Digital Dystopia*

Jean Tirole†

January 8, 2020

Abstract: While data collection and artificial intelligence hold the promise of a sharp reduction in incivilities, they also open the door to mass surveillance by private platforms and governments. After analysing the welfare impact of transparency, whether suffered or voluntary, the paper shows how an expansion of the public sphere may lead to a disintegration of the social fabric in the private sphere, with potentially negative consequences.

The paper then brings to light how political authorities can enlist a social rating to control society without engaging in brutal repression or misinformation. To this end they optimally bundle the individual's political attitudes and social graph with her overall social behavior, so as to mobilize individuals and platforms to enforce the political will. Finally, the paper uses a similar argument to show that democratic countries may be concerned with private platforms in the same way autocratic ones may be wary of public platforms.

Keywords: Social behavior, data integration, social score, platforms, 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, Roland Bénabou, Aimé Bierdel, Erik Brynjolfsson, Paul-Henri Moisson, Charles Pébereau, and participants at the 2019 Luohan Academy conference on Privacy and Data Governance, the NBER summer institute (IT & digitization and IO groups), and at seminars at MIT (MIT Initiative on the Digital Economy) Northwestern, and TSE 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 issue 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.

This paper is best viewed as an exercise in (social) science fiction. Indeed, I do not advance that 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 framework. This framework is borrowed from the literature, even though its interpretation and a couple of results obtained in Section 2 are new. Economic agents are engaged in stable and/or transient relationships and care about their social image. Stable relationships characterize family, friendship, village or employee bonds. Transient relationships capture matching through platforms, independent contracting or large-city interactions. The agents' very desire to project a good image may be harnessed to enhance trust in society.

An agent's behavior may become known to others in two ways: through direct experience of interacting with the agent, and 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 blacklisted or not). "Silo information" or "data islands" prevail in the absence of such a social score; "data integration" corresponds to the public disclosure of the social score. The release of a social score boosts image concerns and thereby prosocial behavior: Stable partners get a clearer assessment of the agent's motivation if they imperfectly observe or interprete the meaning of the agent's behavior in their own interaction with her; and the agent's reputation is extended to future new partners.

There however may be under- or over-signaling, even when social image is a mere positional good (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 and relations are stable; social scoring then reduces welfare. For large externalities and/or with transient relationships, the building of trust through social scoring is desirable. Finally, Section 2 shows that transparency has the opposite effect of reducing prosocial contributions when an individual's overall altruism (or proclivity for doing good) is common knowledge, but not the individual's relative empathy for agents in her social graph. Under such "unknown relative empathy", transparency prevents the individual from pretending to have more "good friends" than she really has, thereby extinguishing her social image concerns. Because social ratings are motivated by a heterogeneity in absolute empathy, we focus in the rest of the paper on the unknown absolute empathy case. Thus, the core of the paper extends the basic framework with unknown proclivity for doing good to study some further and potentially problematic aspects of data integration.

Section 2 then exploits the distinction between private and public spheres. Behaviors in the private sphere are observed solely through direct interaction, as they cannot be reliably rated or their public disclosure would make the individual or the audience uncomfortable. Behaviors in the public sphere by contrast are the object of public disclosure through a rating. We obtain two insights. First, individuals behave better in the public sphere than in the "all public" or "all private" benchmarks; the converse holds for behavior in the private sphere; this implies that prosocial activities, regardless of their overall level, are misallocated, with too much attention paid to the public sphere. Second the public sphere crowds out the private sphere. An expansion in the public sphere (due, say, to technological change) reduces prosociality in both spheres and may even reduce overall prosociality. The overall picture is one of public sphere dominance and disintegration of the social fabric in the private sphere.

Section 3 analyzes how the state can leverage social sanctions to suppress dissent, or more generally to force citizens to adopt societal or religious attitudes that it favors.[1] It generalizes the model of Section 2 by adding another decision for each individual: oppose or accommodate the state. Each agent's type is now two-dimensional. Besides their prosocial proclivity (intrinsic motivation to do good in bilateral interactions), agents differ in their psychological cost of toeing the line. When interacting with others, agents care 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 conformity (lack of dissent) in society. A more autocratic regime puts more weight on the latter than a less autocratic one. We compare behavior when the social rating mixes behavior in social interactions together with the dissent/accommodate choice (bundling) 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.

The main insights go as follows: 1) Bundling prosocial activities and compliance with

[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.

the government's objective into a single score awakens citizens' interest in the score.[2] 2) The state 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 much higher in a society of strangers[3] (of transient relationships) than in a tight knit society of stable relationships. Indeed, it may have no ability to strengthen its hold in the latter society. The intuition behind this result is that in a tight-knit-relationships society, agents have information about each other that unbundles the information supplied by the state. 4) The state must eliminate competition from independent, privately-provided social ratings. 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 political views from their ratings. This competition would lead to de facto unbundling, with no-one paying attention to the government's social score. 5) Transparency/citizens' awareness about the way the social score is computed (together with opaqueness about its components) is essential to the latter's effectiveness.

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 make citizens afraid of being seen in company of dissidents or mere citizens whose lifestyle is frowned upon by the regime. Face recognition and artificial intelligence applied to surveilled social life, communications and social network activities have substantially reduced the state's cost of drawing an accurate social graph of relationships among its citizens.

Section 4 studies how 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 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 citizens sever beneficial ties with others. It thus appeals to autocratic regimes as it reinforces the state's grip.

Section 5 turns to less autocratic regimes. It argues that, while autocratic countries may be wary of public platforms, democratic ones may, to the contrary, be concerned with private ones. Its framework is largely a relabeling of the one of Section 3, thereby economizing on modeling and formal analysis. 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 in Sections 3 and 4, she can grant a favor to the platform (refraining from asking for tougher antitrust or privacy regulation

[2]According to Samantha Hoffman (cited in Hvistendahl 2017) "social credit ideally requires both coercive aspects and nicer aspects, like providing social services and solving real problems".

[3]To borrow Paul Seabright (2010)'s terminology.

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.

Thereby formalizing the notion of media political power, we 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 6 concludes with policy implications and alleys for future research.

Motivation: the advent of social scoring

The much-discussed Chinese social credit system, which is due to be rolled out in 2020, illustrates the potential problems.[4] The following discussion is subject to caution, though, as the terms of this 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 current experimentation 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.[5] 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 (of individual or children) to the best schools or universities.

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

[4]The national social credit system was launched in 2014. It was preceded by local experiments starting in the late 2000s, and of course draws its technological features from the scoring systems developed by the large tech companies.

[5]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.

enforcement. There are two answers. First, we will show 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. Second, and in line with the time-honored debate between Huxley and Orwell on social control,[6] 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 (inefficient and corrupt courts, cost of imprisonment...). 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 such alternative punishments are available, the manipulation of social ratings described below can still strengthen the state's enforcement capacity and be an effective instrument.[7]

Interestingly, and in line with the gist of the paper, similar considerations arise in the private sector. 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. 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.[8]

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.[9] Data also enable sellers to

[6]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".]

[7]In the case of China, the inefficiency of courts in enforcing law was certainly one of the drivers of the social credit project. As Rogier Creemers in a MERICS interview (August 21, 2018) states: "It was found that the existing legal system did not have the wherewithal to provide a sufficient deterrent against [problems of legal compliance]. So the social credit system is in many ways a sort of moralistic, paternalistic system that essentially acts as an amplifier on existing laws and regulations to ensure that people who behave in a sincere and trustworthy way in society are incentivized to do so and people who do not are disincentivised from doing so." See also Ohlberg et al (2017) and Dai (2018).

[8]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 members 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.

[9]A recent entry in this literature is Bonatti-Cisternas (2019), in which firms can prevail themselves

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). Our emphasis on the use of data integration to enlist community enforcement is, to the best of our knowledge, novel.

The paper is also related to the large literature on community enforcement.[10] 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 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 we stress dysfunctional features of such enforcement.

Image concerns feature prominently in a number of theoretical and empirical contributions.[11] This paper uses the Bénabou-Tirole (2006) model of image concerns. As we will later discuss, that literature has brought to light the possibility of under- and over-signaling. The new aspect in the analysis of Section 2 lies in their interpretation and a couple of new results (on unknown relative vs. absolute empathy, and on public and private spheres). The existing literature further supplies the building block for the study of the strategic use of social scoring by the public and private sectors (Sections 3 through 5).

Finally, the paper has implications for certification and auditing more generally. The bundling strategy emphasized here could be applied to the mixture of true reporting with bribes (consulting contracts . . .), with the same need for transparency as to how the grade is actually computed.

---

of a score aggregating the consumer's purchase history, leading to a ratchet effect. 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.

[10] Initiated by Rosenthal (1979), Kandori (1992) and Ellison (1994). See Acemoglu-Wolitzky (2016) and Clark et al (2019) for recent contributions to the literature.

[11] 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 et al (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.[12]

Drivers of social behavior. Individuals or agents are labeled by i, $j \ldots$ Agent i interacts with a continuum of other agents $j \in [0,1]$. The analysis will be simplified by the appeal to the law of large numbers. But interacting with all agents, or even a countable number of them (a measure zero subset) is unneeded: the number of interactions could be finite.

In each interaction, agent i decides to be prosocial ($a_{ij} = 1$) or not ($a_{ij} = 0$). Being prosocial generates an externality $e > 0$ on agent j (the counterparty)[13] and involves private cost c for individual i.

Individuals are heterogenous with respect to their desire to do good. Namely, their intrinsic motivation to do good is ve, where v is distributed according to smooth cumulative distribution $F(v)$ and density $f(v)$ on $[0,1]$, 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.[14] Agent i may care about her reputation with the agents she interacts with, as well as with new agents she will encounter in the future.

Indeed, individual i's social interaction set tomorrow may be composed of the same agents she is currently interacting with ("stable relationships"), of yet unknown agents ("transient relationships"), or of both. Family, friendship, neighborhood and some work relationships are usually stable, while other meeting or work relationships (say, involving

[12] See e.g. Bénabou-Tirole (2006, 2011). The better-measurement effect discussed in Section 2.3 is a direct consequence of the treatment of "excuses" in Bénabou-Tirole (2006) and Bénabou et al (2018).

[13] Equivalently, the externality could be on society as a whole.

[14] 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 $\Pr(ve_2 \geq c) = ve_2$. 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 don't have access to agents with a better reputation). Anticipating a bit, those who have chosen $a_{ij} \equiv 1$ 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^*:

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

foreign trade, platform or large-city interactions) are more akin to the pattern observed in a society of strangers.

Information. We consider two types of information:

- *Silo information.* Agent j interacting with agent i observes signal s_{ij}, equal to a_{ij} with probability $\alpha \in (0,1]$ and nothing ($\emptyset$) with probability $1-\alpha$. That is, agent j may fail to observe,[15] or comprehend the implications of a_{ij}. This information structure is the minimal information structure for interacting agent j. For other (currently non-interacting) agents, the minimal information structure is $\emptyset$ (no information).
- *Social score.* Let $a_i \equiv \alpha \int_0^1 a_{ij} dj \in [0,\alpha]$ denote agent i's observed prosocial behavior when individual actions are integrated into a social score. Individual i's social score is binary and takes value $s_i = 1$ if $a_i = \alpha$ and $s_i = 0$ otherwise.

Definition. Under *data islands* or *silo information*, agents receive the minimal information about agent i's behavior. Under *data integration* or *transparency*, they further receive the common information s_i.

Thus, agent j who interacted with agent i has information about agent i: $I_{ij} = \{s_{ij}\}$ under silo information and $I_{ij} = \{s_{ij}, s_i\}$ under transparency. Similarly, an agent who did not interact with agent i has only the public information, namely $I_i = \{\emptyset\}$ (no information) under silo information and $I_i = \{s_i\}$ under transparency.

Note that we 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). We will later note that this may not always be feasible and will introduce the notion of a "private sphere".

Payoff functions. Individual i with type v_i has payoff function

$$u_i = \int_0^1 [(v_i e - c)a_{ij} + \mu \hat{v}_i(I_{ij})]dj + \nu \hat{v}_i(I_i),$$

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 stable social network, $\nu = 0$. By contrast, on a sharing platform, $\mu = 0$ to the extent that the individual will in the future interact with new agents.

Strategies. Our analysis focuses on *pure, symmetric strategies*. A strategy for agent i is pure if $a_{ij} \in \{0,1\}$ for all j. It is symmetric (or uniform or nondiscriminatory) if for j and k in $[0,1]$, $a_{ij} = a_{ik}$. There will always exist an equilibrium in pure and symmetric strategies in the environments considered in this paper.

[15] The externality may be backloaded.

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 (see footnote 14), this gain has no social value and reputation is a "positional good": the agent's gain is another agent's loss. We will therefore define welfare as[16]

$$W \equiv \int_{i\in[0,1]} \left[\int_{j\in[0,1]} [(v_i e - c) + e] a_{ij} dj \right] di \tag{1}$$

In general, the release of a social score may eliminate 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: learning the other person's characteristics or behavior may alter attitudes; 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 (1). This different expression would not affect the key drivers of our 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 we will therefore adopt (1) as the expression of welfare.

2.2 The data islands benchmark

Because of single crossing, agent i selects $a_{ij} = 1$ if and only if $v_i \geq v^*$. The cutoff v^*, if interior, is given by

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

where

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

For a uniform distribution of v on $[0,1]$, $\Delta(v^*) = 1/2$ for all v^*. More generally, 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 $\Delta(1) = 1 - \bar{v}$.[17] In

[16] The expression of W in equation (1) incorporates warm glows ($v_i e$) into the principal's welfare function. Diamond (2006) discusses the pros and cons of doing so. Our results do not hinge on this specific formulation of W, which only affects the definition of regions in which there is an over- or under-provision of prosocial behavior.

[17] 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).

the downward-sloping part, one must assume that image concerns are not so strong as to preclude uniqueness of the cutoff (and therefore of the social norm); we will maintain such assumptions throughout the analysis. We will further make the convention that $v^* = 1$ if $e - c + \mu\alpha\Delta(1) \leq 0$ and $v^* = 0$ if $-c + \mu\alpha\Delta(0) \geq 0$.[18]

Comparison with the social optimum. Let us index the socially optimal behavior with a "hat". From the expression of W, we see that agent i should choose $a_{ij} = 1$ for all j if $v_i \geq \hat{v}$ (and $a_{ij} = 0$ for all j otherwise), where

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

There is underprovision (resp. overprovision) if $v^{SO} < v^*$ (resp. $v^{SO} > v^*$). Underprovision therefore corresponds to $e > e^s \equiv \mu\alpha\Delta(v^*)$; for instance, for a uniform distribution $e^s = \mu\alpha/2$.

Proposition 1 *(silo reputations)*

(i) Under data islands, there exists an interior[19] *equilibrium in which individual i picks $a_{ij} = 1$ for all j if $v_i > v^*$ and $a_{ij} = 0$ for all j if $v_i < v^*$, where*

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

(ii) There is underprovision of prosocial behavior in a data islands economy if

$$e > e^s \equiv \mu\alpha\Delta(v^*), \tag{4}$$

and overprovision if this inequality is reversed.

This proposition checks for our model the standard result according to which there is underprovision for large externalities and overprovision for small ones.[20] 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 (as illustrated by Lacie in the series *Black Mirror*[21]).

[18] A corner solution at $v^* = 0$ (resp. $v^* = 1$) exists if and only if $\mu\alpha\bar{v} \geq c$ (resp. $\mu\alpha(1 - \bar{v}) \leq c - e$). Thus, the condition $\mu\alpha(1 - \bar{v}) + e > c > \mu\alpha\bar{v}$ (in the case of a uniform distribution $\frac{\mu\alpha}{2} + e > c > \frac{\mu\alpha}{2}$) is sufficient for the existence of an interior equilibrium (and uniqueness under the D1 refinement).

[19] See the conditions for interiority in footnote 18.

[20] Eg. Acquisti et al (2016), Ali-Bénabou (2019), Bénabou-Tirole (2006) and Daugherty-Reinganum (2010). The differentiation of (4) with respect to e yields:

$$\frac{d}{de}(e - \mu\alpha\Delta(v^*)) = 1 + \mu\alpha\frac{\Delta'(v^*)v^*}{e + \mu\alpha\Delta'(v^*)} > 0$$

if $e + \mu\alpha\Delta'(v^*)(1 + v^*) > 0$, which is trivially satisfied in the uniform case ($\Delta' \equiv 0$).

[21] "Nosedive", season 3, episode 1. Another instance of over-signaling occurs when people feel compelled to wish "happy birthday" to Facebook "friends" they hardly know (and accept them as "friends" in the first place).

2.3 Data islands vs data integration

Next, assume that the individual's overall social behavior is made public (transparency). As we will note, this amounts to the publication of a *social score*.

Proposition 2 *(social score)*

(i) Under data integration (transparency), there exists an interior[22] *equilibrium in which individual i picks $a_{ij} = 1$ for all j if $v_i \geq v^*$ and $a_{ij} = 0$ for all j if $v_i < v^*$, where*

$$v^* e - c + (\mu + \nu)\Delta(v^*) = 0. \tag{5}$$

(ii) There is underprovision of prosocial behavior in a data integration economy if

$$e > e^t \equiv (\mu + \nu)\Delta(v^*), \tag{6}$$

and overprovision if this inequality is reversed.

For instance, for a uniform distribution $e^t = (\mu + \nu)/2$. More generally, unless $\alpha = 1$ and $\nu = 0$, compliance is higher when the social score is released. Moving from silo reputations to a social score has two effects, as the comparison between (3) and (5) shows:

(i) Better measurement. Stable partners imperfectly observe (or interpret) the agent's contribution if $\alpha < 1$. The agent is made more accountable by broader behavioral observability.

(ii) Extension of reputation. New partners are informed of the agent's behavior, increasing the agent's incentive to behave.[23]

Optimality. The impact of an increase in incentives (that associated here with the release of a social score) hinges on whether the agent faces too little or too much incentives in the first place. The welfare comparison is found in Table 1.

[22] The analysis of interior and corner equilibria is the same as that in Section 2.2, replacing $\mu\alpha$ by $(\mu + \nu)$.

[23] This second effect is similar to that created by an increase in audience size in Ali-Bénabou (2019). The latter paper also studies the noisy observation of actions, and relates such imperfect measurement to the effect of scaling up or down the size of audience.

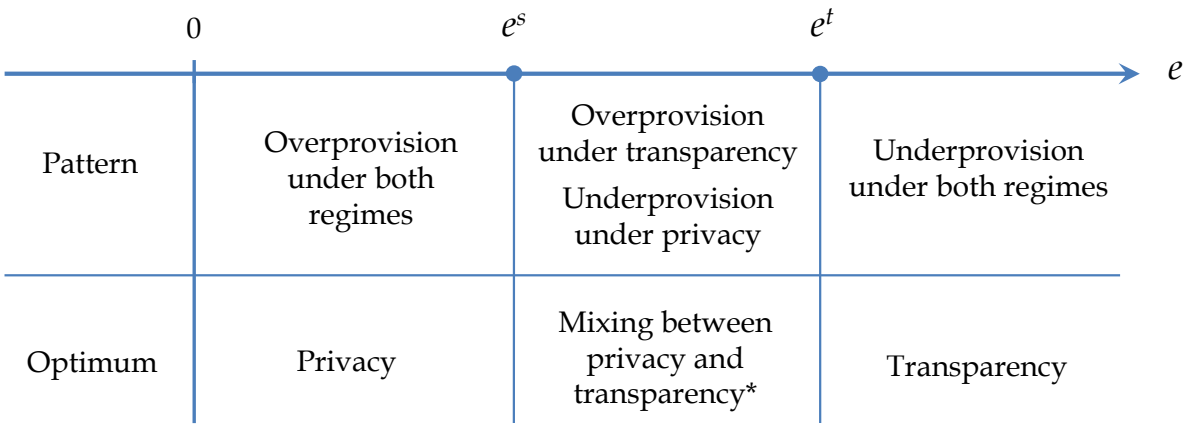

Table 1: Welfare properties of equilibrium

"s" stands for "silo" and "t" for "transparent."
*Probability that ratings will be aggregated into a social rating.

2.4 Discussion

2.4.1 Endogenous transparency ("nothing to hide") and its limits

Section 2.3 implicitly assumed that transparency comes about exogenously (from a platform or a government). Alternatively, agents may choose to disclose their behavior (perhaps through a platform). They may feel compelled to signal that they have "nothing to hide".[24] We augment the silo-reputation game by allowing the agent to reveal $\{s_{ij}\}_{j\in[0,1]}$ or a subset of those signals after taking action $\{a_{ij}\}_{j\in[0,1]}$.

A large literature, starting with Grossman-Hart (1980) and Milgrom (1981), suggests that such disclosure incentives are likely to lead to unravelling, that is, in our context the de-facto, voluntary creation of a social score. This literature however presumes that agents can disclose their *type* (here, v_i) rather than signals about their *actions* (here, $\{s_{ij}\}$). And indeed there exist perfect Bayesian equilibria of our game in which unraveling does not occur. The broad logic of the disclosure literature however applies, in the following sense:

Proposition 3 *(nothing to hide) Full disclosure, leading to the social score equilibrium described in Proposition 2, is an equilibrium of the disclosure game.*

The Appendix shows that full disclosure is the unique equilibrium if one applies a slight variant of Banks and Sobel's D1 refinement.[25]

Discussion. We however note four limits (outside the current model) to endogenous transparency:

[24]This signaling motive is a constant source of inspiration for dystopian fiction, as in *The Circle* or *Black Mirror*.

[25]It requires that if a set of types gain equally from a deviation, then the relative beliefs in that set remain the same under that deviation.

(i) Signal extraction. As emphasized in Bénabou-Tirole (2006), agent i's disclosing favorable information may backfire as it may also signal high image concerns (high μ_i, ν_i) when the latter are unknown. The audience then does not know whether agent i contributed because of a high v_i or because of high μ_i, ν_i.

(ii) Collusion. The disclosure of the $\{s_{ij}\}$ may operate through the ratings of the agents individual i interacts with. To the extent that these ratings flow to their final audience through agent i and thus are not anonymized, there are some reasons to suspect that they might by manipulated.[26]

(iii) Environment-specific information. The audience may not be aware of the level of externalities that were the objects of private interactions. Then the disclosure of $\{a_{ij}\}$ and of $\{s_{ij}\}$ substantially differ in their informational content, as the latter will in general reflect the magnitude of the externality.

(iv) Naivete. As is well-known, the presence of naive agents - taking the lack of disclosure at face value, without realizing that only low-contribution agents have something to hide- would generate less-than-full unraveling.

2.4.2 Signaling overall proclivity for doing good vs. playing favorites

In the model, an individual's type refers to her overall altruism or proclivity for doing good ("unknown absolute empathy"). While this is the right focus when it comes to modeling social ratings, this approach ignores the possibility that we do not value everyone equally.[27] The polar case of preferences posits a distribution of counterparty-specific altruism $v_z \in [0,1]$ where, w.l.o.g., v_z is increasing in $z \in [0,1]$. Let $\bar{v} \equiv \int_0^1 v_{\tilde{z}} d\tilde{z}$. Abusing notation, these preferences give rise to a distribution $F(v)$ over bilateral preferences, and we will denote $\Delta(v_z) \equiv M^+(v_z) - M^-(v_z)$. The function v_z is common knowledge and so there is no asymmetry of information about the individual's overall altruism; by contrast, individual i's specific rank order $j \to \omega_i(j) = z$ is private information and all permutations are equally likely. Thus individual i experiences empathy $v_{\omega_i(j)}$ for individual j. We call this case "unknown relative empathy". Note, first, that there cannot be any extension-of-reputation effect under unknown relative empathy. So, there will be no term in ν in the following expressions.

We saw that transparency increases prosocial contributions when the proclivity for doing good is unknown. This is not so when the distribution of this proclivity, rather than the proclivity itself, is unknown. To show this, let us look for equilibria in which individual i adopts a strategy that favours the agents she cares most about: $a_{ij} = 1 \Leftrightarrow v_{\omega_i(j)} \geq v^*$ for some v^*, or equivalently $a_{ij} = 1 \Leftrightarrow \omega_i(j) \geq z^*$ for some z^*, where $v_{z^*} = v^*$.[28]

[26] At the extreme, they might even reflect interactions that did not take place.

[27] The following follows closely the analysis of pandering in Maskin-Tirole (2019), and casts it in the context of social relationships rather than in that of an election, yielding new insights.

[28] Selecting this equilibrium requires a refinement (namely D1).

Under *silo reputations*, individual i maximizes over cutoff z

$$\int_z^1 \left[v_{\tilde{z}}e - c + \mu\left[\alpha M^+(v^*) + (1-\alpha)\bar{v}\right]\right] d\tilde{z} + z\mu\left[\alpha M^-(v^*) + (1-\alpha)\bar{v}\right],$$

yielding the same condition for the signaling intensity as for an unknown overall altruism. For example, for an interior equilibrium:

$$v^*e - c + \mu\alpha\Delta(v^*) = 0.$$

Suppose next that individual i's behavior is *transparent*, so that all counterparties j learn the fraction of prosocial acts individual i engages in. Individual i then solves

$$\int_z^1 \left[v_{\tilde{z}}e - c + \mu[\alpha M^+(v_z) + (1-\alpha)\bar{v}]\right] d\tilde{z} + z\mu[\alpha M^-(v_z) + (1-\alpha)\bar{v}] = \int_z^1 (v_{\tilde{z}}e - c)d\tilde{z} + \mu\bar{v}$$

where we use the martingale property for beliefs ($(1-z)M^+(v_z) + zM^-(v_z) = \bar{v}$). Hence, the optimal cutoff is given by (for an interior solution)

$$v^*e - c = 0. \tag{7}$$

Transparency prevents the individual from pretending to have more "good friends" than she really has (true good friends are those for whom $v_{\omega_i(j)} \geq c/e$).

Proposition 4 *(unknown relative empathy): Transparency reduces prosocial behavior when the individual's relative (rather than absolute) empathy is unknown.*

Comparison with the social optimum. As in the unknown absolute empathy case, let e^s and e^t denote the levels of externality for which equilibrium behavior is socially optimal under silo reputations and transparency, respectively: $e^s \equiv \mu\alpha\Delta(v^*)$, where $v^*e^s - c + \mu\alpha\Delta(v^*) = 0$. And $e^t \equiv 0$. Therefore if $e \geq e^s$, transparency is strictly suboptimal. There is a cutoff externality, which is strictly lower than e^s, such that transparency is optimal if and only if the externality lies below that threshold.

2.4.3 From there on

The rest of the paper builds on the framework developed in Section 2.1 to investigate less familiar themes. To simplify the expressions without loss of insights, we will assume away the better-measurement effect by positing perfect observation of the counterparty's action:

Assumption 1 *(perfect observability in bilateral relationships):* $\alpha = 1$

2.5 Public sphere dominance

We so far have assumed that either privacy obtains and so reputations are silo ones, or that behaviors are all public/transparent; and when comparing the two, we posited that all behaviors are potentially public. In practice, though, one can distinguish between private and public spheres. Some behaviors are bound to remain in the private sphere, either because their public disclosure would make the individual or the audience uncomfortable, or because they are unobservable by third parties and furthermore cannot be reliably rated: Outsiders may be unable to ascertain whether a rating within a maintained relationship (or to the contrary following an acrimonious separation) is genuine.[29] Other behaviors by contrast lend themselves to being shared in the public sphere if the individual, her environment or society decide to disclose them. This section focuses on the mutual interdependence between the private and public spheres, and on how an expansion in the public sphere impacts overall behavior and welfare.

Suppose that a fraction t of individual i's activities is transparent, while a fraction $s = 1 - t$ is private. In practice, this fraction t may be affected by the technological evolution (cameras, social networks, cheap data storage, artificial intelligence,...) as well as the social pressure for transparency.[30] In the "all public" ($t = 1$) case, the cutoff v^* over which behavior is prosocial is given by $v^*e - c + (\mu + \nu)\Delta(v^*) = 0$. The "all private" ($t = 0$) cutoff is given by $v^*e - c + \mu\Delta(v^*) = 0$. The two cutoffs coincide if and only if $\nu = 0$ (stable relationships).

For $t \in (0, 1)$, we again look for pure-, symmetric-strategy equilibria. The natural generalization of our pure-and-symmetric-strategy focus consists in uniform strategies within each sphere: Let $a^t \in \{0, 1\}$ denote the behavior in the public sphere ("t" stands for "transparent") and $a^s \in \{0, 1\}$ that in the private sphere ("s" stands for "silo").

Agent i chooses $(a^t, a^s) \in \{0, 1\}^2$ so as to solve:

$$\max_{(a^t,a^s)\in\{0,1\}^2} \left\{(v_i e - c)(ta^t + sa^s) + (\mu t + \nu)\hat{v}^t(a^t) + \mu s\hat{v}^s(a^t, a^s)\right\}$$

where $\hat{v}^t$ and $\hat{v}^s$ are the reputations in the public and private spheres. Let v^t and v^s denote the cutoffs in the public and private spheres: $a^r = 1$ if and only if $v \geq v^r$ where $r \in \{s, t\}$.

[29] Note that reputation still matters under such circumstances (to tame an hostile individual or to cement a friendship).

[30] It would be desirable to dig deeper into the microfondations of the distinction between public and private lives. Suppose for instance that the distinction is based on the reliability of ratings. If so, our writing of agent i's objective function presumes that a fraction t of stable relationships gives rise to reliable ratings, while other stable relationships do not. For example, in organizations the social behavior of employees may be evaluated by their colleagues or superiors, through comparative evaluations to curb biases toward grade inflation; by contrast, the reliability of ratings in close friendship or intimate relationships is questionable. Alternative assumptions may be entertained to derive variants of our analysis. For example, one might take the point of view that stable relationships are necessarily part of the private sphere while transient ones belong to the public sphere. The insights would then be very similar, with possible differences regarding comparative statics with respect to t.

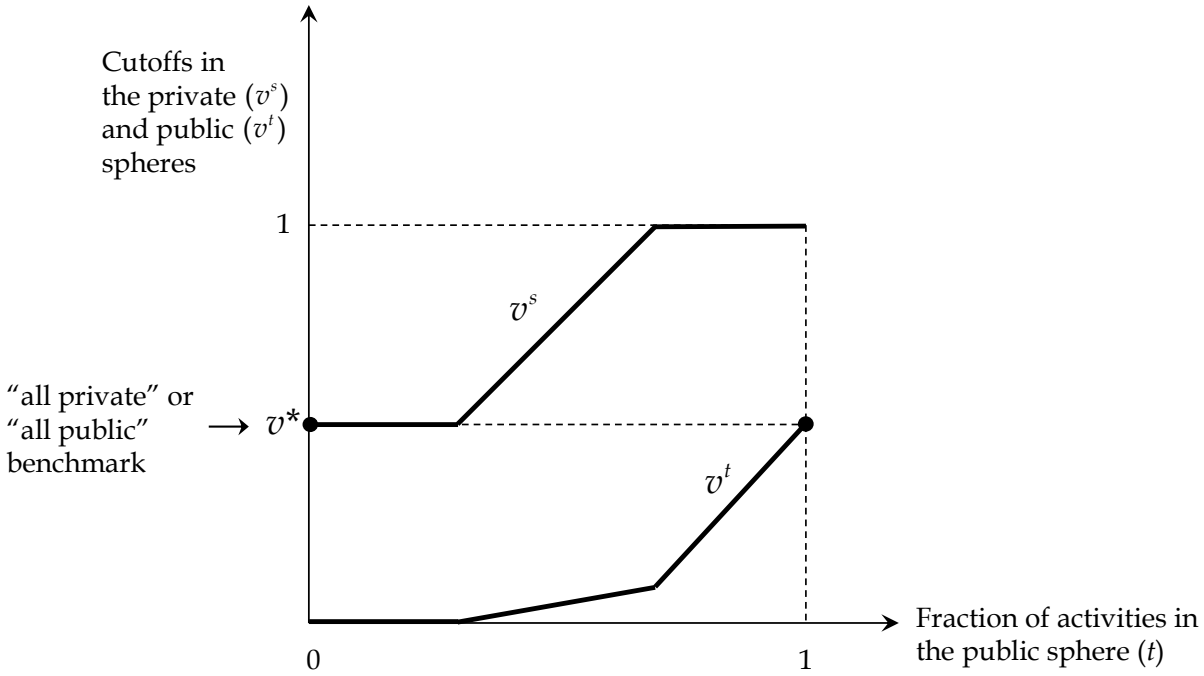

Figure 1: Equilibrium contributions under stable relationships ($\nu = 0$).

Proposition 5 *(public sphere dominance)*

(i) There exists an equilibrium satisfying $v^t < v^s$ and $v^t \leq v^* \leq v^s$. The co-existence of a public and a private spheres implies a misallocation of contributions between the two ($v^t < v^s$).

(ii) When the density f is non-increasing, t, this equilibrium is the unique such equilibrium. v^t and v^s are almost everywhere differentiable in t and

$$\frac{dv^t}{dt} \geq 0 \quad \textit{and} \quad \frac{dv^s}{dt} \geq 0 \qquad \textit{a.e.}$$

(iii) When the density f is single-peaked, multiple equilibria may coexist for a small enough public sphere. The monotonicity of v^t and v^s in t however still applies to stable equilibria.

(iv) An expansion of the public sphere may reduce the total contribution ($\bar{a}(t)$ may decrease with t).

For a narrow public sphere, signaling in the public sphere is cheap and, at least for $\nu > 0$, $v^t = 0$. As t grows, though, signaling in the public sphere becomes more expensive, and this cost effect (weakly) reduces contributions in the public sphere.

Part (i) of Proposition 5 shows that *effort is misallocated*: Some types contribute in the public sphere while higher types do not contribute in the private sphere. The

excessive attention to public behavior leads to a disintegration of the social fabric in the private one. But the split between private and public sphere also affects the total level of contributions:[31]

$$\bar{a}(t) \equiv (1-t)[1-F(v^s)] + t[1-F(v^t)]$$

(with $\bar{a}(0) = \bar{a}(1) = 1 - F(v^*)$ when $\nu = 0$). Hence, whenever differentiable,

$$\frac{d\bar{a}}{dt} = [F(v^s) - F(v^t)] - (1-t)f(v^s)\frac{dv^s}{dt} - tf(v^t)\frac{dv^t}{dt} \tag{8}$$

The first term in the RHS of (8) captures a substitution effect: Contributions are higher in the public sphere and so an extension of the public sphere raises the overall level of contributions. The other two terms capture the observation that contributions in both spheres decline with an expansion of the public sphere. The overall effect is in general ambiguous.[32]

These results may shed light on Stuart Russell's observation[33] that *"[under a system of intrusive monitoring and coercion] outward harmony masking inner misery is hardly an ideal state. Every act of kindness ceases to be an act of kindness and becomes instead an act of personal score maximization and is perceived as such by the recipient."* The second statement can be expressed mathematically as saying that when technology increases image concerns so "prosocial behavior" becomes more frequent, the glory attached to it ($M^+(v^*)$) decreases and truly generous motives pale relative to personal score maximization (the ratio of intrinsic motivation over image concerns decreases). More specific to this section is a tentative interpretation of the first sentence, which may be understood as a deterioration of behavior in the private sphere as technology expands the public sphere.

3 Leveraging social sanctions to consolidate political power

Let us now introduce a state eager to suppress political dissent or more generally to promote some form of compliance, and ask ourselves: can such a government use a social

[31] Welfare is given by $W(t) = t\int_{v^t}^{1}(ve - c + e)dF(v) + s\int_{v^s}^{1}(ve - c + e)dF(v)$.

And so $\frac{dW}{dt} = \left[\int_{v^t}^{v^s}(v - \hat{v})dF(v)\right]e - f(v^t)\frac{dv^t}{dt}t(v^t e - c + e) - f(v^s)\frac{dv^s}{dt}s(v^s e - c + e)$.

[32] Consider the range in which there is no contribution in the private sphere (t is high enough so that $v^s = 1$), though. In this range (when it exists),

$$v^t e - c + \frac{\mu + \nu}{t}\Delta(v^t) = 0,$$

and so

$$\frac{d\bar{a}}{dt} = 1 - F(v^t) - tf(v^t)\frac{dv^t}{dt} = 1 - \frac{c}{e} \quad \text{in the case of the uniform distribution.}$$

So, if $c > e$ the total contribution decreases with t when t is such that $v^s = 1$.

[33] (2019), page 106. Needless to say, the interpretation is mine and need not be the author's.

rating scheme in order to consolidate its power? To study this question, we 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. Here the government's only instrument is the control of the flow of information.

There are indeed concerns that autocratic regimes might use ratings to target dissidents,[34] 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. Autocratic regimes may also promote religious fervour or more generally want to force citizens to conform to their agenda.

Agent i now takes two actions:

1. An anti- or pro-social action $\{a_{ij}\}_{j\in[0,1]}$, where, as earlier, $a_{ij} \in \{0,1\}$ and is symmetric ($a_{ij} = a_{ik}$ for all (j,k)).
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, etc.

The agent's type is two-dimensional. As earlier, v_i, drawn from $F(\cdot)$ with strictly positive density 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 toeing the line, θ_i, distributed according to smooth cumulative distribution $G(\cdot)$ with density $g(\cdot)$. For simplicity, v_i and θ_i are independent.

As earlier, action $\{a_{ij}\}$ is observed by counterparties, but not by future new partners. We assume for the moment that b_i is only observed by the state. We will later note that little (nothing if supp $G = \mathbb{R}^+$) is changed if b_i is observed.

Next, we posit that in their bilateral relationships, agents put more weight on their partner's prosociality than on her attitudes toward the state's agenda; we capture this in a stark way by assuming that agent i care solely about her reputation(s) regarding her prosocial type. Implicitly, other agents stress her reliability and are indifferent to her feelings toward the government or her personal tastes. Thus, agent i's objective is

$$u_i = \int_0^1 [(v_i e - c)a_{ij} + \mu \hat{v}_i(I_{ij})]dj + \nu \hat{v}_i(I_i) - \theta_i b_i,$$

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

Government's objective function. To express the government's concern about dissent, let its objective function be

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

[34] 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."

When $\gamma = 0$, the government is benevolent (internalizes only the agents' welfare W). The higher γ, the more autocratic the government.

Our 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. In this example and more broadly in environments where dissent exhibits a strength in numbers, the second term in the government's objective function could 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 4.

Unbundling benchmark. We first look at the case in which the government releases agent i's behavior in the two realms. Because the θ_i-reputation is irrelevant in private relationships,

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

And agent i chooses $a_i = 1 = a_{ij}$ for all j if and only if

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

$\Longleftrightarrow$

$$v_i e - c + (\mu + \nu)\Delta(v_u^*) \geq 0$$

where v_u^* is the cutoff under unbundling (to be taken equal to 1 if it exceeds 1, or to 0 if it is negative); let $\Delta_u \equiv \Delta(v_u^*)$.

Proposition 6 *(unbundling). When the government separately releases behaviors (a_i, b_i) in the two domains, then each individual is confronted with two distinct decision problems:*

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

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

The outcome is the same as if the state released only $\{a_i\}$.

Bundling. We 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. It conditions a good rating not only on a good social behavior, but also on toeing the line:

$$\begin{cases} 1, & \text{with associated reputation } \hat{v}_1, \text{ if } a_{ij} = 1 \text{ for all } j \text{ and } b_i = 1 \\ 0, & \text{with associated reputation } \hat{v}_0, \text{ otherwise.} \end{cases}$$

We consider sequentially the cases of transient and stable relationships.

3.1 Transient relationships

Let us assume, first, that $\mu = 0$ and $\nu > 0$. For expositional simplicity, we further assume that $c \geq e$. This assumption implies that image concerns are required in order to generate prosocial behavior.[35]

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

$$\begin{cases} a_i = b_i = 1 & \text{iff } v_i e - c + \nu(\hat{v}_1 - \hat{v}_0) \geq \begin{cases} \theta_i & \text{if } \theta_i \geq 0 \\ 0 & \text{if } \theta_i < 0 \end{cases} \\ a_i = b_i = 0 & \text{iff } v_i e - c + \nu(\hat{v}_1 - \hat{v}_0) < \theta_i \text{ and } \theta_i > 0 \\ a_i = 0, b_i = 1 & \text{iff } v_i e - c + \nu(\hat{v}_1 - \hat{v}_0) < 0 \text{ and } \theta_i \leq 0 \end{cases} \tag{9}$$

Let $v_b^*(\theta_i)$ denote the cutoff under bundling for a given θ_i (with again the convention that it is 1 if the solution to (9) with equality exceeds 1, and 0 if the solution is negative). This threshold is weakly increasing, as depicted in Figure 2.

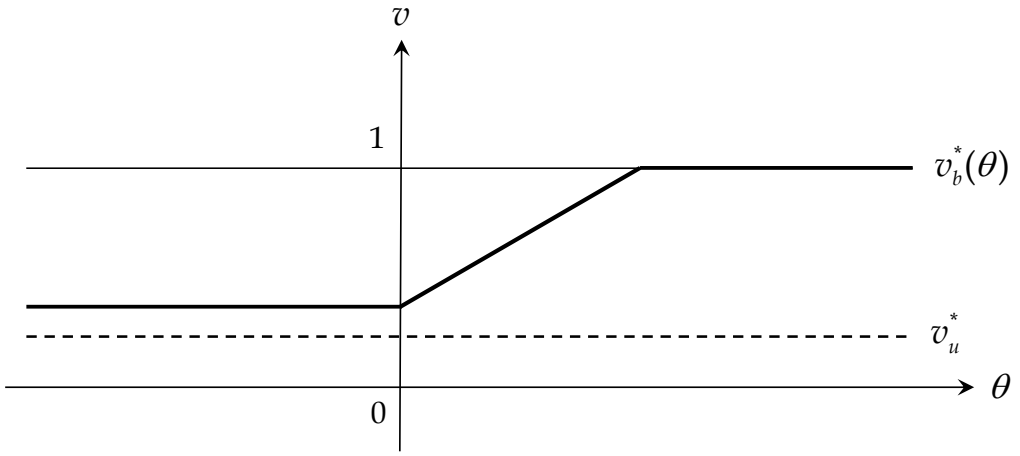

Figure 2: behavior under bundling and unbundling

Let $g_1(\theta)$ and $g_0(\theta)$ denote the conditional densities.[36] $g_1(\theta)/g(\theta)$ and $g_0(\theta)/g(\theta)$ are weakly decreasing and increasing in θ, respectively.

[35] As $ve - c \leq 0$ for all $v \in [0, 1]$.

[36] $g_1(\theta) = \dfrac{g(\theta)[1 - F(v_b^*(\theta))]}{\int g(\tilde{\theta})[1 - F(v_b^*(\tilde{\theta}))]d\tilde{\theta}}$ and $g_0(\theta) = \dfrac{g(\theta)F[v_b^*(\theta)]}{\int g(\tilde{\theta})F(v_b^*(\tilde{\theta}))d\tilde{\theta}}$.

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 equilibrium is particularly easy to characterize in the case of a norm ($\Delta' \leq 0$): image concerns are reduced by bundling and the provision of prosocial behavior is smaller across the board (for all θ). The intuition goes as follows:

(i) The cost θ of toeing the line (when positive) acts as an "excuse" for not contributing. Indeed for high θ, the conditional reputation[37] when not acting prosocially is the prior mean $\bar{v}$, the highest possible reputation in the absence of contribution.

(ii) This cost 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 reduces the fraction of dissenters from $1-G(0)$ to $\int_0^\infty g(\theta)F(v_b^*(\theta))d\theta$. A revealed preference argument implies that a more autocratic ruler is therefore 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 $W \equiv E[(ve - c + e)a(v,\theta) - \theta b(v,\theta)]$; 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 behavior with respect to identity ($b_i = 1$ when $\theta_i > 0$). So $W_u > W_b$. We collect these results and further characterizations in the next proposition:

Proposition 7 *(bundling under transient relationships). Consider a society with transient relationships ($\mu = 0 < \nu$) and assume that $\Delta' \leq 0$. Under bundling, there exists an equilibrium satisfying:*

(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 $v_b^*(\theta) > v_u^*$ and depicted in Figure 2: All types θ behave less prosocially). All equilibria satisfy this property when the distribution F is uniform (so $\Delta' \equiv 0$).

(ii) The social 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) There is (weakly) too much use of bundling (as $W_u > W_b$) if there is underprovision of prosocial behavior under unbundling (i.e. $e \geq \nu\Delta(v_u^*)$).

[37] Recall that θ is not observed by the agent's "audience". But a lack of prosocial behavior might come from a strong aversion to being the line.

This behavior illustrates the trade-off faced by the government: Bundling reduces dissent, but imposes collateral damages on private relationships by reducing pro-social behaivor. In the end, the resolution of this trade-off hinges on how autocratic the regime is (γ).

Example. Suppose that G puts weight on two types, θ_L (probability ρ) and θ_H (probability $1-\rho$). 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 ρ) (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 toe the line, and less likely to be attributed to low ethical standards); and (b) makes the loss of prosocial behavior of high-identity types more pregnant. The implication is that, assuming an underprovision of prosocial behavior, bundling is optimal for the government if there is not too much potential dissent (the people are expected to be docile).[38]

Discussion

(a) *The need to centralize social ratings*

Consider an autocratic government with $\gamma \geq \gamma^*$. To accomplish its goal, it must not share its prerogative. For, suppose that the private sector could collect the same data and publicly issue 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. This competition would lead to de facto unbundling, and no-one would pay attention to the government's social score.[39]

(b) *Commitment*

A common justification given for being wary of state-controlled social scores is their

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

- a loss on θ_H types equal to $(1-\rho)[\int_{v_u^*}^1 (ve - c + e)dF(v)]$
- a net compliance gain on the θ_L types equal to $\rho[[1-F(v_b^*)]\gamma - \int_{v_u^*}^{v_b^*}(ve - c + e)dF(v)]$.

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

[39]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."

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 contribution 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 that the method of computation is disclosed and common knowledge.

Contrary to what is occasionally asserted, it may actually be essential that the algorithm be transparent. For, suppose that the government does not commit to a method of computation and decides ex post on the rating to be given to each agent. The government may for instance take revenge against, and give a low rating to (perhaps a fraction of) citizens having expressed dissent ($b_i = 0$).[40] 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, suppose that the government's objective is

$$W_G = W + \gamma E[b_i] - \varepsilon \int [\nu \hat{v}(v,\theta)]\xi(\theta)g(\theta)f(v)d\theta dv, \tag{10}$$

where ε can be arbitrarily small, 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[41] and (iii) the choice of b is identical to that under unbundling.

To show this, suppose that the government ignores a in its construction of its social score (so the ratings, $\hat{v}_{b=1}$ and $\hat{v}_{b=0}$, depend 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 $\hat{v}_{\max} = \hat{v}_{\min} = \bar{v}$.

Proposition 8 *(time-consistent ratings). If the algorithm computing the social score is itself opaque and the government has a distaste for opponents (as expressed in (10)), then there exists an equilibrium in which the social score is ineffective and the outcome is the same as under unbundling.*

[40] The social welfare function ($W_g = W + \gamma E[b_i]$) is silent on the government's preferences once actions have been selected. We can for example presume 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 ($b_i = 0$). For $\rho > 0$, those who have chosen $\{a_i = 0, b_i = 1\}$ are those who are most supportive of the state (who have the lowest expected θ_i). They have a low v_i. So the state can hardly reward them with a better reputation.

[41] As a consequence, if $c \geq e$, all choose $a = 0$.

(c) *Observable pro/anti government action* (b_i)

Suppose now that an individual's choice of b_i (but not that of a_i) is observed by the audience.

Note, first, that the analysis is unchanged if $\theta_i \geq 0$ for all i (i.e. support $G = \mathbb{R}^+$); for, in the case of unobservable b_i studied so far, there were then 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 rating.

If the support of G includes negative values of θ_i as well, the analysis must be amended, but retains its main features. 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; 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 $\nu(\hat{v}_{00} - \hat{v}_{01}) + \theta^* = 0$. We claim that $\theta^* < 0$. Indeed, the corresponding cutoffs satisfy,[42] 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. 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.

(d) *Caring about identity*

The 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 not only about appearing a desirable match, but also about the identity of her future matches. To fathom potential implications of this remark, suppose that image concerns can be summarized by $\nu[v - \kappa d(\theta_i, \theta)]$, where $\{v, \theta\}$ is the type of the (typical) partner she will be matched with, d measures a distance in the identity space, and ν embodies discounting, number of future partners and the importance attached to their attributes (so the model studied previously is a special case, with $\kappa = 0$ and matches determined by assortative matching: see footnote 14). Note first that absolutely nothing is changed in the unbundling case: from assortative matching, contributors match with partners of average type $M^+(v^*_u)$, and non-contributors with partners of average type $M^-(v^*_u)$; furthermore, it is incentive compatible within these two populations to announce truthfully one's identity and so $d(\theta_i, \theta) = 0$. The analysis is much more complex under bundling, as announcing one's identity truthfully is not incentive compatible: Regardless of one's prosocial choice, there is an incentive to appear

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

as a dissenter so as to manufacture an excuse (if $a_i = 0$) or claim merit (if $a_i = 1$). In general there will be some mismatch, which represents another cost of bundling.[43]

(e) *Social rating popularity*

We observed that the possibility of bundling, if employed, reduces social welfare. The actual impact of bundling is of course type-specific; high-θ types are relatively (weakly) 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.[44] When introducing a social score that allows for bundling, the government will accordingly stress the benefits in terms of bridging the trust gap, especially if suspicions among citizens and between individuals and businesses are running high.

3.2 Stable relationships

Suppose now that relationships are sustained rather than transient ($\mu > 0$, $\nu = 0$). We argue that the state will find it much more difficult to leverage even a monopoly provision in social scoring to consolidate political power in a society of tight knit relationships than in a society of strangers. 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 9 *(ineffectiveness of bundling in a tight knit 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 outcome is the same as when there is no social rating. This equilibrium is robust to D1. The bundling equilibrium akin to that under transient relationships is not robust to D1.*

We only sketch the proof. Agent j's posterior belief about i is $\hat{v}_{ij} = M^+(v^*)$ if $a_{ij} = 1$ and $\hat{v}_{ij} = M^-(v^*)$ if $a_{ij} = 0$, regardless of what the state reports, where $v^*e - c + \mu\Delta(v^*) = 0$. 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 $\theta_i > 0$.

This equilibrium is trivially robust to refinements as all behaviors $\{a_i, b_i\} \in \{0, 1\}^2$ are equilibrium-path behaviors. By contrast, the social scoring equilibrium of Proposition

[43] I have not solved for equilibrium. The analysis is feasible for two levels of identity, but seems rather intricate with a continuum of identities.

[44] To see this, let E denote the aggregate externality (a minorant of E is $G(0)[1 - F(v_b^*(0))]e$); then if $E \geq \mu\bar{v}$, everyone gains, as individuals receive payoff $\mu\bar{v}$ in the absence of social rating.

7 (replacing μ by ν) is not robust. The behavior $\{a_i = 1, b_i = 0\}$ is off the equilibrium path. The type that would benefit most from a deviation to the off-path behavior can be shown to be $v_i = 1$ and θ_i sufficiently large that $v_i = 1$ does not contribute under bundling.[45] With such beliefs, this type is better off deviating.

Remark. An imperfect observability of bilateral behavior ($\alpha < 1$) may reinstate a role for social ratings, bringing the analysis closer to that for a society of strangers (Section 3.1). Similarly, when the types v_i and θ_i are correlated, b_i is informative given a_i (this can be seen easily in the case of unbundling). The broad picture therefore is that bundling is less effective, but not necessarily inoperative, under stable relationships.

3.3 Divisive issues

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, we 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 that majority members do not interact with minority members' and so are just concerned with the minority members toeing the line:

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

Here γ reflects the majority's pure aversion to the minority living according to its preferences. 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 $\nu\hat{v}_i$ (relationships are transient). Assuming as in the rest of this section 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 3.1. Because we assume that the majority cares only about the minority's toeing the line and does not interact with it, it bundles for

[45] More formally all θ_i such that $v_b^*(\theta_i) = 1$.

all $\gamma > 0$.[46] More generally, if the majority puts some weight on the minority's welfare, bundling occurs for γ above some threshold $\gamma^* > 0$.

Observation (divisive issues). The insights of this section apply to environments in which a political majority disapproves of a minority's behavior or expression of opinion. In particular, bundling will be observed when the majority is sufficiently averse to minority's preferred behavior relative to their prosocial behavior.

4 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 in totalitarian regimes, as people are afraid of being seen in company of dissidents or mere citizens whose lifestyle is frowned upon by the regime.[47] 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 induce yet another social pressure -ostracism- on citizens to toe the line.[48] To see how this can work, consider the following, *sequential* choice variant of the model of section 3, and depicted in figure 3.

(1) At "stage 1", agents pick their actions $\{b_i\}$. Agent i's choice is observed by the state as well as the other agents whom she will potentially interact with at stage 2

[46] We may reprove the oppression of minorities by majorities either on the ground that the benefit of oppression for the majority is smaller than the cost suffered by the minority, or on the rationale of insurance behind the veil of ignorance.

[47] 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"

[48] While we stress ostracism between citizens, we 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.

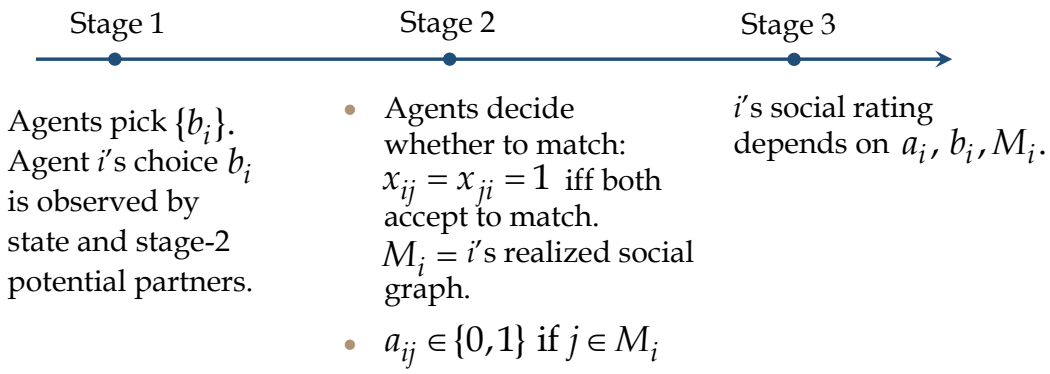

Figure 3: timing (guilt by association)

(but not by her stage-3 audience).[49]

(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 ("it takes two to tango"). Let $x_{ij} = x_{ji} = 1$ if the i-j match is realized and $x_{ij} = x_{ji} = 0$ otherwise.[50] If matched, they pick actions a_{ij} (which, we again assume, is the same for all matched partners). The state observes the actual matches. Let M_i denote individual i's realized (as opposed to potential) matching set or social graph.

(3) The state issues a binary social rating for each individual i 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 (receives reputation $\hat{v}_i = \hat{v}_0$) 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).[51]

Agent i receives rating $\hat{v}_1$ otherwise.[52]

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

[50] We 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.

[51] 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:

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

Assuming that an individual can be tainted directly or indirectly streches the simultaneity assumption somewhat. One can think of this assumption as a stability requirement: If at stage 2, an individual i who chose $b_i = 1$ at stage 1 matches with an individual j having chosen $b_j = 0$, then all agents k such that $x_{ik} = 1$ would discontinue their relationship with i so as to avoid being tainted.

[52] We can assume that the future partners in this transient-relationships environment do not observe b_i.

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. We will label this policy "social-graph-inclusive bundling" or "all-inclusive bundling", as opposed to the "simple bundling" policy studied in Section 3 and the "unbundling" policy of Sections 2 and 3.

The payoff function of individual i is

$$u_i = \int_j x_{ij}[(v_i e - c)a_{ij} + e a_{ji} + \varepsilon]dj - \theta_i b_i + \nu \hat{v}_i,$$

where $\varepsilon > 0$ is a fixed benefit per interaction.[53] We still assume that v_i and θ_i are uncorrelated and that interactions are transient (hence the use of the notation ν for image concerns), as Sections 3.1 and 3.2 have shown that transient relationships make bundling particularly powerful.

We 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 receiving a low social score dissent. As in Section 3, we further assume that $c \geq e$ for expositional simplicity. This assumption guarantees that an individual without image concerns will not choose $a_i = 1$. 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 $a_i \equiv 0$.

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, θ) really has only three choices, depicted in Figure 4.

(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

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

(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:

$$u_i^2 = X(ve - c + \varepsilon) + X_1 e - \theta + \nu \hat{v}_1.$$

(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

$$u_i^3 = \varepsilon + X_1 e - \theta + \nu \hat{v}_0.$$

[53] 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 ε 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.

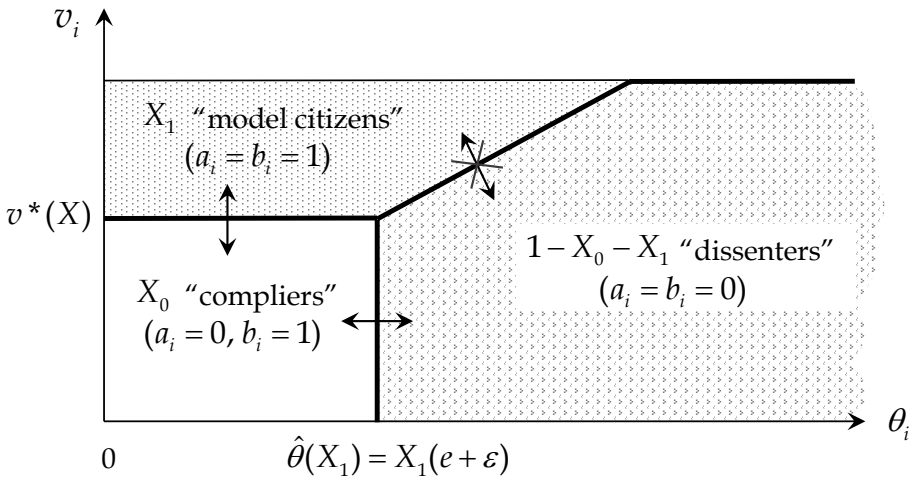

Figure 4: behavior under guilt by association

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

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

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

$$X_1(e + \varepsilon) \geq \theta.$$

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

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

Letting

$$v^*(X) \equiv \max\left\{\min\left\{\frac{c}{e} + \frac{(1-X)\varepsilon - \nu(\hat{v}_1 - \hat{v}_0)}{Xe}, 1\right\}, 0\right\}$$

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

$$\hat{\theta}(X_1) \equiv X_1(e + \varepsilon),$$

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). 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 ε plays no role. Furthermore, because we 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

$$v^*(\theta)e - c + \nu(\hat{v}_1 - \hat{v}_0) - \theta = 0 \tag{11}$$

where $\hat{v}_1$ and $\hat{v}_0$ are computed as in Section 3.1.

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 state:

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

$$X_1(1 - X)(2\varepsilon).$$

(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) *Less dissent.* $E[b_i] = X$ is higher when the social graph is used in the social score provided that[54]

$$G(\hat{\theta}(X_1)) + \int_{\hat{\theta}(X_1)}^{\infty} [1 - F(v^*(X, X_1, \theta))]dG(\theta) \geq \int_0^{\infty} [1 - F(v^*(\theta))]dG(\theta) \tag{12}$$

[54]A sufficient condition for (12) for F uniform to be satisfied is

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

Proposition 10 *(social graph) Guilt by association makes high-score agents ostracize low-score ones. 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: Incorporating the social graph into the social score is more appealing to an autocratic ruler provided that (12) is satisfied (as in Example 1 below).*

Example 1. Suppose that G puts weight only on two types θ_L (probability ρ) and θ_H (probability $1-\rho$). Identity θ_H is strong enough that the individual always picks $a_i = b_i = 0$ regardless of v_i. By contrast, we 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

$$v_b^* e - c + \nu\Delta_b = \theta_L,$$

where

$$\Delta_b(\rho, v_b^*) \equiv M^+(v_b^*) - \left[\frac{\rho F(v_b^*)M^-(v_b^*) + (1-\rho)\bar{v}}{\rho F(v_b^*) + 1 - \rho}\right].$$

When tainting is added to bundling and type θ_L chooses to comply ($b_i = 1$ for all v_i), the cutoff $v_{b,t}^*$ is given by

$$\rho(v_{b,t}^* e - c) + \nu\Delta_b(\rho, v_{b,t}^*) = 0$$

provided that

$$(2\rho - 1)\varepsilon + \rho[1 - F(v_{b,t}^*)]e \geq \theta_L.$$

Tainting raises $E[b_i]$ from $\rho[1 - F(v_b^*)]$ to ρ. It thus appeals more to a more autocratic ruler.

Example 2. Let us return to the continuum-of-types case with θ distributed on $[0, \theta_{\max}]$, and look at some simple equilibria with an amorphous population ($X = 1$) and an all-dissenter one ($X = 0$), respectively.

Suppose, first, that

$$\theta_{\max} \leq [1 - F(v_u^*)](e + \varepsilon) \tag{13}$$

where, as earlier, $v_u^* e - c + \nu\Delta(v_u^*) \equiv 0$. We 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 (13) guarantees that the individual does not gain from dissenting and thereby being ostracized by model citizens.[55]

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 $\varepsilon + \nu\bar{v}$ from her equilibrium behavior

[55] For $v_i \leq v_u^*$

$$[1 - F(v_u^*)]e + \varepsilon + \nu M^-(v_u^*) - \theta \geq F(v_u^*)\varepsilon + \nu M^-(v_u^*)$$

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

($\hat{v}_0 = \bar{v}$). If she picks ($a_i = 0, b_i = 1$), her utility is lower for all θ, as we already noted: $\varepsilon - \theta + \nu\hat{v}_0 = \varepsilon - \theta + \nu\bar{v} < \varepsilon + \nu\bar{v}$. Picking ($a_i = 1, b_i = 0$) yields $v_i e - c + \varepsilon + \nu\hat{v}_0 < \varepsilon + \nu\bar{v}$. Finally obtaining reputation $\hat{v}_1$ requires isolation[56] and yields at most $\nu\hat{v}_1 \leq \nu \cdot 1$. So if

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

all dissent. This example also illustrates the possibility of multiple equilibria due to endogenous network externalities, as its condition of existence is compatible with that, (13), of the amorphous equilibrium.

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

$$\varepsilon - \theta_i + \max\{v_i e - c + \nu M^+(v_u^*), \nu M^-(v_u^*)\}.$$

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, we 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 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.

5 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. We show this by using a framework that is a relabeling of the one in 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.[57]

[56] 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 $b_i = 1$.

[57] 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 force official's compliance.

There is one private platform –or equivalently an arbitrary number of private platforms controling access to "unique viewers".[58] 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.[59]

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 reputation $\hat{\theta}_i$.[60] 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 duality 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 $E[W]$.[61]

Proposition 11 *(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 rating to suppress citizen's dissent.*

6 Concluding remarks

The main insights were summarized in the introduction. In these concluding remarks, we therefore focus on 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

[58] 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.

[59] 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).

[60] This strong assumption is made only for convenience and can be relaxed.

[61] 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.

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 still to be identified, but the analysis in this paper suggests leaving out information about divisive issues- in particular those from which the state, a majority or a platform could derive gains from-, and about the social graph. It also offers to monitor platforms' foray into political coverage unless platform regulation is performed by one or several entirely independent agencies.

Other challenges concern the weights to be put on behaviors we deem worthy of inclusion into such a score (imagine a ruler who is much more preoccupied with jaywalking than with corruption),[62] and how to account for the imperfect reliability of ratings or more generally observability of individual behaviors. Our study of the private and public spheres only touched on the latter issue. 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 shun others. Like in Section 4, but through a different channel, the social fabric and its valuable relationships may be destroyed.

We treated the "government" as a unitary actor. We 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 of principals with heterogenous goals[63] and 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 even have a social score in the first place, or to the contrary the various dimensions of our lifes should remain segmented, with for instances one's credit history affecting only one's ability to obtain future credit or enter financial transactions more broadly. It is interesting to note in this respect that law obeys mostly a silo, case-by-case approach and not social score precepts. We leave these

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

[63]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.

issues and the many other important questions associated with a principle-based design of privacy law for future research.

References

Acemoglu, D., and A. Wolitzky (2016), "Sustaining Cooperation: Community Enforcement vs. Specialized Enforcement," mimeo MIT.

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 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.

Banks, J., and J. Sobel (1987), "Equilibrium Selection in Signaling Games," *Econometrica*, 55(3): 647–661. DOI: 10.2307/1913604.

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., and J. Tirole (2012), "Laws and Norms", mimeo.

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

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 (2019) "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.

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

Cho, I.K., and J. Sobel (1990), "Strategic Stability and Uniqueness in Signaling Games," *Journal of Economic Theory*, 50: 381–413.

Chorzempa, M., Triolo, P. and S. Sacks (2018), "China's Social Credit System: A Mark of Progress or a Threat to Privacy?," Peterson Institute for International Economics, Policy Brief 18-14.

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.

Diamond, P. (2006), "Optimal Tax Treatment of Private Contributions for Public Goods with and without Warm Glow Preferences," *Journal of Public Economics*, 90(4-5): 897–919.

Eggers, D. (2013), *The Circle: A Novel,* Alfred A. Knopf.

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.

Goldfarb, A., and C. E. Tucker (2011), "Privacy Regulation and Online Advertising," *Management Science*, 57(1): 57–71.

Grossman, S., and O. Hart (1980), "Disclosure Laws and Take-Over Bids," *Journal of Finance*, 35: 323–334.

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

Hvistendahl, M. (2017), "Inside China's Vast New Experiment in Social Ranking," *Wired* December 14.

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.

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.

Maskin, E., and J. Tirole (2019), "Pandering and Pork Barrel Politics", *Journal of Public Economics*, 176: 79–93.

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

Milgrom, P. (1981), "Rational Expectations, Information Acquisition, and Competitive Bidding," *Econometrica* 49(4): 921–943.

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

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.

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

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

Appendix

Proof of Proposition 3.

Example of non-unraveling. Let $d_i \in \{D, ND\}$ denote the strategy of disclosing or not disclosing one's social score. And consider the following pure, symmetric strategies in which there is no disclosure and $a_{ij} = 1$ if and only if $v_i \geq v^*$ where $v^*e - c + \mu\alpha\Delta(v^*) = 0$. Disclosure not being on the equilibrium path, let other agents (interacting with j or not) formulate very pessimistic beliefs $\hat{v} = 0$ in case of disclosure. Then no type of agent i wants to deviate from non-disclosure.

Refined equilibrium. Suppose again that there is no disclosure. Equilibrium utilities for types below and above v^*, respectively, are:

$$U_0 \equiv \mu[\alpha M^-(v^*) + (1-\alpha)\bar{v}] + \nu\bar{v}$$

and

$$U_1(v) \equiv ve - c + \mu[\alpha M^+(v^*) + (1-\alpha)\bar{v}] + \nu\bar{v}$$

where $v^*e - c + \mu\alpha\Delta(v^*) = 0$. Now consider a particular deviation to $a_{ij} = 1$ for all j and disclosure of $a_i (= \alpha)$. Suppose that this deviation gives rise to posterior beliefs $\{\hat{v}_{ij}\}_{j\in[0,1]}$ and $\hat{v}_i$ for interacting and non-interacting agents, respectively. Then the gain (or loss if negative) from the deviation is:

$$\begin{cases} \mu\left[\int_0^1 \hat{v}_{ij}dj - \alpha M^+(v^*) - (1-\alpha)\bar{v}\right] + \nu(\hat{v}_i - \bar{v}) \ \text{ for } v_i \geq v^* \\ (v_i e - c) + \mu\left[\int_0^1 \hat{v}_{ij}dj - \alpha M^-(v^*) - (1-\alpha)\bar{v}\right] + \nu(\hat{v}_i - \bar{v}) \\ \qquad = \mu\left[\int_0^1 \hat{v}_{ij}dj - \alpha M^+(v^*) - (1-\alpha)\bar{v}\right] + \nu(\hat{v}_i - \bar{v}) - (v^* - v_i)e \ \text{ for } v_i < v^*. \end{cases}$$

Thus the set of posterior beliefs that make type v_i better off relative to her equilibrium payoff is the same for all types above v^* and is larger than the same set for types lower than v^*. Thus according to D1, types below v^* should be ruled out when observing this deviation and forming posterior beliefs. If we further keep invariant the conditional beliefs of those ($v_i \geq v^*$) who equally gain from all favorable interpretations, then $\hat{v} = M^+(v^*)$ and types $v_i \geq v^*$ gain $[\mu(1-\alpha) + \nu][M^+(v^*) - \bar{v}]$. ■

Public and private spheres

Lemma 1 *A sufficient condition for there to always be more contributions in the public sphere ($v^s \geq v^t$) in any equilibrium is that the density of the type distribution be non-increasing (e.g. uniform).*

Two remarks are in order. First, the sufficient condition in Lemma 1 is much stronger than needed. It can be relaxed more, the higher the relative size of the private sphere

(s/t) and the fraction of new partners (ν/μ). Second, when the density is not non-increasing, image concerns may increase with the level of prosocial contributions, and multiple equilibria may arise; we will return to this point in due time.

Proof of Lemma 1. Suppose to the contrary that $v^s < v^t$ and let $M(v_0, v_1)$ denote the mean of v over the interval $[v_0, v_1)$.

Behavior in the private sphere, being unobservable except to the counterparty, does not impact the reputation in the public sphere. So, for any $v_i \in [v^s, v^t)$,

$$s\left[(v_i e - c) + \mu[M(v^s, v^t) - M^-(v^s)]\right] \geq 0.$$

Similarly the fact that in this interval, agents do not want to contribute publicly implies that:

$$t(v_i e - c) + \mu s[M^+(v^t) - M(v^s, v^t)] + (\mu t + \nu)[M^+(v^t) - M^-(v^t)] \leq 0.$$

These two inequalities are inconsistent if

$$M(v^s, v^t) - M^-(v^s) < \left[1 + \frac{\nu}{\mu t}\right][M^+(v^t) - M^-(v^t)] + \frac{s}{t}[M^+(v^t) - M(v^s, v^t)].$$

The condition is satisfied in particular (for $s > 0$) if for $v^s < v^t$

$$M^+(v^t) - M^-(v^t) \geq M(v^s, v^t) - M^-(v^s). \tag{6.1}$$

Inequality (6.1) is satisfied at $v^s = v^t$ (since $M^+(v^t) \geq v^t$). Furthermore, applying Jewitt's lemma on $[0, v^t]$, $M(v^s, v^t) - M^-(v^s)$ is non-decreasing if the density f is non-increasing. ■

We now look for equilibrium conditions with $v^t \leq v^s$ (there are more contributions in the public sphere). As we will see, such an equilibrium exists, regardless of whether the density condition in Lemma 1 is satisfied).

(i) Private sphere:

$$\text{Either} \quad v^s e - c + \mu[M^+(v^s) - M(v^t, v^s)] = 0 \quad (v^s < 1) \tag{6.2a}$$

$$\text{or} \quad e - c + \mu[1 - M^+(v^t)] \leq 0 \quad (v^s = 1). \tag{6.2b}$$

(ii) Public sphere:

$$\text{Either} \quad t[v^t e - c] + (\mu t + \nu)\Delta(v^t) + \mu s[M(v^t, v^s) - M^-(v^t)] = 0 \quad (v^t > 0) \tag{6.3a}$$

$$\text{or} \quad t(-c) + (\mu t + \nu)\Delta(0) + \mu s[M^-(v^s)] \geq 0 \quad (v^t = 0). \tag{6.3b}$$

The expressions for interior equilibria reflect the trade-off between the warm-glow and cost attached to contributing and the image benefits in the public and privates spheres. The corner conditions can be derived in either of two ways: (a) as limits of interior solutions, or (b) applying the D1 refinement to compute posterior beliefs when a behavior has zero probability.

Stable relationships. To facilitate the comparison with Section 2, let us first assume that relationships are stable ($\nu = 0$). Given our assumption that counterparties observe behavior perfectly ($\alpha = 1$), Propositions 1 and 2 imply that "all silo" and "all transparent" deliver the same behavior, with a cutoff v^*, which we assume interior and therefore given by

$$v^* e - c + \mu\Delta(v^*) = 0. \tag{6.4}$$

Let us rewrite (8) and (9) in the following manner:

$$\text{Either} \qquad v^s e - c + \mu\Delta(v^s) - \mu[M(v^t, v^s) - M^-(v^s)] = 0 \quad (v^s < 1) \tag{6.5a}$$

$$\text{or} \qquad e - c + \mu[1 - M^+(v^t)] \leq 0 \quad (v^s = 1). \tag{6.5b}$$

And (using $s = 1 - t$)

$$\text{Either} \quad v^t e - c + \mu\Delta(v^t) + \mu\Big(\frac{1-t}{t}\Big)[M(v^t, v^s) - M^-(v^t)] = 0 \quad (v^t > 0) \tag{6.6a}$$

$$\text{or} \qquad -c + \mu\bar{v} + \mu\Big(\frac{1-t}{t}\Big)M^-(v^s) \geq 0 \quad (v^t = 0). \tag{6.6b}$$

The comparison of (6.5a) and (6.6a) with (6.4) is instructive. Unless $v^t = 0$, reputation concerns in the private sphere are weaker than in the all-silo or all-transparent cases, as contributions in the public sphere dampen the negative impact of the absence of contribution in the private sphere (see the last term in the LHS of (6.5a)). Conversely, unless $t = 1$, spillovers into the private sphere raise reputation concerns in the public sphere as demonstrated by the last term in the LHS of (6.6a).

The equilibrium levels of contributions, as defined by v^s and v^t in the private and public spheres, are depicted in Figure 1, assuming that

$$e - c + \mu[1 - M^+(v^*)] \leq 0,$$

in which case $v^s = 1$ and $v^t = v^*$ for t sufficiently large.

Condition (11) implies that image concerns are weak in the private sphere: having contributed in the public sphere already sets the individual apart from the chaff, reducing the image benefit from a pro-social behavior in the private sphere[64]. Furthermore, behavior is less pro-social in the private sphere when it is less prosocial in the public sphere, as having contributed in the public sphere is more of a mark of distinction.

[64]This result is closely related to Adriani and Sonderegger (2019)'s general theme that thinner tails decrease signaling incentives. Indeed, they note that truncating a distribution reduces signaling concerns. Here the truncation operates through the release of behavior in the public sphere.

Condition (12) shows that, in contrast, signaling concerns in the public sphere are magnified by the presence of a private sphere. The additional term is proportional to $\mu\frac{1-t}{t}$, and thus is particularly large when the public sphere is small ("cheap signaling"). This means that one cannot guarantee that (6.6a) has a unique solution, unless $f' \leq 0$ and so the additional term is non-increasing with prosocial behavior in the public sphere. Nonetheless, it can be shown that the prosocial behavior in the public and private spheres is decreasing in t at stable equilibria.

Proof of Proposition 7(i).

Consider an arbitrary image benefit $\Delta_b \in [0, \Delta_u]$ (from an equilibrium behavior as depicted in Figure 2). 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)[65] 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

$$\Delta_b = \Delta_u + \int \left[\frac{g_1(\theta, \Delta_b)}{g_0(\theta, \Delta_b)} - 1\right] M^-(v_b^*(\theta, \Delta_b))g_0(\theta, \Delta_b)d\theta \leq \Delta_u.$$

[65] $g_1(\theta, \Delta_b) = \frac{g(\theta)[1 - F(v_b^*(\theta, \Delta_b))]}{\int g(\theta)[1 - F(v_b^*(\theta, \Delta_b))]d\theta}$ and $g_0(\theta, \Delta_b) = \frac{g(\theta)F[v_b^*(\theta, \Delta_b)]}{\int g(\theta)F(v_b^*(\theta, \Delta_b))d\theta}$.

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

■