forthcoming, *American Economic Review*

Digital Dystopia*

Jean Tirole†

December 17, 2020

Abstract: Autocratic regimes, democratic majorities, private platforms and religious or professional organizations can achieve social control by managing the flow of information about individuals' behavior. Bundling the agents' political, organizational or religious attitudes with information about their prosocial conduct makes them care about behaviors that they otherwise would not. The incorporation of the individuals' social graph in their social score further promotes soft control but destroys the social fabric. Both bundling and guilt by association are most effective in a society that has weak ties and is politically docile.

Keywords: Social behavior, social score, platforms, strong and weak ties, social graph, mass surveillance, divisive issues, community enforcement.

JEL numbers: D64, D80, K38.

*This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 669217 - ERC MARKLIM). Jean Tirole acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010. The author gratefully acknowledges the financial support of the TSE Digital Center (the list of sponsors is available at https://www.tse-fr.eu/digital). Daron Acemoglu, Amirreza Ahmadzadeh, Roland Bénabou, Aimé Bierdel, Erik Brynjolfsson, Sylvain Chassang, Bin Cheng, Johannes Hörner, Paul-Henri Moisson, Charles Pébereau, two anonymous referees, and participants at conferences (Luohan Academy conference on privacy and data governance, IT & digitization and IO groups at NBER summer institute, 13th Toulouse conference on the digital economy, privacy conference at Princeton University), and at seminars at MIT, Northwestern, Tehran IAS, TSE and University of Auckland provided helpful comments.

†Toulouse School of Economics (TSE) and Institute for Advanced Study in Toulouse (IAST).

1 Introduction

How transparent should our life be to others? Modern societies are struggling with this question as connected objects, social networks, ratings, artificial intelligence, facial recognition, cheap computer power and various other innovations make it increasingly easy to collect, store and analyze personal data.

On the one hand, these developments hold the promise of a more civilized society, in which incivilities, corruption, fraud, and more generally non-compliance with the laws and norms we deem essential for a successful living together would be a memory of the pre-big-data past. On the other hand, citizens and human rights courts fret over mass surveillance by powerful players engaging in the collection of bulk data in shrouded secrecy; they are concerned that platforms and governments might hold and integrate too much information about what defines us as individuals. This paper attempts to give content to, and shed light on the two sides of the argument, emphasizing the excesses that may result from an unfettered usage of data integration.

Although examples of existing applications to the non-digital world will be provided, the paper is best viewed as an exercise in (social) science fiction. Indeed, I do not advance that, so far, data integration by private platforms or governments has extensively led to dystopic outcomes. Rather, at this junction at which the new technology comes to maturity and given the grave perils posed by such prospects, it is important to understand the channels through which a dystopic society might come about, so as to better design legal and constitutional safeguards.

Section 2 sets up the benchmark. Its framework is borrowed from the literature. Economic agents are engaged in weak or strong ties (transient or stable) relationships and care about their social image. Strong ties characterize family, friendship, village or employee relationships. Weak ties capture matching through platforms, independent contracting or large-city interactions. The agents' very desire to project a good image of themselves may be harnessed to enhance trust in society.

An agent's prosocial behavior may become known to others in two ways: Directly through interacting with the agent, and indirectly through a publicly disclosed rating or social score that encapsulates the agent's behaviors with various agents or in various contexts. This social score is assumed to take a binary form (e.g. the agent is publicly blacklisted or not). "Silo information" prevails in the absence of such a social score. The release of a social score boosts image concerns and thereby prosocial behavior.

There may be under- or over-provision of prosocial behavior, due to two opposite externalities: A classic externality which we associate with the very notion of "prosocial behavior"; and an image externality as social prestige is relative and so agents acquire social esteem at the expense of others. Over-signaling of the kind envisioned in some dystopian movies and books occurs if and only if the prosocial externality is small; social scoring then reduces welfare. For large externalities, social scoring is desirable. The rest

of the paper accordingly focuses on the latter configuration, in line with the standard argument brought in support of social scoring.

The novel analysis begins in Section 3, which analyzes how the state can leverage social sanctions to suppress dissent, or more generally to force citizens to adopt political, societal or religious attitudes that it favors.[1] It generalizes the model of Section 2 by adding another decision for each individual: Dissent or comply (/accommodate the state/toe the line). Each agent's type is now two-dimensional. Besides their prosocial proclivity (intrinsic motivation to do good), agents differ in their psychological cost of compliance. When interacting with others, agents care (mainly) about their reputation with respect to the first dimension. By contrast, the state's objective function is a convex combination of agents' welfare and the extent of compliance (lack of dissent) in society. A more autocratic regime puts more weight on the latter than a less autocratic one.

I compare behavior when the social rating bundles behavior in social interactions together with the dissent/comply choice and when the two are unbundled, to see if the state can and/or chooses to leverage the social score to strengthen its hold on society. In the tradition of information design, I assume that the state can commit to the scoring rule.

The main insights go as follows: 1) Bundling prosocial activities and compliance with the government's objective into a single score exploits agents' interest in each other's score to promote political compliance. 2) The government builds such dual-purpose social scores if and only if it is sufficiently autocratic. 3) Its ability to enforce compliance with its objective through bundling is higher in a weak-ties society than in a strong-ties society. The intuition behind this result is that in a society with strong ties, agents have information about each other that unbundles the information supplied by the government. 4) The government must eliminate competition from independent, privately-provided social ratings. Whenever economic agents are more interested in the social reliability of their partners than in their political, religious or societal views, private platforms would expunge any information about these views from their ratings, de facto unbundling the information. 5) Transparency/citizens' awareness about the way the social score is computed (together with opaqueness about its components) is key to the latter's effectiveness.

Section 4 develops various extensions and reinterpretations of the analysis of Section 3. After demonstrating the robustness of the results to the observability by other agents of the agent's pro/anti-government action, it introduces non-image sanctions. While unavailable to democratic regimes subject to no-discrimination rules, to platforms or to religious or professional organizations, economic sanctions (e.g. price premia on specific services or outright prohibition of purchase) constitute a second pillar in Chinese social credit scoring pilots. The optimal scheme in general involves a mix of image and eco-

[1] Of course, the government's goals need not be stated so bluntly. Behaviors "damaging the dignity or interests of the state or divulging any state secret", "spreading fake news", "fabricating and spreading rumors", or "participating in cult organizations" can be interpreted sufficiently liberally so as to encompass various behaviors that are frowned-upon by the government.

nomic sanctions. The characterization of the bundling strategy is very similar to that in the absence of economic sanctions.

Section 4 then turns to social scoring by actors that are not autocratic governments. First, it notes that a majority in a democratic regime may employ similar techniques to contain a minority's practice of a behavior it reproves. Second, it shows how private platforms may use bundling to subvert democracy. The framework here is a relabeling of the previous one. The "citizens" become the "officials", who have image concerns as they seek re-election. Instead of the platform rating citizens, it "rates" the officials. Concretely, such ratings may take the form of selective disclosure of facts or opinions about politicians, which change the electorate's beliefs about the quality or the congruence of these politicians. An official's decision is again two-dimensional. First, she can strive to serve the citizens or not, the counterpart of the prosocial decision in the basic model. Second, and the counterpart to accommodating the autocratic regime, she can grant a favor to the platform (refraining from asking for tougher antitrust or privacy regulation enforcement or tax collection, subsidizing the media, relaxing online media's editorial responsibility or liability for security breaches) or not. The officials face an (heterogeneous) psychological cost for kowtowing to the platform. I show that private platforms can bundle information about elected officials so as to obtain favors from them, in the same way a state-controlled platform can leverage the social score to suppress dissent.

Section 5 considers the linear-quadratic Gaussian version of the model. In this continuous model, some amount of bundling is always optimal. Compliance receives more weight in the computation of the social score if there is more heterogeneity in pro-social inclination and if individuals are more similar with respect to their aversion to toeing the line. Section 5 also allows types to be correlated. For instance, political compliance to an autocracy might suggest non-prosociality and be socially sanctioned; conversely, for a good cause (say, the environment), support for the government's policies might magnify the reputational gains attached to taking pro-social actions. Compliance increases with the degree of correlation (whether positive or negative). The weight put on compliance hinges on the extent of correlation. While a positive weight incentivizes compliance, signaling concerns call for reducing the dispersion in the scores and thereby creating a rat race for reputation; this second effect in turn suggests a negative weight for negative correlations, and indeed the state's optimal weight on compliance turns negative for a correlation below $-1/\sqrt{2}$.

One of the most problematic aspects of mass surveillance is the coloring of a person's perception by the company she keeps. Guilt by association makes citizens afraid of being seen in company of dissidents or mere citizens whose lifestyle is frowned upon by the regime. Facial recognition and artificial intelligence applied to the surveillance of social life, communications and social network activities have substantially reduced the government's cost of drawing an accurate social graph of relationships among its citizens. Section 6 studies how a government can make use of social graphs by allowing relationships with someone on a blacklist to taint the reputation of those who otherwise would not be on the

black list. Such tainting can induce yet another social pressure -ostracism- on citizens to toe the line. Embodying an individual's social graph into her social score also generates costs, most prominently the destruction of the social fabric: As was observed (with much more primitive surveillance techniques) in East Germany before the reunification, citizens sever beneficial ties with others. Embodying the social graph into the social score appeals to autocratic regimes as it reinforces the state's grip.

I consider a social rating system in which mingling with dissenters is tantamount to dissenting oneself. To avoid being tainted by dissenting friends and family, individuals must therefore ostracize them. Assuming that no-one likes to toe the government's line (but that the intensity of this aversion is heterogenous across individuals), guilt by association, like bundling, is ineffective in a society of strong ties. No-one complies, and people form their own opinion about others without paying attention to the social rating. A society of weak ties exhibits a very different pattern. First, "model citizens", who both comply and behave prosocially, ostracize dissenters so as to keep their favorable rating. A class of "compliers" emerges, who do not obtain a good rating but refrain from dissenting by fear of being ostracized by model citizens. Second, in an example, I show that guilt by association is more effective in taming the population when its initial propensity to dissent is limited (the population is "docile"). Third, there may be multiple equilibria if the value of social links is high: if more agents dissent and fewer behave as model citizens, the ostracizers (dissenters) incur a larger (smaller) social cost of foregone social opportunities; and having dissented becomes a more likely excuse for a low social score. This generates self-fulfilling prophecies. Fourth, even strong ties can be broken by a guilt-by-association rating policy in the context of a mixture of strong and weak ties in society. Finally, I conclude the section with a discussion of the use of guilt-by-association strategies by religious and other organizations.

Section 7 concludes with policy implications and alleys for future research.

Motivation: The advent of social scoring

The much-discussed Chinese social credit system illustrates potential problems. Due to be rolled out in 2020, it was launched in 2014 and was preceded by local experiments starting in the late 2000s. It draws its technological features from the scoring systems developed by the large tech companies. The following discussion is subject to caution, though, as the terms of social scoring are not cast in stone and current pilots may differ from the future implementation anyway. Also, this project is probably not a Chinese idiosyncrasy; while China has a technological lead in the associated technologies and a conducive political situation, social scoring will likely tempt other governments in the near future.

The social score that each individual will receive will embody a variety of criteria; these might include for example credit history, tax compliance, good deeds, environmentally friendly behavior, traffic violations, fraudulent behavior in markets, the spreading of "fake news" and the posting of "inappropriate posts" (whatever this may be interpreted as

meaning), the individual's social graph, personal traits, political or religious opinions, etc.

An individual's social score will be publicly available (and casual evidence on the current experimentations shows that individuals with a favorable score do share it with their relationships anyway) and consequential in two ways. First, it will elicit social sanctions and stigmatization (the modern version of the pillory) as well as social rewards.[2] Second, it will incentivize non-governmental actors to alter their customer relationships to account for the individual's social score; for instance, a bad rating might generate restrictions on access to discounts on purchases, employment, transportation, visas abroad, or access to the best schools or universities.

Soft control vs brute force

An interesting question arises as to why a Leviathan with enough leverage to sustain a law that creates individual social scores does not employ more traditional compliance policies such as brute force and imprisonment instead of bundling and eliciting community enforcement.

In line with the debate between Huxley and Orwell on social control,[3] even an autocratic government may find social scoring an attractive way of ensuring compliance. Traditional repression is rather costly when it extends beyond a small minority; non-social punishments (jail, fines...) are expensive (cost of imprisonment, court inefficiency and corruption[4]...) or require unavailable information. Furthermore, the autocratic government cannot use an iron fist without facing an international opprobrium, especially if punishments are related to political dissent. So, even if alternative punishments are available, the manipulation of social ratings described below can still strengthen the state's enforcement capacity and be an effective instrument.

Of course, I do not expect soft control to always substitute for brute force. Indeed, the paper offers a few clues as to when bundling and/or guilt by association strategies

[2] A vivid illustration of this is the displaying of blacklisted individuals on large LED screens in the streets in some experiments. Key, though, is the wide availability of individuals' ratings. The enlisting of social sanctions by the state is of course not specific to China. For example, under many US states' "Megan's laws", sex offenders' personal information is available on public websites for use by employers and communities. But the scale of China's project, as well as the efficacy of the technology involved, are unprecedented.

[3] The quest for low-cost, long-lasting social control policies is illustrated by Aldous Huxley's October 1949 letter to George Orwell commenting on the latter's dystopian masterpiece, *Nineteen Eighty-Four*: *"Whether in actual fact the policy of the boot-on-the-face can go on indefinitely seems doubtful. My own belief is that the ruling oligarchy will find less arduous and wasteful ways of governing and of satisfying its lust for power, and these ways will resemble those which I described in Brave New World."* [Huxley of course had other instruments (infant conditioning and narco-hypnosis) in mind, and could not have anticipated the emergence of online interactions, servers, AI and facial recognition, but the recent developments fit well with his overall vision. The broader emphasis on soft control of citizens dates back to at least Tocqueville (1838)'s concern that democracies may degenerate into "soft despotism".]

[4] In the case of China, the inefficiency of courts in enforcing law was certainly one of the drivers of the social credit project: See Creemers (MERICS interview, August 21, 2018), Ohlberg et al (2017) and Dai (2018).

may fail to achieve societal control. A society built on strong ties is more resilient to these two strategies. Soft societal control is also less effective when a substantial fraction of the population faces a high compliance cost: the associated increase in the fraction of dissenters (a) provides a better excuse for a low score ("I received a low score because I dissented, not because I am not prosocial") and (b) makes ostracism more costly to the ostracizers and less costly to the dissenters. So soft control is more likely to prevail in weak-ties, docile societies.[5]

Other examples of bundling and/or guilt by association strategies

The second answer to the question of why soft control may advantageously substitute for brute force is simply that the principal may not have coercive power and thus must incentivize agents through the flow of information released about them. Section 3 notes that the underlying logic may be harnessed not only by autocratic governments, but also by entities with limited coercive power: A majority in a more democratic regime or a private platform.

Interestingly, and in line with the gist of the paper, Booking.com's default ranking of hotels embodies in its algorithm not only customer-relevant information such as the ratings by past customers, but also whether the hotel pays its fees to Booking.com on time, an information that is much more relevant to Booking than to the customer.[6] Put differently, Booking.com uses bundling to discipline hotels. In principle, the platform could unbundle (not use this information to rank hotels) and charge penalties for late payments of fees. But this may be an imperfect instrument, both because it is costly to enforce those payments in court and because such penalties are presumably levied on already fragile suppliers.[7]

Bundling and guilt by association strategies are also used by non-governmental organizations that have little or no coercive power. Such organizations are tempted to use their members' image concerns to discipline them, as I will later illustrate with the case of religious communities.

Finally, the paper's core ideas have implications for certification and auditing. The bundling strategy emphasized here could be applied to the mixing of true reporting with "bribes" paid to the certifier (consulting contracts...), with the same need for mutual understanding as to how the grade is actually computed. Whether the market for auditing

[5]It is interesting to note that Uyghurs in China form a strong-ties society and also face a higher cost of compliance with the government's goals than the rest of Chinese society. There is also a political economy constraint to any repressive policy. Such policies must not create too many losers in the population; but Uygurs have little political clout (whether at home or even in Muslim countries).

[6]Booking.com terms state that "On-time payment of commission by accommodations and the commission percentage are also included in the algorithm of the Default Ranking".

[7]Some analogies here: Banks' deposit insurance fees are not risk-based because of the fear that risk-adjusted fees would compound the difficulties faced by distressed banks. And, while Europe's member states in principle pay penalties when they violate their budget and debt caps (under the Maastricht treaty and its updated versions), these penalties are never enforced.

services will take care of this distortion, or else should be regulated is an interesting object of study.

Related literature

The main focus of the economics literature on privacy, nicely reviewed in Acquisti et al (2016), has been the ability of platforms to use data collection and resale to achieve more profitable second- and third-degree price discrimination.[8] Data also enable sellers to target their ads to the intensity of the match value and buyers to reduce their search costs (of course targeted ads may occasionally raise privacy concerns). My emphasis on the use of data integration to enlist community enforcement is, to the best of my knowledge, novel.

Like in Kamenica-Gentzkow (2011) and Rayo-Segal (2010)'s pioneering analyses of Bayesian persuasion and the broader information design literature, players are Bayesian and the sender commits to a disclosure policy (see Proposition 4 on what happens otherwise in my model). A commitment to communicate only a coarse version of the signal in general benefits the sender (a related literature in Computer Science goes under the name of "strategic classification"; see Hardt et al 2016). This pooling, recently studied in multi-dimensional type spaces by Ball (2020) and Frankel-Kartik (2019a) may take the form of an intermediary who filters the information to reduce the impact of the signal on the sender's reward and thereby the distortion (Ball) or a mere commitment by the receiver to an allocation rule that is not ex-post optimal (Frankel-Kartik). Like in these papers and in Bergemann et al. (2015) analysis of aggregate information collection from multiple agents and Bonatti-Cisternas (2020)'s model of scoring-induced ratcheting, the type multidimensionality forces us to restrict attention to simple rules, in my context either a blacklist or a linear scoring rule. I provide the first analysis of social-score bundling, its limits and its welfare impact; incentives induced by guilt by association also are new.

Information design's commitment assumption is reasonable in the context of social scoring. In some of the Chinese pilots or in the case of Booking, the principal discloses the method of computation of the social score; agents can then observe whether the algorithm is indeed employed. Second, when the principal is a religious or professional order, a platform or an accounting firm, agents may personally or by word of mouth learn "how it works"; social learning make them realize that a compliance dimension is embodied in their overall assessment or treatment (naming and shaming, temporary exclusion, excommunication...). Such social learning may also be important when there is explicit disclosure, but the meaning of some terms ("fake news") is subject to interpretation.

[8] Zuboff, in her wider-audience essay (2018), goes beyond the issue of capture of "behavioral surplus" and insists on the loss of agency created by platforms' nudges, enticements, and exploitation of consumers' compulsive nature and habituation.

The paper is related to the large literature on community enforcement.[9] It differs from it both in terms of modeling (through the use of type-based reputation and image concerns instead of a repeated-game approach) and, more importantly, in its focus. First, while that literature unveils the informational and matching conditions under which cooperation can be sustained when relationships are transient, this paper emphasizes how platforms, organizations and governments can employ data integration to further their own goals. Second and relatedly, the repeated-game literature mostly posits benefits from community enforcement (and accordingly focuses on equilibria that exhibit a high level of enforcement), while I stress dysfunctional features of such enforcement.

Image concerns feature prominently in a number of theoretical and empirical contributions.[10] This paper uses the Bénabou-Tirole (2006, 2011) model of image concerns. As I will later discuss, that literature has brought to light the possibility of under- and oversignaling. The existing literature supplies the building block for the study of the strategic use of social scoring by the public and private sectors (Sections 3 through 6).

My study of guilt by association involves citizens selecting in their potential social graph; they may ostracize other citizens whom ceteris paribus they would wish to interact with, but who would confer upon them a bad reputation. This tainting or contagion effect is also present in Peski-Szentes (2013), although in a very different manner. Unlike in my paper, partners' type/group is payoff irrelevant in the latter two-group paper; yet, individuals may not associate with members of the other group by fear that members of their own group ostracize them in the future. Everyone receives (incomplete) information about the type of partners their prospective match has had in the past. This repeated game always has an efficient full-matching equilibrium in which reputation and group belonging play no role, but may have another one in which people do not mingle much across groups. Another paper in which reputations affect matching patterns is my paper on collective reputations (Tirole 1996). There, an individual's type (reliability) is payoff-relevant, and potential partners imperfectly observe her track record. The collective reputation of a group results from its members' behaviors and in turn behaviors are shaped by the group's reputation. The paper characterizes the joint dynamics of individual and collective reputations, with a unique path from any initial condition but a good and a bad steady states. An important difference of the current model with these two models is that reputations are the outcome of information design rather than of exogenously imperfect observability of past behavior.

[9] Initiated by Rosenthal (1979), Kandori (1992) and Ellison (1994). See Acemoglu-Wolitzky (2020) and Clark et al (2019) for recent contributions to the literature. Ali and Miller (2016) study the incentive to disclose a deviation in a bilateral relationship to other agents outside the relationship. Such disclosure is important to trigger multilateral punishments and to make use of the defaulter's overall reputational capital; but it may not be incentive compatible as it destroys some of the benefits from the leveraging. Temporary ostracism generally dominates a permanent one.

[10] On the theory side, contributions include for example Bénabou et al (2018), Bénabou-Tirole (2006, 2011), Bernheim (1994) and Ellingsen-Johannesson (2008). On the empirical front, e.g. Ariely et al (2009), Besley et al (2015), Bursztyn-Jensen (2017), Bursztyn et al (2018), Chen (2017), DellaVigna et al (2012), Karing (2019), Jia-Persson (2017) and Mellström-Johannesson (2008). On both sides the literature is too large to be given proper credit here.

2 The calculus of social approval

2.1 The framework

The model posits that an individual's social behavior results from her intrinsic motivation to do good for others, her cost of doing so, and finally her desire to project a good image of herself.

Drivers of social behavior. Relationships are exogenous. An agent i's action, a_i, is observed by another agent, j (who may equivalently stand for a group of agents, a_i then being a behavioral pattern). Agent i decides to be prosocial ($a_i = 1$) or not ($a_i = 0$). Being prosocial generates an externality $e > 0$ on agent j or on third parties (as in the case of greenhouse gas emissions), and involves private cost c for individual i. The relationship between i and j is a strong tie one in that i cares about the image she projects on j (say, because she will interact again with j in the future).

Individuals are heterogenous with respect to their desire to do good and agent i's intrinsic motivation is unknown to others. Namely, agent i's intrinsic motivation to do good is $v_i e$, where v_i is distributed according to smooth cumulative distribution $F(v_i)$ and density $f(v_i)$ on $[0, 1]$ (the agent does not put more weight on others' utility than on her own), with mean $\bar{v}$. Individual i's intrinsic motivation, v_i, is known to her, but not to others.

Behaviors are driven not only by intrinsic motivation and cost, but also by the desire to look prosocial; that is, individual i cares about others' posterior beliefs $\hat{v}_i$ about her type. This demand for a good reputation may be associated with pure image concerns. Alternatively, a good reputation allows the individual to take advantage of assortative matching to derive future benefits: See the online Appendix. Agent i cares about her reputation with the strong-tie agent j, as well as with other agents she will encounter in the future but who will not have directly observed her behavior. Family, friendship, neighborhood and some work relationships are usually stable, while other social, commercial or work relationships (say, involving foreign trade, platform or large-city interactions) are more akin to the pattern observed in a "society of strangers" (Seabright 2010).

Information. I consider two structures of information:

- *Silo information.* Agent j observes a_i. This information structure is the minimal information structure for interacting agent j. For other (currently non-interacting) agents, the minimal information structure is $\varnothing$ (no information).

- *Social scoring.* Under social scoring, strong and weak ties learn i's social score. In this economy, the social score is simply the action: $s_i = a_i$ (I adopt the convention that $s_i = \varnothing$ in the absence of social scoring). It therefore conveys no new information to j. More generally, the social score would convey extra information even to strong-

tie relationships if either observation was noisy or delayed, or if there were multiple actions.[11] Both can easily be accommodated within the framework.

Strong-tie partner j's information structure about i is $I_{ij} \equiv \{a_i, s_i\}$. Other future partners have information $I_i \equiv \{s_i\}$ about i.

Note that I implicitly assume that agents provide a truthful rating or bring evidence about the behavior of those with whom they interact (or alternatively that the incivilities toward them or toward society as a whole are recorded through cameras equipped with facial recognition).

Payoff functions. Individual i puts weight μ on her reputation vis-à-vis the strong-ties relationship(s) and ν on that vis-à-vis weak ties (here captured by other future relationships). She has payoff function

$$u_i = (v_i e - c)a_i + \mu \hat{v}_i(I_{ij}) + \nu \hat{v}_i(I_i),$$

where $\hat{v}_i(I_{ij})$ and $\hat{v}_i(I_i)$ are the posterior expectations of v_i conditional on informations I_{ij} and I_i, respectively. The intensities μ and ν of social image concerns, which are assumed to be common knowledge, reflect the stability or transience of relationships. In a strong-ties economy, $\nu = 0$. By contrast, on a sharing platform, $\mu = 0$ to the extent that the individual will in the future interact with new agents.

Welfare. The exact definition of social welfare hinges on why agents value their reputations vis-à-vis the agents they are interacting with (μ) as well as new partners (ν). If a gain in reputation is valued either for pure image concerns or because of assortative matching (as described in the online Appendix), this gain has no social value and reputation is a "positional good": An agent's gain is another agent's loss. Although this is not required for the results,[12] I will define welfare related to agent i's decision assuming that image is a positional good:

$$W = E\big[[(v_i e - c) + e]a_i\big]. \tag{1}$$

[11] For instance, if i interacts with strong-tie individuals $j \in J_i$, then i's social score might be i's average behavior: $a_i = \frac{\Sigma_{j \in J_i} a_{ij}}{|J_i|}$ if J_i is finite and $a_i = \int_0^1 a_{ij} dj$ if $J_i = [0,1]$. Actions could also be weighted according to their "importance" (in Chinese pilots, the social score is a weighted average of measured actions).

[12] In general, the release of a social score may avert future matches that deliver a negative joint surplus or, to the contrary, prevent matches that would have created a positive joint surplus. If the reputation mechanism serves to exclude undesirable agents from future interactions, it per se can add social value over and beyond the expression of W in (1). Conversely, information disclosure may rekindle prejudices or encourage discrimination: A racist may refuse to rent an apartment to a member of a minority; the gay, the rich or the member of a religious or ethnic minority may be victims of acts of aggression, etc.

These considerations would lead to the addition of an extra term in the expression of W in equation (1). This different expression would change the boundary between the regions of under- and over-provision of prosocial behavior, but it would not affect the key drivers of my analysis: Individual behavior would still be driven by the desire to build a good reputation; and, anticipating a bit, those variants would alter the welfare cost of bundling ruler-relevant information with actual pro-social behavior and of using the individual's social graph, but not the political benefit obtained through this bundling, delivering similar effects and comparative statics. For expositional simplicity I will therefore adopt (1) as the expression of welfare.

Thus, from the social point of view agent i should choose $a_i = 1$ if $v_i \geq v^{SO}$ (and $a_i = 0$ for all j otherwise), where

$$v^{SO}e - c + e = 0.$$

2.2 Silo information vs. social scoring

Let $\varphi = \mu$ under silo information and $\varphi = \mu + \nu$ under transparency denote the intensities of image concerns without and with social scoring. Because of single crossing, agent i selects $a_i = 1$ if and only if $v_i \geq v^*$ for some threshold v^*. The cutoff if interior, solves

$$v^*e - c + \varphi\Delta(v^*) = 0 \tag{2}$$

where

$$\Delta(v^*) \equiv M^+(v^*) - M^-(v^*) \equiv E[v|v \geq v^*] - E[v|v < v^*].$$

The function Δ measures the difference between the glory attached to behaving prosocially and the stigma incurred when being selfish. When $\Delta' < 0$, an increase in prosocial behavior (a reduction in v^*) raises the agent's reputational incentive. Prosocial behavior is subject to strategic complementarities: It is a norm. And conversely when $\Delta' > 0$.[13]

When $\Delta' < 0$, one must assume that image concerns are not so strong as to preclude uniqueness of the cutoff; I will assume a unique equilibrium (here guaranteed by $e+\varphi\Delta' > 0$) throughout the analysis. I will further adopt the convention that, for intensity of image concerns φ (here $\varphi = \mu$), $v^* = 1$ if $e - c + \varphi\Delta(1) \leq 0$ and $v^* = 0$ if $-c + \varphi\Delta(0) \geq 0$.[14] My theory will not hinge on whether a norm or anti-norm prevails.

Let v^s ("s" stands for "silo") denote the cutoff when $\varphi = \mu$; similarly, $v^t < v^s$ is the cutoff under transparency (the intensity of image concerns is $\varphi = \mu+\nu$). For $v^* \in \{v^s, v^t\}$, there is underprovision (resp. overprovision) if $v^{SO} < v^*$ (resp. $v^{SO} > v^*$). The welfare impact of a release of a social score hinges on whether the agent faces too little or too much incentives in the first place.

Proposition 1 *(impact of social scoring) Let $\varphi = \mu$ under silo information and $\varphi = \mu+\nu$ under transparency denote the intensities of image concerns without and with social scoring.*

[13]The theoretical and empirical literatures have looked at the determinants of the existence of a norm; among these is the shape of the probability distribution over types. Jewitt (2004)'s lemma indicates that (a) if the density f is everywhere increasing, then $\Delta' < 0$; (b) if it is everywhere decreasing, $\Delta' > 0$; and (c) if f is single-peaked, Δ is first decreasing in v^* from $\Delta(0) = \bar{v}$ and then increasing in v^* to $\Delta(1) = 1 - \bar{v}$. When the distribution is single-peaked, the minimum of Δ in general is not reached at the mode of the distribution, unless the distribution is symmetrical (Harbaugh-Rasmusen 2018). See Adriani-Sonderegger (2019) for a much broader discussion of the properties of the Δ function.

[14]A corner solution at $v^* = 0$ (resp. $v^* = 1$) if and only if $\varphi\bar{v} \geq c$ (resp. $\varphi(1 - \bar{v}) \leq c - e$). Thus, the condition $\varphi(1 - \bar{v}) + e > c > \varphi\bar{v}$ is sufficient for the existence of an interior equilibrium (and uniqueness under the D1 refinement).

(i) In equilibrium individual i picks $a_i = 1$ if $v_i > v^*$ and $a_i = 0$ if $v_i < v^*$, where v^* solves

$$v^* e - c + \varphi\Delta(v^*) = 0 \tag{3}$$

if interior, and $v^* = 0$ (resp. $v^* = 1$) if and only if $\varphi\bar{v} \geq c$ (resp. $\varphi(1-\bar{v}) \leq c - e$).

(ii) There is underprovision of prosocial behavior if and only if

$$e > \varphi\Delta(v^*). \tag{4}$$

Proposition 1 checks for this model the standard result according to which there is underprovision for large externalities and overprovision for small ones.[15] The imperfect internalization of the externality is a driver of underprovision, while the desire to gain social recognition may lead to oversignaling for minor externalities.[16]

3 Leveraging social sanctions to consolidate political power

3.1 Broadening of the social score

Let us now introduce a government eager to suppress political dissent or more generally to promote some form of social compliance, and ask ourselves: Can such a government use a social rating scheme in order to consolidate its power? To study this question, I isolate the impact of such a rating by abstracting from policies that are usually associated with an Orwellian state: Brutality, misinformation and denial of truth. In my Huxleyian, soft control world, the government's only instrument is the design of the flow of information.

There are indeed concerns that autocratic regimes might use social scoring to target dissidents,[17] defense lawyers, journalists, or mere individuals who have the misfortune to read the "wrong" books or have tastes that differ from the officially prescribed ones.

[15] At least if $e + 2\varphi\Delta' > 0$ a slightly stronger assumption than the one I made. See e.g. Acquisti et al (2016), Ali-Bénabou (2019), Bénabou-Tirole (2006) and Daugherty-Reinganum (2010). The differentiation of (4) with respect to e yields: $\frac{d}{de}(e - \varphi\Delta(v^*)) = 1 + \varphi\frac{\Delta'(v^*)v^*}{e+\varphi\Delta'(v^*)} > 0$ if $e + \varphi\Delta'(v^*)(1 + v^*) > 0$, which is trivially satisfied in the uniform case ($\Delta' \equiv 0$) or an anti-norm ($\Delta' > 0$). The result that the release of the social score generates more prosocial behavior is similar to that created by an increase in audience size in Ali-Bénabou (2019). The latter also studies the noisy observation of actions, and relates such imperfect measurement to the effect of scaling up or down the size of audience.

[16] As illustrated by Lacie in the series *Black Mirror* ("Nosedive", season 3, episode 1), who is condemned to constantly smile and put up a good front. Other illustrations of oversignaling include wishing "happy birthday" to Facebook "friends" one hardly knows (and accepting them as "friends" in the first place); and creating and maintaining flattering profiles of oneself on Tinder and Facebook.

[17] As Dai (2018) argues, *"As the comprehensive reputation scoring schemes adopted in the Suining and Qingzhen illustrate, authorities in China may in particular feel little constrained from attempting to use negative reputation scoring to restrain local residents from exercising their rights in making online complaints, filing petitions or even public protests."*

Similarly, countries with a state religion, especially theocratic ones, may use social scoring to promote religious fervour.

Agent i now takes two actions:

1. An anti- or pro-social action $a_i \in \{0, 1\}$.

2. An anti- or pro-government action $b_i \in \{0, 1\}$. Behavior $b_i = 0$ is to be interpreted as not toeing the party line, dissenting, exhibiting disapproved tastes, lacking religious fervour or patriotism, etc. Behavior $b_i = 1$ means "compliance".

The agent's type is two-dimensional. As earlier, v_i, drawn from $F(\cdot)$ with strictly positive density $f(\cdot)$ on $[0, 1]$, indexes the agent's intrinsic motivation to do good in her bilateral interactions. But the agent is also characterized by a (positive or negative) psychological cost of compliance, θ_i, distributed according to smooth cumulative distribution $G(\cdot)$ with density $g(\cdot)$. Types v_i and θ_i are independent (see Section 5 for the case of correlation).

As earlier, action a_i is learned only by j and the government. I assume that b_i is observed only by the government. I will later note that little (nothing if $\operatorname{supp} G = \mathbb{R}^+$) is changed if b_i is observed also by other agents.

Next, I posit that in forming opinions about others, agents put more weight on their prosociality than on their attitude towards the state's agenda; I capture this in a stark way by assuming that agent i cares solely about her reputation(s) regarding her prosocial type.[18] Implicitly, other agents stress her reliability and are indifferent to her personal tastes concerning the government's agenda. Thus, agent i's objective is

$$u_i = (v_i e - c)a_i + \mu \hat{v}_i(I_{ij}) + \nu \hat{v}_i(I_i) - \theta_i b_i,$$

where, as earlier, I_i is the public information, and I_{ij} combines the public information with the observation of i's prosocial behavior in the bilateral $\{i, j\}$ relationship.

Government's objective function. To express the government's concern about dissent, let its objective function be a convex combination of welfare and a political objective:

$$V = W + \gamma E[b_i], \quad \text{where} \quad \gamma \geq 0. \tag{5}$$

When $\gamma = 0$, the government is benevolent (internalizes only the agents' welfare W). The higher γ is, the more "autocratic" the government.[19] Because the model applies

[18] This assumption is also relaxed in the linear-quadratic Gaussian model in Section 5.

[19] The results do not hinge on the exact functional form for the government's maximand (here a weighted average of citizens' welfare and of the number of dissenting acts). The key feature is that the government puts more weight than citizens themselves on some type of behavior- here compliance with the government's own objective. For instance, King et al (2013) argue that the Chinese government's main concern is to prevent collective expression; the paper finds that some forms of small, isolated protests and of criticism of party officials (in particular local ones) are tolerated by the censorship, while anything that could generate a collective action is not (similarly, the Qin et al 2017 and 2018 papers show that

to a variety of organizations, "autocratic" will more generally refer to an "autocratic management"; an "autocratic organization" is an organization whose leadership attaches a high value to internal discipline/respect for the leadership/adherence to the official line.

Expectations are now taken over the joint distribution of (v_i, θ_i). The government's partial internalization of agents' welfare may obey one of four possible rationales: A true empathy, concerns about legacy, a fear of rebellion and upheaval, and electioneering.

Unbundling benchmark. I start with the straightforward case in which the government releases agent i's behavior in the two realms. Because the θ_i-reputation is irrelevant in private relationships,

$$b_i = 1 \quad \text{iff} \quad \theta_i \leq 0.$$

Because θ_i and v_i are independently distributed, agent i chooses $a_i = 1$ if and only if

$$v_i e - c + (\mu + \nu)[E(v_i|a_i = 1) - E[v_i|a_i = 0]] \geq 0,$$

so the cutoff, v^u, if interior, is given by

$$v^u e - c + (\mu + \nu)\Delta(v^u) = 0,$$

(this cutoff is to be taken equal to 1 if the solution to $v^* e - c + (\mu + \nu)\Delta(v^*) = 0$ exceeds 1, or to 0 if it is negative); let $\Delta^u \equiv \Delta(v^u)$.

Proposition 2 *(unbundling). When the government separately releases behaviors (a_i, b_i) in the two domains, then the individual solves two distinct decision problems:*

(i) $b_i = 1$ *iff* $\theta_i \leq 0$*;*

(ii) $a_i = 1$ *iff* $v_i \geq v^u$ *where* $v^u e - c + (\mu + \nu)\Delta(v^u) = 0$.

Behavior is the same as if the government released only $\{a_i\}$*; and so,* $v^u = v^t$.

Bundling. I next assume that the government has monopoly over the provision of a social score and bundles the two informations about behavior by granting one of two ratings (a blacklist system). It conditions a good rating not only on a good social behavior, but also on compliance:

$$\begin{cases} 1, & \text{with associated reputation } \hat{v}_1, \text{ if } a_i = b_i = 1 \\ 0, & \text{with associated reputation } \hat{v}_0, \text{ otherwise.} \end{cases}$$

I consider sequentially the cases of strong and weak ties.

---

the central government tolerates some forms of microblogging so as to predict protests and strikes and to learn about local officials' corruption). In this example, and more broadly in environments where dissent exhibits a strength in numbers, the second term in the government's objective function might well be a convex, rather than a linear function of $E[b_i]$, and one might conjecture that social graphs would receive even more attention than predicted in Section 6.

Bundling in a society with strong ties

Suppose that relationships are sustained rather than transient ($\mu > 0 = \nu$). I argue that the state will find it difficult, even impossible, to leverage social scoring to consolidate political power in a strong-ties society. The rationale for this claim is that, in a tight-knit-relationships society, agents have information about each other that acts as a counterweight for the information supplied by the state. Indeed we have:

Proposition 3 *(ineffectiveness of bundling in a strong-ties society). When relationships are sustained ($\mu > 0 = \nu$), the state cannot leverage a monopoly position on social ratings in order to consolidate political power: There exists an equilibrium whose pro-social and dissent behaviors are the same as in Proposition 2.*

I only sketch the proof. Agent j's posterior belief about i is $\hat{v}_{ij} = M^+(v^u)$ if $a_i = 1$ and $\hat{v}_{ij} = M^-(v^u)$ if $a_i = 0$, regardless of what the government reports, where $v^u e - c + \mu\Delta(v^u) = 0$. Because agent j is uninterested in $\hat{\theta}_i$, the bilateral behavior contains all information about i that agent j wants to know. Any social rating is superfluous. So $b_i = 1$ if $\theta_i < 0$ and $b_i = 0$ if $\theta_i > 0$.

While I do not have a counterexample to uniqueness (that would satisfy the uniqueness assumption in the model of Section 2), the equilibrium selected in Proposition 3 is in the spirit of Markovian equilibria, with a coarsening of strategies to let a depend only on v and b on θ, reflecting the separability of the payoff function.

Remark. An imperfect observability of bilateral behavior or an heterogeneity of behaviors by agent i's within her strong-ties social group would reinstate a role for social ratings, bringing the analysis closer to that for a weak-ties society (Section 3.2). Similarly, when the types v_i and θ_i are correlated, b_i is informative given a_i (see Section 5). The broader picture therefore is that bundling is less effective, but not necessarily inoperative, in a society with strong ties.

3.2 Society with weak ties

Let us now assume that $\mu = 0$ and $\nu > 0$. For expositional simplicity, let us further assume that $c \geq e$. This assumption implies that image concerns are required in order to generate prosocial behavior.[20]

Agent i's utility under bundling is

$$u_i \equiv (v_i e - c)a_i - \theta_i b_i + \nu a_i b_i(\hat{v}_1 - \hat{v}_0) + \nu\hat{v}_0.$$

Because of the assumption that image concerns are required to generate prosocial behavior, the pattern ($b_i = 0$ and $a_i = 1$) is ruled out, and only three possible behavioral

[20] As $ve - c \leq 0$ for all $v \in [0, 1]$. This assumption has the extreme implication that a dissenter does not behave prosocially. The important feature for the theory is that the dissenters' loss of image concerns reduces their incentives to behave prosocially. Therefore, the assumption is there mainly for expositional simplicity.

patterns emerge in equilibrium. Furthermore, when $a_i = 0$, $b_i = 1$ if and only if $\theta_i \leq 0$. So

$$\begin{cases} a_i = b_i = 1 & \text{iff } v_i e - c + \nu(\hat{v}_1 - \hat{v}_0) \geq \begin{cases} \theta_i & \text{if } \theta_i \geq 0 \\ 0 & \text{if } \theta_i < 0 \end{cases} \\ a_i = b_i = 0 & \text{iff } v_i e - c + \nu(\hat{v}_1 - \hat{v}_0) < \theta_i \text{ and } \theta_i > 0 \\ a_i = 0, b_i = 1 & \text{iff } v_i e - c + \nu(\hat{v}_1 - \hat{v}_0) < 0 \text{ and } \theta_i \leq 0 \end{cases} \tag{6}$$

Let $v^b(\theta_i)$ denote the cutoff under bundling for a given θ_i (with again the convention that it is equal to 1 if the solution to (6) with equality exceeds 1, and to 0 if the solution is negative). This threshold is weakly increasing, as depicted in Figure 1.

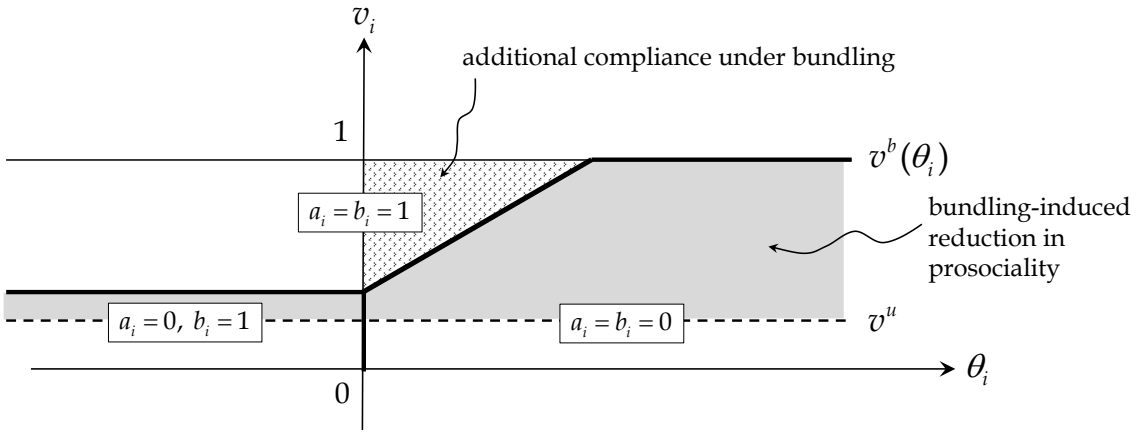

Figure 1: behavior under bundling and unbundling

[The dotted area represents types who increase their compliance under bundling and the shaded area types who reduce their prosocial behavior. Other types' behavior is unchanged.]

Let $g_1(\theta)$ and $g_0(\theta)$ denote the conditional densities.[21] $g_1(\theta)/g(\theta)$ and $g_0(\theta)/g(\theta)$ are weakly decreasing and increasing in θ, respectively.

The image gain can be written as

$$\Delta^b \equiv \hat{v}_1 - \hat{v}_0 = \int [g_1(\theta) M^+(v^b(\theta)) - g_0(\theta) M^-(v^b(\theta))] d\theta.$$

The Appendix shows, in the case of a norm ($\Delta' \leq 0$) or as long as Δ' is positive but "not too high", the existence of an equilibrium with the features that image concerns incentives are reduced by bundling and the provision of prosocial behavior is smaller across the board (for all θ_i). With a uniform density ($\Delta' = 0$), the equilibrium necessarily satisfies these properties. The intuition for these results, depicted in Figure 1, goes as follows:

[21] Conditional on rating 1 or 0, the posterior distributions are $g_1(\theta) = \dfrac{g(\theta)[1 - F(v^b(\theta))]}{\int g(\tilde{\theta})[1 - F(v^b(\tilde{\theta}))] d\tilde{\theta}}$ and $g_0(\theta) = \dfrac{g(\theta) F[v^b(\theta)]}{\int g(\tilde{\theta}) F(v^b(\tilde{\theta})) d\tilde{\theta}}$ respectively.

(i) The cost θ_i of compliance (when positive) acts as an "excuse" for not contributing. Indeed for high θ_i, the conditional reputation[22] when not acting prosocially is the prior mean $\bar{v}$, the highest possible reputation in the absence of contribution.

(ii) A positive cost θ_i raises the cost of obtaining a good rating, and thus reduces the incentive for prosocial behavior. In the presence of a norm (i.e. strategic complementarities: $\Delta' \leq 0$), the lower contribution is self-reinforcing.

Note that bundling increases the fraction of compliers $E[b_i]$ from $G(0)$ to $1-\int_0^\infty g(\theta)F(v^b(\theta))d\theta$. A revealed preference argument (V depends on γ only through the additive term $\gamma E[b_i]$) therefore implies that a more autocratic ruler is more likely to bundle.

Finally, the reduction in prosociality (for all θ) is costly whenever there is under-signaling when the ruler unbundles. To see this, let

$$W \equiv E[(ve - c + e)a(v, \theta) - \theta b(v, \theta)];$$

denote the agents' welfare; it takes value W^u under unbundling and W^b under bundling; bundling generates two inefficiencies: (a) the loss of valuable prosocial contributions: $v^u e - c + e \geq 0 \Rightarrow ve - c + e > 0$ for all $v > v^u$; and (b) counterattitudinal behaviors with respect to identity ($b_i = 1$ when $\theta_i > 0$). So $W^u > W^b$. I collect these results and further characterizations in the next proposition:

Proposition 4 *(bundling under weak-ties relationships). Consider a weak-ties society ($\mu = 0 < \nu$) and assume that $\Delta' \leq 0$. Under bundling, there exists an equilibrium satisfying (all equilibria satisfy these properties when the distribution F is uniform, so $\Delta' \equiv 0$):*

(i) Image concerns are reduced relative to unbundling: $\Delta^b < \Delta^u$, and the prosocial contribution is lower as well (the equilibrium behavior is given by $v^b(\theta) > v^u$ for all θ and depicted in Figure 1: All types θ behave less prosocially).

(ii) The prosocial contribution $\bar{a}(\theta) \equiv 1 - F(v^b(\theta))$ is decreasing in θ.

(iii) There is less dissent ($E[b_i]$ is higher) under bundling than under unbundling; accordingly, there exists $\gamma^* > 0$ such that the government chooses to bundle if and only if $\gamma \geq \gamma^*$.

(iv) Suppose there is underprovision of prosocial behavior under unbundling (i.e. $e \geq \nu\Delta(v^u)$). Then bundling (which occurs whenever $\gamma \geq \gamma^*$) is socially strictly suboptimal ($W^u > W^b$) for two reasons: It discourages prosocial behavior (a-dimension) and it induces counterattitudinal behavior with respect to identity (b-dimension).

[22] Recall that θ_i is not observed by the agent's "audience". The point is that a lack of prosocial behavior might come from a strong aversion to toeing the line and so receives a lower stigma than under unbundling.

This behavior illustrates the trade-off faced by the government: Bundling reduces dissent; but it imposes collateral damages on private relationships by reducing pro-social behavior and it forces a fraction of citizens to adopt counterattitudinal behaviors. In the end, the resolution of this trade-off hinges on how autocratic the regime is (i.e. on γ).

Example 1. Suppose that G puts weight on two types, θ_L (probability κ) and θ_H (probability $1-\kappa$), with $0 < \theta_L < \theta_H$. Identity θ_H is strong enough that under bundling the individual picks $a_i = b_i = 0$ regardless of v_i. By contrast, the cutoff v^b is interior for identity θ_L. Straightforward computations show that an increase in the propensity to rebel (a decrease in κ) (a) reduces type θ_L's prosocial behavior (it supplies a better excuse for not contributing: With a stronger overall identity, the absence of contribution is more likely to be associated with a strong aversion to compliance, and less likely to be attributed to low ethical standards); and (b) makes the loss of prosocial behavior of high-identity types more costly. The implication is that, assuming an underprovision of prosocial behavior, *bundling is optimal for the government if κ is large enough, i.e. there is not too much potential for dissent (the people are expected to be docile).*[23]

Remark (Social score popularity). I observed that bundling reduces aggregate social welfare. The individual impact of bundling is of course type-specific; high-θ types are more affected by bundling. Furthermore, the popularity of the social score will depend on the benchmark ingrained in the citizens' mind. A social score using bundling may be preferable to no social score at all. To see this, let us maintain the simplifying assumption that image concerns are needed to generate prosocial behavior ($c \geq e$); then, when relationships are transient, society is not self-regulated as it exhibits no prosocial behavior in the absence of ratings. The introduction of a social score with bundling benefits everyone in society if it generates enough prosocial behavior.[24] When introducing a social score that allows for bundling, the government will accordingly stress the benefits in terms of bridging the trust gap among citizens and between individuals and businesses.

3.3 Implications

(a) *The need to centralize social ratings*

[23]The cutoff v^b is given by $v^b e - c + \nu[M^+(v^b) - \frac{\kappa F(v^b)M^-(v^b)+(1-\kappa)\bar{v}}{\kappa F(v^b)+(1-\kappa)}] = \theta_L$, and is a decreasing function of κ. To prove the result, note that bundling generates:

- a loss on θ_H types equal to $(1-\kappa)[\int_{v^u}^1 (ve - c + e)dF(v)]$, where $v^u e - c + \nu\Delta(v^u) = 0$,
- a net compliance gain on the θ_L types equal to $\kappa[[1-F(v^b)]\gamma - \int_{v^u}^{v^b}(ve - c + e)dF(v)]$.

Assuming underprovision of prosocial behavior ($e \geq \nu\Delta(v^u)$), bundling is optimal if and only if $\kappa \geq \kappa^*$ for some $\kappa^* > 0$.

[24]To see this, let $\mathcal{E}$ denote the aggregate externality (a minorant of $\mathcal{E}$ is $G(0)[1-F(v^b(0))]e$); then if $\mathcal{E} \geq \nu\bar{v}$, everyone gains, as individuals receive payoff $\nu\bar{v}$ in the absence of social rating and so $\nu\bar{v}$ is a majorant on the loss of image when a social score is introduced.

Consider an autocratic government with $\gamma \geq \gamma^*$. For bundling to accomplish its purpose, the government must not share its prerogative. Suppose a contrario that the private sector has access to the same data and publicly issues social scores. Because economic agents are interested in the social reliability of their partners, but not in whether these partners' tastes fit with the government's views, private platforms would expunge any information about b_i from their ratings.[25] This would lead to de facto unbundling, and no-one would pay attention to the government's social score.[26] More generally, the government does not need to have a full monopoly over scores. Inattention and network externalities might imply that although multiple scoring systems are running side by side, different communities might look exclusively at some scores while other individuals might utilize others.

(b) *The need for commitment*

A common justification given for being wary of state-controlled social scores is their "opaqueness". Note that the scheme considered here is opaque in one sense and completely transparent in another. It is opaque in that the state bundles an agent's various dimensions of social activity into a single score; the private contributions must not be identifiable -in a statistical sense- from the social score or other data sources readily available to the agents. It is transparent in the sense of information design: The method of computation is disclosed and common knowledge.

Contrary to what is occasionally asserted, it is here essential that the algorithm be transparent (or, equivalently, that agents learn how it works). For, suppose that the government does not commit to a method of computation and decides ex post on the score to be given to each agent. The government may for instance take revenge against, and give a low score to (perhaps a fraction of) citizens having expressed dissent ($b_i = 0$). But this time-consistent behavior completely defeats the purpose, as no-one looks at the ratings. It is precisely because the social score sufficiently embodies useful elements (the value of a_i) that it is effective.

To be more formal, perturb slightly the government's objective function into:[27]

$$V = W + \gamma E[b_i] - \varepsilon \int [\nu \hat{v}(v, \theta)] \xi(\theta) g(\theta) f(v) d\theta dv, \tag{7}$$

[25] Even if they cared about knowing θ_i, the private sector would still have an incentive to unbundle the score to meet the audience's demand.

[26] I do not know whether this reasoning is a driver for the lack of permanent license for the private credit evaluation systems in China, but it certainly is consistent with it. In any case, as Dai (2018) recognizes, there is today a private sector demand for unbundling in China: *"Blacklists such as that on judgment defaulters indeed could be of genuine interest to private sector players. But other lists, which proliferate nowadays, could be deemed as mostly noises. For example, compared with a red list of "honest and trustworthy" individuals and firms that government actors desire to praise and promote, the market likely would find it much more useful to have direct access to the transactional and behavioral records underlying such evaluation."*

[27] The benchmark social welfare function ($V = W + \gamma E[b_i]$) is silent on the government's preferences once actions have been selected. I presume quasi-lexicographic preferences in which ex post the government puts higher weight on low-θ agents and therefore allocates good reputations to those who have selected $b_i = 1$ rather than to those who have expressed dissent ($b_i = 0$).

where ε is arbitrarily small but positive, ξ is a strictly increasing function of θ (the government is hostile to opponents; for example $\xi(\theta) = \theta$), and $\hat{v}(v, \theta)$ is, by an abuse of notation, the equilibrium reputation of type (v, θ). The claim is that there exists an equilibrium in which (i) the ex-post rating depends only on the choice of b, (ii)) the rating is uninformative about v (and so, because $c \geq e$, all choose $a_i = 0$), and (iii) the choice of b is identical to that under unbundling.[28] Intuitively, in the absence of commitment, the agents' ratings reflects solely the government's empathy or animosity toward the agents and not their prosocial track record. It is therefore ignored by the agents' audience.

Proposition 5 *(time-consistent ratings). If the algorithm computing the social score is unobserved by agents and the government has (even a slight) distaste for opponents (as expressed in (7)), then there exists an equilibrium in which the social score is uninformative about v and the outcome in both dimensions (choice of a and b) is the same as in the absence of social score.*

4 Extensions and reinterpretations

(a) *Extension 1: Observable compliance* (b_i)

Suppose now that an individual's choice of b_i (but not that of a_i) is observed by the audience. The following observation, proved in the online Appendix, shows that the direct observability of one's compliance choice b_i has no impact on the equilibrium if no-one in the population enjoys toeing the line, and only a minor impact in the general case. Section 6 will assume that b_i is directly observable by prospective partners, and so this robustness result will also prove useful in this respect.

Observation. Suppose that b_i is observed by peers. The analysis is literally unchanged if $\theta_i \geq 0$ for all i (i.e. $\operatorname{supp} G = \mathbb{R}^+$). When the support of G includes negative values of b_i, the analysis is qualitatively unchanged, except that for θ_i negative but above some threshold, the individual chooses $b_i = 0$ when picking $a_i = 0$: Dissenting provides an excuse for the low social score.

The observation is straightforward when the support of G is $\mathbb{R}^+$: in the case of unobservable b_i studied so far, there were only two equilibrium behaviors, $a_i = b_i = 1$ and $a_i = b_i = 0$. Therefore, observing b_i contained no information that was not already in the social score. The equilibrium characterized in Proposition 4 is still an equilibrium.

(b) *Extension 2: Non-image sanctions*

[28] To show this, suppose that the government ignores a in its construction of its social score (so the rating $\hat{v}$ depends only on b), all agents choose a so as to maximize $(ve - c)a$, and b so as to maximize $b[-\theta + \nu\hat{v}_{b=1} - \nu\hat{v}_{b=0}]$. Then $b = 0$ if and only if θ lies beyond some threshold θ^*, while a is uninformative about θ. And so $\hat{v}_{b=1} = \hat{v}_{b=0} = \bar{v}$, and the threshold is $\theta^* = 0$. Let ξ_{ab} denote the expectation of $\xi(\theta)$ conditional on (a, b). Because ξ is a strictly increasing function of θ, $\xi_{01} = \xi_{11} < \xi_{00} = \xi_{10}$. Because $W + \gamma E[b_i]$ is sunk when the government picks ratings, the government picks the highest possible rating, $\hat{v}_{\max}$, when $b = 1$ and the lowest one, $\hat{v}_{\min}$, when $b = 0$. But then $\hat{v}_{\max} = \hat{v}_{\min} = \bar{v}$.

The focus on image concerns so far is justified by the fact that humans have evolved as a deeply social species, with the corollary of strong reputation concerns. The susceptibility of these image concerns to be exploited by governments, politicians, platforms and citizens has been demonstrated throughout history by widespread practices such as naming and shaming[29], the pillory, gossiping and social networks, ratings on platforms and today the social score. Humiliations, jail and other shame-inducing sanctions are generally viewed as more expressive sanctions than monetary fines by both lawmakers and victims who seek redress through acknowledgment of guilt rather than compensation. And guilt by association terrifies us precisely because we attach a high importance to our social relationships.

As I earlier discussed though, social scoring is not the only instrument at the disposal of autocratic governments. "Non-image or material sanctions" include economic sanctions, fines and jail (which are not available in the other applications of the model: religious and other organizations, platforms, and democracies with constitutional provisions against discrimination). And one might wonder whether social scoring would still be used once these alternative sanctions are factored in. Revealed preference (existence of social scores, public shaming and guilt by association) suggests that this is the case, but it is interesting to understand why and whether the key insights might be altered by the presence of alternative sanctions. Along the lines of Huxley's criticism of Orwell, I submit that the mix of sanctions will reflect cost-benefit considerations. The economic inefficiency of jail and the difficulty in levying fines under asymmetric information and risk aversion are no longer to be demonstrated, so the following discussion will focus on the "economic sanctions" (higher prices for some goods and services) that form the second pillar of the social score set up in China.

A couple of further remarks are in order. First, material sanctions often themselves leverage image concerns; this is most obvious in the case of jail. The inability of blacklisted citizens to travel first class in trains and airplanes in some Chinese pilots is as much a status as a comfort sanction. Second, economic sanctions in China piggyback on the same social score that is used for image sanctions (they are the B2C complement to C2C sanctions). Third, arbitrage restricts the set of economic sanctions to goods and services that are nominative and so not easily transferred: passport to travel abroad, transportation, hotel...

The online Appendix briefly studies non-image sanctions and shows that the characterization of the bundling case is very similar to that in the absence of economic penalties. Second, the deadweight loss (and for non-nominative goods the impossibility to discriminate) implies that in general image sanctions will be used even when economic sanctions are available. Third, for material sanctions not too large at least, compliance is higher under bundling, making the bundling strategy more attractive to more autocratic regimes.

[29] *"Ignominy is universally acknowledged to be a worse punishment than death"* (Benjamin Rush, "An Enquiry into the Effects of Public Punishments upon Criminals, and upon Society" Society for Promoting Political Enquiries, Convened at the House of Benjamin Franklin, Esq. In Philadelphia, March 9th, 1787.)

(c) *Reinterpretation 1: Divisive issues in a democratic society*

The same logic can be applied to a democracy in which a majority expresses a strong hostility towards certain minority opinions or behaviors (sexual orientation, abortion, politics, religion...). In this interpretation, $b_i = 0$ corresponds to (possibly secretly) practicing one's minority faith or politics, living according to one's majority-reproved sexual preferences, etc. Minority member i has a distaste $\theta_i > 0$ for kowtowing to the majority's preferences, potentially generating behavior $b_i = 0$ that is reproved by the majority. In the following, I will assume that whether an agent is part of the minority or the majority is common knowledge.

When the "ruler" is de facto a subclass of citizens (the majority), a number of modeling questions arise, such as: Do majority and minority agents interact (in which case bundling, by discouraging prosocial contributions, may exert negative externalities on the majority)? Do agents view externalities on in-group members as having the same value as externalities on out-group ones? Let us sidestep those issues by positing ghettoisation: majority members do not interact with minority members and are just concerned with the minority members toeing the line ($\gamma = +\infty$):

$$V = \max\{\gamma E_{\theta_i \geq 0}[b_i]\}.$$

Minority members are characterized by their prosocial type v_i and the intensity of their identity $\theta_i > 0$.

Suppose that minority member i has image concerns $\nu\hat{v}_i$ (weak-ties society). Assuming again that the prosocial and identity types v_i and θ_i are independent and that image concerns are necessary to generate prosocial behavior ($c \geq e$), the minority member chooses $a_i = b_i = 1$ (reputation $\hat{v}_1$) over $a_i = b_i = 0$ (reputation $\hat{v}_0$) if and only if

$$v_i e - c - \theta_i + \nu(\hat{v}_1 - \hat{v}_0) \geq 0.$$

The analysis is identical with that in Section 3.2. Because the majority is assumed to care only about the minority's toeing the line and not to interact with it, it bundles for all $\gamma > 0$. More generally, if the majority puts some weight on the minority's welfare or bears some of the cost from the reduction in prosocial behavior, bundling occurs for γ above some threshold $\gamma^* > 0$.

Observation (divisive issues). The insights of Section 3 also apply to democracies in which a political majority disapproves of a minority's behavior or expression of opinion.

(d) *Reinterpretation 2: Corporate political clout and the subversion of democracy*

While autocratic countries should be wary of public platforms, democratic ones may, to the contrary, be concerned with private ones. This can be shown by using a framework that is a relabeling of the one of Section 3: Instead of the platform rating citizens, it "rates" officials in government. Concretely, such ratings may take the form of selective disclosure of facts or opinions about politicians, that change the electorate's beliefs about

the quality or the congruence of these politicians. To envision how this might work, the reader may have in mind that the platform can disclose only a subset (or none) of the actions undertaken by the official to the benefit of the community.[30]

There is one private platform –or equivalently an arbitrary number of private platforms controling access to "unique viewers".[31] The platform's viewers are also voters.

Official i selects two actions: First, $a_i \in \{0,1\}$ is an action affecting, perhaps with a lag, the welfare of citizens; $a_i = 1$ adds e to their welfare. The official's intrinsic motivation for picking $a_i = 1$ is $v_i e - c$. The official also cares about her reputation vis-à-vis the electorate, $\hat{v}_i$, as construed by the platform. Let $\nu\hat{v}_i$ denote this component of the official's utility, where ν here captures her re-election concerns.[32]

The official can also grant favors to the platform ($b_i = 1$) or not ($b_i = 0$). Such favors may include refraining from asking for tougher antitrust enforcement or tax collection, subsidising the media, relaxing online media's editorial responsibility, etc. Politician i has distaste $\theta_i \geq 0$, distributed according to $G(\theta_i)$, for kowtowing to the platform. For simplicity, let us assume that the citizens do not care about the value of θ_i. The platform reports good news about the politician (who then has reputation $\hat{v}_1$) if and only if $a_i = b_i = 1$.

To complete the perfect isomorphism with the model of Section 3, let the platform's utility be an increasing function of $E[b_i]$ and possibly incorporate elements of its customers' utility W for attractiveness reasons.[33]

Observation (private platforms' political clout). Private platforms can bundle information about elected officials so as to obtain favors from them, in the same way a state-controlled platform can leverage the social score to suppress citizen's dissent.[34]

Reinterpretation 3: incentivizing public good provision

[30] Conversely, the platform could disclose embarrassing details about the official (private conversation, browsing history, personal lifestyle, stance on divisive issues...). The modeling of such "negative disclosures" differs slightly from that of the concealment of "positive actions", but again such reports can be combined with bundling to induce official's compliance.

[31] What matters is not the platform's market share per se. Rather, it is the possibility that viewers do not receive disconfirming news from elsewhere.

[32] Thus ν reflects the benefits from reelection. The implicit assumption here is that a better reputation for public service increases the probability of reelection (here in a linear way, as obtains in a standard Hotelling differentiation model augmented with vertical-reputation attributes). One could also add voters who are well-aware of the official's policy track record; those would be the counterpart of strong ties in Section 3 (and would correspond to intensity μ of image concerns).

[33] To see the correspondence between selective release of information and the report of a $\hat{v}_i$, suppose that the platform fails to report good actions by the official either when the later picks $b_i = 0$, or when $a_i = 0$ (or both). This reporting indeed leads to a binary rating.

[34] A literature (reviewed in Prat 2018) analyzes how media owners can manipulate news through selective disclosure for their own goals. Prat (2018) characterizes the maximal influence of such a media owner when Bayesian voters a) have subjective beliefs on the probability that media are biased, and b) have a bounded capacity to absorb information from the various sources they follow. Although related through the theme of selective disclosure and manipulation of the democratic process, none of the papers in this vast literature to the best of my knowledge studies bundling as a strategy to discipline politicians.

So far, I have set the state up as the bad guy. The same model can be reinterpreted with the state as the good guy trying to incentivize contributions to the public good. Agents are motivated by appearing empathic or loyal to the in-group they mingle with, but do not care about being seen as good overall citizens. This reinterpretation thus involves a narrow altruism (agents may be loyal to the in-group but have little empathy for society as a whole).

Formally, suppose that $a_i = 1$ corresponds to agent i being nice to her in-group partner while $b_i = 1$ implies a contribution to a global public good, thereby exerting a positive externality γ on the rest of society. As before, each agent has two types; v_i is her empathy for in-group partners; $\theta_i \in \mathbb{R}^+$ is now her cost of contributing to the public good. The individual cares only about her reputation for in-group loyalty ($\hat{v}_i$) but not about appearing to be a good citizen (more generally, loyalty to the in-group looms larger than appearing concerned about one's use of public funds or contribution to global warming). The state's objective function is still given by equation (5), so only the interpretation differs: the state is more, rather than less benevolent than the agents.

There is no contribution to the public good ($E[b_i] = 0$) under unbundling (as $\theta_i \in \mathbb{R}^+$). Proposition 4 characterizes the outcome under bundling. Bundling leads to less in-group solidarity (i.e. in-group prosociality) but generates some supply of the public good. The government chooses to bundle if and only if the public good externality exceeds the threshold given in Proposition 4 ($\gamma \geq \gamma^*$). The benevolent government's choice between bundling and unbundling necessarily maximizes agents' welfare.

5 Linear-quadratic Gaussian model

I now study the version of the model in which preferences are linear-quadratic and the distribution of the two types (prosociality v_i, taste for (non) compliance θ_i) is Gaussian. The continuous version of the model will show that some amount of bundling is always desirable for the government. More importantly, it will allow us to study the impact of the heterogeneity of types (does equilibrium expected compliance increase or decrease with the heterogeneity of prosociality and that of the taste for compliance?) and that of the correlation between the two types.

As discussed in the introduction, a correlation between types v_i and θ_i (for example, political compliance to an autocracy may signal non-prosociality; conversely, supporting a cause promoted by the state and perceived as good by peers is in itself a good signal about v_i, over and beyond what the individual does for the cause) may change the weight put on compliance and alter the state's ability to profit from bundling. To identify these

impacts I explore a linear-quadratic, Gaussian version of the model:[35]

$$\begin{pmatrix} v_i \\ \theta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{v} \\ \bar{\theta} \end{pmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_\theta \\ \rho\sigma_v\sigma_\theta & \sigma_\theta^2 \end{bmatrix} \right)$$

where $\bar{v}$ and $\bar{\theta}$ are the prior means, $\sigma_\cdot^2$ are the variances and $\rho \in [-1, +1]$ measures the correlation between the two variables. The actions a and b are now continuous choices in $\mathbb{R}$. The agent's utility function writes[36]

$$u_i = \left[v_i a - \theta_i b - \left(\frac{a^2 + b^2}{2}\right)\right] + \nu\hat{v}(s),$$

where the quadratic term stands for the cost of selecting actions, $\nu > 0$ and $\hat{v}(s)$ is the posterior belief upon the disclosure of social score

$$s = \alpha a + \beta b,$$

where only the ratio α/β matters for outcomes. The individual's audience learns only her social score. Under unbundling (a and b are disclosed separately), the mean choices are $E[a] = \bar{v} + \nu$ and $E[b] = -\bar{\theta}$. Bundling occurs when only s is disclosed and $\beta \neq 0$ (when $\beta = 0$, the outcome is the same as under unbundling).

The following results are derived in the Appendix:

Proposition 6 *(heterogeneity and bundling) Assume a linear-quadratic Gaussian model with social score $s = \alpha a + \beta b$, and that the state's objective is to maximize compliance $E[b_i]$.*

(i) The state generically benefits from bundling ($\beta \neq 0$).

(ii) Compliance $E[b_i]$

- *grows with the agent's concern about appearing prosocial (with ν)*

[35] Linear-quadratic Gaussian models of signaling have a long tradition: See, e.g., Prendergast-Topel (1996), Fischer-Verrecchia (2000), Bénabou-Tirole (2006) or more recently Bergemann et al (2020), Frankel-Kartik (2019b) and Ball (2020). In the signaling game studied in Frankel-Kartik for instance, the sender also has a two-dimensional type, a "natural action" (say, an innate attractiveness or ability to reimburse loans) and a "gaming ability", and wants to influence the receiver's perception of, say, her natural action. Under an assumption on how the marginal signaling cost varies with the two types and the intensity of signaling, they characterize the informativeness of equilibrium signals, using the weak set-order to reflect equilibrium multiplicity. When stakes (image concerns in this paper's framework) increase, the equilibrium is less informative; indeed the equilibrium is almost uninformative for high stakes. So, for instance, adding new observers generates negative externalities on observers who already had access to the signal.

[36] Appendix C further relaxes the assumption that the agent cares only about her reputation along the v_i dimension, and not about the θ_i one. The payoff function is then

$$u_i = \left[v_i a - \theta_i b - \left(\frac{a^2 + b^2}{2}\right)\right] + \nu\hat{v}(s) + \xi\hat{\theta}(s).$$

- *increases (decreases) with the heterogeneity in prosociality (in the taste for compliance).*

Proposition 7 *(impact of correlation) Assume a linear-quadratic Gaussian model with social score $s = \alpha a + \beta b$, and that the state's objective is to maximize compliance $E[b_i]$.*

(i) Compliance $E[b_i]$ increases with the absolute correlation ($|\rho|$).

(ii) The compliance-maximizing weight β put on b, normalizing $\alpha = 1$, is negative for $\rho \in [-1, -\frac{1}{\sqrt{2}})$ and positive otherwise.

(iii) Prosocial behavior $E[a_i]$ strictly increases with ρ from $-\infty$ for $\rho = -1$ to $+\infty$ for $\rho = +1$.

(iv) Prosociality and welfare. Assume prosociality is suboptimal under unbundling ($e > \nu$). A sufficient condition for unbundling to increase $E[a_i]$ and a fortiori welfare ($W^u > W^b$) is that $\rho < \frac{1}{\sqrt{2}}$ or that ρ is close to 1 (furthermore, $W^u > W^b$ for all ρ when $e < \bar{e}$ for some $\bar{e} > \nu$).

To grasp the intuition for the effect of taste heterogeneity, let us consider the special case of no correlation ($\rho = 0$). Then, at the compliance-maximizing social score:

$$\frac{\beta}{\alpha} = \frac{\sigma_v}{\sigma_\theta} \text{ and } E[a_i] = \bar{v} + \frac{\nu}{2} \text{ and } E[b_i] = -\bar{\theta} + \frac{\nu}{2}\frac{\sigma_v}{\sigma_\theta}. \tag{8}$$

First, note the signal-jamming effect of bundling: The score s is a noisier measure of a when β increases from 0. The mean level of prosociality becomes $\bar{v} + \nu/2 < \bar{v} + \nu$. Second, consider the impact of β on incentives. When there is a lot of heterogeneity along the θ dimension, the social score becomes a very noisy signal of v_i, making the individual less concerned about her social score; increasing β would then be counterproductive as this would make the individual even less concerned about the social score and further deprive the state of its leverage on the b-behavior. Despite this counter-adjustment, the state cannot induce much compliance if σ_θ is large. Conversely, when individuals differ substantially in their prosociality, signaling concerns are important, and the state can take advantage of them by raising β. The gains from bundling are then high.

Next, consider the impact of correlation. For an arbitrary $\rho \in [-1, +1]$, expressions in (8) generalize to:

$$\frac{\beta}{\alpha} = \frac{\sigma_v}{\sigma_\theta}\Big[\frac{1}{\rho + \sqrt{1-\rho^2}}\Big] \text{ and } E[a_i] = \bar{v} + \frac{\nu}{2}\Big[1 + \frac{\rho}{\sqrt{1-\rho^2}}\Big] \text{ and } E[b_i] = -\bar{\theta} + \frac{\nu}{2}\frac{\sigma_v}{\sigma_\theta}\Big(\frac{1}{\sqrt{1-\rho^2}}\Big). \tag{9}$$

The weight β (normalizing $\alpha = 1$) again reflects the relative taste heterogeneity and the intensity of signaling concerns. It is negative for correlation $\rho \in \big[-1, -\frac{1}{\sqrt{2}}\big)$ and positive for correlation $\rho \in \big(-\frac{1}{\sqrt{2}}, +1\big]$. It is determined by two forces.

First, (normalizing $\alpha = 1$) a positive β directly incentivizes compliance (i.e. increases b), provided a high signal is still interpreted as signaling prosociality. Second, the choice

of β also affects the audience's signal extraction problem. Ceteris paribus, the state wants to "homogenize" social scores so that a small increase in b reveals a lot about v. When $\rho > 0$, i.e. when prosociality correlates positively with dissent propensity, the dispersion in the score is reduced by keeping $\beta > 0$. By contrast, when $\rho < 0$, i.e. when prosociality correlates negatively with dissent propensity, the homogenization of social scores suggests picking $\beta < 0$, while the direct incentivization calls for $\beta > 0$. For high negative correlation and normalizing $\sigma_v = \sigma_\theta$ to set aside the relative heterogeneity effect, the state can induce extremely high levels of compliance by selecting $\beta = -(1+\varepsilon)$ with ε positive and small. Then, in the absence of signaling concerns, the scores of all agents are almost equal, creating a tough competition to build a reputation for prosociality: For $\theta = v$, $s = -\varepsilon v + \text{constant}$, and so $d\hat{v}/ds = -1/\varepsilon$ and the marginal reputation gain when raising b is equal to $\nu(1+\varepsilon)/\varepsilon$. Symmetrically, for almost perfect positive correlation, then the optimum is to choose $\beta = 1 - \varepsilon$ for ε positive and small, yielding $s = \varepsilon v + \text{constant}$ and a high sensitivity of updating to the score.

The state's use of the correlation thus creates a "rat race" for reputation and always raises compliance above its no-correlation bundling level. As for welfare, bundling always creates an upward distortion in the individual choice of b. For no or negative correlation, bundling further reduces prosociality. For positive correlation, bundling has a positive impact on prosociality; the comparison between the prosocial benefits and the distortion of behavior along the compliance dimension is summarized in part (iv) of the Proposition 7.

6 Guilt by association: Leveraging the social graph

One of the most problematic aspects of mass surveillance is the coloring of a person's perception by the company she keeps. Guilt by association has historically done substantial harm to the social fabric under totalitarian regimes, as people are afraid of being seen in company of dissidents or mere citizens whose lifestyle is frowned upon by the regime.[37] Face recognition and artificial intelligence applied to surveilled communications and social network activities today substantially reduce the state's cost of drawing an accurate social graph of relationships among its citizens.

States can make use of social graphs by allowing relationships with someone on a blacklist to taint the reputation of those who a priori would not be. Such tainting can

[37]Paul Seabright in *The Company of Strangers* argues that institutions such as markets, cities, money and the banking system allowed the enlargement of the circle of trust well beyond kinship or a very small tribe. He studies how humans developed the ability to trust strangers to meet their most basic needs. In contrast, with very rudimentary means, the Stasi managed to break the social fabric of the GDR and reverse the historical evolution: Friends, colleagues, family, even spouses and children were no longer part of the individual's circle of trust. Today some servers and artificial intelligence suffice to accomplish this task. Accordingly, Russell (2019) coined the expression "automated Stasi".

induce yet another social pressure -ostracism- on citizens to toe the line.[38] To see how this can work, consider the following, *sequential* choice variant of the model of Section 3, and depicted in Figure 2. Because agent i's social graph will be endogenous, I assume that she has a set of potential partners J_i and that she chooses a subset of J_i to interact with. I will further assume that $i \in [0,1]$ and that $J_i = [0,1]$, so as to shorten the exposition. The timing is sequential. First, agents choose their b action, which is observable by other agents. Then agents choose with whom they match (it takes two to tango)[39] and then select actions $a_{ij} \in \{0,1\}$ for the partners j they have selected (and been selected by). Action $a_{ij} = 1$ exerts a positive externality e on agent j.

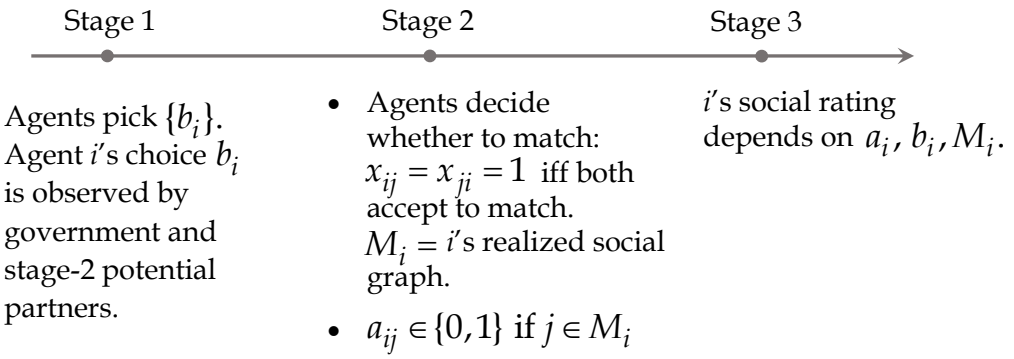

Figure 2: timing (guilt by association)

(1) At "stage 1", agents make their compliance choices $\{b_i\}$. Agent i's choice is observed by the state as well as the other agents whom she will potentially interact with at stage 2 (but not by her stage-3 audience).[40]

(2) At "stage 2", each pair of potentially matched agents i, j decides whether to actually match. A match is formed if and only if both consent to it. Let $x_{ij} = x_{ji} = 1$ if the i-j match is realized and $x_{ij} = x_{ji} = 0$ otherwise.[41] Let M_i denote individual i's realized (as opposed to potential) matching set or social graph. If matched, they

[38] While I stress ostracism between citizens, I later note that the same insights also apply to B2C. They also apply to B2B relationships, a relevant feature for the Chinese corporate social credit system (see "China to impose "social credit" system on foreign companies", *Financial Times*, August 27, 2019): A foreign company has been warned that its partner's rating by customs authorities would affect its rating; similarly, foreign companies that are perceived to run counter the government's views on politically sensitive issues may in the future be blacklisted and therefore ostracized by domestic business partners. Note also that we focus on the government's use of the social graph. Private platforms of course may also consider such use. For instance, in 2012 Facebook obtained a patent for a method of credit assessment that could reflect the credit scores of people in the individual's social network. An individual's Zhima credit score already embodies the scores of their friends.

[39] This decision is a no-brainer if, as in Section 3, scores do not depend on the social graph: The agents benefit from the relationships and accept them all (out-of-equilibrium beliefs specify that someone who turns down relationships are deemed a-social: $\hat{v}_i = \hat{v}_0$).

[40] Allowing b_i to be observed by the stage-3 audience does not alter the insights.

[41] Let us rule out weakly dominated strategies in which a party refuses a match only because she expects that the other party will refuse as well. So a match forms if both so desire.

pick actions a_{ij} and a_{ji}. I will focus on equilibria in which $a_{ij} = a_i$ is the same for all matched partners $j \in M_i$ (recall that v_i is a parameter of overall prosociality). The government observes the actual matches and the actions.[42]

(3) The government issues a binary social rating for each individual i ($s_i \in \{0, 1\}$) on the basis of her action $\{a_i, b_i\}$ as well as her social graph M_i: Agent i is put on the blacklist ($s_i = 0$, inducing reputation $\hat{v}_i = \hat{v}_0$ with weak ties) if

- either she picked $a_i = 0$ in her realized relationships (a-social behavior)
- or $b_i = 0$ (dissent)
- or else $a_i = b_i = 1$, but there exists $j \in M_i$ such that $b_j = 0$ (tainting).[43]

Agent i receives score $s_i = 1$, inducing reputation $\hat{v}_1$ with weak ties, otherwise.

This form of social scoring captures in the starkest form the idea of social graph tainting: The individual's social relations contaminate her social score. I will label this policy "social-graph-inclusive bundling" or "all-inclusive bundling", as opposed to the "simple bundling" and "unbundling" policies of Sections 3 and 4.

As earlier, agent i's prosociality and cost of compliance are denoted v_i and θ_i; the two parameters are independent, for simplicity. The payoff function of individual i is

$$u_i = \int_j \left[x_{ij}[(v_i e - c)a_{ij} + e a_{ji} + \mathfrak{b}] + \mu \hat{v}_i(I_{ij})\right] dj - \theta_i b_i + \nu \hat{v}_i(I_i),$$

where $\mathfrak{b} > 0$ is a fixed benefit per interaction.[44] The weak-ties term $\nu \hat{v}_i(I_i)$ reflects the image concerns vis-à-vis new partners tomorrow, where $I_i = s_i$ is the public-information binary social score described at stage (3) of the timing. The strong-tie term $\mu \hat{v}_i(I_{ij})$ stands for the image concerns vis-à-vis of strong-tie partner j, who observes $I_{ij} = \{a_{ij}, s_i\}$.[45]

[42] Either directly (AI, facial recognition) or indirectly through ratings. It does not matter whether the government observes all actions or a subset of actions.

[43] Alternatively, one could allow tainting to be "viral" by defining the "extended" (or "direct and indirect") matching set or social graph $\hat{M}_i$ as being the set of individuals with whom i is matched directly or indirectly:

$$\hat{M}_i = \left\{j | \exists \{k_1, ..., k_n\} \text{ st } k_1 = i \text{ and } k_n = j, \text{ and } x_{k_m, k_{m+1}} = 1 \text{ for all } m \in \{1, n-1\}\right\}.$$

The decision problem of an individual who can be tainted directly or indirectly is particularly complex under the simultaneity assumption, as it requires anticipating others' matching choices.

[44] This fixed benefit had not yet been introduced, as it plays no role unless the number of an individual's relationships is endogenous. The term $\mathfrak{b}$ will capture the loss of social well-being when relationships are severed. I am agnostic as to the specific form this loss may take. Besides the obvious interpretation as a forfeiture of rewarding human relationships, it may capture the social cost associated with the emergence of yet another form of tribalism (to use an expression due to Jonathan Haidt), this one based on differences in social status attached to the social score.

[45] Agent j also observes whether i accepted to interact with her. This information actually is redundant in the equilibrium studied below. Anticipating on the latter, the choice of b_i together with the social score s_i reveal whether i is a model citizen, a dissenter or a complier, so knowledge of the matching behavior conveys no new information on the equilibrium path. One can assume for example that agent j does not change beliefs off the equilibrium path when player i accepts a match that she was not supposed to, or conversely. That is, strong-tie beliefs depend only on I_{ij}.

I assume that $\theta_i > 0$ for almost all i ($G(0) = 0$) and that $G(\theta) > 0$ for all $\theta > 0$. This ensures that all individuals who behave a-socially ($a_i = 0$) also dissent ($b_i = 0$). As in Section 3, I further assume that $c \geq e$ for expositional simplicity. This assumption guarantees that an individual without image concerns will not choose $a_i = 1$. This will *de facto* include dissenters and their matched relationships. Thus, if either $b_i = 0$ or there exists $j \in M_i$ such that $b_j = 0$ or both, and so $\hat{v}_i = \hat{v}_0$, then $a_i \equiv 0$.

Strong ties

Proposition 8 *(ineffectiveness of guilt by association in a strong-ties society) In a strong-ties society ($\mu > 0 = \nu$), there exists an equilibrium in which all relationships form, all agents dissent ($E[b_i] = 0$) and behavior is the unbundling behavior of Proposition 2: $a_{ij} = 1$ iff $v_i \geq v^u$, where $v^u e - c + \mu\Delta(v^u) = 0$.*

As was the case for Proposition 3, the proof is straightforward. When assessing the prosociality of their partner, an agent j bases her judgment solely on agent i's selected action a_{ij}. This, together with the fact that there is no equilibrium ostracism (all relationships form), implies that all agents dissent ($b_i = 0$). Agent i selects $a_{ij} = 1$ iff $v_i \geq v^u$, where $v^u e - c + \mu\Delta(v^u) = 0$.

Weak ties

Let X denote the fraction of agents who pick $b_i = 1$. A fraction X_1 pick $\{a_i = b_i = 1\}$ and a fraction X_0 pick $\{a_i = 0,\ b_i = 1\}$. So $X = X_1 + X_0$. All individuals using a strategy leading to reputation $\hat{v}_0$ are willing to match with everybody. By contrast, those choosing $a_i = b_i = 1$ do not want to be tainted by partners having chosen $b_j = 0$.

Thus individual i with type (v_i, θ_i) really has only three choices, depicted in Figure 3.

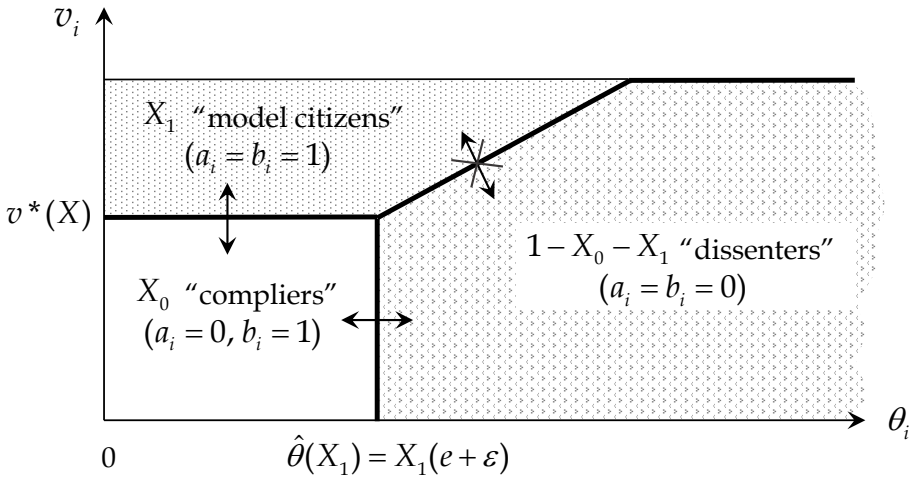

Figure 3: behavior under guilt by association

(1) *Dissenters* pick $b_i = 0$, accept getting a low rating $\hat{v}_i = \hat{v}_0$, match with all potential partners who accept to match with them, and select $a_i \equiv 0$. This strategy yields

payoff

$$u_i^1 = (1 - X_1)\mathfrak{b} + \nu\hat{v}_0$$

(2) *Model citizens* pick $b_i = 1$, go for the high rating $\hat{v}_i = \hat{v}_1$, match only with individuals who have picked $b_j \equiv 1$, and then select $a_i \equiv 1$. This strategy yields:

$$u_i^2 = (X_0 + X_1)(v_i e - c + \mathfrak{b}) + X_1 e - \theta_i + \nu\hat{v}_1.$$

(3) *Compliers* pick $b_i = 1$, match with every one, select $a_i \equiv 0$ and obtain the low rating $\hat{v}_0$. This strategy yields

$$u_i^3 = \mathfrak{b} + X_1 e - \theta_i + \nu\hat{v}_0.$$

Proposition 9 *(guild by association with weak ties) Assume that the social score includes the individual's social graph.*

(i) Guilt by association makes high-score agents ostracize dissenters.

(ii) Social-graph-inclusive bundling becomes more attractive relative to unbundling as the ruler becomes more autocratic. By contrast, it is a priori unclear whether the attractiveness of social-graph-inclusive bundling relative to simple bundling increases with autocratic proclivity, although it does so in Example 2 below.

Example 2 (impact of guilt by association). Example 2 adds guilt by association to Example 1. Suppose that G puts weight only on two types θ_L (probability κ) and θ_H (probability $1-\kappa$). Identity θ_H is strong enough that the individual always picks $a_i = b_i = 0$ regardless of v_i. By contrast, let us look for an interior cutoff for type θ_L. Let Δ^b denote the image gain from $a_i = b_i = 1$ under bundling, but no tainting. The cutoff v^b is given by

$$v^b e - c + \nu\Delta^b(\kappa, v^b) = \theta_L,$$

where

$$\Delta^b(\kappa, v^b) \equiv M^+(v^b) - \left[\frac{\kappa F(v^b)M^-(v^b) + (1-\kappa)\bar{v}}{\kappa F(v^b) + 1 - \kappa}\right].$$

Note that $\Delta^b(\kappa, v^b) < \Delta^b(1, v^b) = \Delta(v^b)$ for all v^b: The presence of strong dissenters (who dissent regardless of their prosociality, in proportion $1-\kappa$) serves as an excuse for the absence of prosocial behavior (as $\bar{v} > M^-(v^b)$). When tainting is added to bundling and type θ_L always chooses to comply ($b_i = 1$ for all v_i), the cutoff v^{ga} (where "ga" stands for "guilt by association") is given by

$$\kappa(v^{ga}e - c) + \nu\Delta^b(\kappa, v^{ga}) = 0$$

provided that complying dominates dissenting: $\kappa[\mathfrak{b} + [1 - F(v^{ga})]e] - \theta_L + \nu\hat{v}_0 \geq (1 - \kappa)\mathfrak{b} + \nu\hat{v}_0$, which is the case if the number $(1-\kappa)$ of dissenters is small and θ_L is small enough as well:

$$(2\kappa - 1)\mathfrak{b} + \kappa[1 - F(v^{ga})]e \geq \theta_L.$$

Tainting improves relationships that survive as it reduces the total cost of prosocial behavior ($v^{ga} < v^b$), but destroys a number $\kappa[1 - F(v^{ga})](1 - \kappa)$ of relationships. It also raises $E[b_i]$ from $\kappa[1 - F(v^b)]$ to κ. It thus appeals more to a more autocratic ruler.

Example 3 (multiplicity). Let us return to the continuum-of-types case with θ distributed on $[0, \theta_{\max}]$, and look at two simple equilibria, with an amorphous population ($X = 1$) and an all-dissenter one ($X = 0$) respectively.

Suppose, first, that

$$\theta_{\max} \leq [1 - F(v^u)](e + \mathfrak{b}) \tag{10}$$

where $v^u e - c + \nu\Delta(v^u) \equiv 0$. I claim that there exists an equilibrium in which no one dissents ($b_i = 1$ for all (v_i, θ_i)) and the individual behaves prosocially ($a_i = 1$) if and only if $v_i \geq v^u$. The individual receives reputation $\hat{v}_0 = M^-(v^u)$ in the off-path event in which $b_i = 0$, regardless of a_i. Condition (10) guarantees that the individual does not gain from dissenting and thereby being ostracized by model citizens.[46]

Second, consider an equilibrium in which $X_1 = X_0 = 0$. That is, $a_i = b_i = 0$ for all (v_i, θ_i), implying that the policy completely backfires in terms of both prosocial behavior and compliance. The individual obtains utility $\mathfrak{b} + \nu\bar{v}$ from her equilibrium behavior ($\hat{v}_0 = \bar{v}$). If she picks ($a_i = 0, b_i = 1$), her utility is lower for all θ, as already noted: $\mathfrak{b} - \theta + \nu\hat{v}_0 = \mathfrak{b} - \theta + \nu\bar{v} < \mathfrak{b} + \nu\bar{v}$. Picking ($a_i = 1, b_i = 0$) yields $v_i e - c + \mathfrak{b} + \nu\hat{v}_0 < \mathfrak{b} + \nu\bar{v}$. Finally, obtaining reputation $\hat{v}_1$ requires isolation[47] and yields at most $\nu\hat{v}_1 \leq \nu \cdot 1$. So if

$$\mathfrak{b} \geq \nu(1 - \bar{v}),$$

all dissent. This equilibrium illustrates the possibility of multiple equilibria due to endogenous network externalities, as its condition of existence is compatible with that, (10), of the amorphous equilibrium.

Can we Pareto-rank these two equilibria? In the all-dissent equilibrium, all individuals receive utility $\mathfrak{b} + \nu\bar{v}$. In the amorphous equilibrium, the individual's utility is

$$\mathfrak{b} - \theta_i + \max\{v_i e - c + \nu M^+(v^u), \nu M^-(v^u)\}.$$

It is smaller than in the all-dissent equilibrium for types who choose $a_i = 0$. But for e close to c, types $\{\theta_i \simeq 0, v_i \simeq 1\}$ are better off than in the all-dissent equilibrium (they have an opportunity to signal their proclivity to do good). So, in general, one cannot select between the two equilibria by using Pareto comparisons.

Because $E[b_i] = 0$ under unbundling, there is trivially at least as much compliance with the state's desires ($E[b_i] \geq 0$) under social-graph-inclusive bundling. However, the

[46] For $v_i \leq v^u$,

$$[1 - F(v^u)]e + \mathfrak{b} + \nu M^-(v^u) - \theta_i \geq F(v^u)\mathfrak{b} + \nu M^-(v^u)$$

for all $\theta \leq \theta_{\max}$. And similarly for $v_i \geq v^u$: $ve - c + \nu M^+(v^u) + [e[1 - F(v^u)] + \mathfrak{b}] \geq ve - c + \nu\hat{v}_0 + \mathfrak{b}F(v^u) + \theta_i$, which gives a weaker condition.

[47] As well as ($a_i = b_i = 1$); for the latter to make sense, though, one needs to assume that there is a very small fraction X who actually choose $b_i = 1$.

comparison with simple bundling hinges on the choice of equilibrium under social-graph-inclusive bundling: While $0 < E[b_i] < 1$ under simple bundling, $E[b_i] = 0$ in the all-dissent equilibrium and $E[b_i] = 1$ in the amorphous one.

An interesting extension of this ostracism model would study the dynamics of social networks. One would expect that the set of potential (and not only realized) matching partners would morph over time into something different. This reconfiguration might take the form of a ghettoisation, with the marginalized dissenters regrouping in ostracized communities.

Application to religious and other organizations. My emphasis on Huxleyian soft control suggests that the strategies discussed here may apply to organizations that have no coercive power: professional orders, religious authorities, associations... Such organizations often create discipline through the exclusion of recalcitrant members. To be certain, delegation of state powers sometimes implies that exclusion is accompanied with legal enforcement (a physician kicked out by the medical order can no longer practice). But most often the damage is reputational: Remaining members spontaneously express suspicion vis-à-vis the excluded member and those who do not and keep interacting with the excluded member may themselves be socially excluded or at least ostracized (guilt by association). These organizations' monopoly on membership and information is used to bundle ratings of social proclivities with behaviours that are unrelated (donating money or time to work for the organization, not challenging authorities) but valuable for the organization's executives.

Interestingly, most religions practice excommunication. Excommunicated members occasionally go to courts, arguing that excommunication has made them lose their social environment.[48] There is a strong pressure on members, including spouse and children, to stop socializing with an excommunicated member.[49] Accordingly, excommunication ("a low score") combined with guilt by association has been and still is a major disciplining tool used by religious orders. And crucially, excommunication can stem from a mix of reasons that weak ties (although not strong ones) cannot disentangle; these reasons range from ordinary prosociality (excess drinking, child abuse...) to dissent with the authorities regarding what should be allowed by the religion. In an amusing case heard by Pennsylvania's supreme court (*Bear v. Reformed Mennonite Church*), the plaintiff

[48] See Thiels (2009) for an extensive review of US jurisprudence with regards to excommunication.

[49] For example, Jehovah's Witnesses' long list of serious sins includes "brazen conduct", including association with disfellowshipped non-relatives, "spiritual" association with disfellowshipped relatives, or criticizing a disfellowshipping decision.
Calvin's Genevan consistory summoned and formally rebuked, and then possibly excommunicated Genevans whose social behavior and compliance behavior (dis-respectfulness in church, bearing traces to Roman Catholicism, blasphemy) it did not appreciate. Social sanctions were considered highly effective: *"The Consistory gave [the citizen] one of the harshest punishments at its disposal: it barred him from attendance at the Lord's Supper... The leader of Geneva's Consistory [Calvin] taught that excommunication, imposed as a judgment on behalf of the collective, was the most appropriate form of moral discipline not only because civil justice often failed but also because all people should be conscious of their membership in the body social."* (Valeri 1997).

had been excommunicated because he had criticized the authorities'... excommunication policy.

My analysis shows that such policies are most effective for the religious authorities when the latter bundle prosocial and compliance behaviors in their excommunication decision (and "excommunication" is what weak links are informed of), and when challenges to the authorities remain rather isolated events.

Mixture of weak and strong ties

Like bundling, guilt by association is much more powerful in weak- than in strong-ties societies. The conclusion that strong ties cannot be undone through guilt by association however is too extreme. Indeed, in the more realistic case in which there are both weak and strong ties, model citizens may ostracize dissenters with whom they form a strong tie.[50]

7 Implications and alleys for future research

Social scores have the potential to enhance trust in society; indeed, they have already promoted better behavior on e-commerce and ride-hailing platforms around the world, and slower and more careful driving in some Chinese cities; besides, many countries have long had a credit rating system that financial institutions can use to ward off bad borrowers, and big data analytics have enabled a more inclusive access to funding for Chinese SMEs. But, as we saw, the private interest of those who design such scores may make them socially dysfunctional. A key challenge for our digital society will be to come up with principle-based policy frameworks that discipline governments and private platforms in their integration and disclosure of data about individuals. The exact contours of such disciplined principles are yet to be identified, but the analysis in this paper suggests leaving out information about divisive issues- in particular those from which the government, a majority or a platform could derive gains from-, and about the social graph. It also suggests monitoring platforms' foray into political coverage unless platform regulation is performed by one or several entirely independent agencies.

The paper's main insights were summarized in the introduction. In these concluding remarks, I therefore focus on alleys for future research. The paper indeed is only a first step and leaves open many important and exciting questions, of which I list a few here.

[50] A simple illustration is supplied by Example 2 above. Start from the case in which there are only weak ties: θ_L types comply ($b_i = 1$) and θ_H types do not ($b_i = 0$); and model citizens ostracize dissenters. Now introduce a small fraction of strong ties on top of the existing weak ties, so that equilibrium behavior hardly changes; for each agent, some of the strong ties will be dissenters and some will choose $b_i = 1$ (in the absence of assortative matching for strong ties, proportions of these are κ and $1 - \kappa$, but this assumption is not needed). Model citizens still ostracize dissenters, whether they are weak- or strong ties.

One direction concerns dynamics. I already mentioned the possibility that guilt by association could over time drive a reconfiguration of social networks and lead to ghettoisation, the marginalized dissenters regrouping in ostracized communities. Another extension would focus on institutional ratcheting and diminishing prospects for political transitions; social score design might over time drift toward the promotion of stricter political control along the lines described in this paper, making it harder and harder for a democratic opposition to organize and leaving upheavals that originate inside the regime as the main source of regime switch.

Another extension would add behavioral elements, with again several potential objects of study. I mentioned that an Online Travel Agency may want to use the bundling strategy described in the paper (and sometimes does). While merchants are well-aware that this is happening, consumers may not be and therefore may not discount appropriately the OTA's recommendations. Behavioral neglect would amplify the platform's benefits from bundling. Similarly, citizens- at least those who toe the line- may not fully comprehend the exact meaning of social scores and overestimate the extent to which they reflect prosocial behavior, especially in the early phase of social scoring. Second, while the paper attributed flaws in social scoring to the rulers' self-interest, rulers may also make unintentional mistakes in its design. Such mistakes are made more likely by the many ill-understood design issues, which I now turn to.

One design challenge concerns the weights to be put on behaviors we deem worthy of inclusion into such a score (imagine a ruler who is more preoccupied with jaywalking than with corruption),[51] and how to account for the imperfect reliability of ratings or more generally observability of individual behaviors. Rating subjectivity may originate in (negative or positive) sentiments, prejudices and discrimination, or mere differences in taste (is the driver "friendly" or "talkative"? Is the restaurant "lively" or "noisy"?). While imperfect reliability is an object of attention for existing platforms, their interaction with social scoring raises new ethical concerns.

By positing that anti- and pro-government activities are measured exogenously (a fine assumption when the measure originates in facial recognition or data mining for instance), this paper may also ignore another important cost of bundling in social ratings: The very process of measuring behavior alters the relationship between the "evaluators" and the "evaluatee". The latter is then on guard, fakes opinions or shuns others. Like in Section 6, but through a different channel, the social fabric and its valuable relationships may be destroyed.

I treated the "government" as a unitary actor. I thereby ducked questions about the construction and use of social scores with multiple layers of government, either horizontal (ministries, or like-minded countries in a data-sharing alliance, say) or vertical (central, regional or local governments), and the concomitant questions about the coordination

[51]The "Honest Qingzhen" program attributes a score to individuals according to over 1,000 criteria (Dai 2018).

of principals with heterogenous goals[52] and about the portability of scores. Similarly, "agents" were also modeled as unitary actors. Doing so sidestepped the question of the comparative impacts of household vs. individual social scoring.

Finally, we may wonder whether we should have a social score in the first place. A single social score communicated to everyone may not always be optimal.[53] Information-design theory suggests that different social scores might be communicated to different audiences; similarly, guilt by association might apply to some relationships, but not others (say, as when the government allows dissenters to have access to basic goods, but not to high-end services). Another reason why social scores may not be disseminated ubiquitously is the protection of self-esteem,[54] as platforms such as Tinder realize. I leave these issues and the many other important questions associated with a principle-based design of privacy law for future research.

[52]For instance, some Chinese pilot experiments with social scoring have secured cheap local public goods through "voluntary" work, as when points are awarded for participating in rural services. Such objectives may well receive a lower weight in the central government's objective function. Furthermore, and as demonstrated in this paper, the government's preferred strategy may depend on socio-economic factors that impact the stability of relationships and the propensity to dissent.

[53]In the context of privacy, see Pébereau (2020).

[54]The evidence in Butera et al (2019) suggests that esteem payoffs may be concave.

References

Acemoglu, D., and A. Wolitzky (2020), "Sustaining Cooperation: Community Enforcement vs. Specialized Enforcement," *Journal of the European Economic Association*, 18: 1078–1122.

Acquisti, A., Taylor, C., and L. Wagman (2016), "The Economics of Privacy," *Journal of Economic Literature*, 54: 442–492.

Adriani, F., and S. Sonderegger (2019), "A Theory of Esteem Based Peer Pressure," *Games and Economic Behavior*, 115: 314–335.

Ali, S.N., and D. Miller (2016), "Ostracism and Forgiveness," *American Economic Review*, 106(8): 2329–2348.

Ali, S.N., and R. Bénabou (2019), "Image Versus Information: Changing Societal Norms and Optimal Privacy," forthcoming *AEJ: Micro*.

Ariely, D., Bracha, A. and S. Meier (2009), "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99(1): 544–555.

Ball, I. (2020), "Scoring Strategic Agents," mimeo, Yale University.

Bénabou, R., and J. Tirole (2006), "Incentives and Prosocial Behavior," *American Economic Review*, 96(5): 1652–1678.

Bénabou, R., and J. Tirole (2011), "Identity, Morals and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 126(2): 805–855.

Bénabou, R., Falk, A. and J. Tirole (2018), "Narratives, Imperatives and Moral Reasoning," mimeo.

Bergemann, D., Bonatti, A. and T. Gan (2020), "The Economics of Social Data," Working Paper.

Bergemann, D., Heumann, T. and S. Morris (2015), "Information and Volatility," *Journal of Economic Theory*, 158(B): 427–465.

Bernheim, B. Douglas (1994), "A Theory of Conformity," *Journal of Political Economy*, 102: 841–77.

Besley, T., Jensen, A. and T. Persson (2015), "Norms, Enforcement, and Tax Evasion," CEPR Discussion Paper No DP10372.

Bonatti, A., and G. Cisternas (2020), "Consumer Scores and Price Discrimination," forthcoming, *Review of Economic Studies*.

Bursztyn, L., and R. Jensen (2017), "Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure," *Annual Review of Economics*, 9: 131–153.

Bursztyn, L., Egorov, G. and R. Jensen (2018), "Cool to Be Smart or Smart to Be Cool? Understanding Peer Pressure in Education", *Review of Economic Studies*, forthcoming.

Butera, L., Metcalfe, R., Morrison, W., and D. Taubinsky (2019), "The Deadweight Loss of Social Recognition," mimeo.

Calzolari, A. and A. Pavan (2006), "On the Optimality of Privacy in Sequential Contracting," *Journal of Economic Theory*, 130: 168–204.

Chen, D. (2017), "The Deterrent Effect of the Death Penalty? Evidence from British Commutations During World War I," TSE Working Paper no. 16-706.

Clark, D., Fudenberg, D. and A. Wolitzky (2019), "Robust Cooperation with First-Order Information," mimeo.

Dai, X. (2018), "Toward a Reputation State: The Social Credit System Project of China", SSRN: https://ssrn.com/abstract=3193577 or http://dx.doi.org/ 10.2139/ssrn.3193577.

Daughety, A., and J. Reinganum (2010), "Public Goods, Social Pressure, and the Choice between Privacy and Publicity," *American Economic Journal: Microeconomics*, 2(2): 191–221.

DellaVigna, S., List, J. and U. Malmendier (2012), "Testing for Altruism and Social Pressure in Charitable Giving," *Quarterly Journal of Economics*, 127: 1–56.

Dewatripont, M., Jewitt, I., and J. Tirole (1999), "The Economics of Career Concerns, Part I: Comparing Information Structure," *The Review of Economic Studies*, 66(1): 183–198.

Ellingsen, T. and M. Johannesson (2008), "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98: 990–1008.

Ellison, G. (1994), "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching", *Review of Economic Studies*, 61: 567–588.

Fischer, P. and R. Verrecchia (2000), "Reporting Bias," *Accounting Review*, 75: 229–245.

Frankel, A. and N. Kartik (2019 a), "Improving Information from Manipulable Data," mimeo, University of Chicago and Columbia University.

Frankel, A. and N. Kartik (2019 b), "Muddled Information," *Journal of Political Economy*, 127(4): 1739–1776.

Harbaugh, R., and E. Rasmusen (2018), "Coarse Grades: Informing the Public by Withholding Information," *American Economic Journal: Microeconomics*, 10: 210–23.

Hardt, M., N. Megiddo, C. Papadimitriou, and M. Wootters (2016), "Strategic Classification," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, ACM*, 111–122.

Jewitt, I. (2004), "Notes on the Shape of Distributions," unpublished.

Jia, R., and T. Persson (2017), "Individual vs Social Motives in Identity Choice: Theory and Evidence from China," mimeo.

Kamenica, E., and M. Gentzkow (2011), "Bayesian Persuasion," *American Economic Review*, 101(6): 2590–2615.

Kandori, M. (1992), "Social Norms and Community Enforcement," *Review of Economic Studies*, 59: 63–80.

Karing, A. (2019), "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone," mimeo, UC Berkeley.

King, G., Pan, J. and M. Roberts (2013), "How Censorship in China Allows Government Criticism but Silences Collective Expression," *American Political Science Review*, May 1-18.

Mellström, C. and M. Johannesson (2008), "Crowding Out Blood Donation: Was Titmuss Right?," *Journal of the European Economic Association*, 6: 845-863.

Ohlberg, M., Ahmed, S. and B. Lang (2017), "Central Planning, Local Experiments: The Complex Implementation of China's Social Credit System," MERICS China Monitor.

Pébereau, C. (2020), "None of Your Business! Efficient Disclosure Policies with Heterogeneous Audiences," mimeo, TSE.

Peski, M. and B. Szentes (2013), "Spontaneous Discrimination," *American Economic Review*, 103(6): 2412–2436.

Prat, A. (2018), "Media Power," *Journal of Political Economy*, 126(4): 1747–1783.

Prendergast, C. and R. Topel (1996), "Favoritism in Organizations," *Journal of Political Economy*, 104: 958–78.

Qin, B., Strömberg, D. and Y. Wu (2017), "Why Does China Allow Freer Social Media? Protests versus Surveillance and Propaganda," *Journal of Economic Perspectives*, 31(1): 117–140.

Qin, B., Strömberg, D. and Y. Wu (2018), "Media Bias in China," *American Economic Review*, 108(9): 2442–2476.

Rayo, L. and I. Segal (2010), "Optimal Information Disclosure," *Journal of Political Economy*, 118(5): 949–987.

Rosenthal, R. (1979), "Sequences of Games with Varying Opponents," *Econometrica*, 47: 1353–1366.

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, Penguin Random House.

Seabright, P. (2010), *The Company of Strangers: A Natural History of Economic Life,* second edition, Princeton University Press.

Thiels, Y. (2009), "L'excommunication : une liberté religieuse controversée," *Jurisprudence de Liège, Mons et Bruxelles* (Larcier), 10 April 2009, p. 678-700.

Tirole, J. (1996), "A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)," *Review of Economic Studies*, 63(1): 1–22.

Tocqueville, A. de, (1838), *Democracy in America*, Saunders and Otley (London).

Valeri, M. (1997), "Religion, Discipline, and the Economy in Calvin's Geneva," *The Sixteenth Century Journal*, 28(1): 123–142.

Zuboff, S. (2018), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile.

Appendix

A. Proof of Proposition 4(i)

Consider an arbitrary image benefit $\Delta^b \in [0, \Delta^u]$ (from an equilibrium behavior as depicted in Figure 1). This defines a behavior

$$a_i = 1 \text{ iff } v_i e - c + \nu\Delta^b - \max\{\theta_i, 0\} \geq 0,$$

and a cutoff $v^b(\theta, \Delta^b) \in [0, 1]$ satisfying $v^b(\theta, \Delta^b) \geq v^u$. To this Δ^b one can associate $\tilde{\Delta}^b$ defined (with obvious notation)[55] by

$$\begin{aligned}\tilde{\Delta}^b &\equiv \int \left[g_1(\theta, \Delta^b)[M^-(v^b(\theta, \Delta^b)) + \Delta(v^b(\theta, \Delta^b))] - g_0(\theta, \Delta^b)M^-(v^b(\theta, \Delta^b))\right] d\theta \\ &= \int g_1(\theta, \Delta^b)\Delta(v^b(\theta, \Delta^b))d\theta + \int \left[\frac{g_1(\theta, \Delta^b)}{g_0(\theta, \Delta^b)} - 1\right] M^-(v^b(\theta, \Delta^b))g_0(\theta, \Delta^b)d\theta.\end{aligned}$$

But note that

$$E_{g_0}\left[\frac{g_1}{g_0} - 1\right] = 0$$

and (g_1/g_0 is decreasing in θ while M^- is increasing in θ)

$$\mathrm{cov}_{\mathrm{g_0}}\left(\frac{\mathrm{g_1}}{\mathrm{g_0}} - 1, \mathrm{M^-}\right) \leq 0$$

and so

$\tilde{\Delta}^b \leq \Delta^u$ if $\Delta' \leq 0$ (using the fact that $v^b(\theta, \Delta^b) \geq v^u$ and so $\Delta(v^b(\theta, \Delta^b)) \leq \Delta(v^u)$).

Furthermore, $\tilde{\Delta}^b$ is non-negative:

$$\int g_1(\theta, \Delta^b)M^+(v^b(\theta, \Delta^b))d\theta \geq M^+(v^u) \geq M^-(1) \geq \int g_0(\theta, \Delta^b)M^-(v^b(\theta, \Delta^b)).$$

Brouwer's fixed-point theorem then demonstrates the existence of such an equilibrium.

Finally, if the distribution of v is uniform, $\Delta(v^*)$ is independent of v^* and so

$$\Delta^b = \Delta^u + \int \left[\frac{g_1(\theta, \Delta^b)}{g_0(\theta, \Delta^b)} - 1\right] M^-(v^b(\theta, \Delta^b))g_0(\theta, \Delta^b)d\theta \leq \Delta^u.$$

This implies that all equilibria involve lower image concerns and a lower prosocial contribution under bundling when the distribution of v is uniform.

■

[55] $g_1(\theta, \Delta^b) = \dfrac{g(\theta)[1 - F(v^b(\theta, \Delta^b))]}{\int g(\theta)[1 - F(v^b(\theta, \Delta^b))]d\theta}$ and $g_0(\theta, \Delta^b) = \dfrac{g(\theta)F[v^b(\theta, \Delta^b)]}{\int g(\theta)F(v^b(\theta, \Delta^b))d\theta}$.

B. Proof of Propositions 6 and 7

Assume that[56]

$$u_i = \left[v_i a - \theta_i b - \left(\frac{a^2+b^2}{2}\right)\right] + \nu\hat{v}(s) + \xi\hat{\theta}(s).$$

And look for an equilibrium with $d\hat{v}/ds = \gamma$ and $d\hat{\theta}/ds = \lambda$. Then

$$\begin{cases} a = v + \nu\alpha\gamma + \xi\alpha\lambda \\ b = -\theta + \nu\beta\gamma + \xi\beta\lambda \end{cases}$$

and

$$\begin{cases} \hat{v}(s) = \bar{v} + \dfrac{\text{cov}(v,s)}{\text{var}(s)}[s - E[s]] \\ \hat{\theta}(s) = \bar{\theta} + \dfrac{\text{cov}(\theta,s)}{\text{var}(s)}[s - E[s]]. \end{cases}$$

Substituting, and letting $y \equiv \frac{\sigma_\theta}{\sigma_v}$

$$\begin{cases} \gamma = \dfrac{\alpha - \beta\rho y}{\alpha^2 + \beta^2 y^2 - 2\alpha\beta\rho y} \\ \lambda = \dfrac{\beta y^2 + \alpha\rho y}{\alpha^2 + \beta^2 y^2 - 2\alpha\beta\rho y}. \end{cases}$$

This yields in particular, for $x \equiv \frac{\alpha}{\beta}$

$$E[b] = -\bar{\theta} + \frac{\nu(x - \rho y) + \xi y(x\rho - y)}{x^2 + y^2 - 2x\rho y}.$$

$E[b]$ is maximized for

$$x = \frac{\xi y^2 + \nu\rho y}{\xi\rho y + \nu} + \sqrt{\left(\frac{\xi y^2 + \nu\rho y}{\xi\rho y + \nu}\right)^2 + \frac{y^2[-\xi\rho - 2\nu\rho^2 + \nu]}{\xi\rho y + \nu}}.$$

The results in the Proposition follow from these expressions.

Let us next consider welfare for $\xi = 0$.

Unbundling. Suppose that $\beta = 0$ and $\alpha > 0$. Then

$$a(v) = v + \nu$$

[56]The weight ξ on $\hat{\theta}$ may be positive or negative. For example, if the individual wants to conform to society and puts weight $\xi(\tilde{\theta})$ on reputation vis-à-vis agent $\tilde{\theta}$, where $\xi(\tilde{\theta})$ is increasing and symmetric around the origin, then $\xi = \int_{-\infty}^{+\infty} \xi(\tilde{\theta}) dG(\tilde{\theta})$ (where $G(\cdot)$ is the marginal distribution) and so $\xi > 0$ if and only if $\bar{\theta} > 0$. The earlier assumption that future partners care about v_i but not θ_i considerably simplifies the analysis. It may also be reasonable in some environments; in a well-functioning workplace or on a trading or sharing platform, people care about their colleagues or trading partners being competent, efficient, friendly and obliging (vertical dimensions), regardless of their political opinions or religion (alternatively, asking colleagues about their politics or religion may be frowned upon). This may be less true of some non-work or trade-oriented activities; there, individuals may enjoy the company of like-minded peers. In such an environment, the individual should be concerned also about appearing a desirable match along the θ_i dimension.

and so

$$\bar{a} = \bar{v} + \nu$$

$$W^u = \Big[E\Big[va - \frac{a^2}{2}\Big] + \bar{a}e\Big] + E\Big[\max_b\Big(-\theta b - \frac{b^2}{2}\Big)\Big]$$

with

$$E\Big[va - \frac{a^2}{2}\Big] + \bar{a}e = E\Big[\frac{v^2}{2}\Big] - \frac{\nu^2}{2} + e(\bar{v} + \nu) \text{ and } E\Big(-\theta b - \frac{b^2}{2}\Big) = \frac{\sigma_\theta^2 + \bar{\theta}^2}{2}.$$

Bundling. Under optimal bundling,

$$a(v) = v + \kappa$$

where $\kappa = \frac{\nu}{2}\big(1 + \frac{\rho}{\sqrt{1-\rho^2}}\big)$.

Furthermore

$$W^b = \Big[E\Big[\frac{v^2}{2}\Big] - \frac{\kappa^2}{2} + e(\bar{v} + \kappa)\Big] + \Big[-\frac{\nu^2}{8}\frac{\sigma_v^2}{\sigma_\theta^2}\frac{1}{1-\rho^2} + \frac{\sigma_\theta^2 + \bar{\theta}^2}{2}\Big]$$

since the agent distorts b due to image concerns.

$$W^u - W^b = \Big[\frac{e\nu}{2}\Big(1 - \frac{\rho}{\sqrt{1-\rho^2}}\Big) + \frac{\nu^2}{2}\Big(\frac{1}{4}\Big(1 + \frac{\rho}{\sqrt{1-\rho^2}}\Big)^2 - 1\Big)\Big] + \Big[\frac{\nu^2}{8}\frac{\sigma_v^2}{\sigma_\theta^2}\frac{1}{1-\rho^2}\Big].$$

The first bracket corresponds to the impact of unbundling on the a-behavior and the second one to its impact on the b-behavior (which is more authentic).

So $W^u > W^b$ in particular if

$$(\nu - \kappa)\big[\big(\frac{4e}{\nu} - 3\big) - \frac{\rho}{\sqrt{1-\rho^2}}\big] > 0.$$

This is a sufficient, not a necessary condition. We have: $\nu > \kappa \Leftrightarrow \rho < \frac{1}{\sqrt{2}}$. And because $e > \nu$, $\frac{4e}{\nu} - 3 > 1$. And so $W^u > W^b$ if $\rho < \frac{1}{\sqrt{2}}$. By contrast, let $k \equiv \frac{4e}{\nu} - 3 > 1$ and $\rho^* \equiv \frac{k}{\sqrt{1+k^2}}$. Then the a-dimension yields more welfare under bundling iff $\rho \in \big(\frac{1}{\sqrt{2}}, \rho^*\big)$. But for $\rho \to 1$, the two terms in $1/(1-\rho^2)$, which are dominant, are both positive, and so $W^u - W^b > 0$. Similarly, one can show that for e close to ν, $W^u - W^b > 0$ for all ρ.

This of course is only a sufficient condition for $W^u > W^b$ since $E\big[-\theta b - \frac{b^2}{2}\big] = \frac{\sigma_\theta^2}{2}$ under unbundling, and $= \frac{\sigma_\theta^2}{2} - \frac{1}{8}\frac{\nu^2\sigma_v^2}{(1-\rho^2)\sigma_\theta^2}$ under bundling.

When $\xi \neq 0$, the expressions are more complex. Of particular interest is

$$E[b] = -\bar{\theta} + \frac{\nu}{2}\frac{\sigma_v}{\sigma_\theta}\big[\frac{1}{\sqrt{1-\rho^2}[\sqrt{C^2 + 1 + 2\rho C} + C\sqrt{1-\rho^2}]}\big]$$

where

$$C \equiv \frac{\xi\sigma_\theta}{\nu\sigma_v}$$

- $E[b]$ grows with ν and σ_v for $\rho \geq -\frac{1}{\sqrt{2}}$
- $E[b]$ decreases with ξ and σ_θ for $\rho \geq -\frac{1}{\sqrt{2}}$
- $\beta < 0$ (for $\alpha \equiv 1$) if and only if $\rho < \rho^*$ for some $\rho^* < 0$
- $E[b]$ goes to $+\infty$ when $|\rho|$ goes to 1.

■

C. Proof of Proposition 9

Individual i prefers the second strategy over the first if $u_i^2 - u_i^1 \geq 0$, or

$$X(ve - c) + X_1 e + (2X_1 + X_0 - 1)\mathfrak{b} - \theta + \nu(\hat{v}_1 - \hat{v}_0) > 0.$$

Individual i picks the third strategy over the first if $u_i^3 - u_i^1 > 0$ or

$$X_1(e + \mathfrak{b}) \geq \theta.$$

Note that if there are no model citizens ($X_1 = 0$), there are no compliers either ($X_0 = 0$): The only benefit of complying is to avoid being ostracized by model citizens.

Finally, individual i picks the second strategy over the third if and only if $u_i^2 - u_i^3 > 0$, or

$$X(ve - c) + \nu(\hat{v}_1 - \hat{v}_0) > (1 - X)\mathfrak{b}.$$

Letting

$$\begin{aligned}
v^*(X) &\equiv \max\left\{\min\left\{\frac{c}{e} + \frac{(1-X)\mathfrak{b} - \nu(\hat{v}_1 - \hat{v}_0)}{Xe}, 1\right\}, 0\right\} \\
v^*(X, X_1, \theta) &\equiv \max\left\{\min\left\{\frac{c}{e} + \frac{(1-X)\mathfrak{b} - \nu(\hat{v}_1 - \hat{v}_0)}{Xe} + \frac{\theta - \hat{\theta}(X_1)}{Xe}, 1\right\}, 0\right\} \\
\hat{\theta}(X_1) &\equiv X_1(e + \mathfrak{b}),
\end{aligned}$$

the equilibrium behavior is described by

(a) $\theta \leq \hat{\theta}$: $\begin{cases} a_i = b_i = 1 & \text{if } v \geq v^*(X) \\ a_i = 0 \text{ and } b_i = 1 & \text{if } v < v^*(X) \end{cases}$

(b) $\theta \geq \hat{\theta}$: $\begin{cases} a_i = b_i = 1 & \text{if } v \geq v^*(X,\, X_1,\, \theta) \\ a_i = b_i = 0 & \text{if } v < v^*(X,\, X_1,\, \theta). \end{cases}$

Finally, an equilibrium satisfies:

$$\begin{aligned}
X_0 &= G(\hat{\theta}(X_1))F(v^*(X)) \\
X_1 &= G(\hat{\theta}(X_1))[1 - F(v^*(X))] + \int_{\hat{\theta}(X_1)}^{\infty} [1 - F(v^*(X, X_1, \theta))]dG(\theta) \\
X &= X_0 + X_1.
\end{aligned}$$

Existence of an equilibrium is guaranteed by Brouwer's theorem. Tainting creates endogenous network externalities, and so we cannot guarantee that equilibrium conditions have a unique solution $\{X, X_1, X_0\}$ (as Example 2 below will illustrate). Suppose that more agents decide to behave as model citizens; ostracism of dissenters by model citizens makes being a complier more attractive relative to being a dissenter. Conversely, the increase in the ratio of the number of compliers-cum-model citizens over that of dissenters reduces the occurrence of ostracism and makes it less costly for agents to behave as model citizens. Let us compare behaviors when the social score does and does not embody the social graph:

The impact of guilt by association

Suppose the ruler bundles, but does not allow the social graph to taint reputations. Then the fixed gain from interaction $\mathfrak{b}$ plays no role. Furthermore, because I restricted θ_i to be non-negative, there are only two behavioral patterns $a_i = b_i = 1$ and $a_i = b_i = 0$. For a given θ and an interior solution ($v^*(\theta) \in (0,1)$), the cutoff type is given by

$$v^*(\theta)e - c + \nu(\hat{v}_1 - \hat{v}_0) - \theta = 0 \tag{1}$$

where $\hat{v}_1$ and $\hat{v}_0$ are computed as in Section 3.2.

Let us now look at the choice of whether to augment the social score with social graph data. As earlier, the state has objective function $W + \gamma E[b_i]$ with $E[b_i] = X$. Embodying the social graph into the social score has three welfare effects for the government:

(1) *Looser social fabric.* The ostracization of non-compliant individuals by high-score ones creates a welfare loss equal to

$$X_1(1 - X)(2\mathfrak{b}).$$

(2) *Impact on prosociality.* Regardless of whether including the social graph increases or decreases prosocial behavior, the sign of this effect on the principal's welfare is a priori ambiguous, as it depends on whether there is over- or under-signaling in the first place.

(3) *Impact on dissent.* $E[b_i] = X$ is higher when the social graph is used in the social score provided that[57]

$$G(\hat{\theta}(X_1)) + \int_{\hat{\theta}(X_1)}^{\infty} [1 - F(v^*(X, X_1, \theta))]dG(\theta) \geq \int_0^{\infty} [1 - F(v^*(\theta))]dG(\theta) \tag{2}$$

■

[57]Where $v^*(\theta)$ is the cutoff under simple bundling. A sufficient condition for (2) for F uniform to be satisfied is

$$\frac{\mathfrak{b} + \theta - \nu(\hat{v}_1 - \hat{v}_0)}{\mathfrak{b} + e} \leq \frac{X_1}{1 - X}.$$

ONLINE APPENDIX

Assortative matching: Reputation as a positional good (Section 2)

Consider a future relationship, with partners potentially exercising externality e_2 on the other. Let c_2 denote the date-2 cost of providing this externality, drawn from the uniform distribution on $[0,1]$. So the probability that agent i provides the externality when her type is v is $\Pr(ve_2 \geq c_2) = ve_2$. So, if $\hat{F}_i(v)$ is the posterior distribution on v_i, the expected externality created by agent i is $[\int_0^1 ve_2 d\hat{F}_i(v)]e_2 = \hat{v}_i e_2^2$.

Agents optimally match with agents of the same reputation (they do not have access to agents with a better reputation). Anticipating a bit, those who have chosen $a_i \equiv 1$ choose as partners agents who have done so as well. Letting v^* denote the cutoff under which agents no longer contribute, the total externality enjoyed by all agents is independent of v^*:

$$\left[F(v^*)\left[\frac{\int_0^{v^*} vdF(v)}{F(v^*)}\right] + [1-F(v^*)]\left[\frac{\int_{v^*}^1 vdF(v)}{1-F(v^*)}\right]\right]e_2^2 = \bar{v}e_2^2.$$

Proof of observation on observable compliance (Section 4 (a))

Suppose, first, that $\operatorname{supp} G = \mathbb{R}^+$. In the case of unobservable b_i studied so far, there were only two equilibrium behaviors, $a_i = b_i = 1$ and $a_i = b_i = 0$. Therefore, observing b_i contained no information that was not already in the social score. The equilibrium characterized in Proposition 4 is still an equilibrium.

Suppose next that the support of G includes negative values of θ_i as well. Let $\hat{v}_{00}$ and $\hat{v}_{01}$ denote the reputation following $\{a_i = b_i = 0\}$ and $\{a_i = 0, b_i = 1\}$, respectively (both are associated with rating 0); and let $\hat{v}_1 = \hat{v}_{11}$ be the reputation following $\{a_1 = b_1 = 1\}$. Among those who choose $a_i = 0$, those with $\theta > \theta^*$ choose $b_i = 0$, where $\nu(\hat{v}_{00} - \hat{v}_{01}) + \theta^* = 0$. I claim that $\theta^* < 0$. Indeed, the corresponding cutoffs satisfy,[58] for $\theta \geq \theta^*$, $v_{00}^*(\theta) = v_{01}^* + (\theta - \theta^*)/\mathrm{e}$, and so $\hat{v}_{01} < \hat{v}_{00}$. The intuition behind this result is again that dissenters have an excuse for not engaging in prosocial acts because they cannot obtain a good social rating anyway. The impact of bundling on $E[b_i]$ is less clear than when b_i is unobservable by future partners. As earlier, bundling induces some $\theta_i > 0$ types to choose $b_i = 1$. Types $\{\theta_i \in [\theta^*, 0], v_i < v_{00}^*(\theta)\}$ choose $b_1 = 0$ while they selected $b_i = 1$ in the absence of bundling: They are in search of an excuse. ■

Non image sanctions (Section 4)

To fix ideas, suppose that the state can impose economic sanctions P on blacklisted agents, at deadweight loss $L(P)$ with $L(0) = 0$, $L' > 0$ and $L'' > 0$. Underconsumption underlies this deadweight loss. Consider a social rating in which agents who do not select $a = b = 1$ are blacklisted. Blacklisting implies both an image penalty $\hat{v}_1 - \hat{v}_0$ and an

[58] Existence as earlier follows from Brouwer's fixed-point theorem.

economic penalty P, as is the case in Chinese pilots.[59] The government's welfare becomes

$$V = W - E[L(P)] + \gamma E[b_i].$$

The cutoff $v^b(\theta)$, if interior, is now given by

$$v^b(\theta)e - c + \nu\hat{v}_1 - \max\{\theta, 0\} = \nu\hat{v}_0 - P.$$

Everything is as if the cost of engaging in prosocial behavior were $c - P$ instead of c. This implies that $v^b(\theta)$ has the same shape as in Figure 1 (with a kink at 0), just shifted down. [By contrast under unbundling, the cutoff is still v^u. And $b = 1$ if and only if $\theta \leq P$.]

A complete characterization of economic sanctions lies out of the scope of this paper (for one thing, there is no reason for P to be the same under bundling and unbundling).

[59] In these pilots, the penalty may take the form of non-access for blacklisted agents to "discounts" in some public enterprises. An alternative to this "double whammy" is a separate sanction P when $b = 0$, independently of the choice of a. This separation between the economic and image sanction is difficult to implement, though: As I already noted, agents have an incentive to disclose the punishment P and thereby prove that $b = 0$ to create an excuse for having selected $a = 0$ (or raise the glory for having selected $a = 1$). So, there is unravelling, and the two dimensions of behavior are necessarily intertwined. I consider here the double whammy designed in current pilots.