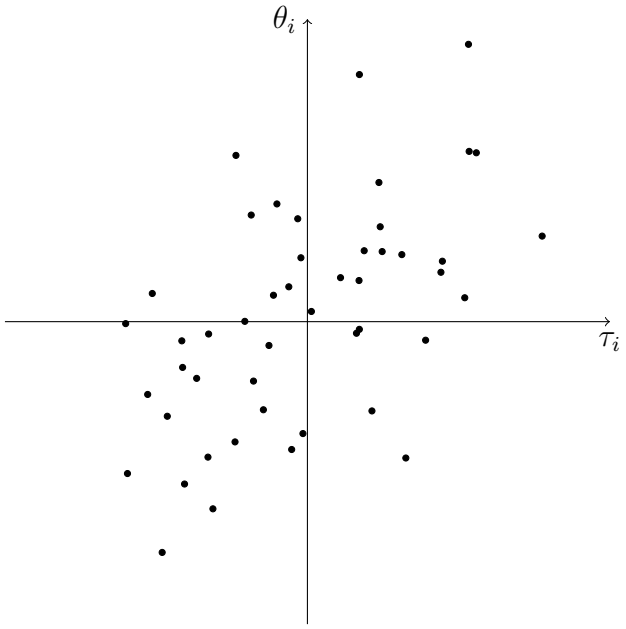


Figure 5.1: An illustration of our model. Each dot represents a citizen. Citizen i's type τ_i and θ_i are correlated (in this example: $r_{\tau\theta} = 0.6$). The OP wants to treat those with $\tau_i > 0$ aggressively and all others mildly, but he cannot observe τ_i. Citizens' policy preferences are given by θ_i; observing the choices of citizens will therefore provide the OP with information about τ_i. "Privacy" is the question whether the OP can or cannot observe an individual's policy choice before deciding how to treat her.

Example 1 (Continued)**.** *Continuing our example, OP might be a potential employer who has to decide whether to hire citizen i (action M) or not to hire i (action A). The employer would prefer to hire i if i is not a drug user and would prefer not to hire i if i is a drug user. The type τ_i would then be binary, i.e. $\tau_i \in \{\underline{\tau}, \bar{\tau}\}$, and would indicate whether*

i is a drug user or not. The first order stochastic dominance assumption on Γ_{θ_i} then simply means that the probability of being a drug user is increasing in θ_i. Hence, τ_i and θ_i are positively correlated which also means that citizens who support drug legalization are relatively more likely to be drug users than citizens opposing legalization. Citizen i prefers to be hired and the disutility of not being hired might be bigger for drug users because their outside options are generally worse.

3 Preliminary Analysis – The Chilling Effect

3.1 OP's Beliefs

We start the analysis with some preliminary results on the citizens' and OP's beliefs and strategies. This will then allow us to establish the chilling effect and analyze its welfare implications.

The payoff of citizen i from the information aggregation stage is $p * \theta_i$. The higher θ_i, the higher is i's benefit from having the policy implemented. Given this structure, it is not surprising that i will use a cutoff strategy: If θ_i is higher than some cutoff/threshold $t(\tau_i)$, i supports the policy and otherwise he does not. In the privacy case, payoffs of the interaction stage do not depend on actions chosen in the information aggregation stage and therefore i will support the policy whenever θ_i is positive.

Lemma 1. *Only cutoff strategies are rationalizable for citizens, i.e. each citizen will choose a cutoff $t(\tau_i)$ and play $p_i = 0$ if $\theta_i < t(\tau_i)$ and $p_i = 1$ if $\theta_i > t(\tau_i)$. In the privacy case, the optimal cutoff $t^p(\tau_i) = 0$.*

Given a cutoff strategy $t(\tau_i)$, we can determine the beliefs of OP in the case without privacy using Bayes' rule as

$$\beta_1(\tau) \equiv prob(\tau_i \leq \tau | p_i = 1) = \frac{\int_{\Re} \int_{\underline{\tau}}^{\tau} \mathbb{1}_{t(\tau_i) \leq \theta_i} \, d\Gamma_{\theta_i}(\tau_i) \, d\Phi(\theta_i)}{\int_{\Re} \int_{\underline{\tau}}^{\bar{\tau}} \mathbb{1}_{t(\tau_i) \leq \theta_i} \, d\Gamma_{\theta_i}(\tau_i) \, d\Phi(\theta_i)} \tag{5.1}$$

$$\beta_0(\tau) \equiv prob(\tau_i \leq \tau | p_i = 0) = \frac{\int_{\Re} \int_{\underline{\tau}}^{\tau} \mathbb{1}_{t(\tau_i) \geq \theta_i} \, d\Gamma_{\theta_i}(\tau_i) \, d\Phi(\theta_i)}{\int_{\Re} \int_{\underline{\tau}}^{\bar{\tau}} \mathbb{1}_{t(\tau_i) \geq \theta_i} \, d\Gamma_{\theta_i}(\tau_i) \, d\Phi(\theta_i)}. \tag{5.2}$$

That is, $\beta_1(\tau)$ is the probability that τ_i is below τ given that i chose $p_i = 1$. These beliefs allow us to define the expected utility of playing A conditional on observing decision p_i and given cutoff strategy $t(\tau_i)$:

$$v_1 = \int_{\Re} \tau \, d\beta_1(\tau) \tag{5.3}$$

$$v_0 = \int_{\Re} \tau \, d\beta_0(\tau). \tag{5.4}$$

The best response of OP to a given cutoff strategy is to choose A against a citizen who chose $p_i = j$ if $v_j > 0$ for $j \in \{0, 1\}$. Otherwise, it is a best response to choose M.[7]

3.2 The Chilling Effect

For the case without privacy, the following lemma states that OP is more likely to play A against citizens who have chosen $p_i = 1$ in the information aggregation stage than against citizens who have chosen $p_i = 0$. Intuitively, citizens with a high θ_i have more to gain from implementing the policy in the information aggregation stage. As θ_i and τ_i are positively correlated, OP is relatively more likely to play A against them.

Lemma 2. *In every perfect Bayesian equilibrium, $v_1 \geq v_0$.*

The previous lemma is the basis of the chilling effect. In equilibrium, OP is more likely to play A against citizen i if this citizen supported the policy, that is, voiced the opinion $p_i = 1$ in the information aggregation stage. For this reason, i is to some degree afraid of supporting the policy. More technically, there are types (θ_i, τ_i) for which a citizen would support the policy in the privacy case but will not support the policy if OP learns p_i before taking his action. The policy decision in the information aggregation stage is therefore biased against the policy in the case without privacy.

There is one minor caveat to this result: If OP's preferences are so strong that he always uses the same action, e.g. OP prefers to play M against both citizens who have played $p_i = 0$ and citizens who have played $p_i = 1$, then no chilling occurs because information on p_i is not relevant for OP's decision and therefore equilibria with and without privacy are identical. Put differently, chilling occurs whenever information about p_i matters for OP's behavior but cannot occur if this information does not affect OP's behavior.

Proposition 1 (Chilling effect)**.** *The equilibrium cutoff for every type τ_i is weakly higher without privacy than in the privacy case. The inequality is strict whenever the absence of privacy changes the equilibrium behavior of OP. The difference of equilibrium cutoff without and with privacy is increasing in τ_i.*

Figure 5.2 illustrates the chilling effect. The effect changes the behavior of citizens with moderate preferences – that is, citizens who are almost indifferent between implementing and not implementing the policy – as it simply shifts the cutoff upwards by a bit. Citizens with a very high preference for the policy will choose $p_i = 1$ with and without privacy and citizens with a very low (that is, negative) preference will choose $p_i = 0$ in both cases. Those that are almost indifferent but support the policy in the

[7]Note that OP's best response does not depend on the number of citizens choosing $p_i = 1$ in the first stage. Intuitively, this information does not contain any information about τ_i (given that p_i is known) because all θ_i and τ_i are independently drawn by assumption.

privacy case are the ones who stop supporting the policy when OP uses information about p_i. In this sense, the citizens who change their behavior do not lose a lot by their behavior change. However, citizens with strong preferences for the policy should be most worried about chilling: They do not change their own behavior but – because chilling changes the behavior of those with more moderate preferences – the policy will with some probability not be implemented without privacy though it would have been implemented with privacy. Those citizens with a strong preference for the policy value the policy most and therefore have all reasons to be worried about other citizens being chilled. In short, privacy changes the behavior of moderate people and protects people with extreme preferences.

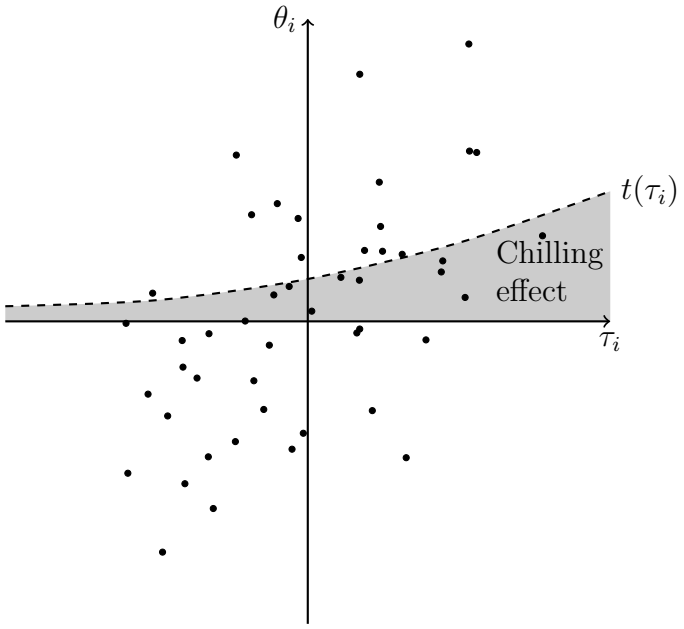


Figure 5.2: An illustration of proposition 1. If decisions are private, all individuals with positive θ_i will support $p = 1$ and all others support $p = 0$. If decisions are public, the OP can use the individuals' decisions to predict their type τ_i. Therefore, some people with relatively low θ_i will misrepresent their preferences to avoid the statistical discrimination. Since the disutility from being treated aggressively rises in τ_i, we get the curve above. Individuals in the grey area are subject to the chilling effect and support $p = 0$ without privacy.

The result that the cutoff is shifting more for citizens with a high type τ_i implies that the cutoffs of higher τ_i are higher. As a consequence, abolishing privacy becomes somewhat less profitable for OP compared to the case where citizens use the same cutoff: The fact that higher τ_i have higher cutoffs reduces the correlation between θ_i and τ_i. This correlation is exactly the reason why discriminating between those citizens who support and those who do not support the policy is beneficial for OP in the first place. Hence, OP's benefits from statistical discrimination are reduced by the chilling effect. This means that an evaluation of whether privacy should be given up will be biased against

privacy if it does not take the behavior change of citizens caused by the removal of privacy into account. The following proposition makes this statement more formally. To do so, we have to add the technical condition that the distribution Γ_0 is symmetric around 0. This ensures that OP does not gain from the fact that all cutoffs increase (while the argument above shows that it is detrimental to OP that cutoffs of higher τ increase by a larger amount). Note that the following proposition does not compare OP's payoffs under privacy and no privacy. In line with the argument above, it compares OP's payoffs without privacy with his payoffs in a hypothetical situation where there is no privacy but citizens use their equilibrium strategies of the privacy case.

Proposition 2. *Assume that the distribution $\Gamma_0(\tau)$ is symmetric around $\tau = 0$.*[8] *OP's payoff without privacy is lower if citizens use the cutoffs $t^{np}(\tau)$ than if they used the cutoffs $t^{p}(\tau) = 0$.*

4 Welfare Analysis

What are the welfare consequences of the chilling effect? It is not hard to see that the chilling effect causes a welfare loss in the information aggregation stage. The bias against the policy means that information is no longer efficiently aggregated and decision 0 is more likely to be taken than optimal. The following lemma states formally that the privacy equilibrium yields a higher expected consumer surplus in the information aggregation stage than the equilibrium without privacy. (We define consumer surplus in the information aggregation stage as $p\sum_{i=1}^{n}\theta_i$.)

Lemma 3. *The cutoff strategy $t^{p}(\tau) = 0$, i.e. the equilibrium strategy in the privacy case, gives a higher expected consumer surplus in the information aggregation stage than any $t^{np}(\tau) > 0$.*

While this shows that individuals are always better off under privacy, this does not allow us to say anything about overall welfare. Without privacy, the OP can adjust his behavior according to people's policy choices p_i and thereby make use of the correlation between θ_i and τ_i to identify individuals with a relatively high τ_i. To avoid case distinctions, we will concentrate in the remainder of this section on the case where OP plays M in the privacy equilibrium, i.e. the unconditional expectation of τ is negative: $\mathbb{E}[\tau] \leq 0$. (This also seems to be the more relevant case in most applications mentioned before.)

Concerning overall welfare, we will derive three strong results that establish sufficient conditions for when privacy is welfare-optimal both for the citizens and the OP. Firstly, if the OP plays a mixed strategy in equilibrium (i.e. he mixes between treating people who choose $p_i = 1$ mildly or aggresively), privacy always provides higher welfare than

[8] An alternative technical condition that is also sufficient for the result to hold is $\mathbb{E}[\tau_i|\theta_i = 0] \geq 0$.

no privacy. This simply follows from the fact that while individuals always lose from lack of privacy, the OP is indifferent between privacy and no privacy if he plays a mixed strategy in the no privacy case.

Secondly, we show that for large n, i.e. if there are many individuals, there exist no equilibria in which the OP plays a pure strategy, and privacy is therefore optimal.

Thirdly, we show the same for large $\delta(\tau)$ – in other words, if the cost of being treated aggressively is very high, then privacy is also welfare optimal.

Proposition 3. *Assume OP plays M in the privacy equilibrium.*
1.) If OP uses a mixed strategy in the equilibrium without privacy, then privacy maximizes welfare.
Assume that (i) δ is differentiable and strictly increasing in τ, i.e. $\delta'(\tau) > 0$ for all $\tau \in [\underline{\tau}, \bar{\tau}]$ and (ii) $\Gamma_\infty = \lim_{\theta_i \to \infty} \Gamma_{\theta_i}$ is a non-degenerate distribution in the sense that $\Gamma_\infty(\tau_i) > 0$ for all $\tau_i > \underline{\tau}$.
2.) Privacy welfare dominates no privacy for large n in the following sense: Compared to the no privacy case, privacy leads to a higher expected consumer surplus and the same expected payoff for OP.
3.) Let the disutility of a citizen facing action A by OP be $r\delta(\tau)$ (instead of $\delta(\tau)$). For r sufficiently large, privacy welfare dominates no privacy.

The intuition behind the second result is that the chilling effect is getting very large if the number of citizens grows. If n is large, each individual citizen only has a small influence on the outcome of the information aggregation stage. This implies that pure strategy separating equilibria no longer exist. Put differently, if OP played A against citizens who chose $p_i = 1$ and played M against citizens who played $p_i = 0$, then citizens would find it optimal to play $p_i = 0$ in order to avoid the aggressive reaction by OP. With a low number of citizens this incentive is countervailed by the downside of playing $p_i = 0$ (when θ_i is positive) which is a worse policy result in the information aggregation stage. As n grows large the own impact on this policy decision is, however, negligible and this negative effect cannot deter citizens from biasing their stated opinion. OP will, therefore, use a mixed strategy in equilibrium. Hence, OP will be indifferent between his two actions and therefore also between privacy and no privacy (otherwise using a mixed strategy would not be optimal). As citizens are clearly worse off without privacy because of the biased information aggregation and the possibly increased probability of A in the interaction stage, the privacy case welfare dominates.

The intuition for the third result is similar: If δ is high, then the benefit from the information aggregation stage is relatively small compared to the potential losses in the interaction stage. Citizens will therefore be chilled a lot if OP plays A against citizens who chose $p_i = 1$. Playing A for sure against those who chose $p_i = 1$ is then no longer a best response. Consequently, OP uses a mixed strategy for r sufficiently high and privacy

welfare dominates.

Note that all the welfare results in proposition 3 are Pareto results *from an ex ante point of view*. That is, privacy makes citizens strictly better off in expectation (i.e. before knowing their type) while the OP is indifferent.

5 Alternative Utility Specifications

In this section, we consider two alternatives to the information aggregation in the first stage modeled so far. First, we consider a setup where citizen i's utility does not depend on choices of other citizens. That is, the first stage decision p_i has nothing to do with information aggregation but is simply a private decision without externalities. Second, we consider a setting in which there is again information aggregation but citizen i's payoff of implementing the policy is given by a common state θ (instead of a personal payoff parameter θ_i). This state, however, is unknown as citizens obtain only noisy private signals of the true state θ. As we will see, similar results to the ones above hold in these setups and some additional insights can be obtained.

5.1 First Stage With Private Decisions Instead of Information Aggregation

We want to consider a setup where individual i's choice (p_i) directly influence his welfare. Our model covers this latter case if we set $n = 1$. This could be the case for people listening to music, attending certain events or meeting certain people, which also is informative about some hidden type. Then p_i has a private consequence, and all of our results (about the existence of the chilling effect) continue to hold. But we can no longer argue that as each individual becomes less and less privotal with larger n, the chilling effect increases and ultimately makes the statistical discrimination ineffective (proposition 3). In our example, this welfare question could be: If a preference for Reggae music is correlated with drug use, should the employer be able to observe, and base his decision on, the music that Alice listens to?

Instead of concentrating on $n = 1$, we will simply assume that every individual's choice directly influences her payoff, so that her payoff from choosing p_i is simply $p_i\theta_i$. That is, we model n citizens but their welfare is independent from one another.[9] In the privacy case, preferences are the same as before. Without privacy, individuals experience a chilling effect that now only depends on the behavior of the OP and the function $\delta(\tau)$, and no longer on their beliefs about the behavior of others.

[9] As this is equivalent to having n of our original models with one citizen each, it is clear that out earlier intermediate results, i.e. lemmas 1–3, still hold.

If the OP plays a mixed strategy in the equilibrium without privacy, we can show just as in proposition 3 above that privacy is pareto optimal, since people are better off with privacy and the OP is indifferent. The interesting case is the one in which OP uses a pure strategy, i.e. plays A against everybody who chooses $p_i = 1$ and M otherwise.

In this case, the OP is better off without privacy (otherwise he would always play M), while individuals are on average worse off. To be able to say anything about welfare, we need to aggregate their respective payoffs.

We can do so on an individual-by-individual basis, that is, we ask: For a given individual, what does this individual lose by losing privacy, and what does the OP gain? This is in line with viewing the OP as a representative of society (for example, the police trying to catch criminals or terrorists), or thinking of the OP as being a group of other players, or even of every individual acting as OP to another individual.

In the case where the OP uses a pure strategy, we can therefore write welfare as[10]

$$\sum_i p_i(\theta_i + \tau_i - \delta(\tau_i)).$$

When is welfare higher without privacy than with privacy? Intuitively, if the correlation between θ_i and τ_i is quite small, then the OP's gain from being able to distinguish individuals according to type is also small, while the individual's loss from not being able to choose their preferred p_i (or being treated aggressively if they do) only depends on $\delta(\tau_i)$. For a given δ, the correlation between θ_i and τ_i would therefore have to be sufficiently high to make no privacy welfare-optimal. Figure 5.3 illustrates this intuition.

If we want to analyze the connection between correlation and δ, we need to restrict the problem by imposing partial orderings of joint distributions, since the set of possible joint distributions is extensive and otherwise intractable. We will therefore make our argument in two different ways. Firstly, we will consider the special case of $\delta(\tau) = \delta$ being a constant. In this case, we can show quite generally that welfare is decreasing in δ and increasing in the correlation between θ_i and τ_i, and that for higher δ, no privacy is only optimal if the correlation is very high.[11] Secondly, we will consider the more general case of $\delta(\tau)$ being an increasing function while restricting the joint distribution of θ_i and τ_i to a family of distributions which are convex combinations of a correlated and an uncorrelated distribution. In both of these cases, we will show that for a given cost function for the individuals (i.e. a given δ), privacy is optimal both for the indidviuals and the OP unless the correlation between θ_i and τ_i is sufficiently high.

Consider first the case in which we only consider constant δ. In this case, we can show quite generally that for a higher δ, the correlation between θ_i and τ_i needs to be

[10] We could use weights to sum up payoffs, but since we have made no assumptions about the magnitude of δ, a scaled δ function would just be another δ function and nothing qualitative would change.

[11] Of course, our result from above still applies: For very large δ, privacy is always welfare optimal.

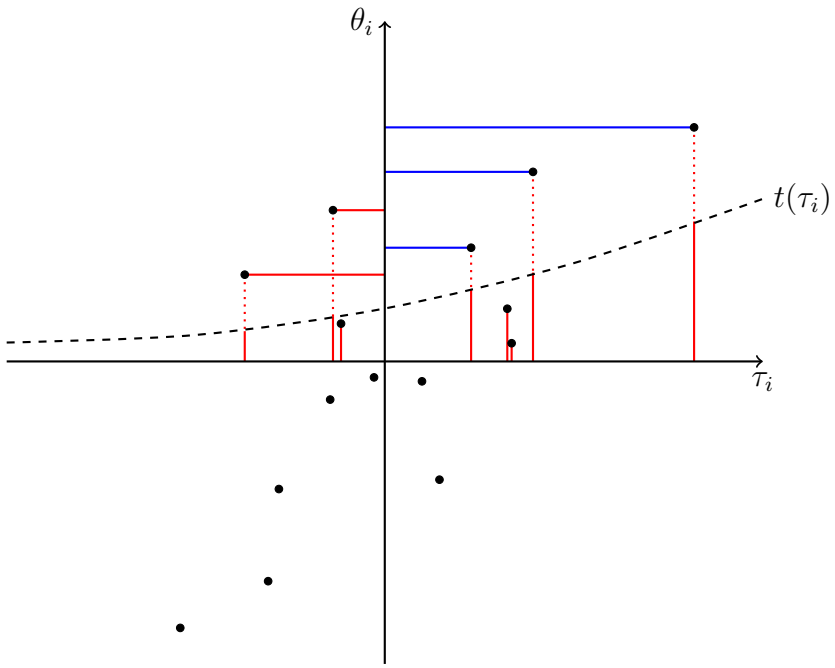


Figure 5.3: Gains (blue) and losses (red) from lack of privacy (compared to privacy). Losses of the individuals are vertical, losses of the OP are horizontal. The sum of the lengths of all blue lines is the overall gain, the sum of all (solid) red lines is the overall loss. Every individual with $\theta_i > 0$ loses either θ_i (if she chooses $p_i = 0$) or $\delta(\tau_i)$ (if she chooses $p_i = 1$ and therefore gets treated aggressively). The OP gets τ_i for every individual who still chooses $p_i = 1$. Intuitively, if we increase correlation between θ_i and τ_i, individuals with $\theta_i > 0$ move to the left (as their expected τ_i increases, which increases the gain of the OP.

higher to make no privacy welfare-optimal. To be able to make this statement, we need a (partial) ordering on the possible joint distributions of θ_i and τ_i. Recall that Γ_{θ_i} is the distribution of τ_i given θ_i; and that we have already assumed that $\Gamma_{\theta_i'}$ first-order stochastically dominates $\Gamma_{\theta_i''}$ if and only if $\theta_i' \geq \theta_i''$. Furthermore, we now assume that $\mathbb{E}[\tau_i|\theta_i = 0] \geq 0$ so that expected τ_i is positive for $\theta_i > 0$ – this guarantees that the OP wants to treat individuals aggressively if their θ_i is positive. We will now say that the correlation is higher in distribution Γ' than in distribution Γ'' if for every $\theta_i > 0$, Γ'_{θ_i} first-order stochastically dominates Γ''_{θ_i}. The following proposition shows that welfare is decreasing in δ and increasing in the correlation between θ_i and τ_i.

Proposition 4. *The welfare difference between no privacy and privacy is decreasing in δ and increasing in Γ.*

We move now to the second case of expressing "weak" correlation (and we allow again for increasing, i.e. non-constant, $\delta(\tau)$). We continue to focus on the interesting case in which – given distributions $\Gamma_{\theta_i}(\tau_i)$ – there is an equilibrium in which OP plays A (M) against those who support (do not support) the policy. We also assume that OP's best response in the privacy case is M. Now consider the marginal distribution of τ_i which we

denote by $\bar{\Gamma}$, that is

$$\bar{\Gamma}(\tau_i) = \int_{\Re} \Gamma_{\theta_i}(\tau_i)\, d\Phi(\theta_i).$$

If for every given θ_i the distribution of τ_i was $\bar{\Gamma}$, then there would be no correlation between θ_i and τ_i and even knowing θ_i directly (instead of p_i) would not yield any benefit for the OP. We will now consider convex combinations of the original distributions Γ_{θ_i} and the distribution $\bar{\Gamma}$. Denote these convex combinations by

$$\Gamma^{\lambda}_{\theta_i}(\tau_i) = \lambda\Gamma_{\theta_i}(\tau_i) + (1-\lambda)\bar{\Gamma}(\tau_i) \qquad \lambda \in [0,1].$$

For $\lambda = 1$ we are in the original problem. Decreasing λ, however, continuously decreases the correlation between θ_i and τ_i. For $\lambda = 0$, there is no correlation between these two variables left. If there is no correlation, then the equilibrium is the same as in the privacy case because OP does not get any information about τ_i from the policy choice of the individuals. Hence, the equilibrium is that OP plays M against everyone and citizens use the cutoff 0 if $\lambda = 0$. This is true regardless of whether there is privacy or not. By continuity, the same is true for low but positive λ. More interestingly, we establish in the following proposition that there is an intermediate range of λ where privacy is strictly welfare superior to no privacy. That is, if correlation is weak privacy welfare dominates (and if the correlation is very weak privacy and no privacy are welfare equivalent).

Proposition 5. *There exist* $0 < \underline{\lambda} < \bar{\lambda} < 1$ *such that*

1. *for* $\lambda \leq \underline{\lambda}$ *privacy and no privacy are welfare equivalent and*
2. *for* $\lambda \in (\underline{\lambda}, \bar{\lambda}]$ *privacy leads to strictly higher welfare than no privacy. The equilibrium for* $\lambda = \bar{\lambda}$ *is in pure strategies.*

5.2 Citizens with Aligned Preferences Under Uncertainty

This subsection considers an alternative model where the private information of citizens in the information aggregation stage is not directly their personal payoff of policy $p = 1$. Instead, citizens all have the same payoff of policy $p = 1$ but each citizen only receives a noisy signal of this payoff. That is, there is an unknown state of the world θ; each citizen has a noisy signal about the state of the world and citizens try to "match the state", i.e. they prefer policy $p = 1$ if the state is positive and $p = 0$ if the state is negative.

This has a striking implication: Lack of privacy makes *every* citizen worse off, since the chilling effect inhibits information aggregation. In our main model, citizens have private preferences over outcomes and therefore some citizens (those with negative θ_i) gain from chilling. Since all citizens now have the same interest – implementing the policy that matches the state – everyone loses from chilling. Hence, our welfare result

for large n in proposition 3 is somewhat stronger in this setting as privacy is now a Pareto improvement not only at the ex ante but even at the interim stage, i.e. after each citizen has observed his signal.

The details of the setting are as follows: The state of the world θ is distributed standard normally and this θ is the payoff consequence of policy $p = 1$ for each citizen. However, the realization of θ is unknown. Each citizen obtains a private signal θ_i which is normally distributed around the true state θ, i.e. $\theta_i \sim N(\theta, \sigma^2)$ where we denote the cdf by $\tilde{\Phi}_\theta$ and the pdf by $\tilde{\phi}_\theta$. All θ_i are assumed to be independent draws from this distribution. The interaction type τ_i of citizen i is drawn from Γ_{θ_i} where again $\Gamma_{\theta_i'}$ is assumed to first order stochastically dominate $\Gamma_{\theta_i''}$ if and only if $\theta_i' > \theta_i''$. This creates a positive correlation between θ_i and τ_i. The interaction stage is exactly the same as in our main model. That is, without privacy a strategy for OP states which of the two actions (A and M) OP plays against a citizen who chose $p_i = 0$ or $p_i = 1$. With privacy, OP only decides which of the two actions he chooses against all citizens. This means that – to keep the setting comparable to the main model – we do not consider strategies (or beliefs) that are contingent upon the number of citizens choosing $p_i = 1$. This is a simplification. However, one can easily imagine settings where OP has to commit to a strategy before he gets to know the citizens' p_is. This is, for example, the case if the interaction is between i and an agent representing OP and p_i is only learned in the interaction. OP then has to instruct the agent in advance how to act.

In the supplementary material to this paper, we provide proofs that are mostly analogous to those of our main model. In particular, the absence of privacy causes a chilling effect and this chilling effect inhibits efficient information aggregation. For large n, there are still only equilibria where the OP mixes, and in any equilibrium where the OP mixes, the OP is indifferent between privacy and no privacy. Now, however, we have Pareto dominance in the sense that every citizen of every type is better off under privacy.

6 Extensions

This section contains two extensions to the main model. First, we show that privacy might have to be mandated, i.e. privacy as an opt in possiblility will lead to the no privacy outcome. Second, we consider the possibility of a defensive action against the OP and use this setup to show that in some scenarios privacy can even make the OP strictly better off.

6.1 Privacy as Opt In

Suppose that citizens have an additional decision in the information aggregation stage: They do not only have to choose p_i but also have to decide whether their choice should

be private or public. OP can observe all public choices but not the private ones – in this case he can only observe that the citizen chose privacy. To isolate the effect of the privacy choice, we will also assume that OP cannot make his behavior contingent on the policy outcome p (which might be realized only at a later point of time). Furthermore, let us assume that M is optimal for OP in the privacy case, i.e. that $\mathbb{E}[\tau] < 0$.

The possibility of hiding one's choice gives rise to multiple equilibria. To see this, consider first an equilibrium in which every citizen always chooses "public" (no matter what θ_i or τ_i). Then the equilibrium of the no privacy case results.[12] Second, consider an equilibrium in which every citizen always chooses "private". This means that we are effectively in the case with privacy. OP's best response is to play M and consequently no citizen has an incentive to deviate.

Naturally, the question arises which of the two equilibria is more robust. We will argue in two different ways that the "always private" equilibrium is not very robust. The reason is an unraveling logic. Citizens who choose $p_i = 0$ are not afraid of making this public as it indicates that their θ_i is low which means that their expected τ_i is also relatively low because of the positive correlation between the two. Given that the expected τ_i is low, OP would therefore still play M against citizens who make a choice $p_i = 0$ public. If, however, everyone who chooses $p_i = 0$ makes this public, then making one's choice private is not different from publically choosing $p_i = 1$.

The simplest way to formalize this intuition is to assume that making one's choice p_i private comes at a small cost $\varepsilon > 0$. In this case, the "all private" equilibrium would only be supported by off equilibrium beliefs such that both $\mathbb{E}[\tau|\text{"public"}, p_i = 0] \geq 0$ and $\mathbb{E}[\tau|\text{"public"}, p_i = 1] \geq 0$ as OP could then threaten to play A against any citizen making his decision public (thereby saving the $\varepsilon > 0$ costs). Given that $\mathbb{E}[\tau] < 0$, these are straightforwardly unreasonable beliefs. In terms of equilibrium refinements, the equilibrium does not satisfy the well known D1 criterion of Banks and Sobel (1987). Roughly speaking, this refinement states the following for our game: Denote by $D(\theta_i, \tau_i)$ the set of OP mixed strategies that are (i) best responses for some OP belief and (ii) would make a deviation by a citizen of type (θ_i, τ_i) profitable. D1 requires that OP's off path beliefs must be zero for type (θ_i', τ_i') if there is a type (θ_i'', τ_i'') such that $D(\theta_i', \tau_i')$ is a strict subset of $D(\theta_i'', \tau_i'')$. It is straightforward to show that the "all private" equilibrium does not satisfy D1. The reason is that the off path beliefs supporting the "all private" equilibrium require that deviations to public stem from citizens with relatively high τ_i no matter whether p_i is zero or one. As δ is increasing in τ_i, there are mixed strategies by OP which would make the deviation profitable for citizens with low τ_i (who are less afraid of action A) but not for citizens with high τ_i. The "all public" equilibrium, on the other hand, satisfies D1.

[12] This equilibrium is supported by the following off equilibrium path belief: if a player chooses "private", OP believes that τ_i is sufficiently high so that A is a best response.

The second way in which the "all private" equilibrium is not robust is the following. Assume that with probability $\varepsilon > 0$ OP has alternative preferences $\tau_i + \varepsilon'$ for playing A. Assume that ε' is such that $\mathbb{E}[\tau] + \varepsilon' > 0$. That is, under the alternative preferences OP plays A given his prior beliefs. Suppose further that these alternative preferences are such that $\mathbb{E}[\tau|\theta_i \leq 0] + \varepsilon' < 0$, i.e. knowing that θ_i is negative OP still best responds by playing M. Again the "always private" equilibrium could then only be sustained by off path beliefs leading to $\mathbb{E}[\tau|$"public", $p_i = 0] + \varepsilon' \geq 0$ and $\mathbb{E}[\tau|$"public", $p_i = 1] + \varepsilon' \geq 0$. As pointed out above, such beliefs are unreasonable and violate the D1 refinement.

6.2 Defensive Actions

Suppose that citizens have the opportunity to take a defending action against being treated aggressively. More precisely, a citizen can take an action D which increases his payoff if OP plays A but decreases his payoff if OP plays M. The defensive action reduces OP's payoff. In our example, Alice could hire a lawyer. Hiring the lawyer is costly but the lawyer will make it harder for the employer to discriminate against Alice. For the employer, dealing with a lawyer is a hassle (whether he discriminates or not) and reduces his payoffs.

What we want to show in this section is that the model can easily be extended in this way and that privacy could lead to (i) OP being *strictly* better off with privacy while (ii) citizens being in expectation strictly better off with privacy. Hence, privacy can be strictly Pareto superior from an ex ante point of view. To this end, it is sufficient to present an example and this is what we are going to do. Suppose $\tau_i \in \{\underline{\tau}, \bar{\tau}\}$, that is, τ_i can have only one of two values. Furthermore, assume that the probability that $\tau_i = \bar{\tau}$ equals

$$\gamma_{\theta_i} = \begin{cases} 0.7 - \frac{0.3}{\theta_i + 1} & \text{if } \theta_i \geq 0 \\ 0.1 - \frac{0.3}{\theta_i - 1} & \text{if } \theta_i < 0. \end{cases}$$

That is, the probability of a high $\bar{\tau}$ is increasing in θ_i and is point symmetric around $(0, 0.4)$. We take $\underline{\tau} = -2$, $\bar{\tau} = 3$ and $\delta(\tau_i)$ as given as in table 5.1.

action/type	$\underline{\tau}$	$\bar{\tau}$
not D	-0.1	-0.125
D	0.0	-0.025

Table 5.1: $-\delta(\tau_i)$ depending on whether the defensive action is taken.

If a citizen takes action D and OP plays M, his payoff is -0.1, that is, the costs of the action D are 0.1. Note that the citizen wants to play D if the chance of A is higher than 1/2. OP payoffs are reduced by 1 if a citizen plays D (for simplicity the payoff reduction is assumed to be independent of OP's action).

Under privacy, it is an equilibrium that every citizen chooses $p_i = 1$ if and only if $\theta_i \geq 0$ while in the second stage OP plays M and no citizen takes the action D. Without privacy, this is no longer an equilibrium as OP prefers to deviate by playing A against all citizens choosing $p_i = 1$: The probability that a citizen is of type $\tau_i = \bar{\tau}$ given $\theta_i \geq 0$ (and therefore $p_i = 1$) is

$$\frac{\int_0^\infty \gamma_{\theta_i}\, d\Phi(\theta_i)}{2} \approx 0.51$$

which implies that OP's best response is A.

Let the probability that alternative 1 is chosen in stage 1 given m citizens supporting it be $q(m/n) = m/n$. Then we get the following equilibrium in the case without privacy: Citizens use cutoff strategies characterized by cutoffs $t(\underline{\tau}) = 0$ and $t(\bar{\tau}) = N{*}0.025$. In the second stage, those citizens that chose $p_i = 1$ will play D. OP plays A against all citizens that chose $p_i = 1$ and M otherwise. To see that this is an equilibrium, note that a citizen of type $(\theta_i, \tau_i) = (0, \underline{\tau})$ is indeed indifferent between choosing $p_i = 0$ and not playing D, which gives a payoff of 0 as OP will play M, and $p_i = 1$ and playing D which also gives a payoff of 0 as OP will then play A. Similarly, a citizen of type $(\theta_i, \tau_i) = (0.025N, \bar{\tau})$ is indifferent between choosing $p_i = 0$ and not playing D and choosing $p_i = 1$ and playing D. The reason is that choosing $p_i = 1$ increases the probability of policy 1 being chosen by $1/N$ and therefore the expected payoff of a citizen with $\theta_i = 0.025N$ by 0.025. However, the down side of choosing $p_i = 1$ is that the payoff in the interaction stage is 0.025 lower as $-\delta(\bar{\tau}) = -0.025$ (when playing D). For OP, the probabilities

$$\begin{aligned} prob(\tau_i = \bar{\tau} | p_i = 0) &= \frac{\int_{-\infty}^{0.025N} \gamma_{\theta_i}\, d\Phi(\theta_i)}{\int_{-\infty}^{0.025N} \gamma_{\theta_i}\, d\Phi(\theta_i) + \int_{-\infty}^{0} 1 - \gamma_{\theta_i}\, d\Phi(\theta_i)} \\ prob(\tau_i = \bar{\tau} | p_i = 1) &= \frac{\int_{0.025N}^{\infty} \gamma_{\theta_i}\, d\Phi(\theta_i)}{\int_{0.025N}^{\infty} \gamma_{\theta_i}\, d\Phi(\theta_i) + \int_{0}^{\infty} 1 - \gamma_{\theta_i}\, d\Phi(\theta_i)} \end{aligned}$$

are such that playing A (M) against those that chose $p_i = 1$ ($p_i = 0$) is optimal, i.e. $prob(\tau_i = \bar{\tau} | p_i = 0) \leq 0.4 \leq prob(\tau_i = \bar{\tau} | p_i = 1)$, if $N \leq 22$.[13]

OP's expected payoffs in the equilibrium without privacy are $-0.043 * N$ while OP profits with privacy are zero. Citizens are strictly worse off if they have $\theta_i > 0$: The reasons are (i) that some are chilled and therefore expect a lower payoff from the information aggregation in stage 1, (ii) those that are not chilled have to endure action A by OP (and have to bear the costs of the defensive action). Citizens with $\theta_i < 0$ benefit from the chilling of other citizens as this chilling implies that their personally preferred alternative is more likely to be implemented. Note, however, that – by the symmetry of the setup – this only offsets the first negative effect on those with $\theta_i > 0$ (in expectation, e.g. behind the veil of ignorance). The second negative effect on those with $\theta_i > 0$ lowers

[13] For $N > 22$, no pure strategy equilibria exist without privacy and OP will therefore be indifferent between privacy and no privacy – cf. proposition 3.

the expected payoff of a citizen.

7 Discussion

7.1 Which Discrimination Should be Permitted: Credit Scores

A bank has to decide to whom to lend. Ideally, it would like to base its decision on the probability that a debtor will repay the loan, but this variable is not directly observable. Instead, the bank can rely on measures that indirectly predict default probability. There are several socioeconomic variables that are easily observed and correlated with default risk, such as national origin, race, gender, age, or place of residence. Using such variables to make credit decisions, and hence treat native-borns, whites or women differently solely because of their identity, is illegal in many countries. In the United States, for example, such "redlining" practices are explicitly outlawed by the Equal Credit Opportunity Act (ECOA) of 1974.

Imagine, however, that the bank starts looking for other pieces of data that can inform its decision and allow it to statistically discriminate among loan applicants. Two such pieces of information are the education level (which can easily be documented by the applicant) and the taste in music (which many millions of people reveal on websites like Facebook, last.fm and similar services). While the former is common practice, the latter is (on purpose) more speculative but not implausible: Facebook owns a patent on aggregating credit scores from the data it collects about its users, and there are many firms that claim to make use of big data to develop more accurate credit scores.[14] We would expect that a preference for some genres of hip hop, since it is correlated with socioeconomic status, can be highly predictive of default risk. The expressed music preference would then be the variable p_i that the bank uses to discriminate between people who do and those who don't get loans, and our model would consequentially predict a chilling effect in which some hip hop fans are held back in their freedom of expression, since they want to improve their credit rating. The individual loss (by not being able to express your own personality) is probably more substantial than the loss in information aggregation here, but it is a welfare loss nonetheless (cf. the results in section 5.1). Note also that those who care too much to be stifled in their music appreciation will find it harder to get a loan, regardless of whether that is justified by their actual creditworthiness or not.

But what is most important is that fans of gangsta rap tend to be similar to each other in many ways, so that the use of inocuous (and predictive) music preference data allows the bank to discriminate based on ethnicity, age and geography without explicitly

[14] One of them, Zest Finance, advertises with the slogan: "All Data is Credit Data." (https://www.zestfinance.com/how-we-do-it.html, retrieved May 2, 2016.)

saying so. This points to a larger question to which our research contributes, but to which we have no definitive answer: What should banks, employers, governments be allowed to discriminate upon? Most people would probably agree that to treat someone better or worse purely because of race or gender is not acceptable (and that contrary to the arguments made by Friedman (1962), such discrimination will not automatically disappear as it can be rational, as pointed out by Arrow (1973) and Phelps (1972)). But demanding that job applicants have a diploma, or giving loans based on past income, is also statistical discrimination: these factors are predictive of whether the employee will be up to the task or the loan will be repaid, but the correlation is less than 1.

Our first extension suggests that an equilibrium where everyone keeps their music preferences secret is not stable (especially if there is some payoff to sharing them). Regulation which prohibits the use of some data for credit decisions, beyond existing laws like the ECOA, could therefore be welfare-enhancing. In particular, recall that our model only requires that some variables are correlated without being causally related. Beside music taste, many other variables are probably correlated with both creditworthiness and race or gender without having any causal relationship with either of these. Clever bankers, or even mindless machine learning algorithms, could pick up on those relationships and use them to improve their credit algorithms, with all the consequences that we have described in our analysis. Our results would therefore strongly support the regulatory use of positive lists, which specify which data can be legally used in credit decisions (as opposed to negative lists, which only specify which data cannot be used).

7.2 "The Tape Has Had Some Chilling Effect": Decision-Making and Transparency

The last decades have seen a move towards transparency in many public bodies – governments, authorities, central banks. But to the extent that the quality of decisions in these institutions depends on aggregating the information of their employees and members, our results suggest that transparency does not necessarily improve welfare.

Consider, for example, the board of a central bank that has to decide on an interest rate change. If the deliberations are private and no minutes are made public, board members express their opinion quite freely.[15] If minutes are later published, however, members will worry about the effect of what they say on their reputation. Assume that board members have different degrees of competence, and that the probability of being wrong about something decreases in one's competence. Board members want to be thought of as competent by the public, their academic colleagues or future employers. Now, a board member considers whether to make an unconventional suggestion. This

[15]This is under our standard assumption that arguing one's viewpoint increases the probability that one's preferred policy will be implemented.

suggestion has some probability of being wrong, and in that case outside observers would adjust their belief of the board member's competence downwards. Publicity can therefore induce him to stay quiet.

This is in line with the results by Meade and Stasavage (2008) and others who examine the effects of a reform introduced in 1993, which mandated that minutes from meetings of the Federal Open Markets Committee (FOMC) of the U.S. Federal Reserve should be published. The reform has significantly increased conformity in the discussion and decreased the number of people who criticized the chairman's proposed interest rate adjustment. There was a strong shift away from free discussion and towards the reading of prepared statements. Thomas Hoenig, president of the Federal Reserve Bank of Kansas City, remarked in a meeting in 1995 that "the tape has had some chilling effect on our discussions. I see a lot more people reading their statements." (Meade and Stasavage, p. 13). Alan Greenspan also warned of this development before the reform was implemented: The FOMC "could not function effectively if participants had to be concerned that their half-thought-through, but nonetheless potentially valuable, notions would soon be made public."(Meade and Stasavage, 2008, p. 12)

Our model therefore suggests that if board members, government ministers or civil servants are worried about how they are being perceived by the outside world, secret meetings can substantially increase the quality of decision making.[16] But this is not universally true, of course. If the correlated types of our model reflect "private interests" (for example, stock holdings by family members) and "policy preferences", privacy would allow the board members to follow their private interests without having anything to fear – which would not improve decision making in the public interest. Privacy is not a panacea, but neither is transparency.

8 Conclusion

Why should an individual care about his or her privacy, why should a society care about the privacy of its members? We have argued that since asymmetric information is a fact of life, questions of privacy are never about whether there should be private information or not, but only how much there should be and how it should be structured. That allows us to answer: Indviduals can worry that information about them could be used "against them", i.e. expose them to discriminative treatment. This result does not require ill will among the discriminator – the discrimination can be perfectly rational, as in the case of the employer trying to distinguish applicants. But it will make it harder for people to choose according to their preferences, and the rational reaction of individuals to having

[16]Consider also the literature on reputation concern and advice, such as Ottaviani and Sørensen (2006), which would also suggest that advisors are more helpful if they are unconcerned about their reputation.

no privacy can impair the ability of a society to efficiently aggregate information. Privacy is not only individually optimal, but also welfare-enhancing.

Our examples show, however, that privacy is not a silver bullet. The solution to problems of "redlining" and new forms of discrimination in lending is not to prohibit borrowers from revealing any information about themselves; and not all goverments would be improved by being able to work in total secrecy. Our analysis allows us to say, however, when privacy is likely to improve welfare. When people's preferred actions under privacy guarantee an outcome that is optimal or close to optimal, the chilling effect decreases welfare. This is the case, for example, when actions have no significant externalities, or when the gains from correctly aggregating information are large.

Apart from the welfare effects, privacy often has a distributive effect: In our main model, there are always people whose preferred policy becomes less likely to be implemented under privacy. (In section 5.2, however, we argue that there can be situations where privacy improves everybody's ouctome.) Others gain: Those who would be subject to the chilling effect without privacy are more likely to get their preferred option with privacy. Moreover, those with strong preferences gain twice from privacy: They are no longer statistically discriminated against, and their preferred option is more likely to be implemented. How should such distributive effects influence whether privacy is implemented? We have no definitive answer, but would like to point out that similar distributive effects arise with free speech: On any single issue, everybody would prefer if those with opposing viewpoints were prohibited from expressing it. Yet in the abstract, most of us would agree that freedom of expression should be universal.

We started this paper by criticizing the "Chicago view" of privacy as inefficient and economically undesirable. But as we have argued that privacy can be fundamental to allowing individuals to freely express themselves, we are returning to an argument by perhaps the most well-known Chicago theorist. Friedman (1962, p. 52), in his discussion of "rules instead of authorities", discusses the question of whether free speech issues should be decided from case to case, or in the abstract. He concludes that:

> When a vote is taken on whether Mr. Jones can speak on the corner, it cannot allow [...] for the fact that a society in which people are not free to speak on the corner without special legislation will be a society in which the development of new ideas, experimentation, change, and the like will all be hampered in a great variety of ways that are obvious to all.

Our analysis suggests that a similar argument can be made about privacy.[17]

[17]It has been pointed out to us that the whistleblower Edward Snowden drew a similar comparison between privacy and free speech in an online debate: "Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say." (https://www.reddit.com/r/IAmA/comments/36ru89/just_days_left_to_kill_mass_surveillance_under/crglgh2, retrieved on July 1, 2016.)

9 Appendix: Proofs

Technical Results

Lemma 4. *Let Φ be the standard normal distribution. Then $\int_{ka-b}^{ka} d\Phi / \int_{ka}^{\infty} d\Phi$ diverges to infinity as $k \to \infty$ for $a, b > 0$.*

Proof of lemma 4: We concentrate on the right tail of the standard normal distribution. If for all $x \in [ka-b, ka]$ and some constant c we have that $\frac{\phi(x)}{\phi(x+b)} \geq c$, then is is also true that

$$\frac{\int_{ka-b}^{ka} d\Phi}{\int_{ka}^{ka+b} d\Phi} \geq c.$$

(This can be seen by noting that the first inequality holds for the range of the integrals of the second inequality.) The pdf of the standard normal distribution is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},$$

and the quotient of $\phi(x)$ and $\phi(x+b)$ is therefore $e^{-\frac{1}{2}(x^2-(x+b)^2)} = e^{xb+\frac{1}{2}b^2}$. For $x \to \infty$, this quotient diverges, and hence $\frac{\int_{ka-b}^{ka} d\Phi}{\int_{ka}^{ka+b} d\Phi}$ diverges for $k \to \infty$. Now note that $\int_{ka}^{\infty} d\Phi = \int_{ka}^{ka+b} d\Phi + \int_{ka+b}^{ka+2b} d\Phi + \ldots$ and that for large k, the quotient between any summand on the RHS and the following summand diverges. This means that the overall sum is smaller than $2\int_{ka}^{ka+b} d\Phi$ as – for k sufficiently high – $\int_{ka}^{ka+b} d\Phi + \int_{ka+b}^{ka+2b} d\Phi + \ldots \leq \int_{ka}^{ka+b} d\Phi \sum_{i=0}^{\infty}(1/2)^i = 2\int_{ka}^{ka+b} d\Phi$. Since we have established above that $\frac{\int_{ka-b}^{ka} d\Phi}{\int_{ka}^{ka+b} d\Phi}$ diverges for large k, that means that $\frac{\int_{ka-b}^{ka} d\Phi}{\int_{ka}^{\infty} d\Phi}$ diverges as well. □

Proofs

Proof of lemma 1: For $\theta_i > (\max_{\tau_i} \delta(\tau_i)) / (min_k\{q(k) - q(k-1) : \ k \in \{1, \ldots, n\}\})$, it is a dominant action to choose $p_i = 1$. Similarly, for $\theta_i < -(\max_{\tau_i} \delta(\tau_i)) / (min_k\{q(k) - q(k-1) : \ k \in \{1, \ldots, n\}\})$, it is a dominant action to choose $p_i = 0$. Write the expected utility

difference of playing $p_i = 1$ and playing $p_i = 0$ as[18]

$$-\delta(\tau_i)\Delta + \theta_i * \sum_{k=1}^{n} ((q(k) - q(k-1)) * prob(k-1)) \tag{5.5}$$

where $prob(k-1)$ is i's belief that exactly $k-1$ other citizens will choose $p_j = 1$ and $\Delta \in [-1, 1]$ is the difference between the (believed) probability that OP plays A when facing a citizen who has played $p_i = 1$ and a citizen who has played $p_i = 0$. Clearly, (5.5) is strictly increasing and continuous in θ_i. As it is optimal to play $p_i = 1$ ($p_i = 0$) if (5.5) is positive (negative), the best response to any given belief is a cutoff strategy where the cutoff is given by the θ_i for which the utility difference above is 0. (Note that the dominance regions above establish that an interior cutoff exists.) Since all best responses are cutoff strategies, all rationalizable actions are cutoff strategies.

In the privacy case, $\Delta = 0$ by definition and therefore (5.5) is zero if and only if $\theta_i = 0$ as the sum is clearly positive (recall that the cumulative distribution function q was strictly increasing by assumption). Consequently, $t^p(\tau_i) = 0$. □

Proof of lemma 2: Suppose $v_1 < v_0$ in equilibrium. In this case, (5.5) is strictly increasing in τ_i as $\Delta < 0$ and therefore $t(\tau_i)$ is strictly decreasing in τ_i.

This implies that we can partition $\Re$ in three intervals $(-\infty, t(\bar{\tau})]$, $(t(\bar{\tau}), t(\underline{\tau})]$, $(t(\underline{\tau}), \infty)$. Denoting the inverse of the equilibrium cutoff t by s, we get

$$\begin{aligned}
v_1 &= \frac{\int_{t(\bar{\tau})}^{t(\underline{\tau})} \int_{s(\theta_i)}^{\bar{\tau}} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i) + \int_{t(\underline{\tau})}^{\infty} \int_{\underline{\tau}}^{\bar{\tau}} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\bar{\tau})}^{t(\underline{\tau})} \int_{s(\theta_i)}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i) + \int_{t(\underline{\tau})}^{\infty} \int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} \\
&\geq \frac{\int_{t(\bar{\tau})}^{\infty} \int_{\underline{\tau}}^{\bar{\tau}} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\bar{\tau})}^{\infty} \int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} \\
&> \frac{\int_{-\infty}^{t(\underline{\tau})} \int_{\underline{\tau}}^{\bar{\tau}} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{-\infty}^{t(\underline{\tau})} \int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} \\
&\geq \frac{\int_{-\infty}^{t(\bar{\tau})} \int_{\underline{\tau}}^{\bar{\tau}} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i) + \int_{t(\bar{\tau})}^{t(\underline{\tau})} \int_{\underline{\tau}}^{s(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{-\infty}^{t(\bar{\tau})} \int_{\underline{\tau}}^{\bar{\tau}} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i) + \int_{t(\bar{\tau})}^{t(\underline{\tau})} \int_{\underline{\tau}}^{s(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} \\
&= v_0
\end{aligned}$$

[18] In principle Δ could depend on the number of citizens choosing $p_i = 1$ in the information aggregation stage. In this case, the expected utility difference is

$$\sum_{k=1}^{n} \{[-\delta(\tau_i)\Delta(k, k-1) + \theta_i] ((q(k) - q(k-1)) * prob(k-1))\}$$

where $\Delta(k, k-1)$ is the difference between the believed probability that OP plays A when facing a citizen who played $p_i = 1$ and k citizens chose 1 and the probability that OP plays A when facing a citizen who played $p_i = 0$ and $k-1$ citizens chose 1. The same argument as below holds: this expression is strictly increasing in θ_i. As will become apparent from (5.1)– (5.4), OP's best response strategy will not depend on the number of citizens choosing 1; see the comment in footnote 7.

where the inequalities use the assumption that $\Gamma_{\theta_i'}$ first order stochastically dominates $\Gamma_{\theta_i''}$ if $\theta_i' > \theta_i''$ and therefore θ_i and τ_i are positively correlated.[19] The result that $v_0 < v_1$ contradicts our initial supposition and therefore $v_1 \geq v_0$ in all equilibria. □

Proof of proposition 1: Consider (5.5) which has to be zero if θ_i equals the equilibrium cutoff level. By lemma 2, $\Delta \geq 0$. In an equilibrium of the privacy case $\Delta = 0$ by assumption and $t^p(\tau_i) = 0$, see lemma 1. In the case without privacy, $t^{np}(\tau_i) < 0$ is impossible as then both terms in (5.5) are negative (recall $\Delta \geq 0$) at $\theta_i = t^{np}(\tau_i)$ with the second term being strictly negative. Consequently, the two terms could not sum to zero. We can therefore conclude that $t^{np}(\tau_i) \geq 0$ which establishes $t^{np}(\tau_i) \geq t^p(\tau_i)$. Note that this inequality is strict if $\Delta^{np} > 0$ as (5.5) would not equal zero for $\theta_i = 0$ and $\Delta > 0$.

Next we have to show that $\Delta > 0$ whenever the equilbirum strategy of OP is influenced by the presence of privacy. By lemma 2, $\Delta \geq 0$. If OP behavior is influenced by the presence of privacy and $\Delta = 0$ then the probability of A has to change in both groups (citizens choosing $p_i = 0$ and citizens choosing $p_i = 1$) by the same amount compared to the privacy case. For concreteness, suppose the probability of A is increased. This implies that in the privacy case the probability of A is less than 1. Consider first the case that OP has strict preferences in the privacy equilibrium which then implies that OP played A with probability 0 in the privacy equilibrium. Consequently, $v_p = \int_{\Re} \int_{\underline{\tau}}^{\bar{\tau}} \tau \; d\Gamma_{\theta_i} \; d\Phi(\theta_i) < 0$. As $\beta_0(\tau)$ and $\beta_1(\tau)$ are obtained by means of Bayesian updating, it is impossible that both $v_0 \geq 0$ and $v_1 \geq 0$. But then it is impossible that playing A against both groups with (the same) positive probability is optimal in the equilibrium without privacy. Second, consider the case where OP is indifferent in the privacy equilibrium and plays A with some probability $\alpha < 1$. Indifference means that $v_p = 0$. If $\Delta = 0$ in the case without privacy, then it is easy to see that $v_1 > v_0$ because of the positive correlation of τ_i and θ_i. But this would imply $v_1 > 0$ and $v_0 < 0$ and therefore $\Delta = 0$ would not be a best response which contradicts that $\Delta = 0$ in equilibrium in the second case. If $\Delta > 0$, however, we already established above that $t^{np}(\tau_i) > t^p(\tau_i)$. The proof for a decrease of the probability of playing A is analogous.

Last we show that the difference $t^{np}(\tau_i) - t^p(\tau_i)$ is increasing in τ_i. As mentioned

[19] To be clear, take the first of the inequalities:

$$\mathbb{E}[\tau|\theta_i > t(\bar{\tau})] = \frac{\int_{t(\bar{\tau})}^{\infty} \mathbb{E}[\tau|\theta_i] \, d\Phi(\theta_i)}{\int_{t(\bar{\tau})}^{\infty} \int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} \leq \frac{\int_{t(\bar{\tau})}^{t(\underline{\tau})} \mathbb{E}[\tau|\theta_i] \int_{z(\theta_i)}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\theta_i + \int_{t(\underline{\tau})}^{\infty} \mathbb{E}[\tau|\theta_i] \, d\Phi(\theta_i)}{\int_{t(\bar{\tau})}^{t(\underline{\tau})} \int_{z(\theta_i)}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i) + \int_{t(\underline{\tau})}^{\infty} d\Phi(\theta_i)}$$

$$\leq \frac{\int_{t(\bar{\tau})}^{t(\underline{\tau})} \mathbb{E}[\tau|\theta_i, \tau \geq z(\theta_i)] \int_{z(\theta_i)}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\theta_i + \int_{t(\underline{\tau})}^{\infty} \mathbb{E}[\tau|\theta_i] \, d\Phi(\theta_i)}{\int_{t(\bar{\tau})}^{t(\underline{\tau})} \int_{z(\theta_i)}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i) + \int_{t(\underline{\tau})}^{\infty} d\Phi(\theta_i)} = v_1$$

where the first inequality holds as $\mathbb{E}[\tau|\theta_i]$ is strictly increasing in θ_i (by the first order stochastic dominance assumption on Γ_{θ_i}) and therefore putting less weight on lower θ_i increases the expectation. The third inequality follows a similar logic and the second one uses that $\mathbb{E}[\tau|\theta_i]$ is strictly increasing in θ_i directly.

above $t^p(\tau_i) = 0$. Using the fact that (5.5) has to be zero at the cutoff, we obtain

$$t^{np}(\tau_i) = \frac{\Delta}{\sum_{k=1}^{n}(q(k) - q(k-1)) * prob(k-1)}\delta(\tau_i). \tag{5.6}$$

Note that $prob(k-1)$ is independent of citizen i's type τ_i as these types are drawn independent from one another. Therefore, the only term in $t^{np}(\tau_i) - t^p(\tau_i)$ which depends on τ_i is $\delta(\tau_i)$ which is increasing by assumption. As the fraction on the righthand side of (5.6) is positive, it follows that $t^{np}(\tau_i) - t^p(\tau_i)$ is increasing in τ_i. □

Proof of proposition 2: We start with the case where OP finds it optimal to play A against all citzens choosing $p_i = 1$ and M against all citizens choosing $p_i = 0$ under both citizen strategies t^{np} and t^p. Recall that OP's payoff is the expected value of τ of all those citizens against which OP plays A. Hence, the payoff difference of OP's payoff between the two scenarios is the expected value of τ in the area between the horizontal axis and t^{np} in figure 5.4 below.

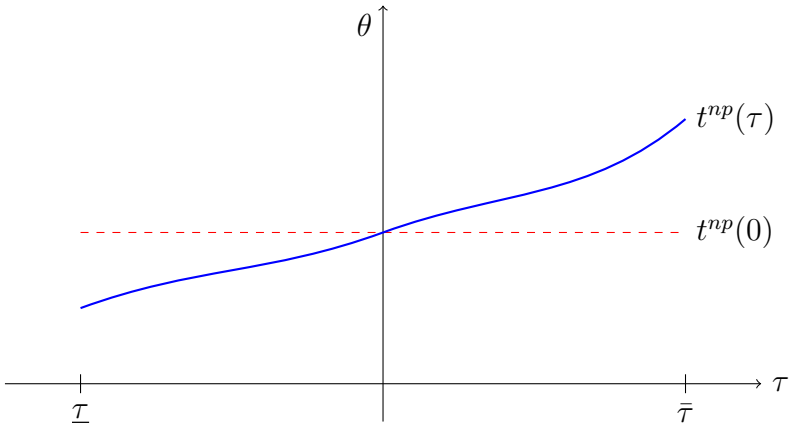


Figure 5.4: Integration range for difference in OP payoff

Denote the inverse function of $t^{np}(\tau)$ as $z(\theta)$. The difference of OP's payoffs between citizens using t^{np} and t^p is

$$\begin{aligned}
&\int_0^{t^{np}(\underline{\tau})}\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_\theta(\tau)d\Phi(\theta) + \int_{t^{np}(\underline{\tau})}^{t^{np}(\bar{\tau})}\int_{z(\theta)}^{\bar{\tau}} \tau\, d\Gamma_\theta(\tau)d\Phi(\theta) \\
&= \int_0^{t^{np}(0)}\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_\theta(\tau)d\Phi(\theta) - \int_{t^{np}(\underline{\tau})}^{t^{np}(0)}\int_{\underline{\tau}}^{z(\theta)} \tau\, d\Gamma_\theta(\tau)d\Phi(\theta) + \int_{t^{np}(0)}^{t^{np}(\bar{\tau})}\int_{z(\theta)}^{\bar{\tau}} \tau\, d\Gamma_\theta(\tau)d\Phi(\theta)
\end{aligned}$$

where the equality simply splits up the integration range which can be easily visualized in figure 5.4. The first of the three double integrals is positive by the following argument: As – by assumption – Γ_0 is symmetric around 0, $\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_0(\tau) = 0$. It follows that $\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_\theta(\tau) > 0$ for all $\theta > 0$ because Γ_θ first order stochastically dominates Γ_0 for all $\theta > 0$. This implies that the first double integral is positive as $t^{np}(0) \geq 0$ by proposition 1. The second double integral is negative as it integrates only over $\tau \leq 0$ and with

the minus sign this second term becomes positive as well. The third double integral is positive as it integrates only over positive τ. Consequently, OP would like to play A against citizens with (τ_i, θ_i) in the area between the horizontal axis and t^{np} which means that OP is better off (given the strategy of playing A if and only if $p_i = 1$) under $t^p(\tau) = 0$ than under t^{np}.

We established that playing A against citizens who play $p_i = 1$ is relatively more attractive if citizens use strategy $t^p(\tau) = 0$ than if they use strategy t^{np}. This implies that whenever OP prefers to play A against citizens who play $p_i = 1$ under t^{np} the same is true under t^p. Hence, we do not have to consider a case where OP plays M against citzens choosing $p_i = 1$ if they use t^p but A if they use t^{np}. In all other cases, OP uses the same action against citizens choosing $p_i = 0$ and against citizens choosing $p_i = 1$. Hence, $t^p = t^{np}$ and OP's payoffs are the same under both strategies (t^p and t^{np}). □

Proof of lemma 3: As the type draws are independent across citizens and as τ is not payoff relevant in the information aggregation stage, it is clear that the consumer surplus optimal cutoff will be independent of τ. Suppose cutoff $t^* \geq 0$ is consumer surplus optimal. A necessary condition for optimality is the following: Say a citizen has type $\theta_i = t^*$, then his vote must be consumer surplus neutral. That is, whether he chooses $p_i = 0$ or $p_i = 1$ must lead to the same expected consumer surplus (conditional on his own type being $\theta_i = t^*$). If this condition was not satisfied, either in- or decreasing t^* will increase expected consumer surplus thereby contradicting the optimality of t^*. We will show that the only t^* satisfying this necessary condition is $t^* = 0$.

The expected difference of consumer surplus when choosing $p_i = 1$ and $p_i = 0$ is (where we write t instead of t^* to shorten notation)

$$t + \sum_{l=0}^{n-1} \binom{n-1}{l} \Phi(t)^l (1 - \Phi(t))^{n-1-l} (q(l+1) - q(l)) \left(l\mathbb{E}[\theta|\theta < t] + (n-1-l)\mathbb{E}[\theta|\theta > t]\right) \tag{5.7}$$

where l is the number of other citizens choosing $p_i = 0$ (according to the cutoff strategy t). First, consider the term $l = (n-1)/2$ (in case n is odd). For this term $l = (n-1-l)$ and as $\mathbb{E}[\theta|\theta < t] + \mathbb{E}[\theta|\theta < t] \geq 0$ by $t \geq 0$, $p_i = 1$ will lead to a higher expected consumer surplus in this case. For $l < (n-1)/2$, we clearly have $l\mathbb{E}[\theta|\theta < t] + (n-1-l)\mathbb{E}[\theta|\theta < t] > 0$ by $t > 0$ and again $p_i = 1$ increases expected consumer surplus. However, for $l > (n-1)/2$ the opposite might be the case. Hence, we have to weigh terms with different l against each other. In particular, we will consider the terms $l > (n-1)/2$ and $n-1-l < (n-1)/2$ jointly. By the assumption that q is point symmetric around $1/2$, $q(l+1) - q(l) = q(n-1-l+1) - q(n-1-l)$. Furthermore, the binomial coefficient is symmetric around the mean which means that also $\binom{n-1}{l} = \binom{n-1}{n-1-l}$. Consequently, we can write the sum of the two terms corresponding to l and $n-1-l$ as follows (using

$z = 2l - n + 1$ and dropping the argument of $\Phi(t)$ to save space)

$$\binom{n-1}{l}\Phi^{n-1-l}(1-\Phi)^{n-1-l}(q(l+1)-q(l))$$
$$\{\mathbb{E}[\theta|\theta<t]\left(l\Phi^z+(n-1-l)(1-\Phi)^z\right)+\mathbb{E}[\theta|\theta>t]\left((n-1-l)\Phi^z+l(1-\Phi)^z\right)\}.$$

We will now argue that the expression in curly brackets (and therefore the whole expression) is positive (for any $l > (n-1)/2$). Note that the only negative term in the curly brackets is $\mathbb{E}[\theta|\theta < t]$. Also recall that $t \geq 0$ and therefore $\Phi \geq 1 - \Phi$. This implies that the term in curly brackets is (weakly) greater than

$$\mathbb{E}[\theta|\theta<t]\left((n-1)\Phi^z\right)+\mathbb{E}[\theta|\theta>t]\left((n-1)(1-\Phi)^z\right)$$

where we increased the weight on the negative term as much as possible (recall that $n-1 \geq l > (n-1)/2$). This means that the term in curly brackets is definitely positive if $\mathbb{E}[\theta|\theta < t]\Phi^z + \mathbb{E}[\theta|\theta > t](1-\Phi)^z \geq 0$. Given that z is an integer between 1 and $n-1$, this inequality is hardest to satisfy for $z = 1$ but there it holds as $\mathbb{E}[\theta|\theta < t]\Phi(t)+\mathbb{E}[\theta|\theta > t](1-\Phi(t)) = 0$ by the definition of conditional expectations.

Now that we know that the expression in curly brackets is positive for all $l > (n-1)/2$ we can conclude that $p_i = 1$ leads to a higher expected consumer surplus than choosing $p_i = 0$ and this is true in a strict sense if $t > 0$: The analysis of the case $l = (n-1)/2$ and the consideration of twin pairs $l > (n-1)/2$ and $n-1-l$ have shown that the sum in (5.7) is positive and adding $t \geq 0$ keeps (5.7) positive (strictly if $t > 0$). It is easy to verify that the sum in (5.7) equals zero if $t = 0$ (in this case $\Phi = 1 - \Phi = 1/2$). This implies that $t = 0$ satisfies the necessary condition for optimality and all other $t > 0$ do not.[20]

Hence, $t = 0$ is optimal if we can verify that an optimal t exists. Note that expected consumer surplus is continuous in t. As it is straightforward that $t \to \infty$ is not optimal, the problem of finding the optimal t can be reduced to maximizing a continuous function over a compact set and a solution exists by the Weierstrass theorem. □

Proof of proposition 3: Let M be optimal for OP in the privacy equilibrium.

1.) Suppose there is a mixed strategy equilibrium in the case without privacy. Then, OP has to play M against both groups with positive probability. If he played A against those who chose $p_i = 1$ for sure and mixed for those who chose $p_i = 0$, then M could not be optimal in the privacy case. Hence, OP can in the case without privacy achieve a payoff equal to his equilibrium payoff by playing M against both groups. Consequently, OP's payoff with and without privacy is the same. Citizens are strictly better off with privacy as (a) there is no chilling effect which means by lemma 3 that expected welfare in

[20] As the setup is symmetric, a similar argument could be made to rule out the optimality of any $t < 0$.

the information aggregation stage is maximized and (b) M will be played with probability 1 against them in the interaction stage.

2.) Now assume that $\delta'(\tau) > 0$. We will show that for n sufficiently high the privacy equilibrium welfare dominates the equilibrium in the case without privacy (or the two are identical).

We are going to make use of the fact that for any $\mu > 0$, we can find an $\hat{n}$ so that for all $n > \hat{n}$, $q((m+1)/n) - q(m/n) < \mu$. This follows from the assumptions that q is strictly increasing and continuously differentiable. Intuitively, q would have to have an infinite slope somewhere for this not to be true.

Now consider (5.5) and suppose $\Delta = 1$. Note that $\sum_{j=1}^{n}((q(j/n) - q((j-1)/n)) * prob(j-1)) < \sum_{j=1}^{n} \mu * prob(j-1)) \leq \mu$, which gets arbitrarily small as n gets large. Consequently, the threshold values become arbitrarily large as n gets large. Note also that using (5.5) we can then write $t(\tau_i) = \delta(\tau_i)/(\sum_{j=1}^{n}((q(j/n)-q((j-1)/n))*prob(j-1))) \geq \delta(\tau_i)/\mu$. Similarly, $t'(\tau_i) = \delta'(\tau_i)/(\sum_{j=1}^{n}((q(j/n)-q((j-1)/n))*prob(j-1))) \geq \delta'(\tau_i)/\mu$. Hence, t is increasing in τ and the slope also becomes arbitrarily large as n increases (as μ can be chosen arbitrarily small for n sufficiently large).

Denoting the inverse of the threshold t by z, we can write

$$v_1 = \frac{\int_{t(\underline{\tau})}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{1 - \Phi(t(\bar{\tau})) + \int_{t(\underline{\tau})}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} + \frac{\mathbb{E}[\tau | \theta_i > t(\bar{\tau})]}{1 + \frac{\int_{t(\underline{\tau})}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{1 - \Phi(t(\bar{\tau}))}}.$$

As z becomes arbitrarily flat for n sufficiently high, we can choose – for n high enough – an $\varepsilon > 0$ such that $\int_{t(\bar{\tau})-\varepsilon}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)/(1-\Phi(t(\bar{\tau}))) > 0.5 \int_{t(\bar{\tau})-\varepsilon}^{t(\bar{\tau})} \int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)/(1-\Phi(t(\bar{\tau})))$. It follows that the second term in v_1 goes to zero as $n \to \infty$ because $\int_{t(\bar{\tau})-\varepsilon}^{t(\bar{\tau})} \int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)/(1 - \Phi(t(\bar{\tau})))$ and therefore its denominator diverges to infinity by lemma 4.

The first term in v_1 converges to something below the unconditional mean of τ which we denote by $\tau^E = \mathbb{E}[\tau]$: For n large, the previous step implies that,

$$\begin{aligned} v_1 &\approx \frac{\frac{\int_{t(\underline{\tau})}^{t(\tau^E)} \int_{\underline{\tau}}^{z(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\underline{\tau})}^{t(\tau^E)} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} + \frac{\int_{t(\tau^E)}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\underline{\tau})}^{t(\tau^E)} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}}{1 + \frac{\int_{t(\tau^E)}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i) + 1 - \Phi(t(\bar{\tau}))}{\int_{t(\underline{\tau})}^{t(\tau^E)} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}} \\ &\leq \frac{\int_{t(\underline{\tau})}^{t(\tau^E)} \int_{\underline{\tau}}^{z(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\underline{\tau})}^{t(\tau^E)} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} + \frac{\int_{t(\tau^E)}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\underline{\tau})}^{t(\tau^E)} \int_{\underline{\tau}}^{z(\theta_i)} d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)} \end{aligned}$$

Note that the first term equals $\mathbb{E}[\tau_i | t(\underline{\tau}) \leq \theta_i \leq t(\tau^E) \ \wedge \ \tau_i \leq z(\theta_i)]$. Clearly, this is below the unconditional mean τ^E. It follows that for a sufficiently small $\varepsilon' > 0$ (and

large n).

$$v_1 \leq \tau^E + \frac{\int_{t(\tau^E)}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\tau^E)-\varepsilon'}^{t(\tau^E)} \int_{\underline{\tau}}^{\bar{\tau}} \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}$$

Note that the same ε' appropriately chosen for some n will also work for higher n (as the density of ϕ thins out for higher θ_i and $t(\tau^E) - t(\underline{\tau})$ is increasing in n). This implies that we can conclude for the limit $n \to \infty$ that

$$v_1 \leq \tau^E + \frac{\int_{t(\tau^E)}^{t(\bar{\tau})} \int_{\underline{\tau}}^{z(\theta_i)} \tau \, d\Gamma_{\theta_i}(\tau) \, d\Phi(\theta_i)}{\int_{t(\tau^E)-\varepsilon'}^{t(\tau^E)} \int_{\underline{\tau}}^{\bar{\tau}} \, d\Gamma_{\infty}(\tau) \, d\Phi(\theta_i)} \leq \tau^E + \frac{\bar{\tau} \int_{t(\tau^E)}^{t(\bar{\tau})} \, d\Phi(\theta_i)}{\int_{t(\tau^E)-\varepsilon'}^{t(\tau^E)} \, d\Phi(\theta_i)} \xrightarrow{n\to\infty} \tau^E$$

where the limit follows from lemma 4 and the above established fact that t goes to infinity as $n \to \infty$. By assumption, OP's best response when facing the unconditional mean τ^E (or a lower τ_i) is M which contradicts the supposition $\Delta = 1$. Hence, $\Delta < 1$ which implies that OP uses a mixed strategy. By the first part of the proposition, privacy then welfare dominates no privacy.

3.) We will show that OP either plays M (independent of p_i) or uses a mixed strategy in the no privacy equilibrium if r is sufficiently high. (1) will then imply (3).

Suppose OP plays a pure strategy in equilibrium. If OP plays M against $p_i = 1$, then – by the assumption that OP plays M in the privacy case – privacy and no privacy case lead to the same equilibrium and the result holds trivially. OP cannot play A against $p_i = 0$: By lemma 2, OP would then also play A against $p_i = 1$. But this is incompatible with Bayesian updating and the assumption that OP plays M in the privacy case. Hence, we only need to consider the case where OP plays M against $p_i = 0$ and A against $p_i = 1$. Consider (5.5) which can be rearranged to get (under the assumption that $\Delta = 1$, i.e. OP plays A against $p_i = 1$ and M against $p_i = 0$)

$$t^{np}(\tau) = \frac{r\delta(\tau)}{\sum_{k=1}^{n}(q(k/n) - q((k-1)/n)) \, prob(k-1)} \geq r\delta(\tau).$$

Hence, t^{np} diverges to ∞ as $r \to \infty$. Furthermore, the slope of t^{np} is linearly growing in r. Hence, the derivative of $t^{np}(\tau)$ also diverges to ∞ as r grows. But then the same steps as in the proof of result (2) above imply that $v_1 \leq \tau^E$, i.e. playing A against $p_i = 1$ is not a best response which contradicts that OP uses the pure strategy corresponding to $\Delta = 1$ in the equilibrium without privacy for r sufficiently large. As – for r sufficiently large – OP uses either mixed strategy in the no privacy equilibrium or plays M regardless of p_i, (1) implies that privacy dominates no privacy. □

Proof of proposition 4: The welfare-difference between no privacy and privacy is given by

$$\int_{\delta}^{\infty} \int_{\underline{\tau}}^{\overline{\tau}} \tau_i d\Gamma_{\theta_i} d\Phi(\theta_i) - \int_{\delta}^{\infty} \delta d\Phi(\theta_i) - \int_{0}^{\delta} \theta_i \Phi(\theta_i). \tag{5.8}$$

If we increase δ, it is clear that the first term weakly decreases, as the area of the integral gets smaller. To see what happens to the second and third term, we can disaggregate them further, assuming that we increase δ by ϵ. Then we get a net effect of

$$-\int_{\delta+\epsilon}^{\infty}\delta d\Phi-\int_{\delta+\epsilon}^{\infty}\epsilon d\Phi-\int_{0}^{\delta+\epsilon}\theta_i\Phi+\int_{\delta}^{\infty}\delta d\Phi+\int_{0}^{\delta}\theta_i\Phi$$
$$=-\int_{\delta+\epsilon}^{\infty}\epsilon d\Phi-\int_{\delta}^{\delta+\epsilon}(\theta_i-\delta)d\Phi,$$

which is negative. Therefore, the overall welfare decreases in δ.

Now consider what happens if we increase Γ. This only has influence on the payoff of the OP; the second and third term in (5.8) above remain unchanged. It follows from our definition that if we replace Γ' with Γ'' and $\Gamma' < \Gamma''$, then $\int_{\underline{\tau}}^{\overline{\tau}}\tau_i d\Gamma''_{\theta_i} > \int_{\underline{\tau}}^{\overline{\tau}}\tau_i d\Gamma'_{\theta_i}$. This means that if we increase Γ, the inner integral in the first term in (5.8) increases, and hence the whole term increases. □

Proof of proposition 5: First consider $\lambda = 0$. Note that the distribution of τ_i under $\bar{\tau}$ is the same as the distribution of τ_i that the OP faces in the privacy case of the original model (with distribution Γ_{θ_i}). As we assumed that OP plays M in the privacy equilibrium, it is clear that the privacy equilibrium is also an equilibrium for $\lambda = 0$. In fact, it is the unique equilibrium: Since M is the best response against the distribution $\bar{\Gamma}$ by assumption, OP has to play M for sure against at least one group of citizens (either those choosing $p_i = 0$ or those choosing $p_i = 1$) by Bayesian updating. Suppose OP played A with positive probability against those who chose $p_i = 1$. Then some citizens with low θ_i would be chilled and play $p_i = 0$. As δ is increasing in τ_i, the best response cutoff would be increasing in τ_i, see (5.6). But then the average τ_i among those choosing $p_i = 1$ is lower than the average τ_i under $\bar{\Gamma}$. Consequently, M is a strict best response by OP because M is a best response against $\bar{\Gamma}$. This contradicts that OP plays A with positive probability.

Note that $\mathbb{E}[\tau_i|\theta_i \geq 0]$ is continuous in λ. Since M is a best response against $\bar{\Gamma}$, that is $\mathbb{E}[\tau_i|\theta_i \geq 0] < 0$ for $\lambda = 0$, the same is true for sufficiently small $\lambda > 0$. Hence, a $\underline{\lambda} > 0$ exists such that for all $\lambda \leq \underline{\lambda}$ the unique equilibrium without privacy is that OP plays M and all citizens use a cutoff of zero. This is equivalent to the privacy equilibrium and therefore privacy and no privacy are welfare equivalent for all $\lambda \leq \underline{\lambda}$. For the result in the proposition, let $\underline{\lambda}$ be the highest λ such that the equilibrium in the no privacy is that OP plays M against citizens choosing $p_i = 1$. Note that $\underline{\lambda} < 1$ as by assumption OP plays A gainst citizens choosing $p_i = 1$ for $\lambda = 1$.

For $\lambda = 1$, the equilibrium of the no privacy case was assumed to be that OP plays A (M) against $p_i = 0$ ($p_i = 1$) in the no privacy case. Denote by λ^* the infimum of all λ for which such an equilibrium exists. Clearly, $\lambda^* \in (\underline{\lambda}, 1)$. Since such an equilibrium

no longer exists for $\lambda < \lambda^*$, it has to hold true that at $\lambda = \lambda^*$ OP is indifferent between playing A and playing M against those playing $p_i = 1$ (for lower λ OP will then prefer to play M as the correlation is too weak and that is why the equilibrium breaks down). Note that the best response cutoffs of the citizens do not depend on λ but only on the OP's strategy. It follows that $\mathbb{E}[\tau_i|\theta_i \geq t^{np}(\tau_i)]$ is continuous in λ for $\lambda \geq \lambda^*$. As the OP is indifferent at λ^*, we have $\mathbb{E}[\tau_i|\theta_i \geq t^{np}(\tau_i)] = 0$ at λ^*. Continuity, implies that $\mathbb{E}[\tau_i|\theta_i \geq t^{np}(\tau_i)]$ is arbitrarily small for λ close but strictly above λ^*. That is, for any $\varepsilon > 0$ there is a $\varepsilon' > 0$ such that imposing privacy leads only to less than ε losses for the OP if $\lambda < \lambda^* + \varepsilon'$. Imposing privacy leads (for $\lambda \in [\lambda^*, \lambda^* + \varepsilon']$) to a discrete increase in citizen welfare for several reasons: First, those choosing $p_i = 1$ no longer face the aggressive response which increases their payoff by $\delta(\tau_i)$. Second, in the privacy case citizens use the cutoff zero instead of $t^{np} > 0$ which leads to a higher surplus in the information aggregation stage. This implies that for $\varepsilon' > 0$ small enough, privacy welfare dominates no privacy for $\lambda \in (\lambda^*, \lambda^* + \varepsilon']$. Let $\bar{\lambda} = \lambda^* + \varepsilon'$. Note that for $\lambda \in (\underline{\lambda}, \bar{\lambda})$ the equilibrium in the no privacy case is necessarily mixed which means implies that privacy is Pareto dominant for these λ, see proposition 3. This establishes the claim. □

9 Appendix: State matching

In this section, we consider a model where the private information of citizens in the information aggregation stage is not directly their personal payoff of policy $p = 1$. Instead citizens have all the same payoff of policy $p = 1$ but each citizen only receives a noisy signal of this payoff. This has a striking implication: Chilling makes every citizen worse off. The reason is that chilling inhibits information aggregation. In the main paper citizens have private preferences over outcomes and therefore some citizens (those with negative θ_i) gain from chilling. Since all citizens have the same interest – implementing the policy if and only if the common payoff consquence is positive, – everyone looses in this setup from chilling.

More precisely, the setting is as follows: The state of the world θ is distributed standard normally and this θ is the payoff consequence of policy $p = 1$ for each citizen. However, the realization of θ is unknown. Each citizen obtains a private signal θ_i which is normally distributed around the true state θ, i.e. $\theta_i \sim N(\theta, \sigma^2)$ where we denote the cdf by $\tilde{\Phi}_\theta$ and the pdf by $\tilde{\phi}_\theta$. All θ_i are assumed to be independent draws from this distribution. The interaction type of citizen i, τ_i, is drawn from Γ_{θ_i} where again $\Gamma_{\theta_i'}$ is assumed to first order stochastically dominate $\Gamma_{\theta_i''}$ if and only if $\theta_i' > \theta_i''$. This creates a positive correlation between θ_i and τ_i. The interaction stage is exactly the same as in the model of the main paper. That is, without privacy a strategy for OP states which of the two actions A and M OP plays against a citizen who chose $p_i = 0$ and which against a citizen who chose $p_i = 1$. With privacy, OP only decides which of the two actions he chooses against all citizens. This means that – to keep the setting comparable to the main paper – we do not consider strategies (or beliefs) that are contingent upon the number of citizens choosing $p_i = 1$. This is a simplification. However, one can easily imagine settings where OP has to commit to a strategy before he gets to know the citizens' p_is. This is, for example, the case if the interaction is between i and an agent representing OP and p_i is only learned in the interaction. OP then has to instruct the agent in advance how to act.

The main change is, therefore, that citizen i's payoff is $\theta p - \mathbb{1}_{s(p_i)=A}\delta(\tau_i)$; that is, θ instead of θ_i enters the utility function. To keep the model tractable, we will assume that $q(m/n) = m/n$.

We first replicate some intermediary results from the main text in this modified setting.

Lemma 5. *For citizens, only cutoff strategies $t(\tau_i)$ are rationalizable. In the privacy case, the optimal cutoff is $t^p(\tau_i) = 0$ for all τ_i.*

Proof. If citizen i receives signal θ_i, he updates his belief α about θ according to

Bayes' rule yielding

$$\alpha(\theta'|\theta_i) = prob(\theta \leq \theta'|\theta_i) = \frac{\int_{-\infty}^{\theta'} \tilde{\phi}_\theta(\theta_i)\, d\Phi(\theta)}{\int_{\Re} \tilde{\phi}_\theta(\theta_i)\, d\Phi(\theta)}.$$

From the normality assumptions, it follows that the pdf of the belief is single peaked with its peak between 0 (the mean of the prior) and θ_i. Furthermore, $\mathbb{E}[\theta|\theta_i] = \int_{\Re} \theta\, d\alpha(\theta|\theta_i)$ is strictly increasing in θ_i with limits $\lim_{\theta_i \to \infty} = \infty$ and $\lim_{\theta_i \to -\infty} = -\infty$. To see this, note that

$$\begin{aligned}
\mathbb{E}[\theta|\theta_i] &= \int_{\Re} \theta \frac{\tilde{\phi}_\theta(\theta_i)\phi(\theta)}{\int_{\Re} \tilde{\phi}_{\hat{\theta}}(\theta_i)\, d\Phi(\hat{\theta})}\, d\theta \\
&= \frac{\int_{\Re} \theta e^{-(\theta_i-\theta)^2/(2\sigma^2)} e^{-\theta^2/2} d\theta}{\int_{\Re} e^{-(\theta_i-\theta)^2/(2\sigma^2)} e^{-\theta^2/2} d\theta} \\
&= \frac{\theta e^{-(-2\theta_i\theta+\theta^2(1+\sigma^2))/(2\sigma^2)} d\theta}{\int_{\Re} e^{-(-2\theta_i\theta+\theta^2(1+\sigma^2))/(2\sigma^2)} d\theta} \\
&= \frac{\frac{1}{\sqrt{2\pi}\sigma/(\sqrt{1+\sigma^2})} \int_{\Re} \theta e^{-\frac{\theta_i^2/(1+\sigma^2)^2 - 2\theta_i\theta/(1+\sigma^2)+\theta^2}{2\sigma^2/(1+\sigma^2)}}\, d\theta}{\frac{1}{\sqrt{2\pi}\sigma/(\sqrt{1+\sigma^2})} \int_{\Re} e^{-\frac{\theta_i^2/(1+\sigma^2)^2 - 2\theta_i\theta/(1+\sigma^2)+\theta^2}{2\sigma^2/(1+\sigma^2)}}\, d\theta} \\
&= \frac{\theta_i}{1+\sigma^2}
\end{aligned}$$

where the last equality holds as the numerator of the second but last line is the expected value of a randome variable distributed $N(\theta_i/(1+\sigma^2), \sigma^2/(1+\sigma^2)^2)$ and the denominator of the second but last line is simply 1 (as it integrates over the density of this random variable).

Citizen i's expected payoff difference between choosing $p_i = 1$ and $p_i = 0$ is[21]

$$-\delta(\tau_i)\Delta + \mathbb{E}[\theta|\theta_i]/n = -\delta(\tau_i)\Delta + \frac{\theta_i}{(1+\sigma^2)n} \tag{5.9}$$

where Δ is again the difference between the probabilities that OP plays A against citizens with $p_i = 1$ and citizens with $p_i = 0$. Clearly, it is optimal to play $p_i = 0$ ($p_i = 1$) for sufficiently low (high) θ_i. (Note that $max_{\tau_i \in [\underline{\tau}, \bar{\tau}]} \delta(\tau_i)$ is bounded.) Furthermore, $\mathbb{E}[\theta|\theta_i]$ is strictly increasing in θ_i which implies that i's best response is a cutoff strategy. Consequently, only cutoff strategies are best responses. The optimal cutoff is given by $t(\tau_i) = (1+\sigma^2)n\delta(\tau_i)\Delta$.

In the privacy case, $\Delta = 0$ and therefore the optimal cutoff is $t^p(\tau_i) = 0$. □

[21] Recall that $q(m/n) = m/n$ which means that i's "influence" on the policy decision is $1/n$.

OP's belief over τ_i given p_i is given by

$$\begin{aligned}\beta_0(\tau') = prob(\tau \leq \tau' | p_i = 0) &= \frac{\int_{\Re}\int_{\Re}\int_{\underline{\tau}}^{\tau'} \mathbb{1}_{t(\tau_i)\geq\theta_i}\, d\Gamma_{\theta_i}(\tau_i)\, d\tilde{\Phi}_\theta(\theta_i)\, d\Phi(\theta)}{\int_{\Re}\int_{\Re}\int_{\underline{\tau}}^{\bar{\tau}} \mathbb{1}_{t(\tau_i)\geq\theta_i}\, d\Gamma_{\theta_i}(\tau_i)\, d\tilde{\Phi}_\theta(\theta_i)\, d\Phi(\theta)} \\ \beta_1(\tau') = prob(\tau \leq \tau' | p_i = 1) &= \frac{\int_{\Re}\int_{\Re}\int_{\underline{\tau}}^{\tau'} \mathbb{1}_{t(\tau_i)\leq\theta_i}\, d\Gamma_{\theta_i}(\tau_i)\, d\tilde{\Phi}_\theta(\theta_i)\, d\Phi(\theta)}{\int_{\Re}\int_{\Re}\int_{\underline{\tau}}^{\bar{\tau}} \mathbb{1}_{t(\tau_i)\leq\theta_i}\, d\Gamma_{\theta_i}(\tau_i)\, d\tilde{\Phi}_\theta(\theta_i)\, d\Phi(\theta)}.\end{aligned}$$

OP's expected utility of playing A against a citizen choosing policy $p_i = 0$ or $p_i = 1$ are then

$$\begin{aligned}v_0 &= \int_{\Re} \tau\, d\beta_0(\tau) \\ v_1 &= \int_{\Re} \tau\, d\beta_1(\tau).\end{aligned}$$

Lemma 6. *In every perfect Bayesian equilibrium (without privacy), $v_1 \geq v_0$.*

Proof. Suppose otherwise. Then $\Delta < 0$ which implies that $t(\tau_i)$ is decreasing. Denote the inverse of t by z. From OP's point of view θ_i is distributed according to the cdf

$$\hat{\Phi}(\theta_i) = \int_{\Re} \tilde{\Phi}_\theta(\theta_i)\, d\Phi(\theta).$$

Using this distribution $\hat{\Phi}$ we can replicate the proof from the main paper one-to-one:

$$\begin{aligned}v_1 &= \frac{\int_{t(\bar{\tau})}^{t(\underline{\tau})}\int_{z(\theta_i)}^{\bar{\tau}} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i) + \int_{t(\underline{\tau})}^{\infty}\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)}{\int_{t(\bar{\tau})}^{t(\underline{\tau})}\int_{z(\theta_i)}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i) + \int_{t(\underline{\tau})}^{\infty}\int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)} \\ &\geq \frac{\int_{t(\bar{\tau})}^{\infty}\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)}{\int_{t(\bar{\tau})}^{\infty}\int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)} \\ &> \frac{\int_{-\infty}^{t(\underline{\tau})}\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)}{\int_{-\infty}^{t(\underline{\tau})}\int_{\underline{\tau}}^{\bar{\tau}} d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)} \\ &\geq \frac{\int_{-\infty}^{t(\bar{\tau})}\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i) + \int_{t(\bar{\tau})}^{t(\underline{\tau})}\int_{\underline{\tau}}^{z(\theta_i)} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)}{\int_{-\infty}^{t(\bar{\tau})}\int_{\underline{\tau}}^{\bar{\tau}} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i) + \int_{t(\bar{\tau})}^{t(\underline{\tau})}\int_{\underline{\tau}}^{z(\theta_i)} \tau\, d\Gamma_{\theta_i}(\tau)\, d\hat{\Phi}(\theta_i)} \\ &= v_0\end{aligned}$$

which contradicts our starting point $v_1 < v_0$. □

The previous result implies that $\Delta \geq 0$ and therefore $t^{np}(\tau_i) = (1+\sigma^2)n\delta(\tau_i)\Delta \geq 0 = t^p(\tau_i)$. We therefore get chilling.

Proposition 6. *The equilibrium cutoff of a type τ_i is higher without privacy than with privacy. If the absence of privacy affects OP's behavior, this relation is strict. The difference of equilibrium cutoffs with and without privacy is increasing in τ_i.*

To establish that this chilling indeed hurts every citizen – as we claimed above – we have to show that the privacy cutoff zero leads to a higher expected consumer surplus than $t^{np}(\tau) > 0$.

Lemma 7. *The cutoff strategy $t^p(\tau) = 0$, i.e. the equilibrium strategy of the privacy case, gives a higher expected consumer surplus in the information aggregation stage than any other $t^{np}(\tau) > 0$.*

Proof. Let $t(\tau)$ be the strategy maximizing expected consumer welfare. Consider citizen i with type $(\theta_i, \tau_i) = (t(\tau'), \tau')$ for some $\tau' \in [\underline{\tau}, \bar{\tau}]$.

Optimality of t requires that expected welfare conditional on i being of type $(t(\tau'), \tau')$ is the same no matter whether i chooses $p_i = 0$ or $p_i = 1$: If this was not the case, say for concreteness $p_i = 1$ lead to a higher expected consumer welfare, then t could not be optimal: As the setup is continuous, it would then also be better for expected consumer surplus if i chose $p_i = 1$ if he was any type in an $\varepsilon > 0$ neighborhood of $(t(\tau'), \tau')$. But as expected welfare is just the expectation of expected welfare conditional on i's type over i's type we get that an alternative strategy t' which is slightly lower than t around τ' leads to higher expected consumer welfare than t. This contradicts the definition of t. Consequently, expected welfare conditional on i being of type $(t(\tau'), \tau')$ has to be the same no matter whether i chooses $p_i = 0$ or $p_i = 1$.

We are now going to show that the just stated (necessary) optimality condition cannot be satisfied for any $t > 0$. However, it is trivially satisfied for t^p by the symmetry of the setup. We focus on citizen i with type $\theta_i = t(\tau_i) > 0$. If citizen i chooses $p_i = 1$ instead of $p_i = 0$, he will increase the probability that $p = 1$ by $1/n$. This can be interpreted as follows: choosing $p_i = 1$ instead of $p_i = 0$ leads with probability $1/n$ to a payoff of θ instead of a payoff of zero (for each citizen). Hence, choosing $p_i = 1$ is best for expected consumer welfare (conditional on i's type) if $\mathbb{E}[\theta|\theta_i] > 0$.[22] As we showed above, $\mathbb{E}[\theta|\theta_i] = \theta_i/(1+\sigma^2)$ which is strictly positive for all $\theta_i > 0$. It follows that $p_i = 1$ leads to strictly higher expected consumer welfare than $p_i = 0$ as $\theta_i > 0$. This contradicts that $t > 0$ maximizes expected consumer surplus. □

The previous results can now be used to obtain a stronger version of our welfare result in the paper. While the paper argued that expected aggregated consumer surplus is higher under privacy if n is large (while OP's payoff is the same with and without privacy), we can now say that the expected utility of each citizen – regardless of his type (θ_i, τ_i) – is higher under privacy for n large. That is, privacy is an interim Pareto improvement here while it was only an ex ante Pareto improvement in the model of the paper.

Proposition 7. *Assume OP plays M in the privacy equilibrium.*
1.) If OP uses a mixed – that is not pure – strategy in the equilibrium without privacy,

[22] Note that conditioning on τ_i is immaterial as τ_i is – given θ_i – not correlated with θ.

then changing to the privacy case increases expected welfare at the interim stage.
2.) Assume that (i) δ is differentiable and strictly increasing in τ, i.e. $\delta'(\tau) > 0$ for all $\tau \in [\underline{\tau}, \bar{\tau}]$ and (ii) $\Gamma_\infty = \lim_{\theta_i \to \infty} \Gamma_{\theta_i}$ is a non-degenerate distribution in the sense that $\Gamma_\infty(\tau_i) > 0$ for all $\tau_i > \underline{\tau}$. Then, privacy welfare dominates no privacy for large n in the following sense: Compared to the no privacy case, privacy leads to a higher expected consumer surplus for each consumer of every type and the same expected payoff for OP if n is sufficiently large.

In order to prove the proposition, we have to first restate the technical result on the limit of tails of Φ that we show in the appendix for $\hat{\Phi}(\theta_i)$.

Lemma 8. *Let $\hat{\Phi}(\theta_i) = \int_\Re \tilde{\Phi}_\theta(\theta_i)\, d\Phi(\theta)$ be the distribution of θ_i from OP's perspective. Then, $\int_{ka-b}^{ka} d\hat{\Phi} / \int_{ka}^{\infty} d\hat{\Phi}$ diverges to infinity as $k \to \infty$ for $a, b > 0$.*

Proof. If we can show that $\hat{\phi}(x)/\hat{\phi}(x+b)$ diverges to infinity as $x \to \infty$ (where $\hat{\phi}$ is the density of $\hat{\Phi}$), then the same proof as in the paper applies. Note that

$$\hat{\phi}(x) = \int_\Re \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}} d\Phi(\theta) = \int_\Re \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}} e^{-\theta^2/2}\, d\theta$$

$$\begin{aligned}
\frac{\hat{\phi}(x)}{\hat{\phi}(x+b)} &= \frac{\int_\Re e^{-(x-\theta)^2/(2\sigma^2)}\, d\Phi(\theta)}{\int_\Re e^{-(x+b-\theta)^2/(2\sigma^2)}\, d\Phi(\theta)} \\
&= \frac{\int_\Re e^{-(x-\theta)^2/(2\sigma^2) - \theta^2/2}\, d\theta}{\int_\Re e^{-(x+b-\theta)^2/(2\sigma^2) - \theta^2/2}\, d\theta} \\
&= \frac{\int_\Re e^{[-(1+\sigma^2)\theta^2 + 2x\theta]/(2\sigma^2)}\, d\theta}{\int_\Re e^{[-2b(x-\theta) - b^2 - (1+\sigma^2)\theta^2 + 2x\theta]/(2\sigma^2)}\, d\theta} \\
&= \frac{\int_\Re e^{-\frac{(\theta - x/(1+\sigma^2))^2}{2\sigma^2/(1+\sigma^2)}}\, d\theta}{\int_\Re e^{(-2b(x-\theta) - b^2)/(2\sigma^2)} e^{-\frac{(\theta - x/(1+\sigma^2))^2}{2\sigma^2/(1+\sigma^2)}}\, d\theta} \\
&= \frac{\int_\Re d\bar{\Phi}(\theta)}{\int_\Re e^{(-2b(x-\theta) - b^2)/(2\sigma^2)}\, d\bar{\Phi}(\theta)}
\end{aligned}$$

where $\bar{\Phi}$ is the cdf of a normal distribution with mean $x/(1+\sigma^2)$ and variance $\sigma^2/(1+\sigma^2)$. As the numerator is 1, the previous expression can be written as

$$\frac{\hat{\phi}(x)}{\hat{\phi}(x+b)} = \frac{1}{\int_\Re e^{-\frac{2b(x-\theta) + b^2}{2\sigma^2}}\, d\bar{\Phi}(\theta)}$$

which diverges to infinity as $x \to \infty$ (because the denominator converges to zero). Given this, the rest of the proof from the main paper goes through one-to-one which implies the lemma. □

Proof of proposition 7: Let M be optimal for OP in the privacy equilibrium.

1.) Suppose there is a mixed strategy equilibrium in the case without privacy. Then, OP has to play M against both groups with positive probability. If he played A against those who chose $p_i = 1$ for sure and mixed for those who chose $p_i = 0$, then M could not be optimal in the privacy case. Hence, OP can in the case without privacy achieve a payoff equal to his equilibrium payoff by playing M against both groups. Consequently, OP's payoff with and without privacy is the same. Citizens are strictly better off with privacy as (a) there is no chilling effect which means by lemma 7 that expected welfare of every consumer (no matter which type) in the information aggregation stage is maximized and (b) M will be played with probability 1 against them in the interaction stage.

2.) Now assume that $\delta'(\tau) > 0$. We will show that for n sufficiently high the privacy equilibrium welfare dominates the equilibrium in the case without privacy (or the two are identical).

We are going to make use of the fact that for any $\mu > 0$, we can find an $\hat{n}$ so that for all $n > \hat{n}$, $q((m+1)/n) - q(m/n) = 1/n < \mu$. Now recall that $t^{np}(\tau_i) = (1+\sigma^2)n\delta(\tau_i)\Delta$. Consequently, the threshold values become arbitrarily large as n gets large (assuming $\Delta = 1$). Note also that t is increasing in τ and the slope also becomes arbitrarily large as n increases. From here, the proof of the main paper applies with $\hat{\Phi}$ in place of Φ. $\square$

Acknowledgements

> Some of you may have had occasion to run into economists and to wonder therefore how they got that way.
>
> *(Tom Lehrer, paraphrased)*

How did I become an economist? Although the question may sound like it is being asked by someone waking up in an unfamiliar place with a terrible hangover, I would encourage you to rather think of a wanderer who looks out on an amazing vista and finds it hard to conceive how he got up here, by foot, from that tiny little dot in the valley where he had breakfast just this morning. Let me retrace the steps.

Throughout my PhD, I could rely on my supervisor, Peter Norman Sørensen. The central experience of being Peter's student was the enormous intellectual freedom that he granted me, based on the idea that everybody has to find their own way of doing economics – but also on an implicit trust that I would be able to do so. This is not to say that Peter was not there when it counted, to provide advice and (adequately dosed) encouragement. But being allowed to find and develop my own research questions (including wandering off on a few tangents) has been an invaluable experience. It means that after four years, I not only have a thesis, but also an idea of the kind of problems that I want to work on in the future, and the ways in which I want to tackle them.

This thesis would look very different had I not had the chance to collaborate with and learn from Christoph Schottmüller, with whom I wrote three papers which are all part of this thesis. I hope that we can continue our (by my standards) enormously fruitful collaboration and our pursuit of becoming the Jagger/Richards of informational microtheory.

Towards the end of my PhD studies, the encouragement and support I received from Marco Ottaviani and Alessandro Pavan (who each hosted me at their respective universities) have been immensely helpful, both for my research and in broadening my academic horizons.

I have learned from countless[23] conversations with colleagues from Copenhagen and all over the world; specific acknowledgements can be found under each paper. Here, I want to give special thanks to Sebastian Barfort, Jeppe Druedahl, Benjamin Falkeborg

[23] I know that this is technically not correct.

and Amalie Sofie Jensen, who were always up for discussions and "pseudo-discussions" (Barfort). Thomas Jensen helped me get my teaching jobs, David Dreyer Lassen helped me get into the PhD program, and Alexander Sebald often came by to ask whether I had any chocolate. I also had the great fortune of having an all-star team consisting of Andreas Madum, Kristoffer Lomholt, Michala Riis-Vestergaard and Edward Webb as teaching assistants during my "reign of terror" (Madum) as lecturer in Micro C. Writing a doctoral thesis is not a team effort, but writing it without colleagues would be like doing gymnastics on your own: Slightly ridiculous, at risk of bodily harm, but most of all: much less fun.

Like everyone else, I have at times benefited from chance and serendipity. In 2007, I met Benjamin Bossan (a biologist!) at a lecture on game theory. Our subsequent friendship and our collaboration on the evolution of social learning (which has been published separately) shaped my thinking, but also allowed me to send costly signals when it came to applying for a PhD scholarship and an academic job.

I was convinced to study economics by Michael Møller (by his example, not by his words – he hoped for a long time that I would do something worthwhile with my life instead). He deserves the credit, and he deserves the blame. Thanks to Michael and Susanne for their friendship and continuous support throughout my years in Denmark.

I dedicate this thesis to my parents and grandparents, without whom I would never have embarked on this academic endeavor or had the stamina and intellectual appetite to get through. I cannot remember my parents ever not supporting or at least tolerating any curiosity or intellectual undertaking of mine. I would probably not have become a social scientist were it not for my grandparents' conviction that building a better society is a possibility and a responsibility.

All this gratitude is not to suggest that I am in any way done. Ultimately, the wanderer looking down into the valley turns around to face a whole massif of peaks, each of them so much more impressive than that molehill upon which he stands now.

Ole Jann

Copenhagen, August 2016

Bibliography

Abreu, D. and M. Brunnermeier (2003). Bubbles and crashes. *Econometrica 71*(1), 173–204.

Acquisti, A. (2010). The economics of personal data and the economics of privacy. Background paper 3, OECD WPISP-WPIE Roundtable.

Acquisti, A., C. R. Taylor, and L. Wagman (2015). The economics of privacy. *Journal of Economic Literature 54*(2), 442–492.

Acquisti, A. and H. R. Varian (2005). Conditioning prices on purchase history. *Marketing Science 24*(3), 367–381.

Ali, S. N. and R. Bénabou (2016). Image versus information: Changing societal norms and optimal privacy. mimeo.

Allen, F., S. Morris, and H. Shin (2006). Beauty contests and iterated expectations in asset markets. *Review of Financial Studies 19 (3)*, 719–752.

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

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton, NJ: Princeton University Press.

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics 4*(6), 1236–1239.

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics 1*(1), 67–96.

Banks, J. S. and J. Sobel (1987). Equilibrium selection in signaling games. *Econometrica 55*(3), 647–661.

Baye, M. R. and J. Morgan (1999). A folk theorem for one-shot Bertrand games. *Economics Letters 65*(1), 59–65.

Benhabib, J. and P. Wang (2015). Private information and sunspots in sequential asset markets. *Journal of Economic Theory 158*, 558–584.

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.

Blume, A. (2003). Bertrand without fudge. *Economics Letters 78*(2), 167–168.

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

Brunnermeier, M. and S. Nagel (2004). Hedge funds and the technology bubble. *Journal of Finance 59 (5)*, 2013 – 2040.

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.

Cespa, G. and X. Vives (2015). The beauty contest and short-term trading. *Journal of Finance 70(5)*, 2099–2153.

Chassang, S. (2008). Uniform selection in global games. *Journal of Economic Theory 139*(1), 222–241.

Chassang, S. and G. P. I. Miquel (2009). Economic shocks and civil war. *Quarterly Journal of Political Science 4*(3), 211–228.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics 23*(4), 493–507.

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.

Daughety, A. F. and J. F. Reinganum (2010). Public goods, social pressure, and the choice between privacy and publicity. *American Economic Journal: Microeconomics 2*(2), 191–221.

De Long, J., A. Shleifer, L. Summers, and R. Waldmann (1990). Noise trader risk in financial markets. *Journal of Political Economy 98*(4), 703–738.

Dow, J. and G. Gorton (1994). Arbitrage chains. *Journal of Finance 49*(3), 819–849.

Economist (2013, November 16). The recorded world: Every step you take. *The Economist*.

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

Farrell, J. and M. Rabin (1996). Cheap talk. *Journal of Economic Perspectives 10*(3), 103–118.

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

Foster, D. P. and R. V. Vohra (1997). Calibrated learning and correlated equilibrium. *Games and Economic Behavior 21*(1), 40–55.

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.

Friedman, M. (1962). *Capitalism and Freedom*. Chicago, IL: University of Chicago Press.

Froot, K. and E. Dabora (1999). How are stock prices affected by the location of trade? *Journal of Financial Economics 53*(2), 189–216.

Fudenberg, D. and D. K. Levine (1999). Conditional universal consistency. *Games and Economic Behavior 29*(1), 104–130.

Galbraith, J. (1954). *The Great Crash 1929*. Mariner Books (reprinted 1997).

Glosten, L. and P. Milgrom (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics 14*(1), 71–100.

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

Greenwald, G. (2014). *No place to hide: Edward Snowden, the NSA, and the US surveillance state*. London, UK: Macmillan.

Hamburger, T. and P. Wallsten (2005, July 24). Parties are tracking your habits. *Los Angeles Times*.

Harsanyi, J. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*. MIT Press.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*(5), 1127–1150.

Hart, S. and D. Schmeidler (1989). Existence of correlated equilibria. *Mathematics of Operations Research 14*(1), 18–25.

Hayek, F. (1945). The use of knowledge in society. *American Economic Review 35(4)*, 519–530.

Heinemann, F. and G. Illing (2002). Speculative attacks: Unique equilibrium and transparency. *Journal of International Economics 58*(2), 429–450.

Hermalin, B. E. and M. L. Katz (2006). Privacy, property rights and efficiency: The economics of privacy as secrecy. *Quantitative Marketing and Economics 4*(3), 209–239.

Hirshleifer, J. (1971). The private and social value of information and the reward to incentive activity. *American Economic Review 61*(4), 561–574.

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.

Hurkens, S. and E. van Damme (2004). Endogenous price leadership. *Games and Economic Behavior 47*(2), 404–420.

Janis, I. (1972). *Victims of Groupthink: a Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin.

Kajii, A. and S. Morris (1997). The robustness of equilibria to incomplete information. *Econometrica 65*(6), 1283–1309.

Kaplan, T. R. and D. Wettstein (2000). The possibility of mixed-strategy equilibria with constant-returns-to-scale technology under Bertrand competition. *Spanish Economic Review 2*(1), 65–71.

Keynes, J. (1936). *The General Theory of Employment, Interest, and Money*. Harcourt (reprinted 1965).

Klemperer, P. (2003). Why every economist should learn some auction theory. In M. Dewatripont, L. Hansen, and S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Invited Lectures to 8th World Congress of the Econometric Society*, pp. 25–55. Cambridge University Press.

Kuran, T. (1997). *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

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

Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica 53*(6), 1315–1336.

Leland, H. (1992). Insider trading: Should it be prohibited? *Journal of Political Economy 100(4)*, 859–887.

Liu, L. (1996). Correlated equilibrium of Cournot oligopoly competition. *Journal of Economic Theory 68*(2), 544–548.

Lorenzoni, G. (2008). Inefficient credit booms. *The Review of Economic Studies 75*(3), 809–833.

Lowenstein, R. (2000). *When genius failed: The rise and fall of Long-Term Capital Management*. Random House Trade Paperbacks.

Martews, A. and C. Tucker (2015). Government surveillance and internet search behavior. mimeo.

Meade, E. E. and D. Stasavage (2008). Publicity of debate and the incentive to dissent: Evidence from the US federal reserve. *Economic Journal 118*(528), 695–717.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry 1*(2), 108–141.

Milgrom, P. and J. Roberts (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica 58*(6), 1255–1277.

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 (2002). Social value of public information. *American Economic Review 92*(5), 1521–1534.

Morris, S. and H. S. 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.

Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America 36*(1), 48–49.

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

Ottaviani, M. and P. N. Sørensen (2006). Reputational cheap talk. *RAND Journal of Economics 37*(1), 155–175.

PEN America (2013). Chilling effects: NSA surveillance drives U.S. writers to self-censor. Report, PEN America.

Perotti, E. and J. Suarez (2002). Last bank standing. what do I gain if you fail? *European Economic Review 46*, 1599–1622.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–661.

Posner, R. A. (1981). The economics of privacy. *American Economic Review 71*(2), 405–409.

Robinson, D., H. Yu, and A. Rieke (2014). Civil rights, big data, and our algorithmic future: A September 2014 report on social justice and technology. Technical report, Upturn.

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

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press.

Schneier, B. (2006, May 18). The eternal value of privacy. https://www.schneier.com/essays/archives/2006/05/the_eternal_value_of.html.

Schneier, B. (2015). *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company.

Shiller, R. (2000). *Irrational Exuberance*. Princeton University Press.

Shleifer, A. and R. Vishny (1997). The limits of arbitrage. *Journal of Finance 52*(1), 35–55.

Shleifer, A. and R. Vishny (2011). Fire sales in finance and macroeconomics. *Journal of Economic Perspectives 25*(1), 29–48.

Shleifer, A. and R. W. Vishny (1992). Liquidation values and debt capacity: A market equilibrium approach. *The Journal of Finance 47*(4), 1343–1366.

Solove, D. J. (2010). *Understanding Privacy*. Cambridge, MA: Harvard University Press.

Stigler, G. J. (1980). An introduction to privacy in economics and politics. *Journal of Legal Studies 9*(4), 623–644.

Vitale, P. (2007). An assessment of some open issues in the analysis of foreign exchange intervention. *International Journal of Finance and Economics 12*, 155–170.

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

Wu, J. (2008). Correlated equilibrium of Bertrand competition. In C. Christos Papadimitriou and S. Zhang (Eds.), *Internet and Network Economics*, Volume 5385, pp. 166–177.

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