University of Nottingham
UK | CHINA | MALAYSIA

NICEP
Nottingham Interdisciplinary Centre for Economic and Political Research

NICEP Working Paper: 2024-04

Segment and rule: Modern censorship in authoritarian regimes

Kun Heo and Antoine Zerbini

Nottingham Interdisciplinary Centre for Economic and Political Research

School of Economics, University of Nottingham, Sir Clive Granger Building,
University Park, Nottingham, NG7 2RD
ISSN 2397-9771

Segment and Rule: Modern Censorship in Authoritarian Regimes[*]

Kun Heo[†] Antoine Zerbini[‡]

April 15, 2024

Abstract

We analyze the incentives of authoritarian regimes to segment access to censored content through technology. Citizens choose whether to pay to access censored online content at a cost fixed by the regime: the firewall. A low firewall segments access and generates more compliance than full censorship – a high firewall – ever could. Regime opponents self-select into consuming censored content, and comply conditional on positive independent reporting. Regime supporters exclusively consume state propaganda, which secures their compliance. This segment-and-rule strategy can be engineered by making local news outlets uninformative, or by affecting the intrinsic benefit from access.

JEL Codes: D72, D82, D83, O33.

Keywords: Censorship, Internet, Information Design, Segmentation.

[*] For invaluable advice and guidance, we thank Stephane Wolton. We thank Emiel Awad, Benjamin Blumenthal, Roel Bos, Luis Bosshart, Martin Castillo Quintana, Anna Denisenko, Torun Dewan, Hanna Folsz, Catherine Hafer, Gleason Judd, Navin Kartik, Dimitri Landa, Elliot Lipnowski, Sergi Martinez, John Marshall, Andrea Mattozzi, Justin Melnick, Giacomo Ponzetto, Carlo Prato, Kris Ramsay, Roxanne Rahnama, Arturas Rozenas and seminar and conference participants at Columbia, EEPGW, FTCP, MPSA, EPSA, NYU, LSE-NYU 2023, Oxford, PSE, Bruneck 2023, VFTW, Bologna, UAB, and Cologne Economics of Media Bias 2024 for helpful feedback and comments. Zerbini gratefully acknowledges support from the Spanish Ministry of Science and Innovation (grants CNS2022-135749 and PID2022-137707NB-I00).

[†] PhD Candidate, New York University. Email: bh1875@nyu.edu.

[‡] PhD Candidate, London School of Economics and Political Science. Email: a.r.zerbini@lse.ac.uk.

1 Introduction

> The internet is uncontrollable. And if the internet is uncontrollable, freedom will win. It's as simple as that.
>
> — Ai Weiwei in 2012 in the Guardian.

The roll-out of the internet across the world was first hailed as a liberating technology for citizens of authoritarian regimes (Diamond, 2010). On top of being able to communicate more easily (Manacorda and Tesei, 2020; Acemoglu, Hassan, and Tahoun, 2018), individual citizens gained the right to decide whether to consume foreign outlets that were previously censored and inaccessible. According to the liberating view, this increased autonomy would then empower citizens in their struggle against authoritarian regimes. This optimism is reflected in the quote of Ai Weiwei, and further substantiated by empirical evidence about the difficulties of authoritarian leaders with online censorship. Indeed, millions bypass censorship firewalls everyday.[1]

We contend that the bypassing of the firewall benefits authoritarian regimes, as long as only a specific segment of the population accesses the uncensored internet. This phenomenon of selective bypassing is not a bug; rather, it is the direct consequence of a strategy of modern and selective censorship. Modern censorship leverages the citizens' ability to choose whether to access banned content in order to make citizens with different political preferences comply with the regime. To that end, the firewall's mild deterrent effect – downloading a VPN suffices to bypass it – serves a dual purpose. On the one hand, it dissuades supporters of the regime from seeking out banned content; their compliance is secured through the propaganda of the state-media. On the other, it does not dissuade opponents of the regime from gaining access to banned foreign outlets. This benefits the regime because opponents do not respond to state-propaganda: they only comply when exposed to positive reporting about the regime from a credible source, such as one that is known to be biased against the regime, and banned domestically. This can be achieved, for instance, if regime opponents are exposed to negative reporting about the performance of a comparable country.[2] This strategy

[1] Empirical evidence suggests that at least 5-10% of internet users have at some point used circumvention softwares to bypass the firewall in China (Chen and Yang, 2019; Hobbs and Roberts, 2018; Shen and Zhang, 2018; Mou, Wu, and Atkin, 2016). Similar evidence has been provided in Iran (Dal and Nisbet, 2022), Russia (Fung, 2022; Xue et al., 2022), Egypt (Lutscher, 2023) or Turkmenistan (Nourin et al., 2023). Further, the very observation of online censorship may reinforce the citizens' incentives to bypass firewalls (Hobbs and Roberts, 2018) and organize political movements (Boxell and Steinert-Threlkeld, 2019; Pan and Siegel, 2020).

[2] Positive and credible reporting need not involve articles from independent outlets that praise the authoritarian

of *segment-and-rule* entrenches authoritarian regimes by garnering more compliance among citizens than full censorship ever could. This argument may explain why the roll-out of the internet is associated with an increase in support for authoritarian regimes, *conditional* on internet censorship (Guriev, Melnikov, and Zhuravskaya, 2021).

Crucially, for segment-and-rule to be possible, regime opponents must have the strongest incentives to bypass the firewall.[3] Empirically, citizens of authoritarian regimes have been shown to bypass the firewall to access a variety of non-informational content, such as, for instance, entertainment broadly defined (Chen and Yang, 2019; Hobbs and Roberts, 2018). Within our theoretical framework, when bypassing the firewall citizens derive *both* an informational benefit and an intrinsic non-informational benefit (entertainment, intrinsic satisfaction to consume independent reporting, etc). While this intrinsic component does not affect beliefs per se, it affects the decision to bypass the firewall, and thus access to information. For segment-and-rule to be possible, equilibrium sorting into access must, we show, take place along the intrinsic dimension, rather than the informational one. To ensure this sorting pattern, we show that authoritarian regimes can deplete their own outlets of informational content – parroting the party line – or affect the intrinsic benefit, for instance through investments in historically revisionist propaganda and selective banning of foreign entertainment.

In turn, selective bypassing of the firewall is not indicative of technological difficulties faced by modern authoritarian regimes nor suggestive of a heightened difficulty to control information flows. Rather, we show that it is symptomatic of a novel form of selective censorship made possible by the agency in information acquisition that citizens gained through the internet. Just as scholars are documenting how modern authoritarian regimes exploit technological change – e.g., AI – for surveillance purposes (Tirole, 2021; Dragu and Lupu, 2021; Beraja et al., 2023; Xu, 2023), we depict a yet bleaker picture: the internet entrenches authoritarian regimes *because* it empowers citizens.

Our theoretical framework focuses on the interactions between a leader (he) and a continuum of heterogeneous citizens (they or she). The leader seeks to maximize compliance in the citizenry.[4]

regime. When US media outlets known to be biased against the CCP report negatively about the situation within the US – e.g., the Opoid crisis, the handling of the pandemic or the discussions around gun violence – this represents positive reporting for the CCP because it paints the CCP in a positive light, *relative* to the US. Huang (2015) and Huang and Yeh (2019) provide empirical evidence of such "relative updating" after exposure to foreign news in China while Chester (2023) documents this "relative argumentation" by CCP controlled outlets.

[3] Suggestive empirical evidence in China (Mou, Wu, and Atkin, 2016) and Iran (Dal and Nisbet, 2022) hints at firewall by-passers being more opposed to the regime and exhibiting lower levels of "political trust".

[4] Compliance captures any action that benefits the leader such as *not* joining an opposition movement, criticizing the regime, protesting, or, actively joining the ruling party. The extent to which non-compliance hurts the regime

Citizens differ in their ex-ante alignment with the regime, which determines their payoff from non-compliance. The more ex-ante aligned with the regime a citizen is, the higher her incentives to comply. In our baseline model, most citizens are "moderates": they are neither too opposed nor too aligned with the regime ex-ante. Citizens do not exactly know whether it is in their best interest to comply and use reports from the state and foreign media to inform their compliance decision. The foreign media targets a foreign audience and is thus non-strategic within the subgame we study; it has some (possibly null) reporting slant *against* the regime. Thus, positive reporting from the foreign media, if any, must be truthful and ensures the compliance of the whole citizenry.

All citizens freely consume the state media. When deciding whether to gain access to the foreign media by bypassing the firewall, a citizen has two considerations in mind. First, gaining access provides an *informational* benefit, in the form of an additional piece of information. Second, gaining access also has benefits and costs that are purely non-informational, or *intrinsic*. This captures any intrinsic difference in how citizens appreciate consuming the foreign media relative to the state media (holding fixed the informational benefit). Importantly, this intrinsic benefit can be correlated with a citizen's political type. In the baseline model such heterogeneity is parameterized with a simple linear functional form and referred to as the *correlation* between a citizen's political type and her intrinsic benefit. At the beginning of the game the leader commits to the reporting strategy of the state media (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019), which he controls, and chooses the common across citizens cost of access to the foreign media. Having observed the state media's report, each citizen decides whether to gain access to the foreign media. Finally, each citizen decides whether to comply.

The heterogeneity of political preferences generates differential incentives to comply across citizens. Thus, providing more information to citizens presents the leader with a trade-off. On the one hand, opponents must have access to credible information from an independent source to comply after positive reporting. On the other, credible independent information requires *all* citizens being exposed to negative reporting often. To resolve this trade-off, the leader would like to pursue a particular type of selective censorship that ensures that while regime opponents do bypass the firewall, "moderates" – which can be convinced by state propaganda – do not.[5] In turn, opponents comply

(e.g., active revolt vs not joining the party) does not matter for any of our core results.

[5] Whether regime supporters gain access does not matter as they do not condition their equilibrium compliance on the foreign media's report.

conditional on positive reporting from a credible foreign media while moderates are kept in the dark and comply: they only consume content from domestic outlets which parrot the party line. We refer to such a strategy as one of *segment-and-rule*.

Crucially, segment-and-rule is not necessarily possible, as it requires a particular form of endogenous sorting into access. To see why, we make two observations. First, the firewall imposes a *common* cost of access. Second, a citizen's political type conditions her informational benefit. Moderate citizens are most unsure as to whether they should comply and would like to inform their compliance decision. In contrast, extremists rarely condition their compliance decision on the outlets' reports. Thus, the value of an additional piece of information is highest for "moderates", which (in the baseline setup) represent the majority of the population. When the intrinsic benefit is common across political lines, the informational benefit is the only source of differential sorting into access. Then, segment-and-rule is impossible because it is moderates, rather than opponents, that have the strongest incentives to bypass the firewall. The leader imposes a high enough cost of access such that no citizen bypasses the firewall. Then, regime supporters comply following positive reporting from the state media while regime opponents never do. With a lower cost of access, the regime could incentivize some opponents and moderates into gaining access. Yet keeping moderates in the dark is a first order concern for the leader, while gambling on the opponents' compliance is a second order one.

In contrast, when the intrinsic benefit is increasing in misalignment with the regime, segment-and-rule is feasible. Even if a citizen's misalignment with the regime reduces her informational benefit from bypassing the firewall, when this misalignment increases her intrinsic benefit sufficiently, the *total* benefit from bypassing the firewall is always increasing in a citizen's type. In turn, an intentionally intermediate cost of access ensures that only opponents bypass the firewall. The state media is kept uninformative to induce the compliance of the base while opponents bypass the firewall and comply following positive reporting of the foreign media.

Crucially, when the intrinsic benefit is not strongly correlated with a citizen's type, so that segment segment-and-rule is not directly feasible, it can be engineered. First, when this correlation is positive but intermediate, the regime depletes local outlets of informational content to increase the informational benefit from gaining access to the foreign media, and ensures that the incentives to gain access are increasing in a citizen's type. This engineering comes at at cost – the regime cannot pick the optimal information structure "freely" – and highlights that the ability to commit to an information structure allows the regime both to tailor their communication to their base *and* to affect the sorting into access to an independent source. Second, when this correlation is low, null or negative, an authoritarian regime may still be able to implement segment-and-rule, provided that the intrinsic benefit captures an entertainment benefit, and that the regime can control access to entertainment. In section 6.3 we show that by selectively banning foreign entertainment that appeals to opponents or producing domestic entertainment that appeals to most except opponents, the regime can engineer the appropriate sorting. This argument can help explain the empirical pattern of domestically available entertainment appealing mostly to citizens aligned with the authoritarian regime, and highlights the instrumental role of entertainment control in censorship and propaganda.[6]

While our focus is on the effect of the internet on censorship and authoritarian rule, we believe the larger theoretical implications of our framework to be useful for applications in other contexts. Our broader point is that a sender – be it an information designer or not – can drastically benefit from the existence of an independent source of information available to the receivers she faces, even if she knows this source to have misaligned interests and to be most likely to reveal "bad" news. This stands in sharp contrast with the common disciplining effect of competition established both theoretically (Battaglini, 2002; Gentzkow and Kamenica, 2017) and empirically in the censorship literature (Galvis, Snyder, and Song, 2016; Qin, Strömberg, and Wu, 2018; Marshall and Kronick, 2022). Crucial to this argument is that (i) the sender faces a continuum of *heterogeneous* receivers and (ii) the sender can ensure that only a specific subset of receiver types are exposed to the independent source. Most interestingly, we show that such tailored communication through segmented access need not require the sender to be able to discriminate on observables as in the literature on private bayesian persuasion (Bardhi and Guo, 2018; Chan et al., 2019; Arieli and Babichenko, 2019), nor an information design approach (see section 6.1).

The rest of the paper is organized as follows. We first review the related literature prior to presenting the baseline setup in section 3. We present equilibrium results in section 4. In section 5

[6] In present-day China, beyond the imposition of a low but positive cost of access via the firewall, the CCP also strategically invests in domestic entertainment such as high profile historical war movies – e.g., *The Battle at Lake Changjin* – or police drama about a corrupt administration – e.g., *The Knockout*; see also Liu and Yao (2023) for empirical evidence of the use of entertainment for propaganda purposes. Further, while high budget and mostly apolitical movies with broad appeal are allowed (e.g., *Jurassic World* or *Transformers*), movies that mostly appeal to opponents are banned (e.g., *Cockroach*, *Eternal Spring* or *Top Gun - Maverick*). Esberg (2020*b*) also documents the censorship of entertainment favoured by regime opponents in the Chilean dictatorship.

we turn to comparative statics and conceptual, empirical and policy implications. We then discuss in section 6 how entertainment can be used strategically by authoritarian regimes, and generalize the baseline framework in various directions (non-information design approach, privately informed regime, access to multiple outlets beyond the firewall, etc.) to highlight the theoretical innovation, prior to concluding in section 7.

2 Related Literature

Our main substantive contribution is to the political economy literature on technological change and modern censorship (Chen and Yang, 2019; Zhuravskaya, Petrova, and Enikolopov, 2020; Egorov and Sonin, 2023). To explain the variation in censorship levels across autocracies, scholars have highlighted that censorship may backfire when it leads to a loss in valuable entertainment (Marshall and Kronick, 2022) or that it may come at an economic cost by making the monitoring of the state apparatus more difficult (Egorov, Guriev, and Sonin, 2009; Lorentzen, 2014).While Edmond (2013) shows in a global game setting that authoritarian regimes may benefit from the roll-out of the internet if the information technology is easily centralized, we argue that authoritarian regimes purposely engage in a strategy of segmented access to banned content, which implies a low cost of access, irrespective of technological capacity. In that sense we provide one micro-foundation of *how* "informational autocracies" (Guriev and Treisman, 2019, 2020) leverage various sources of information to maximize compliance in heterogeneous citizenries. The strategy of segment-and-rule differs from the age-old strategy of *divide-and-rule* because, instead of disrupting coordination among (groups of) citizens (Acemoglu, Robinson, and Verdier, 2004) or elites (Luo and Rozenas, 2023) through the strategic distribution of resources or information, our autocrat exploits the heterogeneity of political preferences, in a setting without any collective action problem. We model both the standard top-down measure of censorship in the literature – how much information flows to citizens from the state outlets (Shadmehr and Bernhardt, 2015; Gehlbach and Sonin, 2014) – and a second bottom-up measure of information acquisition by citizens – the share of who do not gain access to independent banned reporting. We show that in equilibrium these two measures move in opposite directions: modern censorship creates asymmetries across citizens which complicate the empirical task of measuring any change in "censorship".

Theoretically our main contribution lies in the literature on persuasion (Crawford and Sobel, 1982), information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019) and rational inattention (Maćkowiak, Matějka, and Wiederholt, 2023). As in the literature on the persuasion of voting-bodies (Caillaud and Tirole, 2007; Alonso and Câmara, 2016; Awad, 2020), our sender uses an intermediary – an independent source – to maximize compliance among the receivers. In the information design literature on price discrimination (Bergemann, Brooks, and Morris, 2015) and private bayesian persuasion (Bardhi and Guo, 2018; Chan et al., 2019; Arieli and Babichenko, 2019) the designer leverages individual level information, to target her communication to the type of the receiver. Similarly in our setting the sender benefits from targeting communication to the type of the receiver. Crucially however, our sender does not know whom she is facing and thus cannot communicate privately: she can only impose a *common* cost of access and design any *public* experiment. This is not trivial, given that private persuasion through elicitation does not improve on public persuasion in a binary state binary action environment (Kolotilin et al., 2017; Gitmez and Sonin, 2023): our sender leverages the existence of an independent source to avoid the incentive-compatibility constraints imposed by elicitation and targets her communication without any ability to discriminate. Closest to our framework is Matysková and Montes (2023), who also consider a game of bayesian persuasion with rational inattention. They show that the sender's payoff is decreasing in the receiver's cost of information acquisition. By modelling a sender that faces a heterogeneous set of receivers, we instead show the sender's payoff is non-monotonic and single-peaked in the receiver's cost of information acquisition.

As in related works (Gratton and Lee, 2024; Gitmez and Sonin, 2023; Heo and Zerbini, 2023; Gitmez and Molavi, 2023), we model the autocrat's censorship problem using an information design approach. Crucially, our main segmentation argument still stands when the sender only chooses the cost of access, but also when she is endowed with (negative) private information and can communicate via cheap-talk: as both types of senders have aligned incentives to generate a higher posterior belief *and* to pick the segmenting-cost given that belief, this leads to pooling in equilibrium.

3 Model

Consider a game between an authoritarian leader A (he) and a $[0,1]$ continuum of citizens (they or she) indexed by the subscript i.

Citizens' actions and preferences. The citizens choose between two actions: $a_i \in \{0,1\}$. $a_i = 1$ represents compliance with the regime and $a_i = 0$ represents non-compliance. The leader maximizes compliance in the citizenry: his payoff is given by $V(\sigma, c) = \int_0^1 a_i di$.[7] A citizen's payoff from compliance depends on the state of the world $\omega \in \{0,1\}$, where $\Pr(\omega = 1) = p \in (0,1)$. In contrast, a citizen's payoff from non-compliance depends on their private type $\theta_i \in [0,1]$. A citizen payoff from either action is given by

$$u_i(0;\theta_i,\omega) = \theta_i \qquad\qquad u_i(1;\theta_i,\omega) = \omega.$$

We assume that when indifferent between compliance and non-compliance, a citizen complies. Each citizen privately observes her political type θ_i. A high θ_i represents a citizen that is ex-ante opposed to the regime and needs to be convinced that $\omega = 1$ to comply. The continuum of heterogeneous citizens is distributed according to a cdf F, which has a unimodal density f with full support on $[0,1]$. We make two assumptions regarding f. First, there exists a $\theta^\dagger \in (0,1)$ s.t. $F(\theta_i) \geq \theta_i \iff \theta_i \geq \theta^\dagger$. This implies that f has an interior peak $\hat{\theta} \in (0,1)$. Substantively this requires that there are not too many extreme supporters ($\theta_i \approx 0$) or extreme opponents ($\theta_i \approx 1$). Second, we require f to be log-concave.[8]

Media consumption and censorship. There exists two sources of information for the citizens. There is a *state-controlled media* (henceforth, *state media*) whose reporting strategy is chosen by the leader. The leader publicly commits to the pro-regime slant of the state media, $\sigma \in [0,1]$. The whole citizenry observes σ and the realized message $s_{\mathcal{S}} \in \{0,1\}$ from the state media.[9] Given σ, the

[7] We provide sufficient monotonicity conditions for the main qualitative results to be upheld in a game where citizens choose from a continuum of compliance level; see Lemma A.26.

[8] The existence of $\theta^\dagger \in (0,1)$ and log-concavity are sufficient but not necessary. The former rules out a very weak leader ($F(\theta_i) \leq \theta_i \; \forall \theta_i$) and ensures the analysis is smooth. The latter rules out "weird" s-shaped cdfs whose convexity would change many times. We generalize the results to non-unimodal distributions in section 6.5.

[9] We extend the results to a non-information design framework, including one with cheap talk, in section 6.1.

conditional probability that citizens observe $s_{\mathcal{S}} \in \{0, 1\}$ is given by:

$$\Pr[s_{\mathcal{S}} = 0|\omega = 0] = 1 - \sigma \qquad \Pr[s_{\mathcal{S}} = 0|\omega = 1] = 0$$
$$\Pr[s_{\mathcal{S}} = 1|\omega = 0] = \sigma \qquad \Pr[s_{\mathcal{S}} = 1|\omega = 1] = 1.$$

The higher σ is, the more uninformative the state media. Note that the state media does hide good news: given the unimodality of f this is an equilibrium result.[10]

Having observed $s_{\mathcal{S}}$, each citizen decides whether to circumvent the firewall to access the *foreign media*. This outlet has a commonly known exogenous anti-regime slant $\beta \in [0, 1)$ which parametrizes its informativeness. Given β, a primitive, the conditional probability that citizens observe $s_{\mathcal{F}} \in \{0, 1\}$ is given by:

$$\Pr[s_{\mathcal{F}} = 0|\omega = 0] = 1 \qquad \Pr[s_{\mathcal{F}} = 1|\omega = 0] = 0$$
$$\Pr[s_{\mathcal{F}} = 0|\omega = 1] = \beta \qquad \Pr[s_{\mathcal{F}} = 1|\omega = 1] = 1 - \beta.$$

The foreign media never hides true bad news for the regime.

Sorting into access. Circumventing the firewall comes at a cost of access $c \in \mathbb{R}^+$ which the regime chooses at the beginning of the game. Circumventing the firewall benefits the citizens in two ways. First, there is a (weakly) positive informational benefit $b_i(\theta_i, \sigma, s_{\mathcal{S}}, \beta)$. Second, there is a relative intrinsic non-informational benefit $\alpha(\theta_i)$. For ease of exposition we first assume the following functional form for the intrinsic benefit: $\alpha(\theta_i) = z + \gamma * \theta_i$ with $z \in \mathbb{R}$ capturing the common intrinsic benefit and $\gamma \in \mathbb{R}$ the correlation between a citizen's political type θ_i and her intrinsic benefit.[11] The net benefit from gaining access to the foreign media of citizen i is given by:

$$\underbrace{\delta_i(\theta_i, \sigma, s_{\mathcal{S}}, \beta)}_{\text{net benefit}} = \underbrace{\underbrace{b_i(\theta_i, \sigma, s_{\mathcal{S}}, \beta)}_{\text{informational benefit}} + \underbrace{z + \gamma * \theta_i}_{\text{relative intrinsic benefit}}}_{\text{willingness to pay for access}} - \underbrace{c}_{\text{cost of access}} \tag{1}$$

We denote the observed report from the foreign media by $\hat{s}_{\mathcal{F}} \in \{s_{\mathcal{F}}, \emptyset\}$. If a citizen does not gain access to the foreign media, we write $\hat{s}_{\mathcal{F}} = \emptyset$. To ensure that the regime faces a censorship problem,

[10] Heo and Zerbini (2023) show that this assumption is not without loss for different distributional assumptions, e.g., if the citizenry is distributed according to a u-shaped distribution.

[11] We provide general conditions in section 6.4.

we assume that $z \geq 0$ and $\gamma \geq 0$ such that all citizens consume the foreign outlet absent a positive cost of access.[12]

Timing. The sequence of the game is as follows:

1. The leader publicly commits to the reporting slant σ and chooses the common cost of access c.
2. Nature determines ω and privately reveals θ_i to citizen i.
3. Nature generates the state media's report $s_{\mathcal{S}}$ as well as the foreign media's report $s_{\mathcal{F}}$. Each citizen, having observed $s_{\mathcal{S}}$ decides whether to gain access to the foreign outlet; if they do, they observe $s_{\mathcal{F}}$.
4. Each citizen chooses whether to comply. Payoffs are realized. Game ends.

The equilibrium concept is weak Perfect Bayesian Equilibrium.

Comments on the Setup

Interpretation of compliance. Compliance captures any action that benefits the leader, relative to non-compliance, such as not joining a foreign movement, not criticizing the regime nor protesting, or actively supporting the regime by joining the ruling party. There are situations (e.g., the leader is more competent than any challenger, or more resilient to large-scale non compliance) where compliance benefits all, such that the leader's and citizens' incentives align. Importantly, conditional on learning for sure that $\omega = 1$, it is optimal for any citizen to comply, irrespective of one's ex-ante alignment with the regime: that is, we focus on the persuadable citizenry.[13]

Interpretation of the intrinsic benefit. The intrinsic benefit $\alpha(\theta_i)$ captures *any* benefit citizens of authoritarian regimes may derive from gaining access to an uncensored internet, *above and beyond* the informational benefit. Among other things, this captures (i) an entertainment benefit, e.g. from gaining access to censored streaming or social media platforms or foreign sports (Chen and Yang, 2019), (ii) an intrinsic benefit from consuming informational content from an independent source,

[12] Full results are presented in the appendix for any $z \in \mathbb{R}, \gamma \in \mathbb{R}$.

[13] We provide conditions under which the incentives to segment extend to a setting where citizens choose their compliance level from a continuum in Lemma A.26.

(iii) an economic benefit, e.g., for economic elites operating internationally, (iv) a personal benefit independent of entertainment (e.g., accessing banned travel agencies or tourism related websites).

Information flows vs media outlets. We model a dichotomy of *information flows*. First, there are information flows emerging from the set of (generally local) media outlets under the control of the authoritarian regime – the "state media". We assume that all citizens consume the state media, which need not be a TV channel or state-owned newspapers per se. Rather, it captures the general communication of the leader on the true underlying state: e.g., it is the communication of an authoritarian leader about whether the regime is responsible for the economic slow-down, or whether external factors are at play (Rozenas and Stukal, 2019). In short, all citizens are aware of the regime's communication on the relevant policy issue. Second, there are information flows emerging from outlets not directly under the control of the regime – modelled under the umbrella of the "foreign media". These outlets are based outside of the geographical and legal boundaries of the country and outside of the regime's control. They are banned and thus at minima not biased in favor of the regime (we allow for a perfectly unbiased foreign outlet $\beta = 0$). Their reporting slant β is chosen either to maximize revenue or some ideological goal outside of the country – e.g., foreign newspapers (the *Liberty Times*, the *Guardian*, or the *New York Times*) or entertainment outlets (e.g., *HBO* or *Netflix*) – or by starch opponents of the regime who have been banned from the country.[14] We characterize how the informativeness ($(1-\beta) \in (0, 1]$) of the foreign outlet impacts the leader's censorship strategy and overall compliance. As discussed in sections 6.2 and 6.3, our framework allows for micro-foundation of this foreign media umbrella, and exposure to a stream of report.

A mighty leader. Our authoritarian leader can both impose *any* common cost of access to the foreign outlet and choose *any* public reporting strategy for the state media. These assumptions ensure that the incentives of the leader to engage in selective censorship are not driven by the leader's inability to communicate in a particular fashion to the population or to restrain access to the foreign media – a possibility we discuss in section 5.3.[15]

[14] β can be formally micro-founded along the lines of the framework of Gehlbach and Sonin (2014). A foreign outlet targeting a foreign audience chooses β by balancing the two competing goals of (i) garnering advertising revenue by truthfully reporting the state and (ii) impacting the behavior of its audience by slanting the reporting towards the media-preferred citizen behavior.

[15] Endowing with knowledge about the citizen's type – an unrealistic assumption in authoritarian settings (Kuran, 1991) – would allow for the imposition of type-specific costs and private communication and would obviously facilitate the leader's problem.

4 Analysis

4.1 Preliminary Intuition

Consider first citizens who do *not* choose whether to gain access to the foreign media. Absent such agency, a citizen decides whether to comply *given* some reporting slant σ and observed reports $s_{\mathcal{S}}$ and $\hat{s}_{\mathcal{F}}$. When the state media reveals bad news, the state of the world must be $\omega = 0$ and no citizens complies. When the foreign media reveals good news, the state of the world must be $\omega = 1$ and all citizens comply.[16] Following good news from the state media ($s_{\mathcal{S}} = 1$), only citizens sufficiently aligned with the regime comply.

Definition 1. For a given reporting slant σ, $\theta(\hat{s}_{\mathcal{F}}, \sigma) \equiv Pr(\omega = 1 | s_{\mathcal{S}} = 1, \hat{s}_{\mathcal{F}}, \sigma)$ denotes the citizen indifferent between complying and not complying following positive reporting from the state media and some $\hat{s}_{\mathcal{F}}$. Further, we refer to $\theta(\emptyset, \sigma)$ as the *target citizen*.

A citizen complies after one-sided positive reporting if and only if:

$$a_i^*(\theta_i, s_{\mathcal{S}} = 1, \hat{s}_{\mathcal{F}} = \emptyset) = 1 \iff \theta_i \leq \frac{p}{p + (1-p)\sigma} \equiv \theta(\emptyset, \sigma)$$

Similarly, after contradictory reporting ($s_{\mathcal{S}} = 1, s_{\mathcal{F}} = 0$), a citizen complies if and only if

$$a_i^*(\theta_i, s_{\mathcal{S}} = 1, \hat{s}_{\mathcal{F}} = 0) = 1 \iff \theta_i \leq \frac{p\beta}{p\beta + (1-p)\sigma} \equiv \theta(0, \sigma) < \theta(\emptyset, \sigma)$$

The censorship trade-off. In Table 1 we define three important segments within the citizenry and highlight the censorship trade-off faced by the leader when exposure to the foreign outlet is exogenous.

Remark 1. *Given some reporting strategy σ, the range of citizens complying after a given pair of observed media reports $(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}})$ is given by Table 1.*

Opponents only comply after positive reporting from the foreign media. Ideally the leader would ensure that they gain access to the foreign media. However, *conditional compliers* – the "moderates" – comply following positive reporting from the state media that is not contradicted by the

[16] Hereafter, we never condition on $s_{\mathcal{S}} = 0$ or $\hat{s}_{\mathcal{F}} = 1$, since either no one or everyone complies in either of these cases, conditional on observing either of these reports.

Observed Media Reports $(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}})$ Political Segments	$(1,1)$	$(1,\emptyset)$	$(1,0)$	$(0,0)$
Unconditional compliers: $\theta_i \leq \theta(0,\sigma)$	✓	✓	✓	✗
Conditional compliers: $\theta_i \in (\ \theta(0,\sigma), \theta(\emptyset,\sigma)\ ]$	✓	✓	✗	✗
Opponents: $\theta_i > \theta(\emptyset,\sigma)$	✓	✗	✗	✗

Table 1: ✓ represents compliance while ✗ represents non-compliance.

foreign media. Ideally, the leader would prevent conditional compliers from bypassing the firewall by imposing a high cost of access. The leader does not need to censor *unconditional compliers*: they comply even after negative reporting from the foreign media.

This censorship trade-off already suggests the benefit of a strategy of *segment-and-rule* whereby opponents of the regime bypass the firewall, while conditional compliers only consume content from state outlets. In order to understand when such a strategy is feasible in equilibrium, we must first understand how citizens self-select into bypassing censorship or not.

4.2 Endogenous Sorting

We now explain how the two motives of citizens for gaining access – information and the intrinsic benefit – influence their decision to gain access to banned content.

Sorting on the intrinsic dimension. By definition, the common intrinsic benefit z cannot generate heterogeneous sorting into access along political lines. Such heterogeneous sorting can only be generated by the correlation between the intrinsic benefit and politics γ. The stronger this correlation is, the larger the difference in intrinsic benefit between regime supporters (low θ_i) and regime opponents (high θ_i).

Sorting on information. For a given reporting slant of the state media σ, a citizen's ex-ante alignment with the regime determines her willingness to pay for the foreign media's report. The value of information is highest for "moderate" citizens.[17]

Lemma 1. *Suppose that the state media reports positively ($s_{\mathcal{S}} = 1$). Given some reporting strategy σ, the informational benefit from consuming the foreign outlet is*

[17] When the state media reports negatively ($s_{\mathcal{S}} = 0$) the informational benefit from access is null for all citizens.

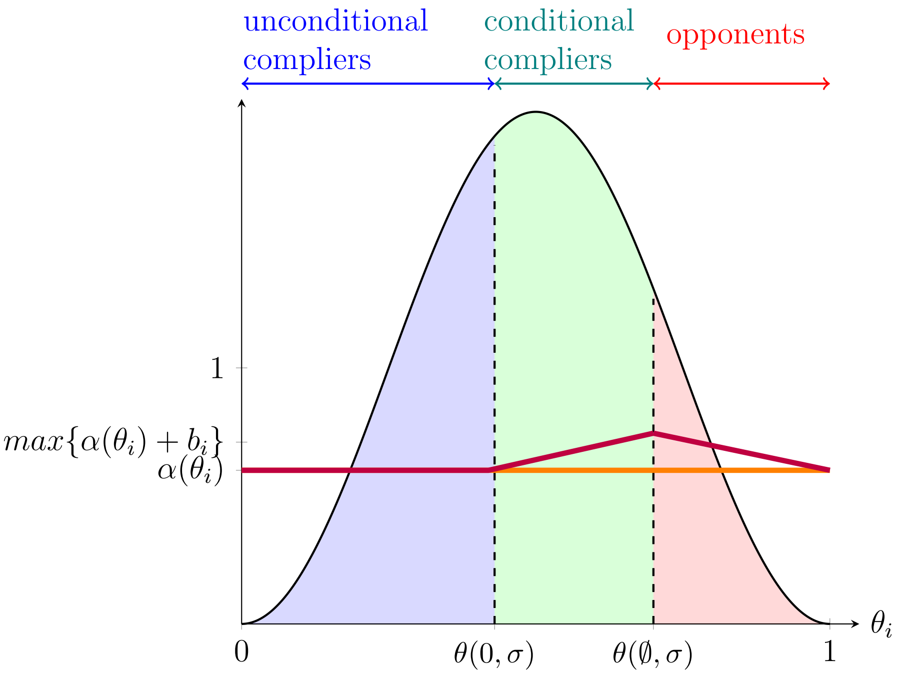

Figure 1: For this illustration, $\beta = 0.31$, $p = .5$ and $f(\theta) = cos(2\pi(\theta - 0.5)) + 1$; $z = 0.6, \gamma = 0, c = 0$. The black lines plots $f(\theta)$. The orange line plots the intrinsic benefit $\alpha(\theta_i)$. The red line plots the maximum willingness to pay to access the foreign media as a function of θ_i, given $\theta(\emptyset, \sigma) = 0.7$ and $s_{\mathcal{S}} = 1$.

- *null for all unconditional compliers* $\theta_i \leq \theta(0, \sigma)$,
- *positive and increasing linearly in* θ_i *for all conditional compliers* $\theta_i \in (\theta(0, \sigma), \theta(\emptyset, \sigma)]$,
- *positive and decreasing linearly in* θ_i *for all opponents* $\theta_i \geq \theta(\emptyset, \sigma)$.

Figure 1 illustrates Lemma 1 in a citizenry where the non-informational benefit is positive ($z > 0$) and constant across citizens ($\gamma = 0$). The citizen indifferent between complying and not complying following positive reporting of the state media – the target citizen $\theta_i = \theta(\emptyset, \sigma)$ – has the highest willingness to pay. The further away a citizen is from the target citizen, the lower her informational benefit. Importantly, the three political segments of Table 1 are defined by the reporting slant of the state media σ. That is, informational sorting is both endogenous to the state media's reporting σ and conditioned by a citizen's political preference θ_i.

We now consider how the strength of the correlation between politics and the intrinsic benefit (γ) determines the type of sorting that takes place in equilibrium, and thus the censorship strategy of modern authoritarian leaders. To refer to the equilibrium reporting slant of the state media in the case of a low, strong and intermediate correlation between politics and the intrinsic benefit (γ) we will use σ^L, σ^S and σ^I; the same superscripts will be used for the equilibrium target citizen.

4.3 Low Correlation: Full Censorship

To understand when segment-and-rule is possible, we make two observations. First, the firewall imposes a *common* cost of access. Second, as long as the state media is not perfectly informative ($\sigma = 0$), the informational benefit is always single-peaked in a citizen's political preference as in Figure 1. Absent sorting on the intrinsic dimension, it is conditional compliers – the "moderates" – who have the strongest incentives to bypass the firewall; not regime opponents. In turn, segment-and-rule is not possible.

Proposition 1. *There exists a unique $\underline{\gamma} \in (0, 1-\beta)$ such that, if the correlation is low ($\gamma \in [0, \underline{\gamma}]$) then,*

- *there exists a unique equilibrium reporting slant $\sigma^* = \sigma^L$ and target citizen $\theta(\emptyset, \sigma^L) = \theta^L$,*
- *the leader imposes the lowest cost of access such that no citizen bypasses the firewall: $c^* = \overline{c}(\theta^L)$.*
- *a citizen complies if and only if $s_{\mathcal{S}} = 1$ and $\theta_i \leq \theta^L$.*

When the correlation between politics and the intrinsic benefit is negative, null or positive but small, the leader maximizes compliance by ensuring that no citizens bypasses the firewall. In turn, conditional and unconditional compliers comply following positive reporting, while opponents *never* comply. This equilibrium strategy is illustrated in the left panel of Figure 2. In principle, the leader could try to censor selectively by imposing a lower cost of access $c' < \overline{c}(\theta^L)$ as in the right panel of Figure 2. A reduction in the cost of access would induce citizens in the blue and red areas to bypass the firewall. This would generate two asymmetric effects.

First, reducing the cost of access asymmetrically affects the *share* of citizens who bypass the firewall on either side of the target citizen. There are more people to the left of the target citizen than to the right. This is because, in equilibrium, the leader always ensures that the share of compliers, conditional on positive reporting, is sufficiently large. Formally, the equilibrium target citizen is always sufficiently ex-ante opposed to the regime: $\theta^L > max\{p, \hat{\theta}\}$. Second, reducing the cost of access asymmetrically affects the decision of citizens on either side of the target citizen,

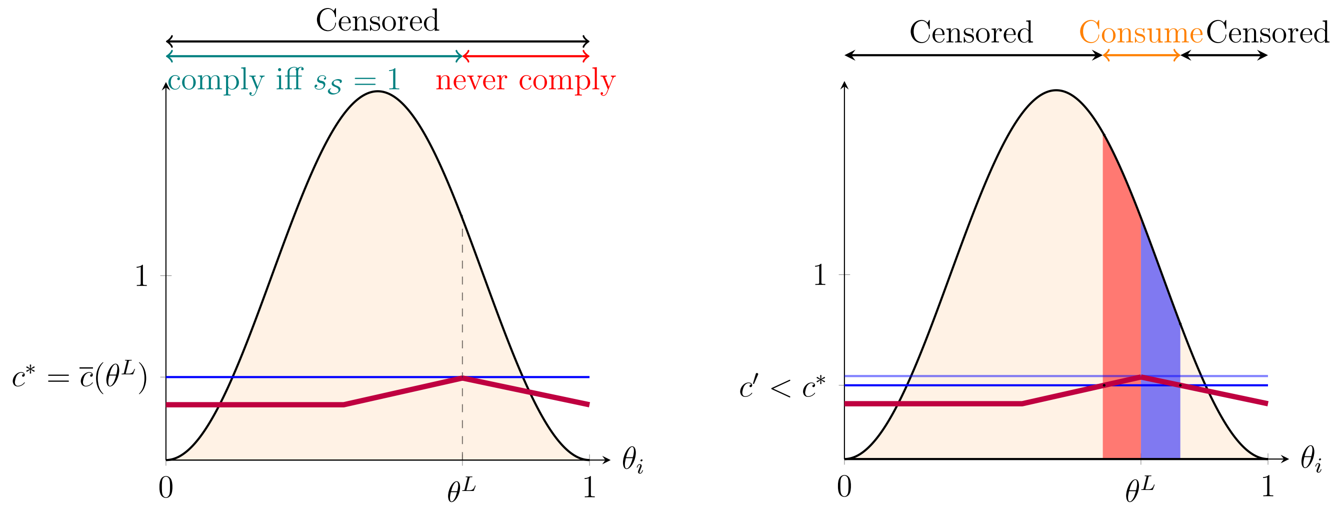

Figure 2: Same parameter values as in Figure 1. The red line plots the willingness to pay for $s_{\mathcal{F}}$. The blue line plots the cost of access. The left panel depicts equilibrium behavior under full censorship. The right panel illustrates the asymmetric effects of an off-path reduction in the cost of access.

conditional on positive reporting from the state media. A citizen to the right of the target citizen changes her decision in favor of the regime only when the state of the world is good – with probability p – *and* the foreign media reports truthfully – with probability $(1 - \beta)$. A citizen to the left of the target citizen changes her decision *against* the regime when the foreign media truthfully reports that the state is bad - with probability $(1-p)$ – but also when they report untruthfully against the regime – with probability $p\beta$. These two forces work in the same direction. A low cost of access makes the authoritarian leader lose more compliance among conditional compliers than he gains among opponents.

4.4 Strong Correlation: Segment-and-Rule

If the intrinsic benefit is sufficietlty positively correlated with a citizen's type (high γ), then the intrinsic benefit becomes the main source of heterogeneous sorting into access: regime opponents have the strongest incentives to bypass the firewall and segment-and-rule is feasible.

Proposition 2. *There exists a unique* $\overline{\gamma} \in [\underline{\gamma}, 1 - \beta)$ *s.t. if* $\gamma \geq \overline{\gamma}$ *then the leader engages in segment-and-rule in equilibrium:*

- *there exists a unique equilibrium reporting slant* $\sigma^* = \sigma^S \in (\sigma^L, 1]$ *and target citizen* $\theta(\emptyset, \sigma^S) = \theta^S < \theta^L$*,*

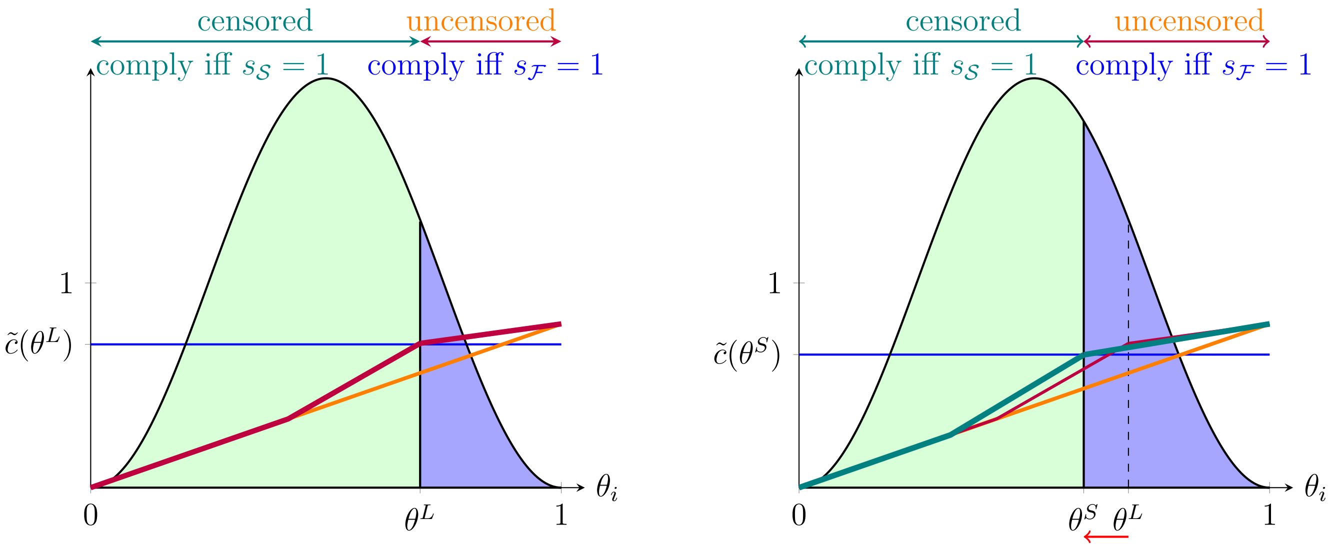

Figure 3: Same parameter values as in Figure 1 with the exception of $z = 0, \gamma = 0.8$. The left panel plots a partial equilibrium picture: the state media reports as if full censorship was implemented. The right panel depicts equilibrium behavior: the state media is less informative than under full censorship.

- *$c^* = \tilde{c}(\theta^S)$ is such that only opponents ($\theta_i > \theta^S$) gain access,*
- *a citizen complies if and only if $s_{\mathcal{S}} = 1$ and either (i) $\theta_i \leq \theta^S$ or (ii) $\theta_i > \theta^S$ and $s_{\mathcal{F}} = 1$,*
- *the level of compliance is maximized and constant in γ.*

When the correlation is strong the regime actively pursues selective censorship through a strategy of segment-and-rule. They pick an intermediate cost of access that is not prohibitively high: it ensures that only opponents bypass the firewall. Importantly, segment-and-rule is feasible through the imposition of an intermediate cost of access *because* the strong correlation generates selective exposure along partisan lines. This is illustrated in the left panel of Figure 3. The strong correlation (the orange line) ensures that the total benefit from bypassing the firewall (the red line) is always – meaning for any $\sigma \in (0, 1]$ – increasing in a citizen's misalignment with the regime.

The advantages of a strategy of segment-and-rule are two-folds. First and foremost, segment-and-rule ensures that only opponents select into bypassing the firewall. Conditional on positive reporting from the foreign outlet – with probability $p(1-\beta)$ – opponents comply, which could not happen under full censorship. Crucially, segment-and-rule delivers more compliance without affecting the behavior of unconditional and conditional compliers (vis-a-vis full censorship). This partial equilibrium logic is illustrated on the left panel of Figure 3. There, the state media is as informative as in the case of full censorship.

Second, segment-and-rule breeds division: it creates a cleavage along political lines between two segments of citizens. Opponents gain access to the foreign media and condition their compliance on its report while the regime's base does not and conditions its compliance on the state media's report. Thus, in equilibrium, the informativeness of the state media only affects the decision of unconditional and conditional compliers.

The state media is then a tool to communicate with the regime's base. It is less informative than in the case of full censorship (Proposition 1), such that the regime's base complies as often as possible while kept in the dark. This is illustrated in the right panel of Figure 3. To provide some intuition, notice that *given* full censorship, the leader faces the following censorship trade-off: increasing the share of compliers ($\uparrow F(\emptyset, \sigma)$) requires making the state media more informative ($\downarrow \sigma$). Formally, the leader faces the following problem:

$$\max_{\sigma} \quad V(\sigma, c = \bar{c}(\theta(\emptyset, \sigma))) \equiv \underbrace{[p + (1-p)\sigma]}_{Pr(\text{positive reporting})} \; F(\theta(\emptyset, \sigma)) \; + \; \underbrace{(1-p)(1-\sigma)}_{Pr(\text{true bad news})} F(0) \tag{2}$$

In contrast, *given* segment-and-rule, the regime secures the compliance of opponents as long the foreign media reports positively, with probability $p(1-\beta)$. They solve the following problem:

$$\max_{\sigma} V(\sigma, c = \tilde{c}(\theta(\emptyset, \sigma)) \equiv \underbrace{[p + (1-p)\sigma]}_{Pr(\text{positive reporting})} \; F(\theta(\emptyset, \sigma)) \; + \; \underbrace{p(1-\beta)}_{Pr(\text{true good news})} \; [F(1) - F(\theta(\emptyset, \sigma))] \tag{3}$$

Thus, to increase the share of compliers $F(\emptyset, \sigma)$ the regime must still make the state more informative ($\downarrow \sigma$). However, in so doing the regime also reduces the benefit from segment-and-rule as the share of opponents decreases ($[F(1) - F(\theta(\emptyset, \sigma))] \downarrow$). Thus, when segment-and-rule is possible the state media is very uninformative so as to ensure the compliance of the regime's base. In some cases it fully parrots the party line: $\sigma^S = 1$ and $\theta^S = p$.[18]

To recap, the leader actively pursues a particular form of selective censorship which involves communicating directly with the regime's base via the state media and indirectly with opponents via the foreign media. Vis-a-vis the low correlation and full censorship case, information flows less freely locally while some citizens bypass the firewall, by design.

[18] E.g., high prior p relative to the shape of the distribution or low reporting slant of the foreign outlet β.

Intermediary Correlation: Engineering Segmentation

A naive intuition may now suggest that a strong correlation between politics and the intrinsic benefit is a pre-requisite for a strategy of segment-and-rule. We now show that when this correlation is neither too strong nor too weak the leader can *engineer* segment-and-rule, by making local outlets even less informative than in the two previous cases.

The informational benefit $b_i(\theta_i, \sigma, 1, \beta)$ is single-peaked and maximized at the target citizen ($\theta_i = \theta(\emptyset, \sigma)$) and decreasing in a citizen's type for opponents. To make segment-and-rule feasible, the leader need only ensure that the total benefit from gaining access is (weakly) increasing in a citizen's type for opponents ($\theta_i > \theta(\emptyset, \sigma)$). The following equation zooms in on this particular segment of the population:

$$\frac{\partial \overbrace{\delta_i(\theta_i, \sigma, \beta, s_{\mathcal{S}} = 1)}^{\text{total benefit}}}{\partial \theta_i} > 0 \iff \underbrace{\frac{\partial \alpha(\theta_i)}{\partial \theta_i}}_{\text{correlation}} = \gamma > \underbrace{\theta(\emptyset, \sigma)(1 - \beta)}_{\text{reduction in } b_i(\cdot) \text{ among opponents}} \tag{4}$$

Equation (4) formalizes the necessary condition for segment-and-rule to be feasible: the total benefit $\delta_i(\theta_i, \sigma, \beta, s_{\mathcal{S}} = 1)$ of opponents must be increasing in a citizen's type θ_i. In the strong correlation case $\gamma \geq \overline{\gamma}$ ensured that the correlation was sufficiently large relative to the informativeness of the foreign media $(1 - \beta)$, such that the regime could pick *any* reporting slant (and thus target citizen $\theta(\emptyset, \sigma)$) and engage in segment-and-rule.

While this is no longer the case, the regime can pick a target citizen sufficiently aligned with the regime ($\theta(\emptyset, \sigma) \downarrow$) and *engineer* segment-and-rule. This requires making the state media less informative ($\sigma \uparrow$). Intuitively, as the state media loses informational content, the value of an additional report increases for any opponent. Thus the informational benefit becomes (relatively) more independent of a citizen's political type; formally the slope of the informational benefit becomes flatter (though still downward sloping). As a result, even a moderate correlation is sufficient for a strategy of segment-and-rule to be implemented.

Proposition 3. *If the correlation is intermediate $\gamma \in [\underline{\gamma}, \overline{\gamma})$ the leader engages in segment-and-rule,*

- *there exists a unique equilibrium reporting slant $\sigma^* = \sigma^I \in [\sigma^S, 1]$ and target citizen $\theta(\emptyset, \sigma^I) = \theta^I \in [p, \theta^S]$,*

- *$c^* = \tilde{c}(\theta^I)$ is such that only opponents ($\theta_i > \theta^I$) gain access,*
- *a citizen complies if and only if $s_{\mathcal{S}} = 1$ and either (i) $\theta_i \leq \theta^I$ or (ii) $\theta_i > \theta^I$ and $s_{\mathcal{F}} = 1$,*
- *the level of compliance is increasing in γ and bounded between*
  - *a lower bound: full-censorship compliance ($\gamma \leq \underline{\gamma}$), and*
  - *an upper bound: partial-censorship without engineering compliance ($\gamma \geq \overline{\gamma}$).*

When the correlation is intermediate the leader engineers segment-and-rule by making the state media parrot the party line even more than in the presence of a strong correlation ($\sigma^I \geq \sigma^S$). Figure 4 illustrates: if the state media is as informative as in the strong correlation case – the green line – then segment-and-rule is impossible as the total benefit is decreasing in a citizen's type for opponents. When the state is less informative – the red line – opponents have more to learn from bypassing the firewall, and segment-and-rule is possible.

Importantly this engineering comes at a cost: it requires making the state media less informative than would otherwise be optimal. Ideally the regime would segment-and-rule by communicating exactly as in the strong correlation case (Proposition 2). The regime must compromise on the reporting slant of the state media so that segment-and-rule is possible. Thus the level of compliance is increasing in the correlation between politics and entertainment. At one extreme ($\gamma = \overline{\gamma}$) segment-and-rule requires no compromise and the level of compliance reaches its upper bound. At the other ($\gamma = \underline{\gamma}$) the leader is indifferent between segment-and-rule and engaging in full censorship. This form of engineering is only observed in some contexts: the $[\underline{\gamma}, \overline{\gamma}]$ interval can be empty.[19]

To recap, modern censorship involves a strategy of segment-and-rule whenever heterogeneous sorting into access occurs sufficiently along the intrinsic dimension ($\gamma \geq \underline{\gamma}$). Then, the state media secures the compliance of the regime's base by parroting the party line while the foreign media inadvertently helps the regime by occasionally persuading opponents to comply.

[19] Engineering is never possible when the prior p is high relative to where most of the mass of citizen lie, such that the state media is already very uninformative when the correlation is strong (i.e. $\sigma^S \approx 1$). We provide precise conditions in the formal appendix.

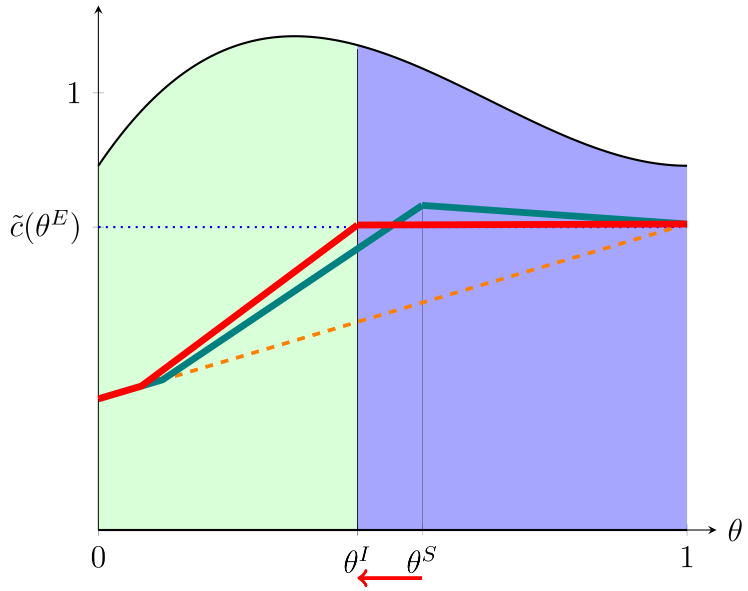

Figure 4: For this illustration, $f(x) = 2x(1-x)^2 + \frac{5}{6}$, $\beta = 0.1$ and $\alpha(\theta_i) = 0.3 + 0.4\theta_i$. The green line plots the willingness to pay conditional on the state media reporting as if there is a strong cleavage ($\gamma \geq \overline{\gamma}$). The red line plots the willingness to pay in equilibrium, given an intermediate correlation ($\gamma \in [\underline{\gamma}, \overline{\gamma}]$) and the equilibrium target citizen θ^I.

5 Comparative Statics and Implications

We now consider to what extent the regime can leverage the citizen's agency in accessing foreign content. To do so, we consider how the informativeness of the banned outlets $(1-\beta)$ affects equilibrium compliance and censorship.[20]

Proposition 4. *For any $\gamma \in (0, \theta^S)$, there exists a unique $\underline{\beta}$ and $\overline{\beta}$ with $0 < \underline{\beta} \leq \overline{\beta} < 1$ such that*

- *the equilibrium reporting slant σ^* is non-monotonic in β: it is constant for any $\beta < \underline{\beta}$, jumps at $\beta = \underline{\beta}$ and is decreasing in β otherwise.*
- *the equilibrium share of citizens who do not bypass the firewall is 1 for any $\beta < \underline{\beta}$, falls down at $\beta = \underline{\beta}$ and is increasing in β otherwise.*
- *the equilibrium compliance is non-monotonic and single-peaked in the informativeness of the foreign media and maximized at $\beta = \overline{\beta}$.*

Fixing some intermediary correlation $\gamma < \overline{\gamma}$, varying the informativeness of the foreign media $(1-\beta)$ affects whether segment-and-rule is feasible, and if so, at what cost, as illustrated in Figure

[20] The same comparative statics are derived with respect to γ and presented in Corollary A.1.

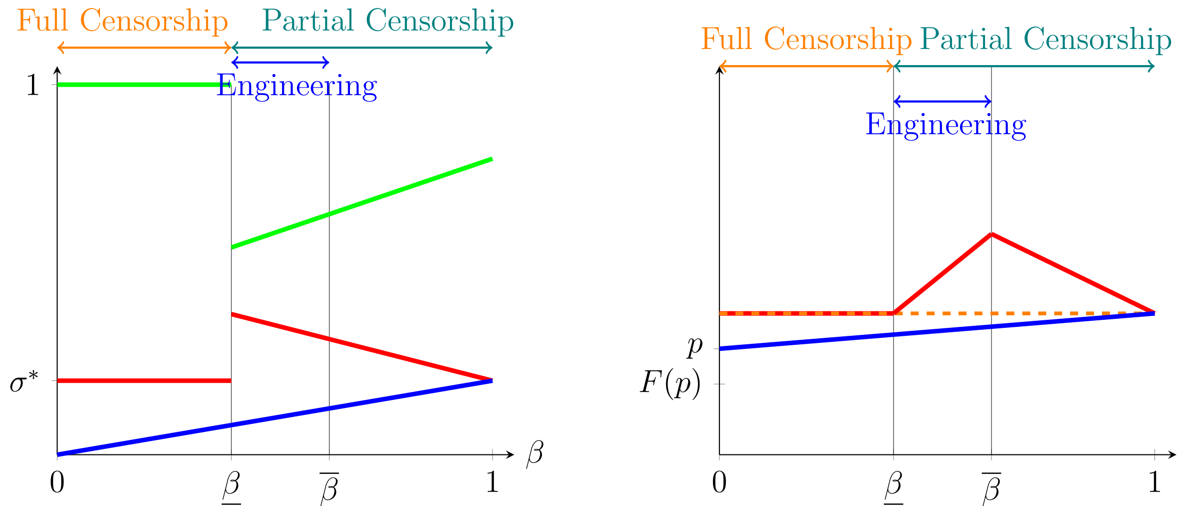

Figure 5: The left panel plots the equilibrium reporting slant (σ^*) absent any censorship (in blue) and given the possibility of manipulating the cost of access (in red), and illustrates the share of citizens who do not bypass the firewall in equilibrium (in green). The right panel illustrates the equilibrium level of compliance absent any censorship (in blue), given full censorship (in dashed orange) and given the possibility of manipulating the cost of access (in red).

5. The more informative the foreign outlet, the more a citizen's type determines her informational benefit (equation 4). Thus if the foreign outlet is too informative ($\beta < \underline{\beta}$) then segment-and-rule is impossible, and full censorship takes place. When segment-and-rule is feasible ($\beta \geq \underline{\beta}$) a clear pattern emerges (left panel of Figure 5). As the foreign outlet becomes less informative ($\beta \uparrow$), the share of citizens who bypass the firewall (green line) and the reporting slant (red line) move in opposite directions. This result is striking in three ways.

5.1 Non-Disciplining Effect of Competition

First, it is surprising from a theoretical perspective. In models of competition between senders the opposition media acts as a *credibility* constraint and disciplines the sender (Battaglini, 2002; Gentzkow and Kamenica, 2017): the more biased the foreign outlet, the more biased the state outlet (Heo and Zerbini, 2023; Zhou and Liu, 2023). This is illustrated by the blue line in the left panel of Figure 5 which plots the state media's reporting slant as a function of the foreign outlet reporting slant, in a hypothetical scenario where censorship is impossible, such as a democratic setting.

In sharp contrast, whenever segment-and-rule takes place in equilibrium ($\beta \geq \underline{\beta}$), the state media becomes less informative as the foreign outlet becomes more informative (the red line on the left panel). To understand this argument, notice that as the foreign media becomes more informative ($\beta \downarrow$), the likelihood of positive reporting that ensures the compliance of opponents increases. By making the state media less informative the leader faces a trade-off. First, his supporters are more likely to comply, since the state media parrots the party line more. Second, his base is smaller, and thus more citizens gain access. As the likelihood of positive reporting from the foreign outlet increases ($\beta \downarrow$), the loss from the smaller size of the leader's base is minimized because opponents comply more often. Thus, as the foreign outlet reports more truthfully, the state media can focus more on ensuring the compliance of its base, the compliers, which is achieved by making the state media less informative.

5.2 Conceptualizing and Empirically Measuring Censorship

Second, we show that in equilibrium a negative association emerges between two empirical measures of censorship – how freely information flows locally among regime controlled outlets, σ^* – and another – the share of citizens who only consume regime controlled outlets. Two implications follow for the literature on censorship. First, formal models that present a unifying framework (Shadmehr and Bernhardt, 2015; Gehlbach and Sonin, 2014) cannot shed light on these asymmetries. When the literature predicts domestic outlets to become less informative ($\sigma^* \uparrow$) such that the regime's base is less informed, then the opponents become more informed by gaining access to banned foreign outlets. Second, the empirical literature must also carefully assess these strategic dynamics: to claim that some shock leads to a change in "censorship" broadly defined, one must (i) measure both dimensions of censorship *and* (ii) obtain some proxy of a citizen's political type, so as to understand who becomes more or less informed.

Lastly, Proposition 4 suggests two potential policy implications for foreign actors interested in weakening an authoritarian leader via strategic investments in their own media landscape.

Implication 1. 1. *Compliance is lowest when banned outlets are most informative ($\beta = 0$) or most uninformative ($\beta = 1$).*

2. *Entertainment content that is polarizing across political lines (high γ) is (i) most likely to be banned and (ii) facilitates segment-and-rule and thus helps the leader achieve higher levels of compliance.*

Informational outlets should avoid being partially biased, in order to attempt to make segment-and-rule impossible ($\beta \leq \underline{\beta}$), or pointless ($\beta = 1$). Further, our framework suggest that strategies of "soft-power" via entertainment could backfire. Outright criticism of the authoritarian regime's culture and norms may backfire by creating content that specifically appeals to regime opponents; we make this point formally in section 6.3.

5.3 Limited Censorship Capacity

So far we have intentionally assumed that the leader faces no censorship capacity constraint, to show that an infinite cost of access would not be optimal even if it was feasible (unless if the correlation is low). In reality the capacity to impose a cost of access varies across regimes. In an extension (see Proposition A.1 in the appendix) we formally consider how a binding constraint – formally $c < \overline{C}$ – on the regime's censorship capacity matters if the correlation is low ($\gamma < \underline{\gamma}$). Then, in equilibrium, if the leader wants to ensure that no citizen bypasses the firewall, but cannot, then the leader imposes the highest possible cost of access.

Then, if the regime has a limited censorship capacity and the correlation is low ($\gamma \leq \underline{\gamma}$), our framework suggests two observationally equivalent explanations for the empirical pattern of selective bypassing of the firewall. To distinguish between them we first characterize the range of citizens that bypass the firewall when the leader would like to censor all citizens, but cannot.

Remark 2. *Suppose that the correlation is low ($\gamma < \underline{\gamma}$) and that full censorship is impossible ($\overline{C} < \bar{c}(\theta^L)$). There exists a unique $\overline{\overline{C}} \in (0, \bar{c}(\theta^L))$ such that the strongest opponent of the regime ($\theta_i = 1$) bypasses the firewall if and only if and only if $\overline{C} \leq \overline{\overline{C}}$.*

If an authoritarian regime can impose a non-negligible cost of access ($\overline{C} > \overline{\overline{C}}$) and aims to minimize the share of citizens bypassing the firewall ($\gamma \leq \underline{\gamma}$), then the range of citizens bypassing it does *not* include the most ex-ante opposed to the regime citizens ($\theta_i = 1$). The following empirical implication follows.

Implication 2. *If the regime's ability to impose a cost of access is not too limited ($\overline{C} \in (\overline{\overline{C}}, \bar{c}(\theta^L))$) and the most extreme opponents of the regime do bypass the firewall then the regime actively pursues partial censorship across citizens.*

Consider present-day China or Iran. Empirical evidence suggests that the regime strongest opponents do bypass the firewall (Shen and Zhang, 2018; Mou, Wu, and Atkin, 2016; Dal and Nisbet, 2022). Further, these regimes clearly possesses the technological capacity to impose a non-trivial cost of access. Our framework suggests that in such settings selective bypassing is a feature of the system, not a bug.

6 Extensions

We now enrich the theoretical framework with extensions that either speak to the robustness of our core theoretical insights or to further implications for the literature on media, censorship, entertainment and technology. framework to study other applications.

6.1 Non-Information Design Approach

The incentives to segment access to an independent source do not rely on the particular technology of information transmission. Only the engineering results of Proposition 3 do rely on the information design approach.

Game without private information. Consider a variant of the baseline game *without* a state-media: the regime need not have any commitment capacity and may thus not be able to credibly communicate. The regime only sets the cost of access to the foreign media.

Proposition 5. There exists a unique $\overline{\gamma}_p = p(1-\beta)$ such that in the unique equilibrium

- If $\gamma \geq \overline{\gamma}_p$ then $c^* = \tilde{c}(p)$. A citizen gains access if and only if $\theta_i \geq p$. All citizens with $\theta_i \leq p$ comply. Citizens with $\theta_i \geq p$ comply if and only if $s_{\mathcal{F}} = 1$.

- If $\gamma < \overline{\gamma}_p$ and $p < \hat{\theta}$ then $c^* \in (0, \overline{c}(p))$. A non-empty interval of citizens gain access.[21] A citizen complies if he does not gain access and is more aligned than the prior citizen ($\theta_i < p$) or if he gains access and observes $s_{\mathcal{F}} = 1$.

- If $\gamma < \overline{\gamma}_p$ and $p \geq \hat{\theta}$ then $c^* = \overline{c}(p)$. No citizen bypasses the firewall and a citizen complies if and only if $\theta_i \leq p$.

[21] A complete statement is provided in Lemma A.21.

When sorting occurs along the intrinsic dimension ($\gamma \geq \overline{\gamma}_p$) then the regime ensures that all the citizens not ex-ante convinced ($\theta_i > p$) gain access to the foreign media. More interestingly, the regime also allows access to the foreign media to some citizens even when segment-and-rule is not feasible ($\gamma < \overline{\gamma}_p$) but the regime is weak ($p < \hat{\theta}$); there the regime generates too little compliance from a full information shutdown and gambles on (unlikely) good news from the foreign media.

Game with private information. Authoritarian regimes may be privately informed about whether it is in the citizens' best interest to comply, which also informs the report from the foreign media they expect citizens to be exposed to. Thus, one may conjecture that the choice of the cost of access could signal the regime's type. We endow the regime with a private type, defined by the realization of their private signal $\hat{\omega} \in \{0, 1\}$ of precision $q \in (\frac{1}{2}, 1)$ that is observed at the beginning of the game. Upon observing this private signal the regime derives a belief $\mu(\hat{\omega}) \in (0, 1)$ about the state of the world and chooses both a cost of access c and a message $m \in [0, 1]$.

Proposition 6. *If $\gamma \geq \overline{\gamma}_p$ then there exists a pooling equilibrium on $c^*(1) = c^*(0) = \tilde{c}(p)$ and $m^*(1) = m^*(0) = m^* \in (0, 1)$ with an off-path belief of $\theta' \in [\mu(0), p]$ for any $c' \neq \tilde{c}(p)$ or $m' \neq m^*$ which sender-dominates all other pooling equilibria and survives both the intuitive and divinity criterion. Further, there exists no separating or semi-separating equilibrium.*

Whenever sorting occurs along the intrinsic dimension, both types of regime pool on the cost that segments access at the "prior-citizen" and on any cheap-talk message that they may send; a strategy of segment-and-rule is still implemented in equilibrium.

To provide some intuition, notice that separation cannot occur because the incentives of both types of regimes are aligned. First, fixing some belief μ of the citizens about the regime's type, any regime wants to pick the optimal segmenting cost $\tilde{c}(\mu)$. Second, both types of regime aim to look like high-types in order to increase compliance through a higher belief μ. More generally, Proposition 6 reinforces our main theoretical result: even a privately informed sender can benefit from the presence of an independent source that is likely to report negatively, provided that she can segment access to it.s

6.2 Consumption Beyond the Firewall

In practice individuals choose whether to download the VPN anticipating that they will get access to some particular content, once past the firewall. First, while some individuals may be information-seekers, others may instead be "intrinsic-content" seekers. Second, citizens may consume more than a single outlet. In this section we explain how these two considerations are embedded in our theoretical framework.

Motivations for access. The citizens' net benefit from gaining access can be re-written as follows

$$\delta_i(\theta_i, \sigma, \beta, s_{\mathcal{S}}, \alpha(\theta_i)) = \rho(b, \theta_i) \times \overbrace{b_i(\theta_i, \sigma, \beta, s_{\mathcal{S}})}^{\text{informational benefit}} + \underbrace{\rho(\alpha(\theta_i), \theta_i) \times \overbrace{\alpha(\theta_i)}^{\text{intrinsic benefit}}}_{\equiv \chi(\theta_i)} - \quad c$$

with $\rho(b, \theta_i)$ and $\rho(\alpha(\theta_i), \theta_i)$ capturing the extent to which an individual is more or of an information or intrinsic-content seeker. Since this involves some comparison across individuals of motivations for gaining access, we normalize $\rho(b, \theta_i) = 1$ while $\rho(\alpha(\theta_i), \theta_i) \in \mathbb{R}$. To illustrate, if *all* citizens are information (respectively, intrinsic-content) seekers then for instance $\rho(\alpha(\theta_i), \theta_i) = 0$ (respectively, $\rho(\alpha(\theta_i), \theta_i) = \rho >>> b_i(\cdot)$). Heterogeneity of intrinsic benefit along political lines could, for instance, be captured by the following functional form: $\rho(\alpha(\theta_i), \theta_i) = 10 \times \theta_i$.

In equilibrium segment-and-rule is feasible and the regime reaches their upper bound payoff if and only if:

$$\frac{\partial \delta_i(\sigma^S, \cdot)}{\partial \theta_i} > 0 \iff \frac{\partial b_i(\sigma^S, \cdot)}{\partial \theta_i} + \frac{\partial \chi_i(\theta_i)}{\partial \theta_i} > 0$$

I.e. the results of the baseline model still apply: it suffices to rewrite the results in terms of conditions on $\chi_i(\theta_i)$ rather than $\alpha(\theta_i)$ to derive the results of Proposition 1 through 4.

Exposure to multiple reports. Once past the firewall citizens could (choose to) observe reports from multiple media outlets who vary in their informativeness. Formally, let us suppose that there are n banned foreign outlets outside of the regime's control and accessible once past the firewall. Each outlet is indexed by $j \in \{1, 2, ...n\}$ and endowed with a reporting slant β_j and intrinsic content parameters $z^j_{\mathcal{F}}$ and $\gamma^j_{\mathcal{F}}$. Each outlet's report is independently drawn. Conditional on a by-passer observing at least one piece of good news from one banned outlet, then this citizen updates that

$\omega = 1$, and complies. Then define $\beta \equiv Pr(s_j = 0 \ \forall j \in \{1, ..., n\})$ and notice that the leader's problem is unaffected vis-a-vis the baseline model.

There are at least three intuitive ways to explicitly model exposure to the stream of signals. First, by-passers may observe *all* the signals from each banned outlet. The ex-ante probability that firewall by-passers comply is then given by $p(1 - \prod_{j=1}^{n} \beta_j)$ and we can here define $\beta \equiv \prod_{j=1}^{n} \beta_j$. Second, it might be that by-passers consume only one report. For instance they may only consume the most informative report. Then it suffices to define $\beta \equiv min_{j \in \{1,...n\}} \beta_j$. Alternatively, a by-passer may be equally likely to observe a single report from any of the n outlets. Then each citizen complies with probability $p(1 - \frac{\sum_{j=1}^{n} \beta_j)}{n})$ and it suffices to define $\beta \equiv \frac{\sum_{j=1}^{n} \beta_j}{n}$. The attention rule could range anywhere between these two extremes without affecting our central message: if opponents *can* be exposed to at least one report from an independent source that sways them into complying, then the sender benefits from the most skeptical receivers self-selecting into gaining access.

6.3 The Instrumental Value of Entertainment Control

So far we have solved for the baseline game, taking the correlation between politics and the intrinsic benefit γ as a primitive. In this section, building on the experimental evidence of Chan et al. (2019) we focus on one interpretation of the intrinsic benefit: the entertainment value from by-passing the firewall. We show how authoritarian regimes can instrumentally control the production and access to entertainment in order to make the control of information flows more efficient.

Investing in domestic entertainment. Rewrite the *relative* intrinsic benefit as follows:

$$\alpha(\theta_i) = \underbrace{z_{\mathcal{F}} + \gamma_{\mathcal{F}} * \theta_i}_{\text{foreign media intrinsic benefit}} - \underbrace{(z_{\mathcal{S}} + \gamma_{\mathcal{S}} * \theta_i)}_{\text{state media intrinsic benefit}}$$

such that $z = z_{\mathcal{F}} - z_{\mathcal{S}}$ and $\gamma = \gamma_{\mathcal{F}} - \gamma_{\mathcal{S}}$. Given some foreign content $(z_{\mathcal{F}}, \gamma_{\mathcal{F}})$, if the regime can manipulate both the quality of local entertainment $(z_{\mathcal{S}})$ and its relative appeal $(\gamma_{\mathcal{S}})$, they can achieve their upper bound compliance payoff. First, they create content which mostly appeals to their base. Formally this requires:

$$\gamma \geq \overline{\gamma} \iff \gamma_{\mathcal{S}} \leq \gamma_{\mathcal{F}} - \underbrace{\theta^S}_{\text{target citizen given strong correlation}} (1 - \beta) \equiv \overline{\gamma}_{\mathcal{S}} \qquad (5)$$

Next, they ensure that only opponents of the regime do gain access to censored content. This can be done by picking the appropriate cost of access c, *or* by picking the appropriate quality of local entertainment $z_{\mathcal{S}}$; $z_{\mathcal{S}}$ and c are substitutable levers. Formally this requires that:[22]

$$EU[\text{ bypass } |\theta_i = \theta^S] = EU[\text{ not bypass } |\theta_i = \theta^S] \iff z_{\mathcal{S}} = \underbrace{b_i(\theta^S, \sigma^S, \beta, s_{\mathcal{S}} = 1)}_{\text{informational benefit}} + z_{\mathcal{F}} + \theta^S(\gamma_{\mathcal{F}} - \overline{\gamma}_{\mathcal{S}}) \tag{6}$$

The minimal level of domestic quality to ensure that the regime can segment access optimally is increasing in both the quality of foreign content ($z_{\mathcal{F}}$) and the cleavage along political lines ($\gamma_{\mathcal{F}} - \overline{\gamma}_{\mathcal{S}}$).

Strategic bans. Chen and Yang (2019) also document that Chinese citizens *may* be exposed to information-intensive content *after* having bypassed the firewall with the goal of consuming low information and high entertainment content (social media, HBO, Youtube, etc.). Our theoretical framework can be interpreted in a similar manner. Then, the regime can think about affecting what entertainment *is* banned, in order to affect sorting into access and make segment-and-rule possible.

Suppose that there are n foreign outlets, each with endowed with a reporting slant β_j and non-informational content parameters $z^j_{\mathcal{F}}$ and $\gamma^j_{\mathcal{F}}$, as in section 6.2. Out of these n outlets, $\tilde{n} < n$ have no informational content ($\beta_j = 1 \forall j \in \{1, 2, ..., \tilde{n}\}$). We explicitly assume that all minimally informational foreign outlets (any outlet with $\beta_j < 1$) are banned and can only be accessed by bypassing the firewall. This is done to focus on how a strong correlation between politics and entertainment can be engineered by strategically banning some of the $\tilde{n}$ purely non-informational outlets. Then the regime selects $k \leq \tilde{n}$ of the non-informational outlets to ban. In turn we define

$$z_{\mathcal{F}} \equiv \sum_{i=1}^{k} \frac{z^i_{\mathcal{F}}}{k}, \quad \gamma_{\mathcal{F}} \equiv \sum_{i=1}^{k} \frac{\gamma^i_{\mathcal{F}}}{k}$$

That is, the entertainment value of the "foreign media" is defined as the average entertainment content value of all the k outlets banned by the regime.[23] Notice that the banning decision only affects the entertainment value from bypassing the firewall. The regime then solves for a dual problem: they look for a list of banned outlets that ensures (i) that $\gamma_{\mathcal{F}} = \gamma_{\mathcal{S}} + \theta^S(1 - \beta) \equiv \overline{\gamma}_o$ such that a strong correlation exists and (ii) that $z_{\mathcal{F}} = z_{\mathcal{S}} - b_i(\theta^s, \cdot) - \theta^S(\overline{\gamma}_o - \gamma_{\mathcal{S}})$ such that only opponents ($\theta_i > \theta^S$) bypass the firewall. These two goals may clash. If such a list does not exist

[22] We assume that the leader pick the lowest $\gamma_{\mathcal{S}}$, in absolute value, whenever indifferent.

[23] For simplicity and without loss we assume equal weights for each banned outlet.

then the regime can also make use of the cost of access c to adjust on the second problem and to instead focus on creating the required correlation by banning only the most polarizing outlets (the highest $\gamma^j_{\mathcal{F}}$).

In the case of present-day China, the CCP appears to be using a combination of (i) a positive cost of access – requiring that citizens invest in some VPN to bypass the firewall – and (ii) strategic investments in polarizing local entertainment – such as *The Battle of Lake Changjin* or *The Knockout*[24] – and (iii) strategic bans of entertainment – e.g., banning *Cockroach*, *Eternal Spring* or *Top Gun - Maverick* while not banning *Jurassic World Fallen Kingdom* or *Transformers Age of Extinction.*

We suggest an alternative explanation for the empirical pattern of domestic entertainment appealing mostly to regime supporters in authoritarian contexts. Domestic entertainment does not appeal to regime supporters because the regime rewards them (Esberg, 2020*a*); rather, domestic entertainment is unappealing to regime opponents so as to ensure that they self-select into consuming banned entertainment, and in turn information.

6.4 Modelling the Intrinsic Benefit

In this section we derive sufficient conditions on the intrinsic benefit $\alpha(\theta_i)$ for the regime to reach their upper bound payoff by segmenting access and setting $\sigma^* = \sigma^S$. Given such an equilibrium reporting slant, no restrictions need be applied to $\alpha(\theta_i)$ for all unconditional compliers ($\theta_i \leq \theta(0, \sigma^S))$) as they never condition their compliance decision on the report of the foreign outlet. Hereafter we focus on rest of the citizenry.

Intuitively, the regime can reach their upper bound equilibrium payoff from segment-and-rule whenever they can find a cost of access that ensures that (i) all *conditional compliers* ($\theta_i \in [\theta(0, \sigma^S), \theta(\emptyset, \sigma^S))$) do not gain access and (ii) *opponents* ($\theta_i \in [\theta(\emptyset, \sigma^S), 1]$) do gain access, *given* the optimal reporting slant σ^S for some distribution of political preferences f. To do so, we introduce two important

[24] In an experimental setting, Yao (2023) finds suggestive evidence that those most ex-ante inclined towards state propaganda in China – possibly regime supporters – are most likely to consume state propaganda. That is, CCP propaganda does appeal mostly to the regime's base.

quantities:

$$\overline{\delta}^{cc} \equiv \quad max_{\theta_i}\ \delta_i\Big(\theta_i \in [\theta(0,\sigma^S),\theta(\emptyset,\sigma^S)],\sigma^S, s_{\mathcal{S}}=1,\beta\Big)+c \tag{7}$$

$$\underline{\delta}^{o} \equiv \quad min_{\theta_i}\ \delta_i\Big(\theta_i \in (\theta(\emptyset,\sigma^S),1],\sigma^S, s_{\mathcal{S}}=1,\beta\Big)+c \tag{8}$$

$\overline{\delta}^{cc}$ denotes the maximal total benefit from gaining access amongst conditional compliers. $\underline{\delta}^{o}$ denotes the minimal total benefit from gaining access amongst opponents.[25]

Proposition 7. For any $\alpha(\theta_i)$ s.t. $\overline{\delta}^{cc} \leq \underline{\delta}^{o}$ the regime sets $\sigma^* = \sigma^S$ and $c^* \in [\overline{\delta}^{cc}, \underline{\delta}^{o}]$ to achieve their upper bound equilibrium payoff.

The straightforward logic of Proposition 7 is illustrated in Figure 6 with an example where the intrinsic benefit is non-monotonic in the misalignment with the regime. Notice that the distribution of political preferences f determines the equilibrium reporting slant and maximal attainable payoff of the regime, but not directly the feasibility of this equilibrium payoff, other than through the reporting slant σ^S. To recap, a strong (quasi-) linear positive association between misalignment with the regime and the intrinsic benefit is sufficient but not necessary for segment-and-rule; this strategy is also feasible when the intrinsic benefit is most enjoyed by both supporters and opponents and less by moderates.

6.5 Distribution of Preferences

The incentives to segment access do not rely on the distributional assumptions made on f: if anything, relaxing unimodality reinforces the incentives to *not* engage in full censorship, unlike in Proposition 1. To simplify exposition we consider as in Proposition 5 a minimalist version of the baseline model: there is no state-media and the regime only selects the cost of access c.[26]

Proposition 8. Consider any f with full support on $[0,1]$.

- If $\gamma \geq \overline{\gamma}_p$, $V(\bar{c}(p)) < V(\tilde{c}(p))$ and $c^* = \tilde{c}(p)$: whenever segment-and-rule is feasible, it dominates full censorship.

[25] Notice that these quantities include the informational benefit, given the optimal reporting slant.

[26] See Lemma A.25 in the appendix for results with the information design approach under different assumptions on f.

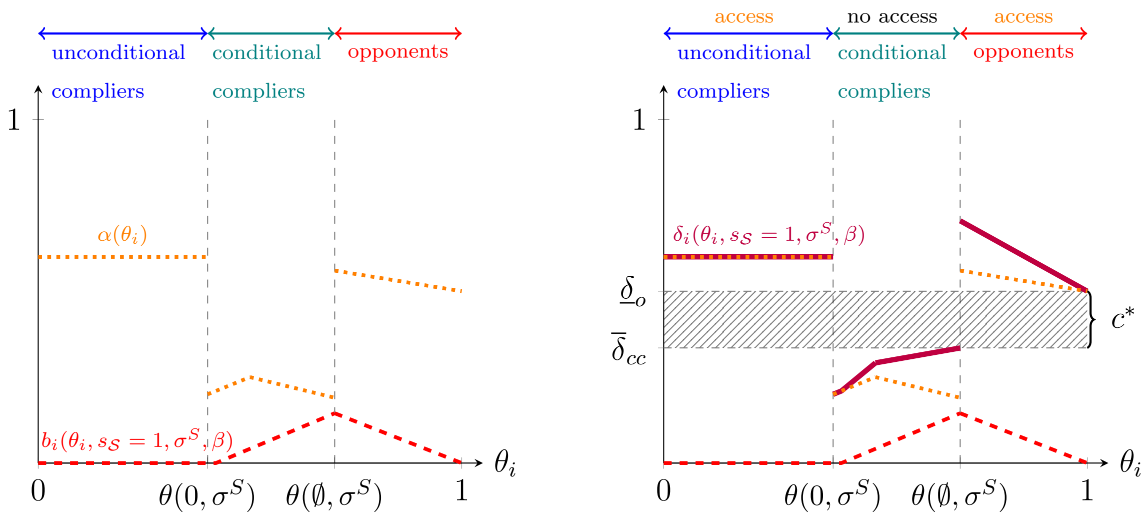

Figure 6: Same parameter values as in Figure 1. Left panel: the dashed red line plots the informational benefit given $s_{\mathcal{S}} = 1$ and $\sigma^* = \sigma^S$ while the dotted orange line plots a given non-monotonic step function intrinsic benefit $\alpha(\theta_i)$ such that unconditional compliers have the highest intrinsic benefit. Right panel: the purple line plots the total willingness to pay for access and the grey dashed area characterizes the set of costs the regime can choose from to secure their upper bound payoff.

- If $\gamma < \overline{\gamma}_p$ there exists pairs of (p, f) such that $V(\overline{c}(p)) < V(c \in [0, \overline{c}(p)))$: when segment-and-rule is not feasible, full censorship may be dominated by partial or no censorship.

When sorting is along the intrinsic dimension ($\gamma \geq \overline{\gamma}_p$) then segment-and-rule is optimal for *any* pair of distribution of preferences f and prior p. Intuitively, all citizens more skeptical than the prior citizens ($\theta_i \geq p$) only ever comply after good news from an independent source. When sorting is along the informational dimension, then segment-and-rule is not feasible ($\gamma < \overline{\gamma}_p$) and yet the regime need not engage in full censorship when f is not unimodal. To build intuition, suppose that sorting happens only along the informational dimension ($\gamma = 0$) and consider a polarized citizenry, such as the one plotted in Figure 7. The regime does not lose anything from letting unconditional compliers ($\theta_i < \theta(0, \sigma = 1) \equiv \theta_\beta$) gain access and can sometimes convince opponents ($\theta_i > \theta(\emptyset, \sigma = 1) = p$) by letting them gain access. Further, given such a distribution and prior pair (f, p) the regime loses at most the compliance of conditional compliers ($\theta_i \in [\theta_\beta, p]$), a rare breed in such a polarized citizenry.

6.6 Additional Extensions

Monitoring through segmentation. A high capacity regime may be able to observe which citizens are using a VPN. In doing so, the regime may be able to gather some information about its

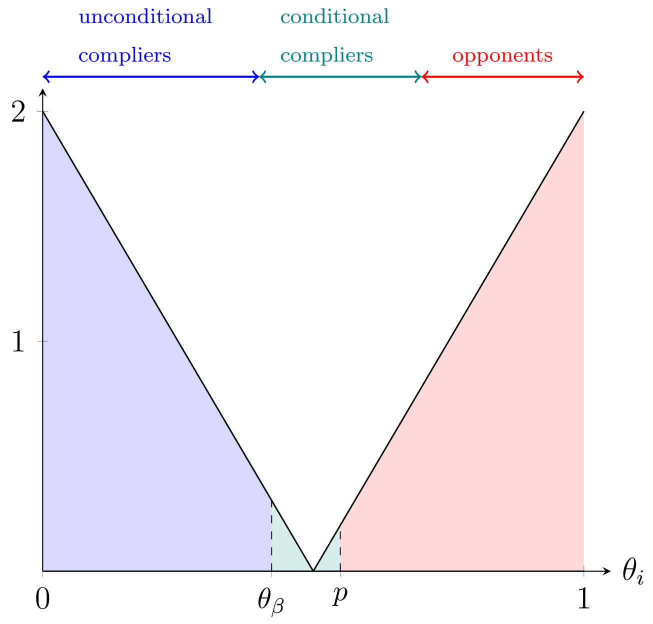

Figure 7: For this figure $f(\theta_i) = 2(1-2\theta_i)\forall\theta_i \in [0, \frac{1}{2})$ and $f(\theta_i) = 2(2\theta_i-1)\forall\theta_i \in [\frac{1}{2}, 1]$, $p = 0.55, \beta = 0.8$. The black lines plots $f(\theta_i)$.

citizens which could be very valuable, given that authoritarian regimes famously struggle to uncover the political leanings of their citizens (Kuran, 1991). In equilibrium the regime can anticipate which segment of citizens a given citizen belongs to if they can observe whether that citizen downloaded a VPN.

In an extension, on top of maximizing compliance, the regime also cares about learning the type of its opponents – for instance because these may be the first-movers in organizing against the regime.[27] Then, citizens know that downloading a VPN risk leading to one's type being uncovered, which could lead to some negative payoff – e.g., imprisonment or threats. We show that endowing the regime with some monitoring capacity reinforces the incentives to segment the population and results in a lower cost of access as citizens internalize the risk from gaining access.

Intertwined levers. In our baseline game, the correlation γ is a primitive and orthogonal to the reporting slant of the state media (σ). One may argue that such a high correlation exists (high γ) *because* the domestic outlets parrot the party line (high σ), which "annoys" opponents and lead to the development of an intrinsic benefit from consuming news from an independent source. We show that if γ is an increasing function of σ then the same qualitative results are derived, as the regime can engineer segmentation mechanically by increasing their reporting slant (Lemma A.15).

[27] See Lemma A.14 in the appendix.

Domestic Segmentation. We consider a game with two domestic outlets and without a foreign media to consider under which conditions segmentation could be achieved locally. We show that, as long as the regime cannot credibly commit to reporting negatively on itself, then all domestic outlets have the same reporting slant and the regime achieves its lower bound payoff of full-censorship.[28] The fact that there are no state-controlled outlets known to be outright critics of the regime they operate under, suggests that to segment access to information, authoritarian regimes must rely on banned independent (foreign outlets).

Non-binary compliance. For tractability reasons we assumed that the citizens' decision to comply is binary. One may be concerned that the attractiveness of a strategy of segment-and-rule relies on this modeling choice: with a continuous compliance level to choose from, as opponents are most of the time exposed to negative news from the foreign outlet, aggregate compliance could be lower under segment-and-rule than full censorship. To address this concern we allow for citizens to choose any compliance level $a_i \in [0,1]$ and derive sufficient conditions for a strategy of segment-and-rule to dominate full censorship, assuming that some equilibrium compliance profile $a_i^*(\theta_i, \mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta))$ exists. If the compliance level (i) decreases in misalignment with the regime, (ii) increases in the belief of a citizen *and* (iii) the usual increasing difference assumption holds – such that the more misaligned a citizen is with the regime, the smaller the marginal effect of a higher belief on ω – then segment-and-rule still dominates full censorship for any reporting slant $\sigma \in [0,1]$ (Lemma A.26).

7 Conclusion

We set out to explain an empirical regularity: despite large investments in censorship capacity, some citizens of authoritarian regimes still bypass firewalls to access an uncensored internet. To do so, we present a model of informational and intrinsic – e.g., entertainment – content control by an authoritarian regime and consumption by a population of heterogeneous citizens. We show that selective bypassing of firewalls to censorship is a feature of the censorship system of modern authoritarians, rather than a bug: it is is symptomatic of a strategy of *segment-and-rule*.

Segment-and-rule leverages both the citizens' agency of information acquisition – gained in the

[28] Formally, as in Heo and Zerbini (2023), the assumption that the regime cannot take over an opposition outlet and still credibly commit to reporting "against itself" – i.e. $Pr(s_{\mathcal{S}} = 1|\omega = 1) = 1$ by assumption – is crucial; see Lemma A.16.

post-internet era – and their heterogeneous political preferences: it keeps the regime's base in the dark and provides opponents with a credible foreign source of information. For this strategy to be feasible, regime opponents must benefit from bypassing the firewall more than regime moderates. Then, we show, to ensure this particular sorting pattern, authoritarian regimes can strategically control access to and production of non-informational content, such as entertainment. This highlights the crucial role of such content: while it does not affect beliefs, it helps "subsidise" the consumption of information from local outlets for the regime's base, and from foreign outlets for regime opponents.

We speak to a growing discussion on the use of modern technologies to facilitate authoritarian control. Scholars have argued that developments in AI could entrench autocrats more than they empower citizens. Because digital surveillance is less intrusive than in-person surveillance it can be rolled out with less resistance (Xu, 2023). This surveillance infrastructure can then facilitate pre-emptive suppression of organized dissent (Dragu and Lupu, 2021) and induce compliance via social-scoring rules (Tirole, 2021). In turn, digital surveillance can help authoritarian leaders reduce their provision of public good (Xu, 2021). To make matters worse, there exists a self-reinforcing dynamic between innovation in AI and the entrenchment of authoritarian regimes (Beraja et al., 2023).

We depict a yet grimmer picture by showing that simpler and cheaper technologies that do not involve any surveillance or data collection can be leveraged by authoritarian leaders. In the context of censorship, a naive intuition would suggest that by empowering citizens with more agency over their content consumption, the internet would have made censorship more difficult and helped citizens bring down authoritarian regimes. Unfortunately, it is precisely this agency gain that made possible a segmentation of the citizenry which authoritarian regimes use to improve their grasp on power. In this respect, the internet entrenched authoritarian regimes.

Bibliography

Acemoglu, Daron, James A Robinson, and Thierry Verdier. 2004. “Kleptocracy and Divide and Rule: A Model of Personal Rule.” *Journal of the European Economic Association* .

Acemoglu, Daron, Tarek A. Hassan, and Ahmed Tahoun. 2018. “The Power of the Street: Evidence from Egypt’s Arab Spring.” *The Review of Financial Studies* 31(1): 1–42.

Alonso, Ricardo, and Odilon Câmara. 2016. “Persuading Voters.” *The American Economic Review* 106(11): 3590–3605. Publisher: American Economic Association.

Arieli, Itai, and Yakov Babichenko. 2019. “Private Bayesian persuasion.” *Journal of Economic Theory* 182: 185–217.

Ashworth, Scott, and Ethan Bueno de Mesquita. 2006. “Monotone comparative statics for models of politics.” *American Journal of Political Science* 50(1): 214–231. Publisher: Wiley Online Library.

Awad, Emiel. 2020. “Persuasive Lobbying with Allied Legislators.” *American Journal of Political Science* 64(4): 938–951. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12523.

Bardhi, Arjada, and Yingni Guo. 2018. “Modes of persuasion toward unanimous consent.” *Theoretical Economics* 13(3): 1111–1149. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/TE2834.

Battaglini, Marco. 2002. “Multiple Referrals and Multidimensional Cheap Talk.” *Econometrica* 70(4): 1379–1401. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00336.

Beraja, Martin, Andrew Kao, David Y Yang, and Noam Yuchtman. 2023. “AI-tocracy*.” *The Quarterly Journal of Economics* 138(3): 1349–1402.

Bergemann, Dirk, and Stephen Morris. 2019. “Information design: A unified perspective.” *Journal of Economic Literature* 57(1): 44–95.

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris. 2015. “The Limits of Price Discrimination.” *American Economic Review* 105(3): 921–57.

Boxell, Levi, and Zachary Steinert-Threlkeld. 2019. “Taxing dissent: The impact of a social media tax in Uganda.” arXiv.

Caillaud, Bernard, and Jean Tirole. 2007. "Consensus Building: How to Persuade a Group." *The American Economic Review* 97(5): 1877–1900. Publisher: American Economic Association.

Chan, Jimmy, Seher Gupta, Fei Li, and Yun Wang. 2019. "Pivotal persuasion." *Journal of Economic Theory* 180: 178–202.

Chen, Yuyu, and David Y. Yang. 2019. "The Impact of Media Censorship: 1984 or Brave New World?" *American Economic Review* 109(6): 2294–2332.

Chester, Patrick J. 2023. "Framing Democracy: Characterizing China's Negative Legitimation Propaganda using Word Embeddings." Unpublished paper, . `https://patrickjchester.com/publication/chprop1/`

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic information transmission." *Econometrica: Journal of the Econometric Society* , 1431–1451. Publisher: JSTOR.

Dal, Aysenur, and Erik C. Nisbet. 2022. "Walking Through Firewalls: Circumventing Censorship of Social Media and Online Content in a Networked Authoritarian Context." *Social Media + Society* 8(4): 20563051221137738. Publisher: SAGE Publications Ltd.

Diamond, Larry. 2010. "Liberation Technology." *Journal of Democracy* .

Dragu, Tiberiu, and Yonatan Lupu. 2021. "Digital Authoritarianism and the Future of Human Rights." *International Organization* 75(4): 991–1017. Publisher: Cambridge University Press.

Edmond, Chris. 2013. "Information Manipulation, Coordination, and Regime Change." *The Review of Economic Studies* 80(4 (285)): 1422–1458. Publisher: [Oxford University Press, The Review of Economic Studies, Ltd.].

Egorov, Georgy, and Konstantin Sonin. 2023. "The Political Economics of Non-democracy." *Journal of Economic Literature* .

Egorov, Georgy, Sergei Guriev, and Konstantin Sonin. 2009. "Why Resource-poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data." *The American Political Science Review* 103(4): 645–668. Publisher: [American Political Science Association, Cambridge University Press].

Esberg, Jane. 2020*a*. "Anticipating Dissent: The Repression of Politicians in Pinochet's Chile.".

Esberg, Jane. 2020*b*. “Censorship as Reward: Evidence from Pop Culture Censorship in Chile.” *American Political Science Review* 114(3): 821–836. Publisher: Cambridge University Press.

Fung, Brian. 2022. “Russian internet users are learning how to beat Putin’s internet crackdown | CNN Business.”.

Galvis, Ángela Fonseca, James M. Snyder, and B. K. Song. 2016. “Newspaper Market Structure and Behavior: Partisan Coverage of Political Scandals in the United States from 1870 to 1910.” *The Journal of Politics* 78(2): 368–381. Publisher: The University of Chicago Press.

Gehlbach, Scott, and Konstantin Sonin. 2014. “Government control of the media.” *Journal of Public Economics* 118: 163–171.

Gentzkow, Matthew, and Emir Kamenica. 2017. “Competition in Persuasion.” *The Review of Economic Studies* 84(1): 300–322.

Gitmez, A. Arda, and Konstantin Sonin. 2023. “The Dictator’s Dilemma: A Theory of Propaganda and Repression.”.

Gitmez, A. Arda, and Pooya Molavi. 2023. “Informational Autocrats, Diverse Societies.” arXiv.

Gratton, Gabriele, and Barton E Lee. 2024. “Liberty, Security, and Accountability: The Rise and Fall of Illiberal Democracies.” *The Review of Economic Studies* 91(1): 340–371.

Guriev, Sergei, and Daniel Treisman. 2019. “Informational autocrats.” *Journal of economic perspectives* 33(4): 100–127. Publisher: American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203-2418.

Guriev, Sergei, and Daniel Treisman. 2020. “A theory of informational autocracy.” *Journal of public economics* 186: 104158. Publisher: Elsevier.

Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya. 2021. “3G Internet and Confidence in Government*.” *The Quarterly Journal of Economics* 136(4): 2533–2613.

Heo, Kun, and Antoine Zerbini. 2023. “Censorship in Large Societies.” Unpublished paper, Working paper.

Hobbs, William, and Margaret E Roberts. 2018. "How Sudden Censorship Can Increase Access to Information.".

Huang, Haifeng. 2015. "International knowledge and domestic evaluations in a changing society: The case of China." *American Political Science Review* 109(3): 613–634. Publisher: Cambridge University Press.

Huang, Haifeng, and Yao-Yuan Yeh. 2019. "Information from abroad: Foreign media, selective exposure and political support in China." *British Journal of Political Science* 49(2): 611–636. Publisher: Cambridge University Press.

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

Kolotilin, Anton, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li. 2017. "Persuasion of a privately informed receiver." *Econometrica* 85(6): 1949–1964. Publisher: Wiley Online Library.

Kuran, Timur. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44(1): 7–48. Publisher: Cambridge University Press.

Liu, Hanzhang, and Linan Yao. 2023. "Entertainment as Trojan Horse: Voluntary Consumption of Propaganda in China." Unpublished paper, . `http://linanyao.com/#research`

Lorentzen, Peter. 2014. "China's Strategic Censorship." *American Journal of Political Science* 58(2): 402–414. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12065.

Luo, Zhaotian, and Arturas Rozenas. 2023. "Ruling the Ruling Coalition: Information Control and Authoritarian Power-Sharing." *Quarterly Journal of Political Science* 18(2): 183–213. Publisher: Now Publishers, Inc.

Lutscher, Philipp M. 2023. "When Censorship Works: Exploring the Resilience of News Websites to Online Censorship." *British Journal of Political Science* , 1–9. Publisher: Cambridge University Press.

Manacorda, Marco, and Andrea Tesei. 2020. "Liberation Technology: Mobile Phones and Political Mobilization in Africa." *Econometrica* 88(2): 533–567. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14392.

Marshall, John, and Dorothy Kronick. 2022. “Collateral Censorship: Theory and Evidence from Venezuela.”.

Matysková, Ludmila, and Alfonso Montes. 2023. “Bayesian persuasion with costly information acquisition.” *Journal of Economic Theory* 211: 105678.

Maćkowiak, Bartosz, Filip Matějka, and Mirko Wiederholt. 2023. “Rational Inattention: A Review.” *Journal of Economic Literature* 61(1): 226–273.

Milgrom, Paul, and Chris Shannon. 1994. “Monotone comparative statics.” *Econometrica: Journal of the Econometric Society* , 157–180. Publisher: JSTOR.

Mou, Yi, Kevin Wu, and David Atkin. 2016. “Understanding the use of circumvention tools to bypass online censorship.” *New Media & Society* 18(5): 837–856. Publisher: SAGE Publications.

Nourin, Sadia, Van Tran, Xi Jiang, Kevin Bock, Nick Feamster, Nguyen Phong Hoang, and Dave Levin. 2023. “Measuring and Evading Turkmenistan’s Internet Censorship: A Case Study in Large-Scale Measurements of a Low-Penetration Country.” Paper presented at the Proceedings of the ACM Web Conference 2023,.

Pan, Jennifer, and Alexandra A. Siegel. 2020. “How Saudi Crackdowns Fail to Silence Online Dissent.” *American Political Science Review* 114(1): 109–125.

Qin, Bei, David Strömberg, and Yanhui Wu. 2018. “Media Bias in China.” *American Economic Review* 108(9): 2442–2476.

Rozenas, Arturas, and Denis Stukal. 2019. “How Autocrats Manipulate Economic News: Evidence from Russia’s State-Controlled Television.” *The Journal of Politics* 81(3): 982–996. Publisher: The University of Chicago Press.

Shadmehr, Mehdi, and Dan Bernhardt. 2015. “State Censorship.” *American Economic Journal: Microeconomics* 7(2): 280–307. Publisher: American Economic Association.

Shen, Fei, and Zhi’an Zhang. 2018. “Do circumvention tools promote democratic values? Exploring the correlates of anticensorship technology adoption in China.” *Journal of Information Technology & Politics* 15(2): 106–121. Publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2018.1449700.

Tirole, Jean. 2021. “Digital Dystopia.” *American Economic Review* 111(6): 2007–2048.

Xu, Xu. 2021. “To Repress or to Co-opt? Authoritarian Control in the Age of Digital Surveillance.” *American Journal of Political Science* 65(2): 309–325. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12514.

Xu, Xu. 2023. “The Unintrusive Nature of Digital Surveillance and Its Social Consequences.”.

Xue, Diwen, Benjamin Mixon-Baca, ValdikSS, Anna Ablove, Beau Kujath, Jedidiah R. Crandall, and Roya Ensafi. 2022. “TSPU: Russia’s decentralized censorship system.” Paper presented at the Proceedings of the 22nd ACM Internet Measurement Conference, Nice France.

Yao, Linan. 2023. “From Screen to Mind: An Experimental Examination of the Power of Propagandist Entertainment on Chinese Public Opinion.”.

Zhou, Congyi, and Liqun Liu. 2023. “How Propaganda and Censorship Cause Global Media Competition.” *Google Docs* .

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. “Political Effects of the Internet and Social Media.” *Annual Review of Economics* 12(1): 415–438. _eprint: https://doi.org/10.1146/annurev-economics-081919-050239.

For Online Publication: Appendix

Notation and Assumptions

- $\hat{\theta}$ denotes the mode of $f(\theta)$
- $\theta^{\dagger}$ is the unique solution to $F(\theta^{\dagger}) = \theta^{\dagger}$ which exists by assumption and must be unique when it does (by the unimodality of $f(\theta)$)
- $\theta^{L} \equiv \arg\max_{\theta} \frac{F(\theta)}{\theta}$ for a unimodal F with some $\theta^{\dagger} \in (0, 1)$.
- $\theta_{\beta} \equiv \frac{p\beta}{p\beta+(1-p)} = Pr(\omega = 1|\sigma = 1, s_{\mathcal{S}} = 1, s_{\mathcal{F}} = 0)$
- $\sigma^{L} \equiv argmax_{\sigma} V(\sigma, \beta, \text{no citizen gains access to } s_{\mathcal{F}})$ denotes the optimal reporting slant conditional on no citizen gaining access.
- $\sigma_{nc} \equiv argmax_{\sigma} V(\sigma, \beta, \text{all citizen gain access to } s_{\mathcal{F}})$ denotes the optimal reporting slant conditional on all citizens gaining access.
- $\pi \equiv Pr(s_{\mathcal{S}} = 1|\omega = 1)$

Assumptions

1. Throughout we assume that $\gamma \geq 0$. For any $\gamma < 0$ full censorship is optimal.
2. Indifference-breaking assumptions:
   (a) Conditional on being indifferent between complying and not doing so after observing $s_{\mathcal{S}} = 1$ (and possibly some other information), a citizen complies.
   (b) Conditional on being indifferent between circumventing the firewall and not doing so, an individual does circumvent the firewall.
   (c) We do not model any intrinsic cost of censorship to the leader and assume that whenever indifferent between a range of cost of access, the leader picks the lowest cost of access within that range.

Baseline Model: General Propositions

In this section we provide complete and more general formal statements of each proposition presented in the paper. Conditional on full censorship occuring in equilibrium (Proposition A.1), we also assume that there exists a possibly binding upper bound on the cost of access that the regime can access, which captures the censorship technological capacity of the regime. Formally, $c \in [0, \overline{C}]$ with $\overline{C} \in \mathbb{R}_+$. That is, Proposition A.1 nests Proposition 1.

To do so we introduce some additional notation and definitions.

Definition 2. The range of citizens gaining access to the foreign outlet is given by $[,]$ with $0 \leq \leq \theta(\emptyset, \sigma) \leq 1$.

In equilibrium we denote these thresholds by and . We now introduce a definition that clarifies which citizens comply with the regime, given some set of reports.

Definition 3. Suppose that $\pi = 1$. For a given reporting strategy σ, $\theta(\hat{s}_{\mathcal{F}} = \emptyset, \sigma) \equiv Pr(\omega = 1 | \hat{s}_{\mathcal{F}} = \emptyset, s_{\mathcal{S}} = 1, \sigma)$ is referred to as the *target citizen*. It denotes the citizen who is indifferent between complying and not doing so after only observing good news from the state media, given some reporting slant σ.

Every citizen consumes the state media, $s_{\mathcal{S}} \in \{0, 1\}$. However, not all citizens necessarily consume the foreign media: abusing notation, $\hat{s}_{\mathcal{F}} \in \{0, 1, \emptyset\}$.

Proposition A.1. *There exists a unique $\underline{\gamma} \in (0, 1-\beta)$ and a unique $\underline{C}(\sigma_{nc}) \in (0, \bar{c}(\theta^L))$ such that if $\gamma \in [0, \underline{\gamma}]$ then in the unique equilibrium,*

- *if $\overline{C} \leq \underline{C}(\sigma_{nc})$ then $c^* = 0$, $\sigma^* = \sigma_{nc}$, $\theta(0, \sigma^*) = \theta^L$ and $= 1$. If $z \geq 0$ then $= 0$, otherwise $= \theta(0, \sigma^*)$. A citizen complies if and only if (i) $s_{\mathcal{F}} = 1$ or (ii) $s_{\mathcal{F}} = 0, s_{\mathcal{S}} = 1, \theta_i \leq \theta(0, \sigma^*)$.*
- *if $\overline{C} > \underline{C}(\sigma_{nc})$*
  - *full censorship takes place whenever possible: $c^* = min\{\bar{c}(\theta^L), \overline{C}\}$. If $\overline{C} < \bar{c}(\theta^L)$ then citizens in the range $[,]$ gain access, with $= \theta^L$ and $\in (\theta(\emptyset, \sigma^*), 1]$. The share of citizens bypassing the firewall is decreasing in $\overline{C}$.*
  - *$\sigma^*(\overline{C}) \in [\sigma_{nc}, \sigma^L]$ and σ^* increases in $\overline{C}$. $\theta(\emptyset, \sigma^*)$ decreases in $\overline{C}$ with $\lim_{\overline{C} \to \bar{c}(\theta^L)} = \theta^L$.*

- *A citizen complies if and only if (i) $\hat{s}_{\mathcal{F}} = 1$ or (ii) $\hat{s}_{\mathcal{F}} = \emptyset, s_{\mathcal{S}} = 1, \theta_i \leq$.*
- *compliance (weakly) increases in the censorship capacity of the regime $\overline{C}$.*

Discussion: the importance of the binding constraint on the cost of access in low correlation citizenries. When the leader's ability to impose a cost of access is too limited ($\overline{C} \leq \underline{C}(\sigma_{nc})$) then it is optimal not to impose any cost of access, simply because increasing the cost of access reduces the share of opponents with access (↓) while not affecting the share of unconditional compliers: formally $\underline{\theta}(\alpha_n, \sigma) < \theta(0, \sigma)$ when the upper bound on the cost of access is too low.
Otherwise, when the leader's ability to impose a cost of access is limited (but not too limited, i.e. $\overline{C} > \underline{C}(\sigma_{nc})$) the regime minimizes the share of citizens bypassing the firewall. This is done via the two levers at their disposal. First, they sets the cost of access as high possible ($c^* = \overline{C}$). Second, they make the state media more informative to reduce the informational benefit from bypassing the firewall. As the regime's ability to impose a cost of access decreases ($\overline{C} \downarrow$) the state media becomes more informative. Formally, among firewall by-passers, the one most aligned with the regime is the same as the one that would just comply after contradictory reporting in the case of no censorship ($= \theta^L$). As the regime's ability to impose a cost of access increases ($\overline{C} \uparrow$), $\sigma^* \uparrow$ and converges towards θ^L. Compliance is thus maximized under full censorship. Put differently, if partial censorship does occur when $\gamma \leq \underline{\gamma}$ in equilibrium it is because it is unavoidable, not because it is actively pursued by the regime.

Proposition A.2. *There exists a unique $\overline{\gamma} \in (\underline{\gamma}, 1 - \beta)$ s.t. if $\gamma \geq \overline{\gamma}$ then in the unique equilibrium:*

- $\sigma^* = \sigma^S \in (\sigma^L, 1]$ *such that* $\theta(\emptyset, \sigma^S) = \theta^S < \theta^L$.
- *censorship is partial even when full censorship is possible. $c^* = \tilde{c}(\theta^S)$ ensures that only citizens more misaligned than θ^S gain access; $= \theta^S$ and $= 1$.*
- *A citizen complies if and only if (i) $\hat{s}_{\mathcal{F}} = 1$ or (ii) $\hat{s}_{\mathcal{F}} = \emptyset, s_{\mathcal{S}} = 1, \theta_i \leq \theta^S$.*
- *the level of compliance is maximized and constant in γ and for any $\gamma \geq \overline{\gamma}$.*

Proposition A.3. *If $\gamma \in [\underline{\gamma}, \overline{\gamma})$ then in the unique equilibrium,*

- $\sigma^* = \sigma^I \in [\sigma^S, 1]$ *such that* $\theta(\emptyset, \sigma^*) = \theta^I \in [p, \theta^S]$.

- *$c^* = \tilde{c}(\theta^I)$ ensures that only citizens $\theta_i \geq \theta^I$ gain access; $= \theta^I$ and $= 1$.*
- *A citizen complies if and only if (i) $\hat{s}_{\mathcal{F}} = 1$ or (ii) $\hat{s}_{\mathcal{F}} = \emptyset, s_{\mathcal{S}} = 1, \theta_i \leq \theta^I$.*
- *compliance is strictly increasing in γ for any $[\underline{\gamma}, \overline{\gamma}]$ and bounded between*
  - *a lower bound: the full-censorship payoff ($\gamma \leq \underline{\gamma}$), and*
  - *an upper bound: the segment-and-rule payoff from the strong correlation case ($\gamma \geq \overline{\gamma}$).*

Finally, this corollary is close in spirit to Proposition 4 and not included in the main text.

Corollary A.1. *For any $\beta \in (0, 1)$, the equilibrium*

1. *reporting slant (σ^*) is non-monotonic in γ: it is constant for any $\gamma < \underline{\gamma}$, jumps at $\gamma = \underline{\gamma}$ and is decreasing in γ otherwise with $\lim_{\gamma \to \infty} \sigma^* > max\{\sigma^* \mid \gamma < \underline{\gamma}\}$*
2. *share of citizens who are censored – do not bypass the firewall – is 1 for any $\gamma < \underline{\gamma}$, falls at $\gamma = \underline{\gamma}$, is increasing in γ (but always strictly below 1) otherwise*
3. *compliance is weakly increasing in γ, strictly so for any $\gamma \in [\underline{\gamma}, \overline{\gamma}]$.*

We now provides proofs for each full proposition and corrolary provided above, as well as any other results presented in the core of the paper.

Proofs

First, we characterize optimal propaganda σ^* given no censorship and full censorship. It is useful to introduce the concept of an upper bound $\overline{C} \in \mathbb{R}_+$ s.t. $c \in [0, \overline{C}]$.

Lemma A.1. *If $\overline{C} = 0$ then there exists a unique $\theta^L \in [max\{p, \hat{\theta}, \theta^\dagger\}, 1)$ and $\sigma_{nc} = min\{\beta \frac{p}{1-p} \frac{1-\theta^L}{\theta^L}, 1\}$ which maximize compliance. Further, compliance increases in β.*
If $\overline{C} > \overline{c}(\theta^L)$ and $c \geq \overline{c}(\theta^L)$ s.t. no citizen gains access then there exists a unique $\theta^L \in [max\{p, \hat{\theta}, \theta^\dagger\}, 1)$ and $\sigma^L = min\{\frac{p}{1-p} \frac{1-\theta^L}{\theta^L}, 1\}$ which maximize compliance.

Proof. Suppose that $\overline{C} = 0$ s.t. $c = 0$ by constraint. The leader's problem is given by

$$\begin{aligned}
&\max_{\sigma} \quad [p + (1-p)\sigma]F(\theta(0,\sigma)) + p(1-\beta)(1 - F(\theta(0,\sigma))) \\
&\max_{\theta(0,\sigma)} \quad p\beta\frac{F(\theta(0,\sigma))}{\theta(0,\sigma)} + p(1-\beta) \\
&(FOC) \quad f(\theta(0,\sigma))\theta(0,\sigma)) - F(\theta(0,\sigma))) = 0
\end{aligned}$$

First, notice that $f(0) \times 0 = F(0) = 0$ and $f(1) \times 1 < F(1) = 1$. Next, notice that, abusing notation, $\frac{\partial}{\partial\theta} f(\theta)\theta = f(\theta) + f'(\theta) \geq \frac{\partial F(\theta)}{\partial\theta} = f(\theta)$ if and only if $\theta \leq \hat{\theta}$. Thus, $f(\theta)\theta > F(\theta)$ for $\theta \leq \hat{\theta}$ and there exists $\theta^L > \hat{\theta}$ such that $f(\theta^L)\theta^L = F(\theta^L)$ by the Intermediate Value Theorem. Further, at θ^L the SOC wrt to θ yields $f'(\theta^L)\theta^L < 0$ since $\theta^L > \hat{\theta}$. Given θ^L, there exists a unique $\sigma^* = \beta\frac{p}{1-p}\frac{1-\theta^L}{\theta^L}$ associated with it. Importantly, it must be the case that the target citizen θ^L can be reached with $\sigma^* \in [0,1]$. If $\theta_\beta > \theta^L$ then $\sigma^* = 1$. Formally, $\sigma^* \equiv \arg\max_\sigma V(\sigma, c = 0)$ is given by

$$\sigma^* = \begin{cases} 1 & \text{if} \quad \theta_\beta \in [\theta^L, 1] \\ \beta\frac{p}{1-p}\frac{1-\theta^L}{\theta^L} & \text{if} \quad \theta_\beta \in [0, \theta^L). \end{cases}$$

Notice that σ^* is increasing in β. Notice that $\sigma^*(c > \overline{c}(\theta^L)) = \sigma^*(c = 0, \beta) = 1$. It follows that compliance increases in β.

Finally, suppose that we allow for $\pi \equiv Pr(s_{\mathcal{S}} = 1|\omega = 1) < 1$. We provide this proof for $c = 0$ which implies the proof holds for $\beta = 1$ which is equivalent to $c > \overline{c}(\theta^L)$. For any $\pi < 1$, write

$$\begin{aligned}
\theta_1^\pi \equiv& \theta(1, 0, \sigma, \pi) = \frac{p\beta\pi}{p\beta\pi + (1-p)\sigma} \\
\theta_0^\pi \equiv& \theta(0, 0, \sigma, \pi) = \frac{p\beta(1-\pi)}{p\beta(1-\pi) + (1-p)(1-\sigma)}
\end{aligned}$$

Note that $\theta(1,\sigma) \geq \theta_1^\pi$ for any $\pi < 1$. Further, $\theta_1^\pi \geq \theta_0^\pi \iff \pi \geq \sigma$. The leader's payoff can be written as

$$V(\sigma, \beta, \pi) = p\Big[\beta\Big[\frac{F(\theta_1^\pi)}{\theta_1^\pi}\pi + \frac{F(\theta_0^\pi)}{\theta_0^\pi}(1-\pi)\Big] + (1-\beta)\Big] \tag{9}$$

We first note that the payoff from true good news from the foreign media is not affected by π nor σ

and can thus be ignored. Observe the following:

$$
\begin{aligned}
V(\sigma, \beta, \pi) &= \pi\Big[p\beta\frac{F(\theta_1^\pi)}{\theta_1^\pi} + (1-\beta)p\Big] + (1-\pi)\Big[p\beta\frac{F(\theta_0^\pi)}{\theta_0^\pi} + (1-\beta)p\Big] \\
&= \pi V(\sigma; \beta, \pi) + (1-\pi)V(1-\sigma; \beta, 1-\pi).
\end{aligned}
$$

Intuition for the formal argument below: one can maximise individually $V(\sigma;\beta,\pi)$ and $V(1-\sigma;\beta,1-\pi)$ by solving for the optimal target citizen θ^L (problem already solved in the baseline game). But then $V(\sigma;\beta,\pi)$ and $V(1-\sigma;\beta,1-\pi)$ cannot be maximised jointly with any $\pi \in (0,1)$.

Formally, we know that $V(\sigma;\beta,\pi)$ attains its unique maximum when $\theta_1^\pi(\sigma) = \theta^L$. Thus, for σ such that $\theta_1^\pi(\sigma) \neq \theta^L$, $V(\sigma;\beta,\pi) < \max_\sigma V(\sigma;\beta,\pi) = V(\sigma^*;\beta,c=0)$. Similarly, $V(1-\sigma;\beta,1-\pi) < \max_\sigma V(1-\sigma;\beta,1-\pi) = V(\sigma^*;\beta,c=0)$ for σ such that $\theta_0^\pi(\sigma) \neq \theta^L$. Therefore, $V(\sigma,\beta,\pi) \leq V(\sigma^*;\beta,c=0)$. □

Denote the maximal payoff from $c = 0$ by $V(\theta^L, c=0) = p[\beta\frac{F(\theta^L)}{\theta^L} + (1-\beta)] \geq p$ where the last inequality follows from the fact that the leader can always ensure a payoff of p by setting $\sigma = 0$. Given some $\theta(\emptyset,\sigma)$, denote the cost of access – assuming it exists – that ensures that only citizens above the target citizen (i.e. $\theta_i \geq \theta(\emptyset,\sigma)$) gain access, by $\tilde{c}(\theta(\emptyset,\sigma))$. Similarly denote the payoff from segment-and-rule (hereafter SAR) assuming it is possible and given some σ by $V(\theta(\emptyset,\sigma),\tilde{c}(\theta(\emptyset,\sigma))) = [p+(1-p)\sigma]F(\theta(\emptyset,\sigma)) + p(1-\beta)[1-F(\theta(\emptyset,\sigma)]$.

Lemma A.2. *Fixing some target citizen $\theta(\emptyset,\sigma)$, whenever SAR is possible, then $c^* = \tilde{c}(\theta(\emptyset,\sigma))$ maximizes compliance. SAR is possible iff $\frac{\partial\delta_i(\theta_i,\sigma,s_{\mathcal{S}})}{\partial\theta_i} \geq 0 \forall\theta_i \in [\theta^{min}, 1] \iff \gamma \geq \theta(\emptyset,\sigma)(1-\beta)$.*

Proof. Given some $\theta(\emptyset,\sigma)$, all $\theta_i \leq \theta(\emptyset,\sigma)$ comply conditional on only observing good news from the state media. Since δ_i is increasing in θ_i, denote by $\tilde{c}$ the unique solution to $\delta_i(\theta_i = \theta(\emptyset,\sigma),\sigma,s_{\mathcal{S}} = 1,\tilde{c}) = 0$.

Note first that any $c' > \tilde{c}$ is dominated by $\tilde{c}$: it only reduces compliance among types $\theta_i > \theta(\emptyset,\sigma)$ and not affect the decision-making of $\theta_i \leq \theta(\emptyset,\sigma)$. Further, any $c' < \tilde{c}$ is dominated by $\tilde{c}$. Given some c', citizens gain access iff $\theta_i > \theta'$ (with $\theta' < \theta(\emptyset,\sigma)$). Then the leader is better off under $\tilde{c}$ iff

$$
(p + (1-p)\sigma - p(1-\beta))[F(\theta(\emptyset,\sigma)) - F(\theta')] > 0
$$

which trivially holds. Thus, *fixing* σ, $c^* = \tilde{c}(\sigma)$.

□

Lemma A.3. *Suppose that SAR is feasible* ($\frac{\partial \delta_i(\theta_i,\sigma,s_S)}{\partial \theta_i} \geq 0 \forall \theta_i \in [max\{p, \hat{\theta}, \theta^\dagger\}, 1]$) *such that* $c^* = \tilde{c}(\theta(\emptyset,\sigma))$. *There exists a unique* $\theta^S \in [p, \theta^L)$ *and* $\sigma^S \in (\sigma^L, 1]$ *which maximizes compliance. Further,* θ^S *increases in* β *and* σ^S *decreases in* β.

Proof. We need only consider $c = \tilde{c}(\theta)$. In turn the leader's problem boils down to

$$\max_{\theta(\emptyset,\sigma)\in(p,1)} [p + (1-p)\sigma]F(\theta(\emptyset,\sigma)) + p(1-\beta)(1 - F(\theta(\emptyset,\sigma))) = p\frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)} + p(1-\beta)(1 - F(\theta(\emptyset,\sigma)))$$

$$(FOC) \quad f(\theta(\emptyset,\sigma))[1 - \theta(\emptyset,\sigma)(1-\beta)] = \frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$$

$$(SOC) \quad \propto f'(\theta(\emptyset,\sigma))(1 - \theta(\emptyset,\sigma)(1-\beta)) = 2f(\theta(\emptyset,\sigma))(1-\beta)$$

First, recall that we assume that $f(1) < 1$ in order to have $\theta^\dagger \in (0,1)$. Then, notice that $f(\theta(\emptyset,\sigma))[1-\theta(\emptyset,\sigma)(1-\beta)]$ is decreasing in $\theta(\emptyset,\sigma)$ forall $\theta(\emptyset,\sigma) > \hat{\theta}$. Also, $\frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$ is single peaked, and maximized at θ^L: $\frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$ increases in θ iff $\theta(\emptyset,\sigma) \leq \theta^L$. Also $f(\theta(\emptyset,\sigma)) \geq \frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$ iff $\theta(\emptyset,\sigma) \leq \theta^L$.
Denote the interior solution to (FOC), if any, by θ^S. Note that $f(\theta^L) = \frac{F(\theta^L)}{\theta^L}$ thus $\theta^S < \theta^L$ and $\sigma^S > \sigma^L$. It must be that $\theta^S < \theta^L$ since $p(1-\beta)[1 - F(\theta(\emptyset,\sigma))]$ is maximized at $\theta(\emptyset,\sigma) = p$ for any $\beta < 1$. There are 2 cases to consider

1. Case 1: $f(\theta(\emptyset,\sigma))[1-\theta(\emptyset,\sigma)(1-\beta)] < \frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)} \forall \theta(\emptyset,\sigma) \in [p,1]$. In this case $\theta^S = p$ and $\sigma^S = 1$. Note that this necessitates $f(p)[1 - p(1-\beta)] < \frac{F(p)}{p}$.

2. Case 2: $f(\theta(\emptyset,\sigma))[1 - \theta(\emptyset,\sigma)(1-\beta)] = \frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$ has 1 or more local argmax on the interval $[p, \theta^L]$.
Step 1: consider some local argmax $\theta^S \geq \hat{\theta}$. *For any* $\theta \in (\theta^S, \theta^L)$, $\frac{F(\theta)}{\theta} - f(\theta)(1-\theta(1-\beta)) > 0$. $\theta^S \geq \hat{\theta}$ implies that $f'(\theta) < 0 \forall \theta > \theta^S$ which proves the claim.
Step 2: consider some local argmax $\theta^S < \hat{\theta}$. *For any* $\theta \in (\theta^S, \theta^L)$, $\frac{F(\theta)}{\theta} - f(\theta)(1-\theta(1-\beta)) > 0$. Suppose not. Then there must exist at least one pair of θ^1 and θ^2 s.t. $\theta^S < \theta^1 < \theta^2 < \hat{\theta}$ where θ^1 is a local argmin and θ^2 a local argmax. Rewrite the SOC evaluated at θ^S as follows:

$$(SOC) \qquad \frac{f'(\theta^S)}{f(\theta^S)} \frac{(1-\theta^S(1-\beta))}{(1-\beta)} < 2$$

(where the inequality follows from θ^S being a local argmax). Notice that $1 - \theta^S(1-\beta)$ decreases in θ^S and so does $\frac{f'(\theta^S)}{f(\theta^S)}$ (weakly) for any log concave f thus no such θ^1 can exist. A contradiction.

Step 3: if some $\theta^S \in (p, \theta^L)$ *exists, then it is unique.* By step 1, if the smallest θ^S is above the mode of f, then it is unique. By step 2, if the smallest θ^S is below the mode of f, then it is unique.

Less stringent assumptions on f. Notice that log-concavity is sufficient but not necessary. It would suffice to require that $\frac{f'(\theta^S)}{f(\theta^S)}\frac{(1-\theta^S(1-\beta))}{(1-\beta)}$ is non-increasing for $\theta \in (p, \hat{\theta})$.

To recap, either a unique interior θ^S exists and is unique, or no such interior θ^S exists and then $\theta^S = p$.

Note that the leader could always pick $\theta = \theta^L$ and $c = \overline{c}(\theta^L)$ in order to attain her ex-ante censorship payoff; since she does not, compliance is strictly higher than under ex-ante censorship.

Finally, observe that as β increases, the LHS of (FOC) increases: $\lim_{\beta\to 1}\theta^S = \theta^L$, and thus σ^S decreases in β.

Finally, notice that it is always optimal to set $\pi = 1$. Suppose not. Then building on the same proof as in Lemma A.1 notice that the compliance level is given by

$$\begin{aligned} V^S(\sigma,\beta,\pi) &= p\Big[\frac{F(\theta_1^\pi)}{\theta_1^\pi}\pi + \frac{F(\theta_0^\pi)}{\theta_0^\pi}(1-\pi) + (1-\beta)(1-F(\theta_1^\pi))\Big] \\ &= \underbrace{p\Big[\frac{F(\theta_1^\pi)}{\theta_1^\pi} + (1-\beta)(1-F(\theta_1^\pi))\Big]}_{V^S(\sigma,\tilde{c}(\theta(\emptyset,\sigma)),\pi=1)} + \underbrace{(1-\pi)\frac{F(\theta_0^\pi)}{\theta_0^\pi} - (1-\pi)\frac{F(\theta_1^\pi)}{\theta_1^\pi}}_{\equiv Z} \end{aligned}$$

Observe first that $Z > 0 \iff \theta^L < \theta_0^\pi < \theta_1^\pi$ (recall that θ^L is the unique argmax of $\frac{F(\theta)}{\theta}$ and $\frac{F(\theta)}{\theta}$ decreases in θ for any $\theta > \theta^L$) and thus we need only consider this case. Then recall from Lemma A.1 that $\frac{F(\theta_1^\pi)}{\theta_1^\pi}\pi + \frac{F(\theta_0^\pi)}{\theta_0^\pi}(1-\pi) < \frac{F(\theta^L)}{\theta^L}$ and thus

$$\begin{aligned} &p\Big[\frac{F(\theta_1^\pi)}{\theta_1^\pi}\pi + \frac{F(\theta_0^\pi)}{\theta_0^\pi}(1-\pi) + (1-\beta)(1-F(\theta_1^\pi))\Big] < p\Big[\frac{F(\theta^L)}{\theta^L} + (1-\beta)(1-F(\theta_1^\pi))\Big] \\ &< p\Big[\frac{F(\theta^L)}{\theta^L} + (1-\beta)(1-F(\theta^L))\Big] < p\Big[\frac{F(\theta^S)}{\theta^S} + (1-\beta)(1-F(\theta^S)\Big] \end{aligned}$$

where the penultimate inequality follows from $\theta_1^\pi > \theta^L$. Thus notice that, for any $\pi \in (\sigma, 1)$ the regime would be better off with $\pi = 1$ and $\sigma = \sigma^L$. The last inequality follows directly from the core part of proof of this very lemma. □

For the next three lemmas we assume that full censorship is feasible ($\overline{C} > \overline{c}(\theta^L)$) and compare the payoffs from segmentation to that of full censorship. Later on show that when segmentation is dominated by full censorship, the regime maximizes compliance by censoring as much as possible ($c^* = \overline{C}$), or not at all ($c^* = 0$).

Lemma A.4. *Assume $\overline{C} > \overline{c}(\theta^L)$. There exists a unique $\overline{\gamma} \in (0, 1-\beta)$ s.t. if $\gamma \geq \overline{\gamma} \equiv \theta^S(1-\beta)$ then $c^* = \tilde{c}(\theta^S)$, only citizens of type $\theta_i \geq \theta^S$ gain access, $\sigma^* = \sigma^S$ and $\theta(\emptyset, \sigma^*) = \theta^S$.*

Proof. Given some target citizen $\theta(\emptyset, \sigma)$, $\frac{\partial \delta_i(\theta_i, \sigma, s_S)}{\partial \theta_i} \geq 0 \iff \gamma \geq \theta(\emptyset, \sigma)(1-\beta)$.
Recall that, given some β and p, compliance is maximized when the leader can pick $\theta(\emptyset, \sigma^*) = \theta^S$. If $\gamma \geq \theta^S(1-\beta)$ the leader can pick any target citizen $\theta(\emptyset, \sigma) \in [\theta^S, 1]$ and ensure that the net benefit from gaining access is (weakly) increasing in θ_i in equilibrium. In turn, by Lemma A.3, it follows that $c^* = \tilde{c}(\theta^S)$ and $\theta(\emptyset, \sigma^*) = \theta^S$.

□

Lemma A.5. *Assume $\overline{C} > \overline{c}(\theta^L)$ and $\gamma \in (p(1-\beta), \overline{\gamma})$. If $\theta^S > p$ then there exists a unique $\theta^I \in (p, \theta^S)$ s.t. $\theta(\emptyset, \sigma)(1-\beta) \geq \gamma \iff \theta(\emptyset, \sigma) \leq \theta^I$. Otherwise no such θ^I exists.*

Proof. Denote by θ^I the largest $\theta(\emptyset, \sigma) \in [p, \theta^S]$ s.t. $\gamma \geq \theta(\emptyset, \sigma)(1-\beta)$. Denote the reporting slant associated with θ^I by σ^I. Notice that $\theta(\emptyset, \sigma)(1-\beta)$ monotonically increases in $\theta(\emptyset, \sigma)$: by the intermediate value theorem such a θ^I must exist and be unique. □

Lemma A.6. *Suppose θ^I exists and assume $\overline{C} > \overline{c}(\theta^L)$ and $\gamma \in (p(1-\beta), \overline{\gamma})$; further, suppose that $p\frac{F(\theta^L)}{\theta^L} > p\frac{F(p)}{p} + p(1-\beta)[1-F(p)]$. Then there exists a unique $\theta^\dagger_w$ s.t. $p\frac{F(\theta^L)}{\theta^L} \geq p\frac{F(\theta^I)}{\theta^I} + p(1-\beta)[1-F(\theta^I)] \iff \theta^I < \theta^\dagger_w$.*

Proof. By assumption $\theta^I < \theta^S$ and by inspection of (FOC) and (SOC), θ^I dominates any $\theta(\emptyset, \sigma) \in (p, \theta^I)$. Further, any $\theta(\emptyset, \sigma) \in (\theta^I, \theta^S)$ is dominated by θ^I and/or θ^L since it does not allow for SAR *and* leads to suboptimal propaganda.
Consider under which conditions is the leader better off generating SAR (choosing θ^I) rather than ensuring his full ex-ante censorship payoff (choosing θ^L). This is the case iff

$$[p + (1-p)\sigma^L]F(\theta^L) < [p + (1-p)\sigma^I]F(\theta^I) + p(1-\beta)[1-F(\theta^I)]$$
$$\iff \frac{F(\theta^L)}{\theta^L} < \frac{F(\theta^I)}{\theta^I} + (1-\beta)[1-F(\theta^I)]$$

where the second line follows from (i) $\theta^L = \frac{p}{p+(1-p)\sigma^L}$ and (ii) $\theta^I = \frac{p}{p+(1-p)\sigma^I}$. Note that by Lemma A.3 $\frac{F(\theta^I)}{\theta^I} + (1-\beta)[1-F(\theta^I)]$ is maximized and single-peaked at $\theta^I = \theta^S$. Thus there are two cases to consider:

1. If $\frac{F(\theta^L)}{\theta^L} \leq \frac{F(p)}{p} + (1-\beta)[1-F(p)]$ then $\theta(\emptyset, \sigma^*) = \theta^I$.

2. If not, then by Lemma A.1 and the intermediate value theorem there exists a unique $\theta^\dagger_w \in (p, \theta^S)$ s.t. if $\theta^I \geq \theta^\dagger_w$ then $\theta(\emptyset, \sigma^*) = \theta^I$, and otherwise $\theta(\emptyset, \sigma^*) = \theta^L$.

Finally note that if $\gamma < p(1-\beta) < \theta^\dagger_w(1-\beta)$, then irrespective of the choice of the target citizen, $\frac{\partial \delta_i}{\partial \theta_i} < 0 \; \forall \theta_i > \theta(\emptyset, \sigma)$.

□

In turn, define $\underline{\gamma} \equiv \theta^\dagger_w(1-\beta)$ as the minimal correlation such that SAR dominates full censorship. Observe that $\underline{\gamma} \in (0, \overline{\gamma})$.

In Lemmas A.7 to A.13 we characterize equilibrium play when the leader's maximal payoff from perfect segmentation is dominated by full censorship: that is, for all such lemmas we assume that $\gamma < \underline{\gamma}$. Within that range, we consider both the cases of a non binding and a binding upper bound on the cost of access: i.e. $\overline{C} < \overline{c}(\theta^L) = z + \gamma\theta^L + b_i(\theta_i = \theta^L | \theta(\emptyset, \sigma) = \theta^L)$ where $b_i(\cdot | \theta(\emptyset, \sigma))$ denotes the informational benefit from consuming $s_{\mathcal{F}}$ for type θ_i, given some target citizen $\theta(\emptyset, \sigma)$.

For the purpose of the following lemmas we write the net common benefit from consuming the foreign media as $\delta_i(\theta_i, \sigma, s_{\mathcal{S}}) = b_i(\theta_i, \sigma, s_{\mathcal{S}}) + \gamma\theta_i + \alpha_n$ where $\alpha_n = z - c$ denotes the net of cost common non informational benefit from consuming the foreign media. Observe that if $\alpha_n \geq 0$, then the entire citizenry consumes the foreign media. We treat α_n as an exogenous parameter, and, in some lemmas, switch from the cost c to the net common benefit α_n.

Hereafter we sometimes use the notation $(\theta(\emptyset, \sigma), c)$ or $(\theta(\emptyset, \sigma), \alpha_n)$ to denote a strategy - a target citizen and a cost of access - for the leader.

Lemma A.7. *Suppose $0 \leq \overline{C} \leq \overline{c}(\theta^L)$. Given some $\theta(\emptyset, \sigma) \in (\theta^\dagger, 1)$ there exists two unique cutoffs $\underline{\theta} \in [0, \theta(\emptyset, \sigma)] = \frac{\beta\theta(\emptyset,\sigma) - \alpha_n}{1-(1-\beta)\theta(\emptyset,\sigma)+\gamma}$ and $\overline{\theta} \in [\theta(\emptyset, \sigma), 1] = \frac{\alpha_n + (1-\beta)\theta(\emptyset,\sigma)}{(1-\beta)\theta(\emptyset,\sigma)-\gamma}$ s.t. only citizens of type $\theta_i \in [\underline{\theta}, \overline{\theta}]$ circumvent the firewall after observing $s_{\mathcal{S}} = 1$.*

Proof. θ_i consumes $s_{\mathcal{F}}$ if and only if it increases her expected payoff:

$$\max_{a_i \in \{0,1\}} E[u_i(a_i|\theta_i, s_{\mathcal{S}}, s_{\mathcal{F}}, \sigma, \beta)] + \alpha_n + \theta_i \gamma \geq \max_{a_i \in \{0,1\}} E[u_i(a_i|\theta_i, s_{\mathcal{S}}, \emptyset)]$$

Suppose that $s_{\mathcal{S}} = 0$. Then there is no informational benefit of consuming $s_{\mathcal{F}}$, so the entire citizenry never complies and consumes $s_{\mathcal{F}}$ iff $\alpha_n \geq 0$.

Recall that $\mu(\cdot, 1) \equiv \frac{p}{p+(1-p)\sigma} = \theta(\emptyset, \sigma)$, $\mu(0,1) \equiv \frac{p\beta}{p\beta+(1-p)\sigma} = \theta(0, \sigma)$. Note that we need only consider $\alpha_n < 0$. For any $\alpha_n \geq 0$ all citizens consume $s_{\mathcal{F}}$ and thus the leader is indifferent between any $c > 0$ s.t. $\alpha_n > 0$ and, by assumption, the leader always picks the lowest cost of access between any cost of access that yields the same level of compliance.

Claim 1: There exists a unique $\underline{\theta}(\alpha_n, \sigma) \in [0, \theta(\emptyset, \sigma)]$, such that a citizen of type $\theta_i \in [0, \theta(\emptyset, \sigma)]$ gains access iff $\theta_i \geq \underline{\theta}(\alpha_n, \sigma)$. Following $s_{\mathcal{S}} = 1$, a citizen gains access iff

$$\underbrace{[\beta\theta(\emptyset, \sigma) + (1 - \theta(\emptyset, \sigma))]}_{Pr(s_{\mathcal{F}}=0|s_{\mathcal{S}}=1)}\theta_i + \underbrace{(1-\beta)\theta(\emptyset, \sigma)}_{Pr(s_{\mathcal{F}}=1|s_{\mathcal{S}}=1)} + \underbrace{\alpha_n + \gamma * \theta_i}_{\text{intrinsic benefit}} \geq \underbrace{\theta(\emptyset, \sigma)}_{\text{payoff from not gaining access and } s_{\mathcal{S}}=1}$$

$$\iff \theta_i \geq min\{\frac{\beta\theta(\emptyset, \sigma) - \alpha_n}{1 - (1-\beta)\theta(\emptyset, \sigma) + \gamma}, \theta(\emptyset, \sigma)\} \equiv \underline{\theta}(\alpha_n, \sigma)$$

We note that for any $\theta_i \in [0, \theta(\emptyset, \sigma)]$, the informational benefit is given by $b_i(\theta_i, \cdot) = (\beta\theta(\emptyset, \sigma) + 1 - \theta(\emptyset, \sigma))\theta_i - \beta\theta(\emptyset, \sigma)$ and $\frac{\partial^2 b_i(\theta_i, \cdot)}{\partial\theta_i \partial\theta(\emptyset, \sigma)} < 0$: the higher $\theta(\emptyset, \sigma)$, the flatter the slope of $b_i(\theta_i, \cdot)$. Note too that $\underline{\theta}(\alpha_n, \sigma) = \theta(\emptyset, \sigma)$ if $\alpha_n \leq \underline{\alpha}_n \equiv -(1-\beta)\theta(\emptyset, \sigma)(1 - \theta(\emptyset, \sigma))$.

Claim 2: There exists a unique $\overline{\theta}(\alpha_n, \sigma) \in [\theta(\emptyset, \sigma), 1]$, such that $\theta_i \in [\theta(\emptyset, \sigma), 1]$ gain access iff $\theta_i \leq \overline{\theta}(\alpha_n, \sigma)$. Citizens with $\theta_i > \theta(\emptyset, \sigma)$ gain access following $s_{\mathcal{S}} = 1$ iff

$$\underbrace{(1-\beta)\theta(\emptyset, \sigma)}_{Pr(s_{\mathcal{F}}=1|s_{\mathcal{S}}=1)} + \underbrace{[\beta\theta(\emptyset, \sigma) + (1 - \theta(\emptyset, \sigma))]}_{Pr(s_{\mathcal{F}}=0|s_{\mathcal{S}}=1)}\theta_i + \underbrace{\alpha_n + \gamma * \theta_i}_{\text{intrinsic benefit}} \geq \underbrace{\theta_i}_{\text{payoff from not gaining access}}$$

$$\iff \theta_i \leq max\{\theta(\emptyset, \sigma), min\{\frac{\alpha_n + (1-\beta)\theta(\emptyset, \sigma)}{(1-\beta)\theta(\emptyset, \sigma) - \gamma}, 1\}\} \equiv \overline{\theta}(\alpha_n, \sigma)$$

We note that for any $\theta_i \in (\theta(\emptyset, \sigma), 1]$ the informational benefit is given by $b_i(\theta_i, \cdot) = (1-\beta)(1 - \theta_i)\theta(\emptyset, \sigma)$ and $\frac{\partial^2 b_i(\theta_i, \cdot)}{\partial\theta_i \partial\theta(\emptyset, \sigma)} < 0$: the higher $\theta(\emptyset, \sigma)$, the steeper the slope of $b_i(\theta_i, \cdot)$. □

Lemma A.8. *Given some $\theta(\emptyset, \sigma)$ (i) there exists a unique $\theta_F^M = min\{\theta_i \in [0,1] : a_1^*(\theta_i, s_{\mathcal{S}} = 1, \hat{s}_{\mathcal{F}}, \sigma) = 1\} = max\{\theta(0, \sigma), \underline{\theta}(\alpha_n, \sigma)\}$ such that all $\theta_i \leq \theta_F^M$ comply conditional on $s_{\mathcal{S}} = 1$, and (ii)* *there exists a unique* $\underline{\underline{\alpha}}_n \in [\underline{\alpha}_n, 0)$ *s.t.* $\underline{\theta}(\alpha_n, \sigma) \leq \theta(0, \sigma)$ *iff* $\alpha_n \geq \underline{\underline{\alpha}}_n$.

Proof. Claim 1: there exists a unique $\theta_F^M = min\{\theta_i \in [0,1] : a_1^*(\theta_i, s_{\mathcal{S}} = 1, \hat{s}_{\mathcal{F}}, \sigma) = 1\} = max\{\theta(0,\sigma), \underline{\theta}(\alpha_n, \sigma)\}$ *such that all* $\theta_i \leq \theta_F^M$ *comply conditional on* $s_{\mathcal{S}} = 1$. There are two cases to consider. Case 1: $\theta(0,\sigma) \leq \underline{\theta}(\alpha_n, \sigma)$. By definition all $\theta_i \leq \underline{\theta}(\alpha_n, \sigma)$ are exposed to $\hat{s}_{\mathcal{F}} = \emptyset$ and thus comply following $s_{\mathcal{S}} = 1$. Case 2: $\theta(0,\sigma) > \underline{\theta}(\alpha_n, \sigma)$. Here all types $\theta_i \in [\underline{\theta}(\alpha_n, \sigma), \theta(0,\sigma)]$ are exposed to $\hat{s}_{\mathcal{F}} \in \{0, 1\}$ but always comply following $s_{\mathcal{S}} = 1$.

Claim 2: there exists a unique $\underline{\underline{\alpha}}_n \in [\underline{\alpha}_n, 0)$ *s.t.* $\underline{\theta}(\alpha_n, \sigma) \leq \theta(0, \sigma)$ *iff* $\alpha_n \geq \underline{\underline{\alpha}}_n$. This follows directly from the facts that (i) $\underline{\theta}(\alpha_n, \sigma)$ decreases in α_n and (ii) for any $\alpha_n \geq 0$ then all citizens gain access ($\underline{\theta}(\alpha_n, \sigma) = 0$) and (iii) for any $\alpha_n \leq \underline{\alpha}_n$ no citizen gains access ($\underline{\theta}(\alpha_n, \sigma) = \theta(\emptyset, \sigma)$) and (iii) $\theta(0,\sigma) \geq \theta_\beta > 0$. We note that this implies the existence of a unique $\underline{C} \in [0, \overline{C}]$ s.t. $\forall c \in [0, \underline{C}]$ then $\underline{\theta}(\alpha_n, \sigma) \leq \theta(0, \sigma)$. □

Lemma A.9. $\theta_F^M = \theta^L > \hat{\theta}$ for any $\alpha_n \in \mathbb{R}$.

Proof. Claim 1: given some $\theta(\emptyset, \sigma)$ *and* $\alpha_n < \underline{\underline{\alpha}}_n$ *then* $\theta_F^M = \underline{\theta}(\alpha_n, \sigma) = \theta^L$. By $\alpha_n < \underline{\underline{\alpha}}_n$ we know that $\underline{\theta}(\alpha_n, \sigma) > \theta(0, \sigma)$. Observe that $\frac{\partial \underline{\theta}(\alpha_n, \sigma)}{\partial \alpha_n} < 0$ and thus $\underline{\theta}(\alpha_n, \sigma)$ reaches its minimum at $\alpha_n = \underline{\underline{\alpha}}_n$. Further, recall that we already pinned down the equilibrium reporting strategy of the state media under no censorship ($\alpha_n \geq \underline{\underline{\alpha}}_n$) and full censorship ($\alpha_n \leq \underline{\alpha}_n$) in Lemma A.1. Thus

$$lim_{\alpha_n \to \underline{\underline{\alpha}}_n} \theta(0, \sigma) = \theta^L = lim_{\alpha_n \to \underline{\underline{\alpha}}_n} \theta(\emptyset, \sigma)$$

This follows from the fact that both with full and no censorship, given some cdf F, the share of compliers conditional on $s_{\mathcal{S}} = 1$ is held constant; only σ^* changes (see Lemma A.1). Thus, by the squeeze theorem, $lim_{\alpha_n \to \underline{\underline{\alpha}}_n} \underline{\theta}(\alpha_n, \sigma) = \theta^L$ and thus $\underline{\theta}(\alpha_n, \sigma) \geq \theta^L$. Importantly, we also know that $lim_{\alpha_n \to \underline{\alpha}_n} \theta(\emptyset, \sigma) = \theta^L$; since $\underline{\theta}(\alpha_n, \sigma)$ is maximized at $\alpha_n = \underline{\alpha}_n$ and since $\underline{\theta}(\alpha_n, \sigma) \leq \theta(\emptyset, \sigma)$, this implies that $\underline{\theta}(\alpha_n, \sigma) \leq \theta^L$. Thus $\underline{\theta}(\alpha_n, \sigma) = \theta^L$. Further we recall that θ^L is on the concave side of F; thus $\underline{\theta}(\alpha_n, \sigma) > \hat{\theta}$.

Claim 2: given some $\theta(\emptyset, \sigma)$ *and* $\alpha_n \geq \underline{\underline{\alpha}}_n$ *then* $\theta_F^M = \theta(0, \sigma) = \theta^L$. If in equilibrium $\alpha_n \geq \underline{\underline{\alpha}}_n$ then it is optimal to set $c^* = 0 \implies \alpha_n \geq 0$. By Lemma A.1 this implies that compliance is maximized with $\sigma = \sigma_{nc}$ s.t. $\theta(0, \sigma_{nc}) = \theta^L$. □

Lemma A.10. If $\alpha_n > \underline{\underline{\alpha}}_n \iff \underline{\theta}(\alpha_n, \sigma) < \theta(0, \sigma)$ compliance increases in α_n and is maximized at

$\alpha_n \geq 0$. Further, $\sigma(\alpha_n) = \arg\max V_w(\sigma; \alpha_n > \underline{\underline{\alpha}}_n, \beta) = \sigma_{nc}$.

Proof. Given $\alpha_n > \underline{\underline{\alpha}}_n \iff \underline{\theta}(\alpha_n, \sigma) < \theta(0, \sigma)$ then compliance is maximized by imposing no cost of access ($\alpha_n \geq 0$) because $\frac{\partial \overline{\theta}(\alpha_n,\sigma)}{\partial \alpha_n} > 0$ while $\frac{\partial \theta(0,\sigma)}{\partial \alpha_n} = 0$. Further, given that it is optimal not to impose any cost of access, we know from Lemma A.1 that compliance is maximized by setting $\sigma(\alpha_n) = \sigma_{nc}$ s.t. $\theta(0, \sigma_{nc}) = \theta^L$. □

Lemma A.11. If $\alpha_n \leq \underline{\underline{\alpha}}_n \iff \underline{\theta}(\alpha_n, \sigma) \geq \theta(0, \sigma)$ compliance decreases in α_n and is maximized at $\alpha_n = \underline{\alpha}_n$. Further, $\sigma(\alpha_n) = \arg\max V_w(\sigma; \alpha_n \leq \underline{\underline{\alpha}}_n, \beta)$ is unique and weakly decreasing in α_n with $\sigma(\alpha_n = \underline{\underline{\alpha}}_n) = \sigma_{nc}$ and $\sigma(\alpha_n = \underline{\alpha}_n) = \sigma_c$.

Proof. Claim 1: Given some $\theta(\emptyset, \sigma)$ and $\alpha_n \leq \underline{\underline{\alpha}}_n$ then compliance weakly decreases in α_n. Suppose that $\gamma = 0$. Observe the following (either algebraically or from Bayes-plausibility)

$$E[\{\underline{\theta}(\alpha_n,\sigma), \overline{\theta}(\alpha_n,\sigma)\}|s_{\mathcal{S}} = 1] = \underbrace{[\theta(\emptyset,\sigma)\beta + 1 - \theta(\emptyset,\sigma)]}_{Pr(\underline{\theta}(\alpha_n,\sigma))} \underline{\theta}(\alpha_n,\sigma) + \underbrace{\theta(\emptyset,\sigma)(1-\beta)}_{Pr(\overline{\theta}(\alpha_n,\sigma))} \overline{\theta}(\alpha_n,\sigma) = \theta(\emptyset,\sigma)$$

Recall that $\hat{\theta} < \underline{\theta}(\alpha_n, \sigma) < \overline{\theta}(\alpha_n, \sigma)$. For any $\alpha_n \in (\underline{\alpha}_n, \underline{\underline{\alpha}}_n)$, conditional on $s_{\mathcal{S}} = 1$, the leader's payoff is given by

$$\begin{aligned} V_w(\theta(\emptyset,\sigma);\alpha_n,\beta) =& [\theta(\emptyset,\sigma) + 1 - \theta(\emptyset,\sigma)]F(\underline{\theta}(\alpha_n,\sigma)) + \theta(\emptyset,\sigma)(1-\beta)[F(\overline{\theta}(\alpha_n,\sigma)) - F(\underline{\theta}(\alpha_n,\sigma))] \\ =& [\theta(\emptyset,\sigma)\beta + 1 - \theta(\emptyset,\sigma)]F(\underline{\theta}(\alpha_n,\sigma)) + \theta(\emptyset,\sigma)(1-\beta)F(\overline{\theta}(\alpha_n,\sigma)) < F(\theta(\emptyset,\sigma)) \end{aligned}$$

Where the inequality follows directly from the facts that (i) $\hat{\theta} < \underline{\theta}(\alpha_n, \sigma) < \overline{\theta}(\alpha_n, \sigma)$ and f is unimodal. Further, since $\frac{\partial \underline{\theta}(\alpha_n,\sigma)}{\partial \alpha_n} < 0$, $\frac{\partial \overline{\theta}(\alpha_n,\sigma)}{\partial \alpha_n} > 0$ and since $lim_{\alpha_n \to \underline{\alpha}_n} \underline{\theta}(\alpha_n, \sigma) = \theta(\emptyset, \sigma)$, $lim_{\alpha_n \to \underline{\alpha}_n} \overline{\theta}(\alpha_n, \sigma) = \theta(\emptyset, \sigma)$ thus $lim_{\alpha_n \to \underline{\alpha}_n} V_w(\theta(\emptyset, \sigma); \alpha_n, \beta) = V_w(\theta(\emptyset, \sigma); \underline{\alpha}_n, \beta)$. Thus, as α_n decreases, $V_w(\theta(\emptyset, \sigma); \alpha_n, \beta)$ increases and converges to the full censorship payoff.[29]

Suppose now that $\gamma \in (0, \underline{\gamma})$: recall that $\underline{\theta}(\alpha_n, \sigma) = \frac{\beta\theta(\emptyset,\sigma) - \alpha_n}{1-(1-\beta)\theta(\emptyset,\sigma)+\gamma}$ and $\overline{\theta}(\alpha_n, \sigma) = \frac{(1-\beta)\theta(\emptyset,\sigma)+\alpha_n}{(1-\beta)\theta(\emptyset,\sigma)-\gamma}$.

[29] Put differently: We can construct a linear function $l_w(\cdot)$ such that $l'_w(\cdot) = \dfrac{F(\overline{\theta}(\alpha_n,\sigma)(\sigma,\alpha_n)) - F(\underline{\theta}(\alpha_n,\sigma)(\sigma,\alpha_n))}{\overline{\theta}(\alpha_n,\sigma)(\sigma,\alpha_n) - \underline{\theta}(\alpha_n,\sigma)(\sigma,\alpha_n)}$, $l_w(\underline{\theta}(\alpha_n,\sigma)) = F(\underline{\theta}(\alpha_n,\sigma))$, $l_w(\overline{\theta}(\alpha_n,\sigma)) = F(\overline{\theta}(\alpha_n,\sigma))$, and $l_w(\theta(\emptyset,\sigma)) = V_w(\theta(\emptyset,\sigma);\alpha_n,\beta)$ because the expectation of a linear function is a linear function of expectation. Notice that holding $\theta(\emptyset,\sigma)$ fixed, as α_n increases, $l_w(\theta(\emptyset,\sigma))$ decreases as its distance to $F(\theta(\emptyset,\sigma))$ increases. In other words, $V_w(\sigma)$ is decreasing in α_n.

Recall that $\underline{\theta}(\alpha_n, \sigma)$ is the unique solution to

$$\theta(\emptyset, \sigma) = \underbrace{(\beta\theta(\emptyset, \sigma) + 1 - \theta(\emptyset, \sigma))}_{Pr(s_{\mathcal{F}}=0|s_{\mathcal{S}}=1)} \underline{\theta}(\alpha_n, \sigma) + \underbrace{(1-\beta)\theta(\emptyset, \sigma)}_{Pr(s_{\mathcal{F}}=1|s_{\mathcal{S}}=1)} \overline{\theta}(\alpha_n, \sigma)$$
$$+ (1-\beta)\theta(\emptyset, \sigma)(1 - \overline{\theta}(\alpha_n, \sigma)) + \gamma\underline{\theta}(\alpha_n, \sigma) + \alpha_n$$

Observe that $1 - \overline{\theta}(\alpha_n, \sigma) = \frac{-\gamma - \alpha_n}{(1-\beta)\theta(\emptyset,\sigma) - \gamma}$ and thus

$$\begin{aligned}(1-\beta)\theta(\emptyset, \sigma)(1 - \overline{\theta}(\alpha_n, \sigma)) + \gamma\underline{\theta}(\alpha_n, \sigma) + \alpha_n = {} & \gamma\underline{\theta}(\alpha_n, \sigma) + \alpha_n + \gamma(1 - \overline{\theta}(\alpha_n, \sigma)) \\ & + [(1-\beta)\theta(\emptyset, \sigma) - \gamma](1 - \overline{\theta}(\alpha_n, \sigma)) \\ = {} & \gamma(1 + \underline{\theta}(\alpha_n, \sigma) - \overline{\theta}(\alpha_n, \sigma)) + \alpha_n - \gamma - \alpha_n \\ = {} & \gamma(\underline{\theta}(\alpha_n, \sigma) - \overline{\theta}(\alpha_n, \sigma)) < 0\end{aligned}$$

Then for any $\gamma \in (0, \underline{\gamma})$,

$$\begin{aligned}E[\underline{\theta}(\alpha_n, \sigma), \overline{\theta}(\alpha_n, \sigma)|s_{\mathcal{S}} = 1] = {} & (\beta\theta(\emptyset, \sigma) + 1 - \theta(\emptyset, \sigma))\underline{\theta}(\alpha_n, \sigma) + [(1-\beta)\theta(\emptyset, \sigma)]\overline{\theta}(\alpha_n, \sigma) \\ = {} & (\beta\theta(\emptyset, \sigma) + 1 - \theta(\emptyset, \sigma) - \textcolor{red}{\gamma})\underline{\theta}(\alpha_n, \sigma) + [(1-\beta)\theta(\emptyset, \sigma) + \textcolor{red}{\gamma}]\overline{\theta}(\alpha_n, \sigma) \\ & + \gamma(\underline{\theta}(\alpha_n, \sigma) - \overline{\theta}(\alpha_n, \sigma)) = \theta(\emptyset, \sigma)\end{aligned}$$

In turn the same proof as in the case of $\gamma = 0$ applies. Lastly notice that for any $\alpha_n \geq \underline{\underline{\alpha}}_n$ the leader's payoff is constant in α_n. Thus the leader's payoff is continuously (weakly) decreasing in α_n.

Claim 2: $\sigma(\alpha_n) = \arg\max V_w(\sigma; \alpha_n \leq \underline{\underline{\alpha}}_n, \beta)$ *is weakly decreasing in* α_n*.* Observe first that given $\alpha_n \geq \underline{\underline{\alpha}}_n$ then $\sigma(\alpha_n) = \sigma^*(c = 0) = \sigma^L$ and similarly, given $\alpha_n \leq \underline{\alpha}$ then $\sigma(\alpha_n) = \sigma^L$.
Consider $\sigma' > \sigma$ and $\alpha_n < \alpha_n^{'}$ s.t. $(\alpha_n, \alpha_n^{'}) \in [\underline{\alpha}_n, 0]$. We want to show that $V_w(\sigma', \alpha'_n; \beta) - V_w(\sigma, \alpha'_n; \beta) \geq 0$ implies $V_w(\sigma', \alpha_n; \beta) - V_w(\sigma, \alpha_n; \beta) \geq 0$, so that σ^* is weakly increasing in $-\alpha_n$, or it is weakly decreasing in α_n by the single-crossing property (Milgrom and Shannon, 1994; Ashworth and Bueno de Mesquita, 2006).

Notice that $V_w(\sigma', \alpha_n'; \beta) - V_w(\sigma, \alpha_n'; \beta) \geq 0$ implies $V_w(\sigma', \alpha_n; \beta) - V_w(\sigma, \alpha_n; \beta) \geq 0$ if

$$
\begin{aligned}
&V_w(\sigma', \alpha_n'; \beta) - V_w(\sigma, \alpha_n'; \beta) \leq V_w(\sigma', \alpha_n; \beta) - V_w(\sigma, \alpha_n; \beta) \\
&\iff V_w(\sigma, \alpha_n; \beta) - V_w(\sigma, \alpha_n'; \beta) \leq V_w(\sigma', \alpha_n; \beta) - V_w(\sigma', \alpha_n'; \beta) \\
&\iff [p\beta + (1-p)\sigma][F(\underline{\theta}(\sigma, \alpha_n)) - F(\underline{\theta}(\sigma, \alpha_n'))] + p(1-\beta)\Big[F(\overline{\theta}(\alpha_n, \sigma)(\sigma, \alpha_n)) - F(\overline{\theta}(\alpha_n, \sigma)(\sigma, \alpha_n'))\Big] \\
&\leq [p\beta + (1-p)\sigma'][F(\underline{\theta}(\sigma', \alpha_n)) - F(\underline{\theta}(\sigma', \alpha_n'))] + p(1-\beta)\Big[F(\overline{\theta}(\alpha_n, \sigma)(\sigma', \alpha_n)) - F(\overline{\theta}(\alpha_n, \sigma)(\sigma', \alpha_n'))\Big].
\end{aligned}
$$

Note that $F(\underline{\theta}(\sigma', \alpha_n)) \leq F(\underline{\theta}(\sigma, \alpha_n))$ for any $\alpha_n < 0$ and

$$
\begin{aligned}
\frac{\partial \underline{\theta}(\sigma, \alpha)}{\partial \theta_c} &= \frac{\beta(1+\gamma) - (1-\beta)\alpha_n}{[1-(1-\beta)\theta_c]^2} \geq 0 \\
\frac{\partial^2 \underline{\theta}(\sigma, \alpha)}{\partial \theta_c \partial \alpha_n} &= \frac{-(1-\beta)\alpha_n}{[1-(1-\beta)\theta_c]^2} \leq 0 \iff \frac{\partial^2 \underline{\theta}(\sigma, \alpha)}{\partial \sigma \partial \alpha_n} \geq 0
\end{aligned}
$$

Thus, if it was the case that f is uniform such $F(\theta) = \theta \forall \theta \in [0, 1]$, the proof would be complete. Since F is concave at any $\underline{\theta}(\sigma, \alpha)$ and since $\frac{\partial^2 \underline{\theta}(\sigma, \alpha)}{\partial \sigma \partial \alpha_n} \geq 0$, then $F(\underline{\theta}(\sigma', \cdot))$ is on a "stiffer" side of F than $F(\underline{\theta}(\sigma, \cdot))$: that is, the fact that the proof is complete for a uniform F implies that the same result necessarily holds for a concave F (at these points): i.e. $F(\underline{\theta}(\sigma', \alpha_n^{'})) - F(\underline{\theta}(\sigma, \alpha_n^{'})) > F(\underline{\theta}(\sigma', \alpha_n)) - F(\underline{\theta}(\sigma, \alpha_n))$.

Note that $F(\overline{\theta}(\alpha_n, \sigma)(\sigma', \alpha_n)) < F(\overline{\theta}(\sigma, \alpha_n))$ for any $\alpha_n < 0$ and

$$
\begin{aligned}
\frac{\partial \overline{\theta}(\sigma, \alpha)}{\partial \theta_c} &= \frac{(1-\beta)[-\alpha_n - \gamma]}{[(1-\beta)\theta(\emptyset, \sigma)]^2} \geq 0 \\
\frac{\partial^2 \overline{\theta}(\sigma, \alpha)}{\partial \theta(\emptyset, \sigma) \partial \alpha_n} &= \frac{-(1-\beta)}{[(1-\beta)\theta(\emptyset, \sigma)]^2} < 0
\end{aligned}
$$

where the first inequality follows from the fact that $-\alpha_n \geq \gamma$ otherwise $\overline{\theta} > 1$.
The second inequality implies that $\frac{\partial^2 \overline{\theta}(\sigma, \alpha)}{\partial \sigma \partial \alpha_n} > 0$. In turn, the same proof as for $\underline{\theta}$ applies and thus $F(\overline{\theta}(\sigma', \alpha_n^{'})) - F(\overline{\theta}(\sigma, \alpha_n^{'})) > F(\underline{\theta}(\sigma', \alpha_n)) - F(\overline{\theta}(\sigma, \alpha_n))$. □

Lemma A.12. *For any $c \leq \overline{C} \in [z, \overline{c}(\theta^L)] \iff \alpha_n \in [\underline{\alpha}_n, 0]$ compliance increases in c (decreases in α_n): $c^* = min\{\overline{C}, \overline{c}(\theta^L)\}$.*

Proof. Step 1: compliance increases in c with a non-binding $\overline{C}$. From Claim 5 of Lemma A.7, we know that for a given $\theta(\emptyset, \sigma)$, a leader would pick the highest c, so his partial equilibrium payoff is

$p\frac{F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$ given $\theta(\emptyset,\sigma)$ and the leader can only chooses c. Also, by Lemma A.11, $\sigma(\alpha_n)$ is weakly increasing in c. We know the following two facts. First, $F(\theta^L)/\theta$ is decreasing in θ for $\theta \geq \theta^L$. Second, we know that $\theta(\emptyset,\sigma) \geq \theta^L$. Therefore, the leader is better off picking the highest c.

Step 2: compliance increases in c with a binding $\overline{C}$. By the law of iterated expectation, we know that the leader's payoff is given by $[p\beta + (1-p)\sigma] \underbrace{\theta^L}_{=\underline{\theta}(\alpha_n,\sigma)} + p(1-\beta)\overline{\theta}(\alpha_n,\sigma) + (1-p)(1-\sigma)\cdot 0 = p$. Furthermore, $[p\beta + (1-p)\sigma]\theta^L + p(1-\beta)\overline{\theta}(\alpha_n,\sigma) = \theta(\emptyset,\sigma)$; the leader's equilibrium payoff can be rewritten as

$$[p + (1-p)\sigma]l_F(\theta(\emptyset,\sigma)) = \frac{l_F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}p$$

where $l_F(\theta(\emptyset,\sigma)) = \beta F(\theta^L) + (1-\beta)F(\overline{\theta}(\alpha_n,\sigma)) = EU_l[\cdot|s_{\mathcal{S}} = 1]$ which is evaluated at $\theta(\emptyset,\sigma)$ since $E[\underline{\theta},\overline{\theta}(\alpha_n,\sigma)|s_{\mathcal{S}} = 1] = \theta(\emptyset,\sigma)$.

We know that $l_F(\theta(\emptyset,\sigma)) < F(\theta(\emptyset,\sigma))$ (by concavity of F on $\theta \in [\theta^L, \overline{\theta}(\alpha_n,\sigma)]$).

Also, $\frac{l_F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$ is decreasing in $\theta(\emptyset,\sigma)$. This follows from two facts. First, $\overline{\theta}(\alpha_n,\sigma) \geq \theta^L$. Second, $F(\cdot)$ is concave on that range. Therefore, $\frac{l_F(\theta(\emptyset,\sigma))}{\theta(\emptyset,\sigma)}$, which is the value of a line segment evaluated at p, between the point $(0,0)$ and a point in another line segment between the points $(\theta^L, F(\theta^L))$ and $(\overline{\theta}(\alpha_n,\sigma), F(\overline{\theta}(\alpha_n,\sigma)))$. By the concavity of F, this value decreases as $\overline{\theta}(\alpha_n,\sigma)$ increases and is maximized at $\theta(\emptyset,\sigma) = \theta^L$, $l_F(\theta(\emptyset,\sigma)) = F(\theta(\emptyset,\sigma))$.[30] Finally, from Lemma **??**, we know that $\theta(\emptyset,\sigma)$ (and thus $dxdx\overline{\theta}(\alpha_n,\sigma)))$ is decreasing in c; therefore, the leader's equilibrium payoff is increasing in c. □

So far we have shown the existence of a unique $\underline{\underline{\alpha}}_n$ *given* some $\theta(\emptyset,\sigma)$. We now show that in equilibrium there exists a unique $\underline{\underline{\alpha}}_n(\sigma_{nc})$ which is the unique solution to $\theta(\emptyset,\sigma_{nc}) = \underline{\theta}(\alpha_n,\sigma_{nc})$.

Lemma A.13. There exists a unique $\underline{\underline{\alpha}}_n(\sigma_{nc}) \in (\underline{\alpha}_n, 0)$ which is the unique solution to $\theta(\emptyset,\sigma_{nc}) = \underline{\theta}(\alpha_n,\sigma_{nc})$. In the unique equilibrium,

- if $\alpha_n > \underline{\underline{\alpha}}_n(\sigma_{nc})$ then $c^* = 0, \sigma^* = \sigma_{nc}$.
- if $\alpha_n \leq \underline{\underline{\alpha}}_n(\sigma_{nc})$ then $c^* = \overline{C}, \sigma^* \in [\sigma_{nc}, \sigma^L]$ and σ^* increases in $\overline{C}$.

Proof. By Lemma A.7 through A.12 we have already characterized equilibrium play for any α_n, given some $\theta(\emptyset,\sigma)$ and $\underline{\underline{\alpha}}_n$. It thus suffices to show that $\underline{\underline{\alpha}}_n(\sigma_{nc})$ exists and is unique.

[30] To provide some intuition, notice that the slope of $l_F(\theta(\emptyset,\sigma))$ decreases as $\overline{\theta}(\alpha_n,\sigma)$ increases, since $\underline{\theta} = \theta^L$ for any $c < \overline{c}(\theta)^*$, by the concavity of F. Recall too that the leader's ex-ante payoff is evaluated at the prior p.

To show existence notice that, $\underline{\underline{\alpha}}_n(\sigma_{nc})$ is the unique solution to $\theta(\emptyset, \sigma_{nc}) = \underline{\theta}(\alpha_n, \sigma_{nc})$ (same intermediate value theorem proof as in Lemma **??**).

To show uniqueness, suppose that there exists some $\alpha' \neq \underline{\underline{\alpha}}_n(\sigma_{nc})$. For any $\alpha_n > \alpha'$ we know that $c^* = 0$ and $\sigma^* = \sigma_{nc}$. For any $\alpha_n \leq \alpha'$ we know that $c^* = \overline{C}$ and $\sigma^* \geq \sigma_{nc}$. At $\alpha_n = \alpha'$ the same compliance level can be reached with maximal censorship ($\sigma^* = \sigma_{nc}, c^* = \overline{C}$) and without censorship ($\sigma^* = \sigma_{nc}, c^* = 0$) since $\underline{\theta}(\alpha', \sigma_{nc}) = \theta(0, \sigma_{nc})$; yet the only solution to $\underline{\theta}(\alpha_n, \sigma_{nc}) = \theta(0, \sigma_{nc})$ is $\underline{\underline{\alpha}}_n(\sigma_{nc})$; a contradiction.

Finally, to go back to the notation from Proposition A.1, notice that $\underline{C}(\sigma_{nc}) = z - \underline{\underline{\alpha}}_n(\sigma_{nc})$. □

Extension: Robustness of Segmentation Incentives

Suppose that the consumption of $s_{\mathcal{F}}$ can reveal θ_i in the following way. With some probability $\rho \in (0, 1)$ the regime can observe which individuals consumed $s_{\mathcal{F}}$ through their use of some software (e.g., a vpn). The regime cannot observe θ_i at the individual level, but knows, in equilibrium, which segment of the population is consuming $s_{\mathcal{F}}$. The leader's payoff can then be written as

$$V_l = V(\sigma, c) + \rho \int_{\theta(l)}^{1} l_i dF(\theta_i) \tag{10}$$

where $l_i = 1$ iff a citizen consumes $s_{\mathcal{F}}$ and $l_i = 0$ otherwise. $\theta(l)$ refers to the minimal level of ex-ante misalignment with the regime of a citizen such that the regime would like to know whom that citizen is (e.g., to engage in preventive repression). For simplicity, we assume that $\theta(l) \geq \theta^L$ s.t. the regime does not have any obvious incentives to generate segmentation *just* for learning incentives.
Ceteris paribus, a citizen does not want the regime to learn her type. Suppose that, if the leader can observe who consumed $s_{\mathcal{F}}$ then the citizen incurs a cost $c_l > 0$ (e.g., if targeted by preventive repression). A citizen's net benefit from consming $s_{\mathcal{F}}$ is then given by

$$\delta_i(\theta_i, \cdot) = b_i + \alpha(\theta_i) - c - c_l * \iota \tag{11}$$

ι can be interpreted as the surveillance capacity of the regime. Denote $\kappa = c + c_l * \iota$ and denote the total expected equilibrium cost to a citizen who bypasses the firewall by κ^*.

Lemma A.14. *Suppose that $c_l * \iota < min\{\tilde{c}(\theta^S), \tilde{c}(\theta^I)\}$. Then Lemma A.1 to Lemma A.12 still apply* *with $c^* = \kappa^* - c_l * \iota$. Further, $V_l^* > V(\sigma^*, c^*)$. The cost of access is strictly lower than in the baseline game.*

Proof. Given some $(\theta(\emptyset, \sigma), c, c_l)$, $c_i = 1 \iff b_i + \alpha(\theta_i) > C$. Observe that the introduction of learning does not affect b_i nor $\alpha(\theta_i)$. All the previous proofs still apply (simply replace c^* by κ^*) with one difference: it must be the case that $c^* \geq 0$. Thus if $c_l * \iota$ is too large, then the equilibrium cost of access could be too large which could reduce compliance w.r.t. the baseline game. Presumably the regime could then pick a lower c_l in order to avoid such a problem. In any case, $c_l * \iota < min\{\tilde{c}(\theta^S), \tilde{c}(\theta^I)\}$ rules this out. □

Extension: Propaganda *Creates* Sorting

In this extension, we entertain the possibility that the regime may *directly* make the consumption of foreign content political by making the state-media parrot the party line. That is, γ is no longer a primitive but rather an increasing function of σ. Formally, the relative entertainment benefit is now given by

$$\alpha_i(\theta_i, \sigma) = z + \theta_i \underbrace{(\eta + \rho\sigma)}_{\equiv \gamma(\sigma)}$$

For simplicity we assume that $\eta = 0$ and focus on the effect of making the state media less informative, that is, ρ. We assume that $\rho > 0$.

Lemma A.15. *There exists a unique $\overline{\rho} \in (0, 1)$ and $\underline{\rho} \in (0, \overline{\rho})$ s.t. in the unique equilibrium*

- *if $\rho > \overline{\rho}$ then $\sigma^* = \sigma^S$ and $c^* = \tilde{c}(\theta^S)$ and $\gamma(\sigma^*) > \overline{\gamma}$*
- *if $\rho \in [\underline{\rho}, \overline{\rho}]$ then $\sigma^* = \sigma^C \in [\sigma^S, max\{\sigma^I\}]$ and $c^* = \tilde{c}(\theta^C)$ and $\gamma(\sigma^*) \in [\underline{\gamma}, \overline{\gamma}]$*
- *if $\rho < \underline{\rho}$ then $\sigma^* = \sigma^L$ and $c^* = \tilde{c}(\theta^L)$ and $\gamma(\sigma^*) < \underline{\gamma}$*

Proof. Recall that $\theta^S = \frac{p}{p+(1-p)\sigma^S}$. Then suppose that

$$\rho\sigma^S \geq \theta^S(1 - \beta) \iff \rho \geq \frac{p}{p + (1-p)\sigma^S} \frac{(1-\beta)}{\sigma^S} \equiv \overline{\rho}$$

Then the regime picks $\sigma^* = \sigma^S$ and $c^* = \tilde{c}(\theta^S)$ and achieve its maximal payoff.

Next, suppose that $\rho < \overline{\rho}$. Recall that $max\{\sigma^I\}$ is such that $V(max\{\sigma^I\}, \tilde{c}(\theta^I)) = V(\sigma^L, \overline{c}(\theta^L))$.

Thus the regime may be able to pick $\sigma \in (\sigma^S, max\{\sigma^I\}]$ such that $\rho\sigma \geq \theta(1-\beta)$. Denote by σ^C the unique solution to $\rho\sigma^C = \theta^C(1-\beta)$. Then, if $\sigma^C \in (\sigma^S, max\{\sigma^I\}]$ the regime picks $\sigma^* = \sigma^C$. Denote the target citizen associated with σ^C by θ^C. In turn, segment-and-rule takes place iff

$$\rho \geq \frac{\theta^C}{\sigma^C}(1-\beta) \iff \rho \geq \frac{p}{p+(1-p)\sigma^C}\frac{(1-\beta)}{\sigma^S} \equiv \underline{\rho}$$

If $\rho < \underline{\rho}$ then $\sigma^* = \sigma^L$ and $c^* = \overline{c}(\theta^L)$.

□

Extension: Domestic Segmentation

Suppose that the leader can control the reporting slant of two domestic outlets and suppose that there is no foreign outlet. Each citizen must consume one outlet and gains access to its signal. An outlet $i \in \{1,2\}$ reporting strategy is given by a pair (π_i, σ_i) with $\pi_i = Pr(s_i = 1|\omega = 1)$.

Lemma A.16.
- *If the regime can pick any experiment for both outlets ($(\pi_i, \sigma_i) \in [0,1]^2$ $\forall i \in \{1,2\}$) then the regime can replicate his segment-and-rule maximal payoff through domestic segmentation with $\sigma_1^* = \sigma^S, \pi_1^* = 1, \sigma_2^* = 0, \pi_2^* = 1-(\sigma^S)^2$.*
- *If the regime is constrained to "regime-credible" experiments ($\pi_i = 1, \sigma_i \in [0,1]$ $\forall i\{1,2\}$) then the regime can do no better than his minimal payoff of full censorship and no domestic segmentation is observed.*

Proof. We assume that when indifferent between two outlets a citizen consumes the most informative outlet.

Case 1: suppose that $\pi_1 = \pi_2 = 1$ and $0 \leq \sigma_2 < \sigma_1 \leq 1$. There exists a unique $\theta_1 = \mu(s_1 = 1) < \theta_2 = \mu(s_2 = 1)$.

- $\theta_i < \theta_2$: payoff from consuming the second outlet is $p + (1-p)(1-\sigma_2)\theta_i$ and from consuming the first outlet it is $p + (1-p)(1-\sigma_1)\theta_i$ (for $\theta_i < \theta_1$) or θ_i (for $\theta_i \in [\theta_1, \theta_2]$); i.e. all consume the most informative outlet.

- $\theta_i > \theta_2$: either way the citizen *never* complies. Thus she is indifferent between either outlet and consumes the most informative outlet.

Hence, all citizens (weakly) go for the most informative outlet and it is without loss to restrict attention to a single domestic outlet.

Case 2: suppose that $\pi_1 = 1, \pi_2 \in [0, 1)$ *and* $\sigma_1 \in (0, 1), \sigma_2 = 0$. There exists a unique $\theta_1 = \mu(s_1 = 1) > \theta_2 = \mu(s_2 = 0)$.

- $\theta_i < \theta_2$ get p from consuming the second outlet and $p + (1 - p)(1 - \sigma_1)\theta_i$ from consuming the first outlet; i.e. all consume the first outlet. Note that this interval is empty if $\pi_2 = 1$.
- $\theta_i \in [\theta_2, \theta_1]$: there exists a unique $\theta^* = \frac{p(1-\pi)}{p(1-\pi)+(1-p)\sigma} \in (\theta_2, \theta_1)$ such that a citizen consumes the first outlet iff $\theta_i \leq \theta^*$.
- $\theta_i > \theta_1$ get θ_i from the first outlet and $p\pi + [1 - p\pi]\theta_i > \theta_i$; i.e. all consume the second outlet.

Then notice that if the regime can pick any $\sigma_1 \in (0, 1)$ and $\pi_2 \in (0, 1)$ they can set $\sigma_1^* = \sigma^S$ and $\pi_2^* = 1 - (\sigma^S)^2$ s.t. $\theta^* = \theta^S$ so that they retrieve their payoff from segment-and-rule in the strong association case ($\gamma \geq \overline{\gamma}$) of the baseline game. Crucially however, this requires the regime to be able to credibly commit to one of its own outlets reporting negatively on the regime: $\sigma_2^* = 0, \pi_2^* < 1$. □

Extension 5: Game Without Bayesian Persuasion

Suppose now that there is no state-media. Further, the regime observes a private signal $\hat{\omega}$ about ω with $Pr(\hat{\omega} = \omega|\omega) = q \geq \frac{1}{2}$.

Results without private information: $q = \frac{1}{2}$

Lemma A.17. *Suppose that $\gamma \geq p(1 - \beta)$. In the unique equilibrium, $c^* = \tilde{c}(p)$. A citizen bypasses the firewall if and only if $\theta_i \geq p$. All citizens with $\theta_i \leq p$ comply. Citizens with $\theta_i \geq p$ comply iff $s_{\mathcal{F}} = 1$.*

Proof. Observe first that since $q = \frac{1}{2}$ the choice of c does not reveal any information about ω to the voter. Further, since $\gamma \geq p(1 - \beta)$, segment-and-rule is always feasible.

In equilibrium the regime's payoff is given by

$$V^*(\tilde{c}(p)) = F(p) + [1 - F(p)]p(1 - \beta)$$

Suppose that the regime sets $c' > \tilde{c}(p)$. Then a citizen bypasses the firewall iff $\theta_i \geq \overline{\theta}'$ with $\overline{\theta}' \in (p, 1]$. All the citizens in the $[p, \overline{\theta}']$ do not gain access and never comply. The regime's payoff is then given by $F(p) + [1 - F(\overline{\theta}')]p(1 - \beta) < V^*(\tilde{c}(p))$.

Suppose that the regime sets $c' < \tilde{c}(p)$. Then a citizen bypasses the firewall iff $\theta_i \geq \overline{\theta}'$ with $\overline{\theta}' \in [0, p)$. The regime's payoff is then given by $F(\overline{\theta}') + [1 - F(\overline{\theta}')]p(1 - \beta) < V^*(\tilde{c}(p))$ where the inequality follows from $1 > p(1 - \beta)$. □

Lemma A.18. *Suppose that $\gamma < p(1 - \beta)$ and $p \geq \hat{\theta}$. In the unique equilibrium, $c^* = \overline{c}(p)$. No citizens bypasses the firewall and a citizen complies iff $\theta_i \leq p$.*

Proof. In the baseline proof with $\gamma \leq \underline{\gamma}$ (Lemma A.7 through A.12) simply input $\sigma = 1$ and $\theta^L = p$ to show that since segment-and-rule is impossible (by $\gamma < p(1-\beta)$) then the regime can do no better than $c^* = \overline{c}(p)$. □

Lemma A.19. *Suppose that $\gamma < p(1 - \beta)$ and $p < \hat{\theta}$. Then $c^* \in (0, \overline{c}(p))$. There exists a unique $\theta_L^* \in (\hat{\theta}, 1)$ s.t. $\theta_L^* = argmax_{\overline{\theta}}\ F(\theta_\beta) + [F(\overline{\theta}) - F(\theta_\beta)]p(1 - \beta)$. There exists a pair of $\underline{\theta}^* \in [\theta_\beta, p)$ and $\overline{\theta}^* \in (p, \theta_L^*)$ which are pinned down by c^*. Further, the level of compliance is bounded as follows: $V^*(c^*) \in (V(c = 0) = F(\theta_\beta) + [1 - F(\theta_\beta)]p(1 - \beta), p)$. A citizen complies if he does not bypass the firewall and is more aligned than the prior citizen ($\theta_i < \underline{\theta}^*$) or if he does and observes $s_{\mathcal{F}} = 1$.*

Proof. Notice first that under $c = 0$ then all citizens with $\theta_i \leq \theta_\beta$ always comply. Hence the lower bound on the equilibrium level of compliance.

Note too that given that $p < \hat{\theta}$ the regime would be better off fully revealing the state to the citizen and deriving a payoff of p but is unable to do so without being able to commit to an experiment; hence the upper bound on the level of compliance.

Next, observe that given some $c < \overline{c}(p)$ then a range of citizens with $\theta_i \in [\underline{\theta}, \overline{\theta}]$ (with $\underline{\theta} \geq \theta_\beta$) bypasses the firewall.

Notice too that as in Lemma A.7, the leader's payoff is always evaluated at the prior p and thus $p < \hat{p}$ any pair of $\{\underline{\theta}, \overline{\theta}\}$ pinned down by $c < \overline{c}(p)$ must improve on the full censorship payoff; hence

$c^* \leq \overline{c}(p)$.

Notice too if the regime could ensure that $\underline{\theta} = \theta_\beta$ and pick any $\overline{\theta} > p$ then there exists a unique $\theta^*_L = argmax_{\overline{\theta}}\ F(\theta_\beta) + [F(\overline{\theta}) - F(\theta_\beta)]p(1-\beta)$ which maximizes compliance by concavifying $F(\theta_i)$ *given* that $\underline{\theta} = \theta_\beta$. Whether there exists a cost $c \in (0, \overline{c}(p))$ such that $\underline{\theta} = \theta_\beta$ *and* $\overline{\theta} = \theta^*_L$ depends on the primitives. The regime picks the cost $c^* \in (0, \overline{c}(p))$ which generates the pair $(\underline{\theta}, \overline{\theta})$ which best concavifies $F(\theta_i)$ with the constraint that $\underline{\theta} \in (\theta_\beta, p)$. Notice that

Note that multiple cost of access can generate multiple pairs of $\{\underline{\theta}, \overline{\theta}\}$ which all lead to the same payoff for the regime. □

Results with private information: $q \in (\frac{1}{2}, 1)$

The regime now is endowed with a private type, defined by the realization of his private signal $\hat{\omega} \in \{0, 1\}$, and derives a belief $\mu(\hat{\omega}) \in (0, 1)$.

Lemma A.20. *If $\gamma \geq p(1-\beta)$ then there exists a pooling equilibrium on $c^*(1) = c^*(0) = \tilde{c}(p)$ with an off-path belief of $\theta' \in [\mu(0), p]$ which sender-dominates all other pooling equilibria and survives both the intuitive and divinity criterion. Further, there exists no separating or semi-separating equilibrium.*

Proof. In the sender-optimal pooling equilibrium a regime of type $\hat{\omega}$ derives a payoff of

$$V(\tilde{c}(p)) = F(p)[1 - \mu(\hat{\omega})(1-\beta)] + \mu(\hat{\omega})(1-\beta)$$

Step 1: there does not exist any separating or semi-separating equilibrium. Suppose not and let us first consider the best such separating equilibrium for both types of regime, with $c(1) = \tilde{c}(\mu(1)), c(0) = \tilde{c}(\mu(0))$. In such an equilibrium a low-type can always profitably deviate to $c'(0) = \tilde{c}(\mu(1))$. More generally, it follows from observation of equation (12) that in any separating equilibrium a low-type has a profitable deviation. The same argument rules out any semi-separating equilibrium. Suppose that a high-type picks some c_1 w.p. 1 while a low-type picks c_1 w.p. $m \in (0, 1)$ and $c_0 \neq c_1$ w.p. $1 - m$. Following c_0 the citizens derive a belief $\mu(c_0) = \mu(0)$ and after c_1 they derive a belief of $\mu(c_1) \in (p, \mu(1))$. In turn, given some m, we focus on the most efficient such semi-separating equilibrium: that is, $c_1 = \tilde{c}(\mu(c_1), c_0 = \tilde{c}(\mu(c_0))$. In turn, notice that a low-type can never be made indifferent and thus always benefits from deviating to $m = 1$. The same argument rules out a semi-separating equilibrium where a high-type is mixing.

Step 2: there exists a pooling equilibrium on $\tilde{c}(p)$ sustained with an off-path belief of $\theta' \in (\mu(0), p)$. Prior to ruling out profitable deviations, recall that $\gamma > p(1-\beta)$ ensures that SAR is feasible in equilibrium with $c^* = \tilde{c}(p)$.

Deviation 0: $c' > \tilde{c}(p)$. Denote by $\overline{\theta}'$ the first citizen to bypass the firewall under the deviation and associated off-path belief $\theta' < p$. Suppose that under such a deviation SAR is still feasible; then all $\theta_i \in (\theta', \overline{\theta}')$ comply w.p. 0 vs w.p. 1 in equilibrium. The ex-ante level of compliance of $\theta_i \in [\overline{\theta}', 1]$ is unaffected by the deviation; i.e. not a profitable deviation.

Deviation 1: $c' = 0$. Without the assumption on the off-path belief, we know that citizens must derive some off-path belief $\theta' \in [\mu(0), \mu(1)]$. The payoff from such a deviation is given by

$$V(c' = 0) = F(\theta')[1 - \mu(\hat{\omega})(1-\beta)] + \mu(\hat{\omega})(1-\beta) \tag{12}$$

Clearly, for *both* types of regime, such a deviation is profitable iff $\theta' > p$. Consider now a pooling equilibrium where the citizens' off-path belief is $\theta' \in [\mu(0), p]$.[31] For any mixed best response (MBR) from the citizens, the sets of MBR for which each type benefits from deviating to $c = 0$ coincide (by equation 12).

Deviation 2: c' $\in (0, \tilde{c}(p))$. Given the off-path belief θ', the best deviation available to the regime is to $c' = \tilde{c}(\theta')$. However, observation of equation (12) implies that the regime is worse off for any $\theta' < p$ relative to its equilibrium payoff (irrespective of the regime's type). Crucially this applies to both type of regimes for any $q \in (1/2, 1)$.

Step 3: Any other pooling equilibrium with $c^* \neq \tilde{c}(p)$ is sender-dominated by the one characterized above. Consider a pooling equilibrium with $c > \tilde{c}(p)$. In such an equilibrium, a range of citizens in $(p, \overline{\theta})$ (with $\overline{\theta} \in (p, 1]$) do not bypass the firewall and thus never comply. Thus, both types of senders would benefit from deviating to $\tilde{c}(p)$, unless if the voter holds some particular off-path belief. Consider a pooling equilibrium with $c < \tilde{c}(p)$. Then there is a range of citizens in $(\underline{\theta}, p)$ (with $\underline{\theta} \in (0, p)$) who comply only conditional on good news from outside, but would always comply in a pooling on $\tilde{c}(p)$ equilibrium. □

[31] Denote $1 - \mu(\hat{\omega})(1-\beta) = A(\hat{\omega})$ and observe that $0 < A(1) < A(0) < 1$. Thus, while both types of incumbent would benefit from deviating to $c = 0$ if the off-path belief was $\theta' > p$, a low-type would benefit the most.

Lemma A.21. *Suppose that* $\gamma \leq \mu(0)(1-\beta)$ *and* $p > \hat{\theta}$. *Then there exists a pooling equilibrium on* $c^*(1) = c^*(0) = \overline{c}(p)$ *with an off-path belief of* $\theta' \in (\mu(0), p)$ *which survives both the intuitive and divinity criterion.*

Proof. Consider such a pooling equilibrium and recall that any off-path belief must be such that $\theta' \in [\mu(0), \mu(1)]$: citizens cannot learn more than what the regime already knows. Then, notice that $\gamma \leq \mu(0)(1-\beta)$ implies that $\gamma \leq \theta'(1-\beta)$ and thus segment-and-rule is impossible. The equilibrium payoff is $F(p)$.

Then consider an off-path belief $\theta' \in [\mu(0), p]$ to any deviation to $c' < \overline{c}(p)$. The leader's payoff following such a deviation is given by

- $F(\theta') < F(p)$ if $c' > \overline{c}(\theta')$
- $F(\underline{\theta})(1-\mu(\hat{\omega})) + F(\overline{\theta})(1-\mu(\hat{\omega})) < F(p)$ if $c' < \overline{c}(\theta')$. Notice that the inequality follows from the fact that the leader's payoff is evaluated at the prior and that under c' the regime's payoff is a convex combination of $F(\theta)$ evaluated at two beliefs that average back to the prior; since $p > \hat{\theta}$ the inequality follows.

Thus either way, given θ' there does not exists any profitable deviation to $c' < \overline{c}(p)$. Further, both types of regime do not benefit from any deviation for any mixed best response of the voter, as long as the off-path belief(s) are below the prior.
Notice too that if $c' > \overline{c}(p)$, with $\theta' \in [\mu(0), p]$, there also cannot exist any profitable deviation since the leader's maximal deviation payoff is given by $F(\theta') < F(p)$. □

We do not claim equilibrium uniqueness when $\gamma < p(1-\beta)$ and have not considered $\gamma \in (\mu(0)(1-\beta), p(1-\beta))$ as for these parameter values a change in the cost of access may affect the informational benefit from bypassing the firewall and in so doing make segment-and-rule possible.

Suppose now that the privately informed regime can send a cheap-talk message $m \in \{0, 1\}$ on top of picking the cost of access.

Lemma A.22. *If* $\gamma \geq p(1-\beta)$ *then there exists a pooling equilibrium on* $c^*(1) = c^*(0) = \tilde{c}(p)$ *and* $m(1) = m(0) = m^*$ *with an off-path belief of* $\theta' \in [\mu(0), p]$ *for any* $c' \neq \tilde{c}(p)$ *or* $m' \neq m^*$ *which sender-dominates all other pooling equilibria and survives both the intuitive and divinity criterion. Further, there exists no separating or semi-separating equilibrium.*

Proof. Step 1: a sender-optimal pooling equilibrium with $c^*(1) = c^*(0) = \tilde{c}(p)$ *and* $m(1) = m(0) = m^*$ *with an off-path belief of* $\theta' \in [\mu(0), p]$ *exists.* By Lemma A.17 we need only show that there does not exist any profitable deviation to m' for either type, *given* $c^*(1) = c^*(0) = \tilde{c}(p)$. Consider a deviation to $m' \neq m^*$; given the off-path belief θ' the regime's deviation payoff is given by equation 12 and thus no profitable deviation exists.

Step 2: no separating or semi-separating equilibrium exists. By Lemma A.17 we need only show that there separation cannot be sustained through m alone, as we already know that it cannot be sustained through the choice of c. Conjecture a separating equilibrium with $c^*(1) = c^*(0) = \tilde{c}(p)$ and $m^*(1) = 1, m^*(0) = 0$. Given such an equilibrium the citizen learn the regime's type; simply substitute $\theta' = \mu(0)$ in equation 12 to notice that a low-type regime has a profitable deviation to $m'(0) = 1$. By the same logic, there does not exist (semi-) separating equilibrium with separation on both c and m.

□

Extension 6: Relaxing the Unimodality of f

We aim to show that the incentives of the regime to use a strategy of segment-and-rule do not rely on the pdf f being unimodal.[32]

Game without BP

As in Extension 5.1, suppose first that the regime cannot commit to a reporting slant and does not have any private information about the state of the world ω. We first trivially show that when segment-and-rule is possible without engineering, it is always optimal.

Lemma A.23. *Suppose that* $\gamma \geq p(1-\beta)$. *Given some pdf* f *with full support on* $[0,1]$, *then* $V(\bar{c}(p)) < V(\tilde{c}(p))$ *and* $c^* = \tilde{c}(p)$.

Proof. Trivially, $V(\bar{c}(p)) = F(p) < F(p) + p(1-\beta)(1-F(p)) = V(\tilde{c}(p))$. □

Next, we show that even when segment-and-rule need not be feasible without engineering, there

[32] See the proof of Lemma A.3 for an explanation of why log-concavity is needed under some conditions to guarantee uniqueness of σ^S when f is unimodal.

exists pairs of prior and distribution (p, f) such that it is optimal for the regime *not* to engage in full censorship.

Lemma A.24. *Suppose that $\gamma < p(1-\beta)$. There exists pairs of (p, f) (with f having full support on $[0,1]$) such that $V(\bar{c}(p)) < V(c \in [0, \bar{c}(p)))$.*

Proof. It suffices to show that there exists a pair (p, f) s.t.

$$V(c=0) = F(\theta_\beta)(1-p(1-\beta)) + p(1-\beta) > F(p) = V(\bar{c}(p)) \tag{13}$$

Consider a symmetric around $\frac{1}{2}$ u-shaped distribution f with a nadir denoted by $\hat{\theta} = \frac{1}{2} = \theta^\dagger$ (where $\theta^\dagger$) denotes the unique solution to $F(\theta^\dagger = \theta^\dagger)$) and suppose that $p > \hat{\theta} = \theta^\dagger > \theta_\beta$. Then notice that F is convex at p and thus (13) holds. □

Game with BP

Given some pdf with full support on $[0,1]$ that is not unimodal, the optimal experiment conditional on full censorship need not involve $\pi^* = Pr(s_{\mathcal{S}} = 1|\omega = 1) = 1$ (see Heo and Zerbini (2023) for a complete formal treatment). Here, as in Heo and Zerbini (2023) we assume that the regime faces a credibility constraint: they cannot credibly commit to hiding good news: i.e. $\pi = 1$ by constraint.

Lemma A.25. *Suppose that $\gamma \geq (1-\beta)$ and that $\pi = 1$. Then, given some pdf f with full support on $[0,1]$ and reporting strategy σ, $V(\bar{c}(\theta(1,\emptyset)), \sigma^*, \pi^*) < V(\tilde{c}(\theta(1,\emptyset)), \sigma^*, \pi^*)$.*

Proof. Trivially, consider some reporting strategy σ and target citizen $\theta(1,\emptyset)$ then

$$\begin{aligned} V(\bar{c}(\theta(1,\emptyset)), \sigma, \pi = 1) =& p\frac{F(\theta(1,\emptyset))}{\theta(1,\emptyset)} < p\frac{F(\theta(1,\emptyset))}{\theta(1,\emptyset)} + [1 - F(\theta(1,\emptyset))]p(1-\beta) \\ =& V(\tilde{c}(\theta(1,\emptyset)), \sigma, \pi = 1) \end{aligned}$$

Notice that this proof does not characterize the optimal experiment given segment-and-rule. It only generalizes the result that full censorship is strictly dominated by a strategy of segment-and-rule when the latter is feasible without engineering *and* the regime faces a credibility constraint. □

Extension 7: Non-Binary Compliance

Let us now assume that $a_i \in [0,1]$. Instead of solving explicitly for a modified version of the game, we fix a reporting slant σ and target citizen θ, and aim to derive under which conditions segment-and-rule dominates full censorship, given some equilibrium compliance profile $a_i^*(\theta_i, \mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta))$. We assume such an equilibrium compliance profile exists and make the following assumptions about it:

1. $a_i^*(\theta_i, \mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta)) = \mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta)$ if $\mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta) = 0$ or 1.

   All (respectively no) citizens provides full compliance (respectively zero compliance) conditional on perfectly knowing that $\omega = 1$ (respectively $\omega = 0$).

2. $\frac{\partial a_i^*(\theta_i, \mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta))}{\partial \theta_i} \leq 0$.

   The more ex-ante misaligned with the regime a citizen, the lower her equilibrium compliance level.

3. $\frac{\partial a_i^*(\theta_i, \mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta))}{\partial \mu} \geq 0$.

   The higher a citizen's belief about ω, the higher her equilibrium compliance level.

4. $\frac{\partial^2 a_i^*(\theta_i, \mu(s_{\mathcal{S}}, \hat{s}_{\mathcal{F}}|\sigma, \beta))}{\partial \mu \partial \theta_i} \leq 0$ and 0 at $\theta_i = 1$.

   The more misaligned a citizen is with the regime, the smaller the marginal effect of a higher posterior on ω.

We focus on the behavior of regime opponents ($\theta_i > \theta$) given some reporting slant σ, and conditional on $s_{\mathcal{S}} = 1$ since otherwise no citizens provide any positive level of compliance. We also assume that segment-and-rule is feasible, e.g., $\gamma > 1 - \beta$.

Lemma A.26. *Suppose that assumptions 1. through 4 hold and that $\gamma > 1 - \beta$. Then, given some $\sigma \in [0,1]$ and $\theta \in [p, 1]$, there exists a unique $\tilde{\theta} \in [\theta, 1)$ s.t. $V(\tilde{c}(\tilde{\theta}), \sigma) > max\{V(\overline{c}(\theta), \sigma), V(\tilde{c}(\theta), \sigma)\}$.*

Proof. Then, define the gain from a strategy of segment-and-rule at the individual level by

$$\begin{aligned}\Delta_s(\theta_i) =& p(1-\beta) * 1 + [1 - p(1-\beta)]a_i^*(\theta_i, \mu(1, 0|\sigma, \beta)) - a_i^*(\theta_i, \mu(1, \emptyset|\sigma, \beta)) \\ =& p(1-\beta)(1 - a_i^*(\theta_i, \mu(1, \emptyset|\sigma, \beta))) - [1 - p(1-\beta)](a_i^*(\theta_i, \mu(1, \emptyset|\sigma, \beta)) - a_i^*(\theta_i, \mu(1, 0|\sigma, \beta))).\end{aligned}$$

Note that $\frac{\partial^2 a_i^*(\theta_i,\mu(s_\mathcal{S},\hat{s}_\mathcal{F}|\sigma,\beta))}{\partial\theta_i\partial\mu} \leq 0$ is sufficient (but not necessary) to ensure that $\frac{\partial\Delta_s(\theta_i)}{\partial\theta_i} > 0$.

Then, if $\Delta_s(\theta) > 0$, then there exists a unique $\tilde{\theta} \in [\theta, 1)$ with the above property. Suppose not. Notice that $\Delta_s(1) \geq 0$ by $\frac{\partial^2 a_i^*(\theta_i,\mu(s_\mathcal{S},\hat{s}_\mathcal{F}|\sigma,\beta))}{\partial\mu\partial\theta_i}(\theta_i = 1) = 0$.[33] Then, by the intermediate value theorem, there exists a unique $\tilde{\theta} \in [\theta, 1)$ s.t. $\Delta_s(\theta_i) \geq 0 \iff \theta_i \geq \tilde{\theta}$. Then, fixing some σ, $V(\tilde{c}(\tilde{\theta}), \sigma) > max\{V(\bar{c}(\theta), \sigma), V(\tilde{c}(\theta), \sigma)\}$. That is, as long as $\Delta_s(\theta) > 0$ – which is always the case when assumptions 1 through 4 hold – the regime can always, fixing a reporting slant σ and target citizen θ, adjusts the cost of access such that a strategy of segment-and-rule dominates a strategy of full censorship. □

Extension 8: Intrinsic Benefit and Feasibility of Segment-and-Rule

So far $\alpha(\theta_i) = z + \gamma \times \theta_i$. Hereafter we derive necessary conditions on $\alpha(\theta_i)$ for the regime to reach their upper bound payoff by segmenting and setting $\sigma^* = \sigma^S$.

Given some $\alpha(\theta_i)$ and $\sigma^* = \sigma^S$ two important cutoffs exists: $\theta(0, \sigma^S)$ and $\theta(\emptyset, \sigma^S)$, and there exists a unique equilibrium informational benefit $b_i(\theta_i, \sigma^S, s_\mathcal{S} = 1, \beta)$.

Forall $\theta_i \leq \theta(0, \sigma^S)$, the regime need not prevent access to the foreign media since exposure to $s_\mathcal{F}$ does not affect the citizens' compliance; no restriction need be applied to $\alpha(\theta_i)$ for any $\theta_i \leq \theta(0, \sigma^S)$. Next, denote

$$\begin{aligned}
\overline{\delta}^{cc} \equiv \quad & max_{\theta_i}\ \delta_i\Big(\theta_i \in [\theta(0,\sigma^S), \theta(\emptyset,\sigma^S)], \sigma^S, s_\mathcal{S} = 1, \beta\Big) \\
\underline{\delta}^{o} \equiv \quad & min_{\theta_i}\ \delta_i\Big(\theta_i \in (\theta(\emptyset,\sigma^S), 1], \sigma^S, s_\mathcal{S} = 1, \beta\Big)
\end{aligned}$$

and similarly denote $\overline{\theta}^{cc}$ as the (largest) solution to $\delta_i(\theta_i \in [\theta(0,\sigma^S), \theta(\emptyset,\sigma^S)], \sigma^S, s_\mathcal{S} = 1, \beta) = \overline{\delta}^{cc}$ and $\underline{\theta}^o$ as the (smallest) solution to $\delta_i(\theta_i \in (\theta(\emptyset,\sigma^S), 1], \sigma^S, s_\mathcal{S} = 1, \beta) = \underline{\delta}^o$.

Lemma A.27. *For any $\alpha(\theta_i)$ s.t. $\alpha(\theta_i)$ s.t. $\overline{\delta}^{cc} \leq \underline{\delta}^o$ the regime sets $\sigma^* = \sigma^S$ and $c^* \in [\tilde{c}(\overline{\theta}^{cc}), \tilde{c}(\underline{\theta}^o)]$.*

Proof. Given some $\alpha(\theta_i)$ s.t. $\overline{\delta}^{cc} \leq \underline{\delta}^o$ there exists a range of costs $c \in [\tilde{c}(\overline{\theta}^{cc}), \tilde{c}(\underline{\theta}^o)]$ that ensure that $\theta_i \in [\theta(0,\sigma^S), \theta(\emptyset,\sigma^S)]$ do not gain access while $\theta_i \in (\theta(\emptyset,\sigma^S), 1]$, *given* $\sigma^* = \sigma^S$. The regime is indifferent between any of these costs, and reaches their upper bound payoff. □

[33] Since this assumption implies that $a^*(\theta_i = 1)$ is independent of $\mu(\cdot)$.