UAB
Universitat Autònoma de Barcelona

Essays on Organizational Design, Digital Platforms, and Media Economics

Manuel Lleonart-Anguix

Universitat Autònoma de Barcelona

Essays on Organizational Design, Digital Platforms, and Media Economics

Author: Manuel Lleonart-Anguix

Advisor: Pau Milán

2026

A dissertation submitted to the Departament d'Economia i d'Història Econòmica in fulfillment of the requirements for the degree of Doctor of Philosophy in Economic Analysis.

A Clara, a Conso y a Manolo.

Acknowledgments

All of this would not have been possible without the immense help of my advisor, Pau Milán. Thank you for showing me what research is about. I hope someday to be as good a researcher, advisor, and teacher as you are. Thanks for understanding me and supporting me during the bad times. I could not think of a better person to advise me through this journey. Thank you for sharing your experience and your perspective on this profession. I feel privileged to always have your support. My sincere thanks also to Francis Bloch for hosting me in Paris and giving me the best advice when I needed it: if it does not work, just begin again. The months I spent with you truly changed my vision about academia. To Mikhail Drugov, for your invaluable support during the market and for all those five minutes that became hours. To all the professors of IDEA. Those who are and those who were: Jordi Caballé, Johannes Gierlinger, Inés Macho, Jordi Massó, Fernando Payro, Chara Papioti, Marta Troya, and many more. I am enormously grateful to Àngels, Mercé and Marta. Thank you for all those days when I entered your office without a clue what to do and left with a smile. You make it so easy to defend the public system that helped many of us reach a high level of education. To Jaime.

Never ever I would dream of doing this journey with other people than my friends at IDEA. To Jacob, I give thanks for meeting you. You are behind many of the happiest moments I had during these years. I am glad to know that Marina and Carlota would make you happy too. To Jou, for making every moment more interesting. For always having an anecdote, a curiosity or a good way of beginning a conversation. To Luis, flatmate, coauthor, colleague, office mate, and above all, a friend. For all the conversations about philosophy, sports, cinema, literature and, sometimes, even economics. To Sergi, for your ability to make me smile under any circumstances and for all your help in this process. You are one of the smartest people I have ever met. Thank you for making me question everything I know. To the good Ph.D. candidates of IDEA, in particular to my office-mates: I

had the immense pleasure of first knowing you as brilliant students, and now as friends for life. To Antonio, without you, this last year of Ph.D. would have been much worse. I hope you find the missing piece. To Nico, always an example to follow and a person to trust. Thanks for remembering me of my worth when I forgot about it. To Lisa and Sofija, I wish all the students were as hardworking and engaged as you were. It has been a pleasure. Finally, to my colleagues of the Ph.D: Alannah, Elif, Gabriela, Proud, Silvestre, Sophie, Tancredi, Tristany and the others.

Of equal importance are the people that I met in seminars and conferences. Thanks to those that decided to work with me these years. In particular, to Miguel. I remember how our first meeting completely changed the view I had about research. Your way of understanding economics and always finding a way to make your research meaningful has been one of the biggest inspirations on the kind of researcher I want to be. To Santiago and Basile, for giving me the chance to work in this new project. It is always scary to explore new avenues but the experience is a pleasure when I can walk with such smart people by my side.

Nada de esto habría sido posible sin la ayuda de Ramón Tamarit. Todavía recuerdo la primera vez que entré en tu despacho y te dije que me gustaba tu asignatura. Gracias por entender mejor que yo lo que quería decir, por recomendarme que fuese a IDEA y por tu ayuda durante estos años; antes como profesor, ahora como amigo. Gracias a Iván, Rubén y Alik por haberme acompañado durante los años de la carrera y hacer este viaje inolvidable.

Imposible habría sido llegar aquí sin el apoyo de mi familia. Gracias a mis padres por creer en mí cuando les dije que quería estudiar matemáticas, por llevarme y traerme a todos los sitios que ha hecho falta, por obligarme a estudiar inglés y por confiar en mi criterio. Siempre he intentado que os sintáis tan orgullosos de mí como yo me siento de vosotros. Gracias a Carmen. En el fondo sigo siendo ese niño pequeño que dibujaba a su hermana en grande a su lado. Gracias a Rocío, por intentar entenderme hasta cuando ni yo tengo claro qué pienso.

Òbviament, i per damunt de tot, gràcies eternes a Clara; per estar al meu costat cuidant-me i fent-me feliç cada dia de la nostra vida, per haver-me ensenyat a dir t'estime. Gràcies per les pizzes, els baos i les paelles, per les vesprades de

sofà i per voler-me quan ni jo ho feia. Mai podré agrair prou que m'hages deixat formar part de la teua vida i del teu poble i que m'hages ensenyat que la felicitat només té sentit quan és compartida. Gràcies a tu ja no tinc por.

Finally, I acknowledge financial support of the Spanish *Ministerio de Ciencia e Innovación* through grant PRE2021-099026.

Preface

This dissertation studies how the design of institutions and algorithms shapes incentives, information, and welfare. The three chapters span organizational economics, platform regulation, and media economics, but share a common concern: when intermediaries—whether supervisors, algorithms, or news outlets—stand between principals and the information they need, the design of the intermediation layer is a first-order problem.

Chapter 1 develops a framework for the optimal design of internal monitoring structures when supervisors themselves can be bribed. In a principal–supervisor–worker hierarchy, the firm jointly chooses the architecture of oversight—how many supervisors per worker and how their spans of control overlap—and the incentive pay needed to keep supervisors honest. Dual monitoring, assigning two independent supervisors to the same worker, can lower the wage required to maintain honesty by raising the cost of collusion. The resulting wage savings can outweigh the extra hiring cost, making redundancy optimal precisely when monitoring technology is effective and collusion pressures are high. The analysis reframes dual monitoring not as bureaucratic waste but as a rational enforcement tool, and provides design rules linking observable features of technology and governance to the optimal degree of overlap.

Chapter 2, joint with Miguel Risco, builds a theoretical model of communication and learning on a social media platform. We characterize the algorithm that an engagement-maximizing platform implements in equilibrium and show that it overexploits similarities between users, locking them in echo chambers. Moreover, learning vanishes as platform size grows large. We then explore alternatives. The reverse-chronological algorithm that platforms reincorporated after the Digital Services Act turns out to be insufficient, so we construct a "breaking-echo-chambers" algorithm that improves learning by promoting opposite viewpoints. Finally, we advocate for horizontal interoperability as a regulatory measure to

align platform incentives with social welfare. By eliminating platform-specific network effects, interoperability incentivizes platforms to adopt algorithms that maximize user well-being.

Chapter 3, joint with Luis Menendez, studies how competition in the news market reshapes the ideological positioning of media outlets. We distinguish three mechanisms through which outlets can bias the news: story selection (omission), narrative tone (framing), and airtime allocation (intensity). We develop a theoretical model that yields predictions about how each margin responds to increased competition, and test them using transcripts from Spanish prime-time television news, exploiting the launch of a new newscast as a source of competitive pressure. Using a Shapley decomposition, we show that framing is the dominant mechanism for every outlet, accounting for at least two-thirds of observed slant.

Contents

Chapter 1

Firm Design when Monitors Can Be Bribed

Manuel Lleonart-Anguix[1]

Abstract

This paper develops a framework to design internal monitoring structures that minimize the cost of enforcing honesty. I study internal monitoring when supervisors themselves can be bribed. The firm jointly chooses the *architecture* of oversight (how many supervisors per worker and how their spans of control overlap) and the *incentives* needed to keep supervisors honest. Dual monitoring—assigning two independent supervisors to the same workers—makes wrongdoing harder to conceal and lowers the wage required to maintain honesty. The analysis shows that redundancy is optimal when monitoring technology is effective and collusion pressures are high—precisely the environments where overlap in monitoring is most valuable, that is, when verifiability is strong and monitors' outside options are weak. By linking organizational structure to the credibility of enforcement, the paper reframes dual monitoring not as waste but as a rational enforcement tool.

[1]Universitat Autònoma de Barcelona and Barcelona School of Economics. Contact: manuel.lleonart@bse.eu. I am grateful to Pau Milán for his invaluable guidance throughout the development of this paper. Part of this research was conducted during my visit to the Paris School of Economics, under the kind invitation of Francis Bloch, to whom I am also deeply indebted. I thank Catherine Bobtcheff, Mikhail Drugov, Inés Macho-Stadler, Antonine Macé, Paul-Henri Moisson, Fernando Payro, David Pérez-Castrillo, Patrick Sewell, Marta Troya, and seminar participants at the BSE Jamboree, the Microlab at UAB, and the ENTER Jamboree at the SSE for their helpful comments and suggestions.

In low- and middle-income countries, weak governance and internal oversight failures cost firms an estimated 2–5% of annual revenue—losses comparable to major tax burdens or trade barriers (ACFE, 2024; World Bank, 2006). Across industries and countries, the presence of multiple supervisors on the same task has typically been explained as a way to reduce mistakes, segregate duties, or combine different skills. In such accounts, duplication arises when tasks are complex or multidimensional—requiring error reduction (Sah and Stiglitz, 1986), segregation of duties (Kobelsky, 2014), and knowledge complementarities (Garicano, 2000; MacDuffie, 1995). These accounts do not predict *dual monitoring* for simple, standardized tasks, where redundancy is usually treated as waste. My argument is different: when monitors can collude with workers, that very possibility can make dual monitoring optimal even in simple tasks. Firms, regulators, and public agencies face this problem daily when they rely on auditors, inspectors, or recruiters who might collude with those they oversee. Modeling incentive pay and oversight structure jointly reveals a different logic. Assigning two independent supervisors to the same worker can be optimal not because the task is complex, but because redundancy lowers the pay needed to keep each supervisor honest. This shifts how we think about assigning supervisors across workers and when overlap should be part of efficient design.

This paper develops a framework to study the optimal design of internal oversight when supervisors themselves can be bribed. I jointly endogenize the monitoring architecture—how many supervisors oversee each worker and how their spans of control (the number of workers assigned to each supervisor) overlap—and the pay needed to keep supervisors honest. Here, overlap means assigning two independent supervisors to the same worker. By incorporating collusion into the joint choice of organizational structure and incentive pay, the model shows that dual monitoring can be optimal even for simple, standardized tasks: by making concealment harder, it reduces the compensation required to keep supervisors honest, and in some environments these wage savings exceed the cost of hiring an additional supervisor. What appears wasteful in theories where redundancy is justified only by task complexity can therefore be a rational response to the possibility that monitors can be bribed. The framework links oversight design to monitoring technology and incentive pay, and develops a practical mechanism to deter collusion: *dual monitoring* with aligned incentives that make bribery

unattractive. It reframes effective oversight as disciplining monitors as well as supervising workers.[2]

The environment is a principal–supervisor–worker hierarchy in the spirit of Tirole (1986), with a continuum of workers performing identical tasks. The key differences are that detection is endogenous to organizational design, and that monitors are strategic agents who may collude both with one another and with workers. Each worker is high- or low-productivity, with type privately known to them. The principal hires one or more supervisors to oversee subsets of workers and report deviations. Detection is imperfect and declines with a supervisor's span of control (attention is scarce). Workers can bribe their monitors to suppress adverse reports. The firm chooses (i) how many monitors per worker and how spans overlap, and (ii) a wage policy that renders accepting bribes unattractive and, at the same time, avoids shirking for workers. Crucially, the frequency and profitability of bribe opportunities depend on the same choices that govern detection: structure and pay are inseparable design margins.

Two levers interact. First, architecture: assigning two independent supervisors to the same worker raises the cost of silence, since a deviator must bribe both. Second, pay: compensation makes acceptance of any bribe unattractive, accounting for how span affects detection and the flow of potential bribes. Compensation can also be tied to the other supervisor's report, creating incentives to report misconduct, although this force is limited because the two monitors may also collude with each other and with the worker. Because structure and pay co-determine both detection and bribe incentives, the monitoring unit operates as an endogenous enforcement technology rather than a fixed input. In the absence of monitoring, the principal must raise the wage for high outputs to avoid shirking.

First, the possibility of collusion overturns the one-monitor-per-worker rule: when baseline verifiability is high or detection decays gradually with span, the firm optimally assigns two independent monitors to the same worker. Second, the monitor's wage is non-monotonic in the span of control: with very small or

[2]Versions of overlapping oversight exist in practice—for example, teacher evaluations by school administrators and external reviewers (District of Columbia Public Schools, 2023), dual authorizations in banking (Basel Committee on Banking Supervision, 2011; Reserve Bank of India, 2011), two-person custody rules (Office of the Comptroller of the Currency, 2012), multiple signatories on reports (Federal Deposit Insurance Corporation, 2014), and overlapping financial regulators (Keightley and Jickling, 2017); international guidelines also endorse "four-eyes" safeguards (OECD, 2018).

very large spans, the expected number of detections—and hence the monitor's expected wage—is low, while at intermediate spans it peaks. Increasing overlap then reduces the honesty premium and can lower per-monitor pay even as the number of monitors rises, so the total wage bill need not increase—and may even fall. This creates a trade-off between paying more for loyalty and hiring more monitors to share oversight: better monitoring technology shifts the optimum toward broader spans, while a greater scope for collusion shifts it toward narrower spans and more redundancy. Third, labor-market conditions shape the case for overlap: when supervisors have weak outside options, most of their pay is devoted to deterring bribery and overlap trims that premium, making dual supervision relatively cheap; when outside options are strong, much of pay is needed simply to retain supervisors, the savings from overlap are small, and the firm favors wider spans without duplication. These patterns link observable features of technology and governance with clear policies about hierarchy width, double monitoring and wage structure and allow me to measure the loses from naïve designs.

By embedding the presence of collusion into the design problem, the model reframes dual monitoring from bureaucratic slack into a practical mechanism to deter collusion. This makes firm design a tool against collusion: use dual monitoring where verifiability is strong and align incentives to eliminate the gains from bribery. The analysis also provides a basis for evaluating anti-corruption policies that strengthen internal governance—through redundancy, span limits, and incentive design—rather than relying solely on external enforcement, and it suggests redundant layers of oversight may persist even when downstream production considerations favor flatter structures, because those layers help discipline monitors and not only workers. Finally, it provides simple design rules for when redundancy in oversight is efficient and when it can be safely reduced, and suggests empirical predictions linking monitoring structures to enforcement strength across countries and sectors.

Literature Review. A central theme in the economics of organizations is that hierarchy shapes how firms maintain control as they grow. The tradition closest to this paper views firms as *monitoring hierarchies*, where the main challenge is the *loss of control* associated with expansion, and compensation must adjust to preserve discipline (Calvo and Wellisz, 1978, 1979; Keren and Levhari, 1983;

Qian, 1994; Williamson, 1967). Later work connects these internal trade-offs to finance and market structure (Acemoglu and Newman, 2002; Chen, 2017; Chen and Suen, 2019). What these models typically abstract from is that layers can *talk*. If a monitor can collude with a worker, the principal must choose not only spans of control but also whether to use *dual monitoring* to deter collusion. In my framework, that choice matters because dual monitoring changes the compensation required to keep monitors honest.[3]

The main message is that, once communication enables collusion, overlapping supervision can be a cost-effective way to sustain discipline. Yet they largely abstract from the possibility that monitors themselves may be corruptible. Once this is recognized, monitoring no longer operates solely as a substitute for wages to discipline workers: it becomes a source of wage premia for monitors themselves, fundamentally altering the design problem at the heart of the firm. Firms must then choose spans, overlap, and incentive schemes not only to manage information frictions but to sustain honest enforcement, which is the gap my analysis addresses.

Building on this foundation, a second set of contributions studies collusion and contracting more directly. Alchian and Demsetz (1972) propose making the supervisor the residual claimant as a solution to the metering problem, but leave open the case where the monitor himself may be corruptible. In the classic principal–supervisor–agent framework, collusion (Compte, 1995; Kofman and Lawarrée, 1996; Laffont and Martimort, 2000; Tirole, 1986) and shirking (Rahman, 2012) undermines incentives and forces the principal to adapt contracts. This perspective naturally links to the theory of common agency: while Bernheim and Whinston (1986) modeled multiple principals bidding for influence, overlapping monitors create a mirror problem in which agents bid for silence. My framework formalizes this logic, making organizational architecture part of the tools that the principal can use to fight collusion. I build on a large *efficiency-wage* literature in which imperfect monitoring leads firms to pay above-market wages to deter shirking (Shapiro and Stiglitz, 1984). I add that rents are shaped not only by imperfect monitoring but also by potential collusion between monitors and workers, and I endogenize firm architecture—span and *dual monitoring*—to show

[3]Related work studies hierarchies as screening devices (Ioannides, 2012; Sah and Stiglitz, 1986) or as knowledge-based organizations (Garicano, 2000; Garicano and Rossi-Hansberg, 2004, 2006).

when structure lowers the required honesty premia.

Finally, a large literature studies corruption and enforcement institutions. Classic work emphasized corruption as a distortionary tax or capture of regulators (Laffont and Tirole, 1991; Rose-Ackerman, 1978; Shleifer and Vishny, 1993), while recent contributions examine collusive networks inside organizations (Drugov, 2010; Fan, 2022). Empirical evidence confirms that oversight design matters: stronger auditor incentives reduce misreporting in India (Duflo et al., 2013), adding a second countersignature lowered procurement prices in Pakistan (Bandiera et al., 2021), and overlapping jurisdictions mitigate capture in administrative law (Gersen, 2007). My analysis draws on these insights by treating "four-eyes" policies as a firm-level analogue to overlapping jurisdictions: costly, but optimal when the risk of bribery is high.

In sum, the existing literatures explain how hierarchies shape incentives, information, and expertise; how contracts adapt to collusion; and how enforcement institutions can be corrupted or strengthened. My contribution is to bring these strands together, modeling the internal governance of firms as a corruption-control technology and showing how redundancy and overlap can sustain honest enforcement even when monitors themselves are at risk of capture.

Road map. The rest of the paper is organized as follows. Section 1.1 introduces the theoretical framework. Section 1.2 analyzes the benchmark without collusion. Section 1.3 studies optimal monitoring under collusion. Section 1.4 analyzes the loses from naïve results. Section 1.5 concludes.

1.1 Model

This paper develops a continuous version of the classic Principal–Supervisor–Agent framework introduced by Tirole (1986), extending it to allow the principal to endogenously choose the monitoring hierarchy within the firm. While detailed descriptions of each agent type follow, it suffices for now to view the principal as the residual claimant of firm profits, the agents as the sole productive units, and the monitors as intermediaries whose role is to report agents' productivity in order to mitigate the informational rents arising from asymmetric information.

The principal lacks both the time and the ability to directly observe agent behavior. Therefore, she can hire monitors to perform this task on her behalf.

Agents There is a continuum of agents, uniformly distributed over the interval $A = [0, 1]$. All agents are ex-ante identical. Let $i \in A$ denote a generic agent. Each agent generates output x_i given by:

$$x_i = \theta_i + e_i,$$

where θ_i represents the agent's productivity and $e_i \geq 0$ denotes effort. Productivity is binary: $\theta_i \in \{\theta_H, \theta_L\}$, with $\theta_H > \theta_L \geq 0$. The productivity type is drawn from a uniform distribution: $q \sim U(0, 1)$, implying that the expected share of high- and low-productivity agents is equal. Productivity realizations are independent across agents.

I define the skill diversity as $\Delta\theta = \theta_H - \theta_L \leq \frac{1}{2}$, such that high-productivity agents can mimic low-productivity ones by exerting strictly positive effort. This assumption is crucial for generating a non-trivial contracting problem under asymmetric information.[4]

Agents observe their productivity before choosing effort. Effort entails a cost given by $c(e_i) = \frac{e_i^2}{2}$. Each agent signs a contract with the principal specifying a wage $W(x_i, r_i)$, contingent on the realized output (observable to all parties) and a report r_i issued by the monitors (defined below). Agents are risk neutral, with utility $U_i(x) = x$, so expected utility is $\mathbb{E}[W(x_i, r_i) - c(e_i)]$. The participation constraint for each agent is:

$$\mathbb{E}[W(x_i, r_i) - c(e_i)] \geq 0. \tag{PC$_W$}$$

Monitors The principal cannot directly observe either agents' effort or their productivity. To address this, she may hire monitors who are tasked with observing the productivity of a subset of agents. Each monitor j observes a subset $I_j \subseteq A$, with $\mu_j = \int_{I_j} di$ denoting monitor j's span of control. As agents are ex-ante homogeneous, the principal is indifferent over which specific subset of agents a

[4] If instead $\Delta\theta > \frac{1}{2}$, mimicking would require high-productivity agents to exert negative effort, violating the constraint $e_i \geq 0$. One could alternatively model agents as being able to destroy output at a cost, i.e., allow $e_i < 0$, but this approach is not pursued here.

monitor observes, conditional on μ_j. Thus, all results will be expressed in terms of the measure μ_j.

For each agent $i \in I_j$, monitor j receives a productivity signal σ_{ij}. With probability $p(\mu_j) = p_0 e^{-\lambda \mu_j}$, the signal is informative, $\sigma_{ij} = \theta_i$, and with probability $1 - p(\mu_j)$, the signal is uninformative: $\sigma_{ij} = \varnothing$. I refer to $p(\mu_j)$ as monitoring efficiency. It captures both the probability of obtaining verifiable information and the legal enforceability of that information. Two primitives summarize the monitoring technology. The first is p_0, the baseline verifiability of a task. It captures how easy it is to detect misbehavior if a monitor focuses exclusively on a single agent. High p_0 corresponds to environments where evidence of misconduct is naturally transparent—e.g. physical theft, clear performance metrics, or verifiable outputs. Low p_0 corresponds to tasks where wrongdoing is inherently hard to spot—e.g. effort in service provision, subtle quality shading, or collusion in procurement. Intuitively, p_0 measures how "detectable" shirking or corruption is, even before span-of-control considerations enter. The second is λ, the span–efficiency parameter. It governs how quickly monitoring effectiveness decays as the span of control widens. A high λ means monitors are quickly overloaded—each additional agent sharply reduces the probability of detection—so oversight is fragile to increases in span. A low λ means monitors scale more smoothly: adding agents dilutes attention only gradually. Intuitively, λ captures the elasticity of monitoring with respect to workload, or how well monitors can multitask without missing misconduct. Together, p_0 and λ define the "quality" of the monitoring technology. p_0 tells us how sharp the monitor's eyes are when focused; λ tells us how blurry those eyes become as the monitor's span grows.

After observing the signal, monitor j sends a report $r_{ij} \in \{\theta_i, \varnothing\}$ to the principal. Since multiple monitors may report on the same agent, I denote the full report vector as $r_i = (r_{ij})_{j \in M}$.

Information is assumed to be hard, i.e., verifiable by the principal at no cost. Thus, a monitor cannot misreport the signal, but can choose to withhold information by reporting $\varnothing$ regardless of the true signal.

Monitors are compensated via contracts specifying payments of the form $\int_{I_j} S(x_i, r_i) di$. They are risk neutral with utility $V(x) = x$, so expected utility is $\mathbb{E}[S(x_i, r_i)]$. Their participation constraint is:

$$\mathbb{E}[S(x_i, r_i)] \geq \bar{V}. \tag{PC$_M$}$$

To avoid degenerate results stemming from the law of large numbers in the continuum framework, I restrict $S(x_i, r_i)$ to be a function of individual-level outcomes rather than aggregate statistics.

Finally, both agents and monitors are subject to limited liability: their wages must be weakly non-negative in all states:

$$S(x_i, r_i) \geq 0. \tag{LL$_M$}$$

Principal The principal is the residual claimant of profits. She designs the wage schemes for both agents $W(x_i, r_i)$ and monitors $S(x_i, r_i)$, and chooses the organizational structure $\Omega = (\mu_j, \gamma)$, where γ denotes the measure of overlap between monitors' spans of control i.e., the extent of *common agency*. She is risk neutral and seeks to maximize expected profits:

$$\mathbb{E}[\Pi(W, S)] = \mathbb{E}\left[\int_0^1 (\theta_i + e_i - W(x_i, r_i) - S(x_i, r_i))\, di\right],$$

where expectations are taken over the distributions of productivity θ_i and monitoring signals σ_{ij}.

The organizational design is defined as $F = (M, \Omega)$, where M is the number of hired monitors, and $\Omega = (\mu_j, \gamma)$ specifies their individual spans of control and overlap. Agents may fall under three monitoring regimes: *unmonitored*, *singly monitored*, or *jointly monitored*.

Given the stochastic nature of both productivity and monitoring, each agent-monitor pair can be in one of four possible states: $(\theta_i, \sigma_i) \in \{(H, H), (H, \varnothing), (L, \varnothing), (L, L)\}$ that correspond to four observable outputs, $X = \{x_{\theta_L,R}, x_{\theta_L,\varnothing}, x_{\theta_H,\varnothing}, x_{\theta_H,R}\}$. For any $x_i \notin X$, the agent chooses the effort that maximizes her expected utility. Under limited liability, the principal can set $W(x_i, r_i) = 0$ for $x_i \notin X$, inducing $e_i = 0$. To ensure incentive compatibility, it suffices to guarantee that:

$$W(x_i, r_i) - \frac{e_i^2}{2} \geq 0, \quad \text{for all } x_i \in X,$$

and $W(x_i, r_i) = 0$ otherwise. These constraints imply that the participation

constraint is automatically satisfied in all realizations, and can therefore be omitted in subsequent analysis.

Timing The timing of the model is as follows:

1. The principal chooses the organizational design $\Omega = (\mu_j, \gamma)$ and offers contracts $W(x_i, r_i)$, $S(x_i, r_i)$ to agents and monitors.
2. Upon acceptance, agents learn their productivity type θ_i, and monitors receive their signals σ_{ij}.
3. Monitors submit reports r_{ij}.
4. Reports arrive before effort so that verifiable audits shape incentives; agents observe the report realization when choosing e_i.
5. Output x_i is realized; payments $W(x_i, r_i)$ and $\int_{I_j} S_j(x_i, r_i)di$ are made; the principal collects the residual profits.

First-Best Benchmark Before analyzing the role of monitors and the constraints introduced by asymmetric information, it is useful to characterize the first-best allocation. Suppose that the principal can directly observe each agent's effort level. In this case, the agent's compensation can be conditioned directly on both effort and productivity, i.e., $W(e_i, \theta_i)$. Since there is no informational asymmetry, monitoring becomes redundant and the optimal organizational design involves no hired monitors: $M = \varnothing$.

Given the ability to observe effort, the principal fully internalizes the cost of effort and solves the following optimization problem:

$$\begin{aligned} \max_{\{W,e\}} \quad & \int_0^1 \mathbb{E}\left[\theta_i + e_i - W(e_i, \theta_i)\right] di \\ \text{s.t.} \quad & W(e_i, \theta_i) \geq \frac{e_i^2}{2}, \quad \forall i \in [0, 1]. \end{aligned}$$

Because effort is observable, the principal can extract all of the agents' rents by setting compensation equal to their disutility of effort, i.e.,

$$W(e_i, \theta_i) = \frac{e_i^2}{2},$$

while choosing the effort level that maximizes net surplus. The optimal choice is $e_i = 1$ for all agents and all productivity types. Thus, in the absence of informational frictions, the principal induces the efficient level of effort and pays each agent a wage equal to their effort cost, independently of their productivity level θ_i. This result hinges on the fact that marginal returns to effort are constant across productivity types.

1.2 Firm design under no collusion

Before studying the problem of the organizational design under collusion it is interesting to understand how the principal designs the monitoring regime and what contracts does he offers when collusion is not possible. This section is structured as follows. First, I will characterize the contracts of the agents when no monitor is hired. Then, I will characterize the contracts of one monitor whose span of control is μ and the contract of the monitored agents. I will show that the contract under states 2 and 3 (no signal is received) correspond with the contract under no monitoring. With this, I will solve the optimal span of control when there is only one monitor in the firm. I do this process for the case where two monitors are hired and crucially show the optimal span of joint monitoring discussing its effects on profits and monitor and agent behavior.

1.2.1 No Monitoring

In the absence of monitoring, the principal cannot distinguish between productivity types beyond the realization of output. Since productivity can be either high or low, the contract must be designed around two observable output levels: $x_{\theta_L,\varnothing}$ and $x_{\theta_H,\varnothing}$, corresponding to low and high productivity types, respectively, when no informative signal is available.

$$x_{\theta_L,\varnothing} = e_{\theta_L,\varnothing} + \theta_L, \qquad x_{\theta_H,\varnothing} = e_{\theta_H,\varnothing} + \theta_H.$$

To induce effort from both types, the principal must design a contract that satisfies the agents' incentive compatibility constraints. First, the utility of each agent type must be non-negative, ensuring participation:

$$W(x_{\theta_L,\varnothing},\varnothing) - \frac{e^2_{\theta_L,\varnothing}}{2} \geq 0, \tag{IC$_{L\varnothing}$}$$

$$W(x_{\theta_H,\varnothing},\varnothing) - \frac{e^2_{\theta_H,\varnothing}}{2} \geq 0. \tag{PC$_H$}$$

Second, the principal must prevent high productivity agents from mimicking the behavior of low productivity agents to reduce the cost of effort. This leads to a third incentive compatibility constraint:

$$W(x_{\theta_H,\varnothing},\varnothing) - \frac{e^2_{\theta_H,\varnothing}}{2} \geq W(x_{\theta_L,\varnothing},\varnothing) - \frac{(e_{\theta_L,\varnothing}-\Delta\theta)^2}{2}, \tag{IC$_{H\varnothing}$}$$

If (IC$_{H\varnothing}$) does not hold, a high productivity agent could choose $\tilde{e} = e_{\theta_L,\varnothing} - \Delta\theta$ to produce $x_{\theta_L,\varnothing}$ and receive $W(x_{\theta_L,\varnothing},\varnothing)$. Notice that constraint (PC$_H$) becomes redundant when both (IC$_{L\varnothing}$) and (IC$_{H\varnothing}$) are met.

Given these constraints, the principal solves the following program:

$$\begin{aligned} \max_{\{W,e\}} \quad & \mathbb{E}\left[\int_0^1 (\theta_i + e_i - W(x_i,r_i))\, di\right] \\ \text{s.t.} \quad & \text{(IC}_{L\varnothing}\text{)},\ \text{(IC}_{H\varnothing}\text{)} \end{aligned}$$

The solution to this standard principal-agent problem is summarized below.

Proposition 1.1 (No Monitoring Contract)**.** *In the absence of a monitor, the contract that solves the principal's problem is given by:*

$$e_{\theta_L,\varnothing} = 1-\Delta\theta, \quad e_{\theta_H,\varnothing} = 1, \quad W(x_{\theta_L,\varnothing},\varnothing) = \frac{(1-\Delta\theta)^2}{2}, \quad W(x_{\theta_H,\varnothing},\varnothing) = \frac{1+2\Delta\theta-3\Delta\theta^2}{2}.$$

Under asymmetric information, the principal must raise the wage for high productivity agents to incentivize effort while reducing the required effort for low productivity agents. This tradeoff discourages high-type agents from mimicking low-type behavior, since doing so would require more effort relative to the output gained. As a result, low productivity agents receive zero surplus (constraint (IC$_{L\varnothing}$) binds), while high productivity agents earn an informational rent due to the need to deter imitation (constraint (IC$_{H\varnothing}$) also binds). These rents can be expressed as:

$$W(x_{\theta_H,\varnothing},\varnothing) - \frac{e^2_{\theta_H,\varnothing}}{2} = \Delta\theta\left(1-\frac{3}{2}\Delta\theta\right).$$

It is easy to see that the rents that high productive agents obtain are increasing in the productivity differences with their low productive counterparts, as $\Delta\theta \leq \frac{1}{2}$. To facilitate comparisons with other monitoring regimes, I define the expected profits under the no-monitoring scenario as $\Pi_{\varnothing M}$. Using the equilibrium contract:

$$\begin{aligned}\Pi_{\varnothing M} &= \int_0^1 \theta_i di + \int_0^1 e_i di - \int_0^1 W(x_i, \varnothing) di \\ &= \frac{1}{2} + \theta_L + \frac{\Delta\theta^2}{2}. \end{aligned} \tag{1.1}$$

This represents the minimum profit the principal can obtain in equilibrium when offering an optimal contract without any monitoring.

1.2.2 Single Monitoring

Assume now that the principal hires a single monitor to observe a subset of agents of size μ. Since there is only one monitor, I omit any subscript on the span of control. In this setting, the principal must offer two types of contracts: one to the monitor that ensures her participation, and another to the agents that deters high-productivity agents from mimicking low-productivity ones, as in previous sections. Moreover, the reported agents need to have positive utility to avoid shirking as in the perfect information scenario.

$$W(x_{\theta_L,\theta_L}, \theta_L) - \frac{e^2_{\theta_L,\theta_L}}{2} \geq 0, \tag{IC$_{LL}$}$$

$$W(x_{\theta_H,\theta_H}, \theta_H) - \frac{e^2_{\theta_H,\theta_H}}{2} \geq 0. \tag{IC$_{HH}$}$$

The firm's expected profits depend on the menu of contracts for the agents, $\mathbf{W}$, the monitor's contract $\mathbf{S}$, and the monitoring structure $\Omega = \{\mu\}$. For the subset of agents not monitored, expected profits are the same as under the no-monitoring regime. For monitored agents, the principal can condition wages on both output and the monitor's report:

$$\begin{aligned}\mathbb{E}\left[\Pi(\Omega, \mathbf{W}, \mathbf{S})\right] =& (1-\mu)\Pi_{\varnothing M} + \frac{1}{2}\mu\,(\theta_H + \theta_L + p(\mu)(e_{\theta_L,\theta_L} + e_{\theta_H,\theta_H} - W(x_{\theta_L,\theta_L}, \theta_L) \\ & - W(x_{\theta_H,\theta_H}, \theta_H) - S(x_{\theta_L,\theta_L}, \theta_L) - S(x_{\theta_H,\theta_H}, \theta_H) + (1-p(\mu))(e_{\theta_L,\varnothing} + e_{\theta_H,\varnothing} \\ & - W(x_{\theta_L,\varnothing}, \varnothing) - W(x_{\theta_H,\varnothing}, \varnothing) - S(x_{\theta_L,\varnothing}, \varnothing) - S(x_{\theta_H,\varnothing}, \varnothing))) \end{aligned} \tag{1.2}$$

Given this setup, the optimal contract for any span of control μ solves the following problem:

$$\max_{\{\mathbf{S},\mathbf{W},\mathbf{e}\}} \mathbb{E}\left[\Pi(\mu, \mathbf{W}, \mathbf{S})\right]$$

s.t.

$$(\text{LL}_M),\ (\text{PC}_M),\ (\text{IC}_{H\varnothing}),\ (\text{IC}_{L\varnothing}),\ (\text{IC}_{LL}),\ (\text{IC}_{HH})$$

In the absence of collusion, the contracts for agents and the monitor are independent, and can therefore be optimized separately. Moreover, the monitor's contract is not necessarily unique: any contract binding her participation constraint and ensures weakly higher payoffs for truthful reporting in both states will be profit-maximizing.

The following proposition characterizes one such optimal contract under imperfect monitoring and no collusion:

Proposition 1.2. *(Contracts under One Monitor) If there is one monitor covering a span μ, firm profits are maximized with the following contracts:*

$$S(x_i, r_i) = \frac{\bar{V}}{\mu}, \quad W(x_{\theta_L,\theta_L}, \theta_L) = W(x_{\theta_H,\theta_H}, \theta_H) = \frac{1}{2}, \quad e_{\theta_L,\theta_L} = e_{\theta_H,\theta_H} = 1,$$

and all other variables are as specified in Proposition 1.1.

Without collusion, all reports are truthful, and the principal only needs to guarantee that the monitor's participation constraint is satisfied. Whenever the principal receives an informative signal about an agent's productivity, she can perfectly infer effort from output. In this case, informational asymmetries are resolved, and the principal implements the first-best contract. In states where $\sigma_i = \theta_i$ (hence $r_i = \theta_i$), the agent's utility can be fully extracted. However, in uninformative states, the principal must offer the contract used in the no-monitoring regime, allowing high-productivity agents to extract rents due to asymmetric information.

Using Proposition 1.2, together with equations (1.1) and (1.2), I can compute expected profits for any $\mu \in [0, 1]$:

$$\mathbb{E}\left[\Pi(\Omega, \mathbf{W}, \mathbf{S})\right] = (1-\mu)\Pi_{\varnothing M} + \mu\frac{\theta_H + \theta_L}{2} + \frac{1}{2}\mu + \mu(1 - p(\mu))\frac{\Delta\theta^2 - \Delta\theta}{2} - \bar{V}.$$

Substituting the expression for $\Pi_{\varnothing M}$ yields:

$$\mathbb{E}\left[\Pi(\mu, \mathbf{W}, \mathbf{S})\right] = \Pi_{\varnothing M} - \bar{V} + \mu p(\mu)\frac{\Delta\theta(1-\Delta\theta)}{2}. \tag{1.3}$$

Equation (1.3) describes the expected profits the principal can attain for any given span of control μ. Importantly, profits depend on μ only through the product $\mu p(\mu)$, where $p(\mu)$ is decreasing in μ. The following proposition establishes the optimal span of control in this case.

Proposition 1.3. *The optimal span of control when $|M| = 1$ and there is no collusion is:*

$$\mu = \min\left\{\frac{1}{\lambda}, 1\right\}.$$

This result shows that the optimal span of control depends solely on the elasticity of the monitoring technology. Intuitively, this reflects decreasing returns to scale in monitoring. When the probability of obtaining a valid signal does not deteriorate too quickly with span of control (i.e., $\lambda \leq 1$), the optimal structure is for a single monitor to oversee the entire firm. If instead, $\lambda > 1$, a part of the workforce would optimally remain unmonitored.

Corollary 1.1. *In the absence of collusion, the principal hires a monitor if and only if:*

$$\begin{cases} \frac{p_0 e^{-1}}{\lambda}\frac{\Delta\theta(1-\Delta\theta)}{2} \geq \bar{V} & \text{if } \lambda \geq 1, \\ p_0 e^{-\lambda}\frac{\Delta\theta(1-\Delta\theta)}{2} \geq \bar{V} & \text{if } \lambda < 1. \end{cases}$$

Both left-hand sides are:

- *Increasing in p_0 (baseline monitoring effectiveness),*
- *Decreasing in λ (monitoring elasticity),*
- *Increasing in $\Delta\theta$ for $\Delta\theta < \frac{1}{2}$.*

The decision to hire a monitor therefore depends not only on the productivity gap and the monitor's reservation utility, but also on the efficiency of the monitoring technology. Specifically, a monitor is more likely to be hired when (i) baseline detection ability is high (p_0 is large), and (ii) monitoring efficiency

declines slowly with span of control (i.e., λ is low). Consequently, tasks with low verifiability—such as creative or knowledge-intensive activities—are less likely to be monitored under this structure. In contrast, when monitoring is scalable and robust to increased span, it is optimal for a single monitor to supervise a large share of the workforce.

1.2.3 Double Monitoring

Assume now that the principal hires two distinct monitors, m_1 and m_2. To define the organizational structure Ω, the principal must choose three variables: the span of control for each monitor, μ_{m_1} and μ_{m_2}, and the extent of joint supervision, μ_2. Note that the form of the monitors' contracts remains unchanged. This is because the only constraint the principal faces when designing these contracts is the monitors' participation constraint. Consequently, for any given span of control μ_{m_i}, offering a constant wage equal to $\frac{\bar{V}}{\mu_{m_i}}$ for any output and report satisfies that constraint. In expectation, each monitor then receives exactly their reservation utility, independent of their span of control, making monitoring as inexpensive as possible. It follows that both monitors are optimally assigned the same span of control, i.e., $\mu_{m_1} = \mu_{m_2}$.

The next proposition shows that the optimal contracts for both the agents and the monitors do not change with the addition of a second monitor, provided collusion is not possible.

Proposition 1.4. *The optimal contract under two monitors is the same as in Proposition 1.2.*

The proof, which is omitted, follows the same logic as that of Proposition 1.2. For notational convenience, I denote the span of control of each monitor by μ and the extent of joint supervision by γ. Since the firm's size is normalized to one, I must have $\mu \in [0, 1]$. The intersection γ must be non-negative and bounded such that it exceeds $2\mu - 1$ and does not exceed μ. Thus, the feasible organizational structure must satisfy the following constraints:

$$\gamma \geq 0,$$
$$\gamma \geq 2\mu - 1,$$
$$\mu - \gamma \geq 0.$$

This allows me to express the expected profits under any monitoring structure $\Omega = \{\mu, \gamma\}$ as:

$$\mathbb{E}(\Pi(\Omega, \mathbf{W}, \mathbf{S})) = (1 - 2\mu + \gamma)\Pi_{\varnothing M} + 2(\mu - \gamma)\Pi_{1M}(\mu, \mathbf{W}, \mathbf{S}) + \gamma\Pi_{2M}(\Omega, \mathbf{W}, \mathbf{S}),$$

where Π_{1M} and Π_{2M} represent the expected profits of the firm under single and joint monitoring, respectively. Substituting the contracts from Propositions 1.2 and 1.4 yields:

$$\mathbb{E}(\Pi(\Omega, \mathbf{W}, \mathbf{S})) = \Pi_{\varnothing M} + \left[2\mu p(\mu) - \gamma p(\mu)^2\right] \frac{\Delta\theta(1 - \Delta\theta)}{2} - 2\bar{V}. \tag{1.4}$$

The profit function when two monitors are employed without collusion consists of three key components: (i) the baseline profits $\Pi_{\varnothing M}$ when no monitors are used; (ii) the cost of hiring the two monitors, $2\bar{V}$; and (iii) the informational gain, captured by the expression $[2\mu p(\mu) - \gamma p(\mu)^2] \frac{\Delta\theta(1-\Delta\theta)}{2}$. The latter reflects the reduction in informational rents earned by high-productivity agents. Importantly, since profits are linearly decreasing in the extent of joint supervision, the following result holds:

Proposition 1.5. *The optimal length of joint supervision is given by:*

$$\gamma = \max\{0, 2\mu - 1\}.$$

This result implies that, regardless of monitoring efficiency or span of control, the principal will always minimize the extent of joint supervision when there is no collusion. The key intuition is that assigning new agents to a monitor expands coverage, whereas increasing joint supervision only duplicates information on workers who are already being monitored. Hence, when collusion is absent, the principal prefers to use monitoring capacity to supervise more agents rather than to generate redundant signals on the same agent.

The next proposition characterizes the optimal span of control as a function of λ:

Proposition 1.6.

- *If* $\lambda \geq 2$*, the span of control is* $\mu = \frac{1}{\lambda}$*.*
- *If* $\lambda \leq 1$*, the optimal span of control is* $\mu = 1$*.*
- *If* $\lambda \in (1, 2)$*:*
  - $\mu \leq \frac{1}{\lambda}$*.*
  - μ *is decreasing in both* λ *and* p_0*.*
  - *If* $\lambda \geq 2\left(1 + W_0\left(-\frac{p_0}{e}\right)\right)$*, then* $\mu = \frac{1}{2}$ *is a local maximum of the profit function, though not necessarily a global one.*[5] *For* p_0 *large enough, other local maxima exist.*

The result highlights a subtle trade-off introduced by overlapping supervision. When the span of control is widened, each monitor covers more agents, and their individual detection probability declines. At the same time, however, the number of monitors observing any given agent may increase, so that detection depends not only on the effectiveness of each monitor but also on the degree of redundancy across them.

This creates what can be called an N–versus–P effect: the probability of detection is shaped both by the number of monitors (N) overseeing a task and by the effectiveness (P) of each individual monitor. As μ expands, P falls because each monitor supervises more agents, but N may rise whenever spans overlap so that two monitors jointly oversee the same agent. For some values of μ, these forces offset each other, and the overall probability of detection may rise or fall depending on which margin dominates. This non-monotonicity is the central insight of the proposition.

I now compare profits under one and two monitors. The conditions under which the principal prefers two monitors are given in the following corollaries.

Corollary 1.2. *If* $\lambda \geq 2$*, the principal prefers to hire two monitors if and only if:*

$$\frac{p_0 e^{-1}}{\lambda} \cdot \frac{\Delta\theta(1 - \Delta\theta)}{2} \geq \bar{V}.$$

[5] W_0 denotes the principal branch of the Lambert W function, defined implicitly by $W(x)e^{W(x)} = x$, with $W_0(-1/e) = -1$ and $W_0(0) = 0$.

Observe that Corollary 1.2 reproduces the same condition as Corollary 1.1. This is because, when the optimal span of control is less than one half, and the principal is willing to hire one monitor to supervise $\frac{1}{\lambda}$ agents for a wage $\bar{V}$, then it is also optimal to hire a second monitor to cover the remainder of the firm. More generally, this suggests that if the condition in the corollary holds and $\lambda \geq n$, it is optimal to hire n monitors. However, this does not necessarily extend to cases where $\mu > \frac{1}{2}$, since the marginal benefit of hiring an additional monitor declines due to the joint supervision effect. The next corollary characterizes the hiring decision as a function of the monitors' reservation utility and monitoring efficiency when $\lambda \leq 1$.

Corollary 1.3. *If $\lambda \leq 1$, the principal hires two monitors if and only if:*

$$p_0 e^{-\lambda}\left(1 - p_0 e^{-\lambda}\right) \cdot \frac{\Delta\theta(1-\Delta\theta)}{2} \geq \bar{V}.$$

This corollary shows that, under both very low and very high monitoring efficiencies, the incentive to hire two monitors is weak. When detection is poor, no monitor is valuable enough relative to their cost. Conversely, when detection is highly effective, a single monitor suffices to obtain truthful and precise information, rendering the second monitor redundant. The incentive to hire two monitors peaks at intermediate signal accuracy—particularly when the monitoring precision is around one half. This reflects a trade-off: one monitor is not sufficiently precise alone, yet signals are still informative enough to justify their cost. These regions are illustrated in Figure 1.1, which maps optimal hiring choices as a function of monitor efficiency and reservation utility. In this case, hiring behavior is monotonic in the cost of monitors: for fixed p_0 and λ, the number of monitors hired decreases as $\bar{V}$ increases.

Finally, I consider the hiring decision under the intermediate case, $\lambda \in (1, 2)$.

Corollary 1.4. *Assume $\lambda \in (1, 2)$. Define*

$$A(\mu_1^*; p_0, \lambda) \equiv 2\mu_1^* p(\mu_1^*) - \mu^2 p(\mu_1^*)^2.$$

$$V_{1M}(p_0, \lambda) \equiv \frac{p_0}{\lambda} e^{-1}\Delta, \quad V_{2M}(p_0, \lambda) \equiv \left(A(\mu_1^*; p_0, \lambda) - \tfrac{p_0}{\lambda} e^{-1}\right)\Delta.$$

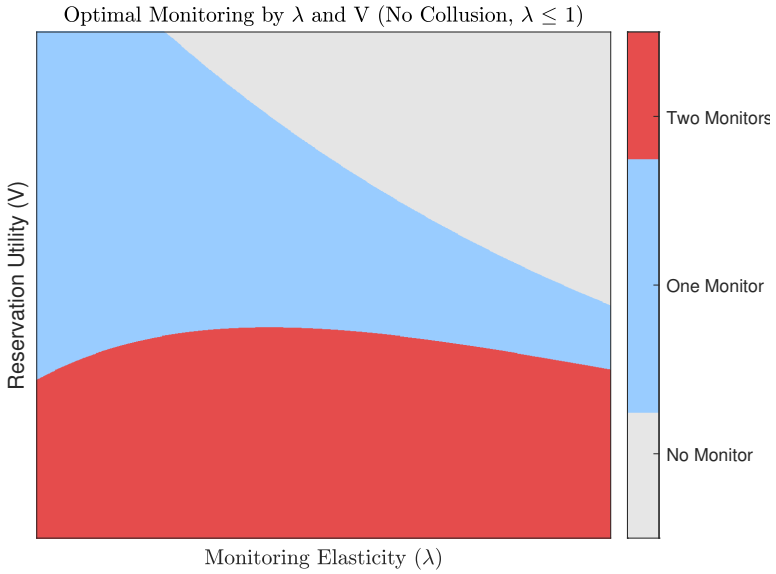


Figure 1.1: Optimal hiring decision for $\lambda \leq 1$ as a function of reservation wages of the monitors $\bar{V}$ and the monitoring elasticity λ. $p_0 = 0.7$. No collusion.

Thus this gives a monotonic ordering for the three hiring conditions.

$$\begin{cases} \bar{V} \leq V_2(p_0, \lambda) & \Rightarrow \textit{two monitors}, \\ V_2(p_0, \lambda) < \bar{V} < V_1(p_0, \lambda) & \Rightarrow \textit{one monitor}, \\ \bar{V} \geq V_1(p_0, \lambda) & \Rightarrow \textit{no monitor}. \end{cases}$$

Moreover, the cutoffs satisfy $\frac{\partial V_{1M}}{\partial p_0}, \frac{\partial V_{2M}}{\partial p_0} > 0$ *and* $\frac{\partial V_{1M}}{\partial \lambda}, \frac{\partial V_{2M}}{\partial \lambda} < 0$*, so higher verifiability expands the hiring regions while faster span decay contracts them.*

The hiring behavior when $\lambda \in (1, 2)$ mirrors the logic of the previous case, but with an additional trade-off. For high monitor reservation wages $\bar{V}$, monitoring is too costly and the principal hires no one. For low values of $\bar{V}$, it is optimal to employ two monitors, while for intermediate values, a single monitor is preferred.

The key difference in this range of λ is that when the span μ_1 exceeds $1/2$, some agents fall under the oversight of *two* monitors (N increases), even as each monitor's detection probability declines with span (P decreases). The net detection term,

$$A(\mu_1) = 2\mu_1 p(\mu_1) - \mu^2 p(\mu_1)^2,$$

captures precisely this tension: redundancy improves coverage but is offset by weaker individual monitors.

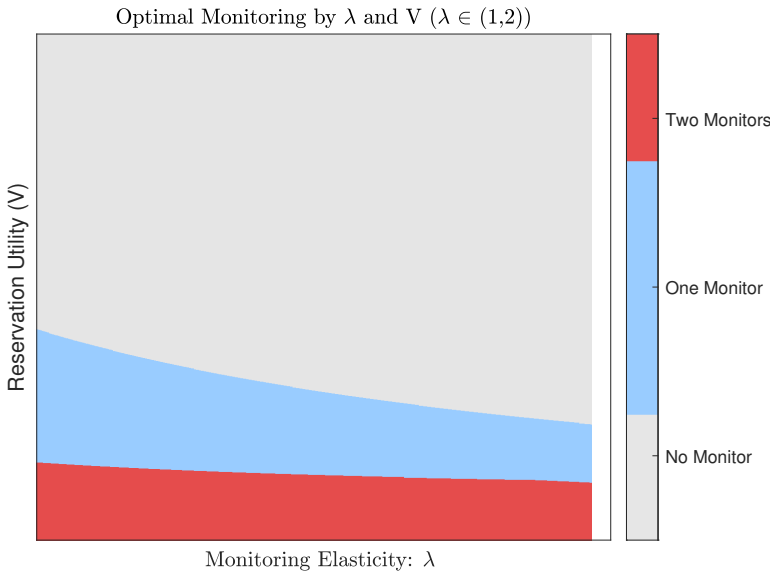


Figure 1.2: Optimal hiring decision for $\lambda \in (1, 2)$ as a function of reservation wages of the monitors $\bar{V}$ and the monitoring elasticity λ. $p_0 = 0.7$.

As Figure 1.2 shows, comparative statics follow directly. Better verifiability ($p_0 \uparrow$) magnifies the returns to redundancy and raises the range of $\bar{V}$ for which two monitors are worthwhile. By contrast, faster decay (λ increases) erodes effectiveness more steeply as spans widen, compressing the profitability of two-monitor structures and pushing the principal toward simpler designs. Figure 1.2 illustrates this dynamic: unlike the $\lambda \leq 1$ case, the condition for hiring two monitors is now monotone in λ, with higher decay always making multiple monitors less attractive.

Finally, the last three corollaries help us understand how different combinations of λ and $\bar{V}$ shape the incentives to implement monitoring. When λ is relatively small, monitoring remains efficient even as the reservation wage $\bar{V}$ increases, making it optimal to hire monitors over a wide range of values for $\bar{V}$. However, as λ increases, the effectiveness of monitoring deteriorates, and the range of reservation wages for which hiring monitors is profitable narrows. In the intermediate case where $\lambda \in (1, 2)$, increases in λ not only reduce monitoring efficiency but also shrink the optimal span of control. This further limits the number of informative signals obtained, making monitoring relatively less attractive and justifiable only for low values of $\bar{V}$. The case $\lambda \geq 2$ is even more restrictive. Here, the reduction in span of control is such that part of the firm may remain entirely unmonitored. This further diminishes the returns to hiring monitors. Importantly, in this range,

the conditions for hiring one or two monitors coincide and become monotonically decreasing in λ. As a result, the principal is willing to pay for monitoring only when $\bar{V}$ is sufficiently low.

In conclusion, as monitoring elasticity λ increases, the set of reservation wages $\bar{V}$ that justify the use of monitoring becomes strictly smaller. Higher values of λ therefore impose tighter constraints on the firm's ability to rely on organizational monitoring as an incentive device.

1.3 Firm Design under Collusion

Assume now that, after each agent learns her productivity and monitors observe her signal, but before any effort is exerted or report is submitted to the principal, the agent and the monitor may enter into a side contract. This side contract specifies a transfer $t(x_i, r_i) \in \mathbb{R}$ from the agent to the monitor, contingent on the realized output and the report sent by the monitor. Crucially, this transfer is not restricted to be positive: a negative transfer corresponds to a payment from the monitor to the agent—for instance, to incentivize a particular effort choice.

Throughout this section, I adopt a simple assumption: regardless of the underlying bargaining process, any subset of players that can achieve a strictly higher joint utility by coordinating their actions will choose to collude.

In this context, the contracts derived in the previous section do not deter collusion. Consider the case in which a high-productivity agent is observed by a monitor who receives a perfectly informative signal. Since the monitor's wage is invariant to whether she reports or not, she is indifferent between the two actions. The agent, however, knows that if she is not reported, she will earn information rents due to asymmetric information. This creates a profitable deviation in which the agent offers a side payment to induce silence from the monitor. To eliminate this incentive, it is necessary that the combined payoff of the agent and the monitor is strictly higher when the report is truthful than when it is withheld. This requirement is central to designing contracts that are robust to collusion.[6]

[6]Importantly, for any contract offered by the principal, there always exists an equilibrium in which monitors truthfully report their signals—i.e., $r_{ij}(\sigma_{ij}, t(x_i, r_i)) = \sigma_{ij}$—and agents do not initiate any side transfers. This non-collusive equilibrium is sustained by the absence of credible bribing opportunities and the monitor's indifference to deviating. However, such an equilibrium may not be unique: other equilibria involving collusion can emerge, depending

When more than one monitor oversees the same agent—as in joint monitoring regimes —collusion is not restricted to agent-monitor pairs. Under certain realizations of the signal and contract structure, one monitor may have incentives to transfer resources to another monitor in exchange for suppressing a report. In this section, I examine such possibilities and study how the threat of collusion—whether vertical (agent-monitor) or horizontal (monitor-monitor)—shapes both the optimal contract design and the internal organization of monitoring within the firm.

As in the previous section, I begin by characterizing the optimal contracts and organizational structure when only one monitor is hired. I then turn to the case of two monitors. Finally, I analyze the principal's incentives to hire monitors under varying parameter configurations, taking into account the constraints imposed by the possibility of collusion.

1.3.1 Single Monitoring under Collusion

Assume now that the principal hires only one monitor. As discussed earlier, the contract from Proposition 1.2 fails to be collusion-free due to incentives for high-productivity agents to collude with the monitor. In particular, if

$$S(x_{\theta_H,\theta_H},\theta_H) \leq S(x_{\theta_H,\varnothing},\varnothing) + W(x_{\theta_H,\varnothing},\varnothing) - \frac{e^2_{\theta_H,\varnothing}}{2},$$

then the monitor and the agent prefer to misreport, violating collusion-freedom. This situation arises, for example, if the monitor receives the same wage regardless of whether she reports the high type or not. Thus, the contract must satisfy the following collusion constraint:

$$S(x_{\theta_H,\theta_H},\theta_H)+W(x_{\theta_H,\theta_H},\theta_H)-c(e_{\theta_H,\theta_H}) \geq S(x_{\theta_H,\varnothing},\varnothing)+W(x_{\theta_H,\varnothing},\varnothing)-c(e_{\theta_H,\varnothing}). \tag{CC$_{HH}$}$$

There is also a second potential collusion channel. Since the monitor's wage depends on both the report and the productivity of the agent, she might have incentives to bribe the agent to reduce effort and mimic a lower type if that

on the strategic incentives embedded in the contract. I define a *collusion-free contract* as a contract menu that induces a unique equilibrium in which no collusion arises, either between agents and monitors or across monitors.

increases her compensation. To prevent this, the contract must also satisfy:

$$S(x_{\theta_H,\varnothing},\varnothing)+W(x_{\theta_H,\varnothing},\varnothing)-c(e_{\theta_H,\varnothing}) \geq S(x_{\theta_L,\varnothing},\varnothing)+W(x_{\theta_L,\varnothing},\varnothing)-c(e_{\theta_L,\varnothing}-\Delta\theta). \tag{CC$_{H\varnothing}$}$$

Note that if the agent's incentive compatibility constraint (IC$_{H\varnothing}$) binds, then (CC$_{H\varnothing}$) reduces to

$$S(x_{\theta_H,\varnothing},\varnothing) \geq S(x_{\theta_L,\varnothing},\varnothing).$$

I will later verify that this is the case in equilibrium. Hence, the contract that maximizes the principal's profits under collusion must satisfy: limited liability for both agents and monitor, the monitor's participation constraint, the agents' incentive compatibility constraints (IC1 for low types and IC3 for separating types), and the collusion constraint (CC$_{HH}$ and CC$_{H\varnothing}$). The principal's problem becomes:

$$\max_{\{\mathbf{S},\mathbf{W},e\}} \mathbb{E}\left[\Pi(\mu,\mathbf{W},\mathbf{S})\right]$$

s.t.

(LL$_M$), (PC$_M$), (IC$_{H\varnothing}$), (IC$_{L\varnothing}$), (IC$_{LL}$), (IC$_{HH}$), (CC$_{HH}$), (CC$_{H\varnothing}$).

The following proposition characterizes the optimal solution to this program.

Proposition 1.7 (Optimal Contract under One Monitor with Collusion)**.** *Suppose that agents and the monitor may collude through side contracts. Then, the profit-maximizing collusion-free contract when one monitor observes a share μ of the agents features:*

- ***Effort:*** *Agents exert first-best effort except for unreported low-productivity agents. In that case, effort is distorted downward:*

$$e_{\theta_L,\theta_L} = e_{\theta_H,\theta_H} = e_{\theta_H,\varnothing} = 1, \quad e_{\theta_L,\varnothing} = \max\left\{1-\frac{1}{1-p(\mu)}\Delta\theta, \frac{2\bar{V}}{\Delta\theta\mu p(\mu)}+\frac{\Delta\theta}{2}\right\}.$$

- ***Agent Wages:*** *The incentive constraints (IC$_{L\varnothing}$) bind for low types and (IC$_{H\varnothing}$) for separating high types. High types reported truthfully may receive*

rents:

$$W(x_{\theta_L,\theta_L},\theta_L)=\frac{1}{2},\quad W(x_{\theta_L,\varnothing},\varnothing)=\frac{e^2_{\theta_L,\varnothing}}{2},$$
$$W(x_{\theta_H,\varnothing},\varnothing)=\frac{1}{2}+\max\left\{W(x_{\theta_L,\varnothing},\varnothing)-\frac{(e_{\theta_L,\varnothing}-\Delta\theta)^2}{2},\ S(x_{\theta_H,\theta_H},\theta_H)\right\},$$
$$W(x_{\theta_H,\theta_H},\theta_H)\geq\frac{1}{2}.$$

- ***Monitor Payments:*** *The monitor is compensated only when she reports a high-productivity agent:*

$$S(x_{\theta_H,\theta_H},\theta_H)=\max\left\{\frac{2\bar{V}}{\mu p(\mu)},\ W(x_{\theta_H,\varnothing},\varnothing)-W(x_{\theta_H,\theta_H},\theta_H)\right\}.$$

All other transfers are zero:

$$S(x_{\theta_L,\theta_L},\theta_L)=S(x_{\theta_L,\varnothing},\varnothing)=S(x_{\theta_H,\varnothing},\varnothing)=0.$$

The key insight is that improving monitor accuracy—a higher detection probability $p(\mu)$—makes collusion more tempting, since supervisors now observe more high-productivity agents in expectation. To deter this, the principal lowers effort incentives for unreported low types, thereby shrinking the rents available for collusion. This mechanism limits bribes but comes at the cost of efficiency, since reducing effort depresses output and may lower total profits. The trade-off is thus between stronger discipline against collusion and weaker productive incentives.[7]

To classify equilibria, let $\bar{V}_1(\mu)$ and $\bar{V}_2(\mu)$ denote the values of the monitor's reservation wage at which its participation constraint binds when (i) the collusion constraint is slack and (ii) the collusion constraint alone is binding, respectively:

$$\bar{V}_1(\mu):=\mu p(\mu)\frac{\Delta\theta}{2}\left(1-\tfrac{3}{2}\Delta\theta\right),\qquad \bar{V}_2(\mu):=\mu p(\mu)\frac{\Delta\theta}{2}\left(1-\Delta\theta\tfrac{3-p(\mu)}{2(1-p(\mu))}\right).$$

These thresholds partition the equilibrium space as follows:

- If $\bar{V}>\bar{V}_1(\mu)$, the monitor's participation constraint is binding while collusion is not a threat, so the non-collusion contract remains optimal.

[7] I restrict attention to cases where $p(\mu)$ is small enough to ensure $e_{\theta_L,\varnothing}\geq 0$. If instead $e_{\theta_L,\varnothing}<0$, unreported low types are excluded from the firm.

- If $\bar{V} < \bar{V}_2(\mu)$, only the collusion constraint binds. The monitor is paid exactly enough to eliminate profitable side-payments:

$$S(x_{\theta_H,\theta_H}, \theta_H) = W(x_{\theta_H,\varnothing}, \varnothing) - W(x_{\theta_H,\theta_H}, \theta_H).$$

- If $\bar{V} \in (\bar{V}_2(\mu), \bar{V}_1(\mu))$, both constraints bind. The optimal contract then allocates fixed rents to the monitor and adjusts agent transfers accordingly:

$$W(x_{\theta_H,\theta_H}, \theta_H) = \tfrac{1}{2}, \quad S(x_{\theta_H,\theta_H}, \theta_H) = \tfrac{2\bar{V}}{\mu p(\mu)}, \quad W(x_{\theta_H,\varnothing}, \varnothing) = \tfrac{1}{2} + S(x_{\theta_H,\theta_H}, \theta_H).$$

This characterization also clarifies how rents are distributed. When $\bar{V} \leq \bar{V}_2(\mu)$, agents and monitors earn equal rents, coinciding with the agent's non-collusion rents. When $\bar{V} \in (\bar{V}_2(\mu), \bar{V}_1(\mu))$, the monitor's participation constraint absorbs all its surplus, leaving collusion rents entirely to the agent.

Importantly, monitor rents depend positively on p_0: better verifiability raises detection probability, which in turn increases the frequency of valuable reports and hence the scope for bribes. By contrast, when p_0 is low, detection is rare, collusion is less attractive, and the principal does not need to transfer as much surplus to the monitor to prevent corruption. Thus, while stronger monitoring improves accuracy, it also magnifies the rents at stake in collusion, tightening incentive constraints and altering who captures surplus in equilibrium.

Optimal Structure under One Monitor and Collusion

Proposition 1.7 allows us to compute the expected profits when only the collusion constraint (CC_{HH}) binds, for any given span of control:

$$\mathbb{E}\left[\Pi(\Omega, \mathbf{W}, \mathbf{S})\right] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{4}\mu\frac{p(\mu)}{1-p(\mu)}, \tag{1.5}$$

and when both the collusion constraint and the monitor's participation constraint bind:

$$\Pi(\Omega, \mathbf{W}, \mathbf{S}) = \Pi_{\varnothing M} + \frac{1}{2}\mu\left[\Delta\theta(1-\Delta\theta) - (1-p(\mu))\frac{1}{2}\left(1-\frac{\Delta\theta}{2}\right)^2\right] - \frac{\bar{V}^2}{\Delta\theta^2\mu p(\mu)^2} + \frac{\bar{V}^2}{\Delta\theta^2\mu p(\mu)} + \frac{\bar{V}}{\Delta\theta p(\mu)} - \frac{3\bar{V}}{2p(\mu)} + \bar{V}\left(\frac{1}{2} - \frac{1}{\Delta\theta}\right).$$

Notice that if $\bar{V} = \bar{V}_2(\mu)$ both expressions coincide, which implies that, fixing μ the profit function is continuous with respect to the reservation utility of the monitor. The same reasoning applies when $\bar{V} = \bar{V}_1$. This makes the profit function continuous for $\bar{V}$. The next proposition characterizes the optimal monitoring span under collusion when $\bar{V} < \bar{V}_2$:

Proposition 1.8. *Assume that in the optimal program,* PC_M *is slack. Then, the optimal span of control* μ *that maximizes profits when one monitor is hired under collusion is implicitly given by:*

$$\mu = \min\left\{\frac{1-p_0e^{-\lambda\mu}}{\lambda}, 1\right\},$$

which has the explicit solution using the main branch of the Lambert W function, W_0*:*

$$\mu = \min\left\{\frac{1+W_0(-p_0e^{-1})}{\lambda}, 1\right\}.$$

This proposition highlights that optimal span of control shrinks in the presence of collusion. The reason is that higher monitoring efficiency—through either larger λ or p_0—increases the exposure of the monitor to high-type agents. This, in turn, raises the incentive and the cost to prevent collusion. The principal internalizes this cost by reducing the number of agents monitored, even if monitoring is costless and reporting is frictionless. In short, under collusion, the cost of a monitor becomes increasing in their span of control.

Corollary 1.5 (Optimal Span of Control under Collusion)**.** *The optimal span of control under collusion is strictly smaller than under non-collusion, except in the special case where both coincide. In particular, the two are equal—and equal to one—if and only if:*

$$\lambda \leq 1 + W_0(-p_0e^{-1}),$$

where $W_0(\cdot)$ *denotes the main branch of the Lambert W function.*

Moreover, the optimal span of control μ^* is strictly decreasing in both the elasticity of monitoring λ and the baseline detection probability p_0. In the limit as p_0 converges to 1, the optimal μ^* converges to zero.

As in the non-collusion case, a higher monitoring elasticity λ reduces the optimal span of control. However, under collusion, even p_0—the baseline precision of monitoring—has a negative effect on μ^*. A higher p_0 increases the chance of collusion occurring and hence raises the compensation needed to deter it. This prompts the principal to reduce μ until the firm re-enters the region where collusion is no longer profitable.

Hiring vs Not Hiring a Monitor under Collusion. Proposition 1.7 also allows me to compare the profitability of hiring a monitor with the outside option of not hiring at all. When only (CC_{HH}) is binding and $\mu \leq 1$, the expected profits can be written as:

$$\mathbb{E}\left[\Pi(\mu, \mathbf{W}, \mathbf{S})\right] = \Pi_{\varnothing M} - \frac{\Delta\theta^2}{4\lambda} W\left(-p_0 e^{-1}\right).$$

As in the non-collusion case, profits fall with λ and rise with p_0 and $\Delta\theta$. Higher baseline probabilities p_0 raise detection rates and thus improve monitoring effectiveness, which boosts profits via (i) an adjustment in the optimal span of control, and (ii) a direct efficiency gain. The role of λ is subtler but follows a similar logic. As $W_0\left(-p_0 e^{-1}\right) < 0$, profits are decreasing in λ; when the elasticity of monitoring decreases, profits increase.

Notice that since $W(-p_0 e^{-1}) \leq 0$, the firm always achieves *strictly higher* profits by hiring a monitor than not, provided that the monitor's cost satisfies $\bar{V} \leq \bar{V}_2$. Assume now that $\lambda \leq 1 + W_0\left(-p_0 e^{-1}\right)$. From the Corollary, we know that $\mu = 1$. Hence, the profit function can be written as,

$$\mathbb{E}\left[\Pi(1, \mathbf{W}, \mathbf{S})\right] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{4} \frac{p_0 e^{-\lambda}}{1 - p_0 e^{-\lambda}}.$$

As the second term is positive, also in this case the principal prefers to hire one monitor than none. As a consequence, the following Corollary.

Corollary 1.6. *When $\bar{V} \leq \bar{V}_2$, the principal always prefers to hire one monitor than none.*

1.3.2 Double Monitoring

When two monitors are employed and agents face different monitoring regimes, the principal must design distinct contracts tailored to each situation. Specifically, agents may fall into one of three regimes: unmonitored, monitored by a single monitor, or jointly monitored. Accordingly, the principal must offer a separate contract for each case. The same applies to monitors: each is given a contract for the region they supervise independently, and a different contract for the region of joint supervision.

If the monitor's participation constraint binds, the principal pays the monitor the minimum required wage. It is more interesting, however, to analyze the case where the participation constraint is slack. In this case, the principal may condition the monitor's wage not only on her own report but also on the report of the other monitor. In the joint monitoring region, the principal can implement cross-payment schemes with penalties ($\beta \geq 0$) and bonuses ($\delta \geq 0$) to incentivize honest reporting. I will index contracts by the number of monitors involved in the supervision.

Hiring two monitors generates two main benefits. First, as in the single-monitoring case, it increases the likelihood of detecting agents' productivity. Second, through cross-monitoring and the associated wage structure, it mitigates collusive incentives and can therefore lower the honesty premium that must be paid to supervisors. The gain from dual monitoring thus operates through both information and incentives. Formally, the firm's expected profit when hiring two monitors is:

$$\mathbb{E}\left[\Pi(\Omega, \mathbf{W}, \mathbf{S})\right] = (1 - 2\mu + \gamma)\Pi_{\varnothing M} + 2(\mu - \gamma)\Pi_{1M} + \gamma\Pi_{2M},$$

where Π_{iM} denotes the profits in the region monitored by i monitors. Note that if $\gamma = 0$, the monitors do not overlap, and the profit function becomes a linear transformation of the one-monitor case. Hence, when there is no overlap, each monitor's span of control is identical to the single-monitor case.

Conversely, when $\gamma = \mu$, both monitors observe the same agents. Then the profit function simplifies to:

$$\mathbb{E}\left[\Pi((\mu, \mu), \mathbf{W}, \mathbf{S})\right] = (1 - \mu)\Pi_{\varnothing M} + \mu\Pi_{2M},$$

which mirrors the one-monitor structure but replaces single-monitor profits with joint-monitor profits.

Before solving the principal's problem, it is crucial to understand the new sources of collusion. In regions with single monitoring, collusion constraints remain unchanged. However, in the jointly monitored region, collusion can occur in multiple ways: between one monitor and the agent, between the two monitors, or among all three parties.

To prevent this, I introduce new collusion constraints. The constraint ensuring that both monitors prefer to report truthfully when they observe the same agent:

$$2S(x_{\theta_H,\theta_H},\theta_H)+W(x_{\theta_H,\theta_H},\theta_H)-c(e_{4,1})\geq 2S(x_{\theta_H,\varnothing},\varnothing)+W(x_{\theta_H,\varnothing},\varnothing)-c(e_{3,1}). \tag{CC_{H2H}}$$

Next, (CC_{H1H}) ensures that a single monitor who observes a high-type agent in the joint region prefers to report alone:

$$S(x_{\theta_H,\theta_H},\theta_H)+\delta+S(x_{\theta_H,\varnothing},\varnothing)-\beta+W(x_{\theta_H,\theta_H},\theta_H)-c(e_{4,1})\geq 2S(x_{\theta_H,\varnothing},\varnothing) +W(x_{\theta_H,\varnothing},\varnothing)-c(e_{3,1}). \tag{CC_{H1H}}$$

Lastly, (CC_M) ensures that the first monitor cannot be bribed into silence when the other monitor stays silent:

$$2S(x_{\theta_H,\theta_H},\theta_H)\geq S(x_{\theta_H,\theta_H},\theta_H)+\delta+S(x_{\theta_H,\varnothing},\varnothing)-\beta. \tag{CC_M}$$

Additionally, high-type agents not reported in the joint monitoring region face an incentive constraint ensuring they do not mimic low types. This constraint is denoted as $(\mathrm{IC}_{H\varnothing})_2$.

Let the expected utility of each monitor be the weighted sum of their expected payoffs across the regions: $\mu-\gamma$ under independent monitoring and γ under joint monitoring. The participation constraint ensures this expected utility exceeds $\bar{V}$. The principal's program is then:

$$\begin{aligned}\max_{\{S,W,e\}}\ & \mathbb{E}\left[\Pi(\Omega,\mathbf{W},\mathbf{S})\right]\\ \text{s.t.}\ & (\mathrm{LL}_M),\ (\mathrm{PC}_M),\ (\mathrm{IC}_{L\varnothing}),\ (\mathrm{IC}_{H\varnothing}),\ (\mathrm{IC}_{H\varnothing})_2,\\ & (\mathrm{CC}_{HH}),\ (\mathrm{CC}_{H2H}),\ (\mathrm{CC}_{H1H}),\ (\mathrm{CC}_M),\ \beta\geq 0\end{aligned}$$

The next proposition characterizes the optimal contract under this setting.

Proposition 1.9. *Suppose that agents and the monitors may collude through side contracts. Then, the profit-maximizing collusion-free contract when* $\Omega = (\mu, \mu_2)$ *and* (PC_M) *is slack, features:*

- ***Effort:*** *Agents exert first-best effort except for unreported low-productivity agents. In that case, effort is distorted downward:*

$$e_{\theta_L,\theta_L,2} = e_{\theta_H,\theta_H,2} = e_{\theta_H,\varnothing,2} = 1, \quad e_{\theta_L,\varnothing,2} = 1 - \frac{\Delta\theta}{(1-p(\mu))^2}.$$

- ***Agent Wages:*** *The incentive constraints* ($\mathrm{IC}_{L\varnothing}$) *bind for low types and* ($\mathrm{IC}_{H\varnothing}$) *for separating high types. High types reported truthfully may receive rents:*

$$\begin{aligned} &W_2(x_{\theta_L,\theta_L},\theta_L) = \frac{1}{2}, \quad W_2(x_{\theta_L,\varnothing},\varnothing) = \frac{e^2_{\theta_L,\varnothing}}{2}, \\ &W_2(x_{\theta_H,\varnothing},\varnothing) = \frac{1}{2} + \Delta\theta - \frac{2+(1-p(\mu))^2}{2(1-p(\mu))^2}\Delta\theta^2, \\ &W_2(x_{\theta_H,\theta_H},\theta_H) \geq \frac{1}{2}. \end{aligned}$$

- ***Monitor Payments:*** *The monitor is compensated only when she truthfully reports a high-productivity agent:*

$$S_2(x_{\theta_H,\theta_H},\theta_H) = \frac{1}{2}\left(W_2(x_{\theta_H,\varnothing},\varnothing) - W_2(x_{\theta_H,\theta_H},\theta_H)\right),$$
$$\delta = S_2(x_{\theta_H,\theta_H},\theta_H).$$

  All other transfers are zero:

$$S_2(x_{\theta_L,\theta_L},\theta_L) = S_2(x_{\theta_L,\varnothing},\varnothing) = S_2(x_{\theta_H,\varnothing},\varnothing) = 0.$$

Because all contracts satisfying the above equality yield the same profit, we can fix $W_2(x_{\theta_H,\theta_H},\theta_H) = \frac{1}{2}$ and let $2S_2 = W_2(x_{\theta_H,\varnothing},\varnothing) - \frac{1}{2}$. As in the previous section, this type of contract allows us to compute the most extreme case of the monitor's wage and check when the participation constraint binds.

Given the results in the previous proposition, I can write the expected profits for general monitoring regime (μ, γ) when the participation constraint is not binding as:

$$\mathbb{E}\left[\Pi(\mu, \mathbf{W}, \mathbf{S})\right] = \Pi_{\varnothing M} + \frac{\Delta\theta^2 p(\mu)}{2(1-p(\mu))}\left[\mu + \gamma \cdot \frac{1}{1-p(\mu)}\right].$$

This function reveals that expected profits increase with both μ and γ, provided the participation constraint is not binding. When it is, profits revert to a different regime where both participation constraint of the monitor and the set of collusion constraints are simultaneously binding. The next proposition formalizes the optimal overlapping monitoring structure in this context.

Notice that, if we set $\gamma = 0$, the previous expression becomes the profits under collusion when only one monitor is hired (1.5). As each monitor takes its own part of the firm, the cross payments are not present in the optimal contract.

Proposition 1.10. *The optimal monitoring structure under collusion with two monitors features a non-trivial overlap in their monitoring responsibilities. Specifically, the optimal level of joint monitoring is given by:*

$$\gamma = \min\left\{\mu, \tilde{\mu}_2\right\},$$

where $\tilde{\mu}_2$ denotes the critical level of joint monitoring that makes the monitor's participation constraint bind, conditional on a given μ.

This result stands in stark contrast to Proposition 1.5, which establishes that, in the absence of collusion, the optimal degree of monitoring overlap should be minimized. In the benchmark non-collusive case, monitors are simply redundant in the intersection, and overlapping responsibilities only duplicate costs without generating any additional incentive or informational benefits. Therefore, the principal finds it optimal to minimize γ as much as possible.

Under collusion, however, the logic reverses. Allowing a non-zero overlap in monitoring coverage provides the principal with a powerful new tool: *cross-monitoring*. When two monitors observe the same agent, the principal can design contracts that depend on both monitors' reports. This allows her to structure cross-rewards that are effective at deterring collusion. These cross-incentives lower

the temptation for either monitor to accept side deals from agents, because the payoff from deviating depends on the independent action of the other monitor.

Furthermore, introducing overlap can reduce the cost of achieving collusion-proofness. When monitors work in isolation, they must be individually compensated to resist collusion, which can become prohibitively expensive, especially when detection probabilities are high. But with overlapping monitoring, *deterrence is shared*: the cost of maintaining truthful behavior is spread across two agents, reducing the required transfer per monitor. As a result, the joint region can generate a net savings in enforcement costs while maintaining or increasing the principal's expected output due to higher detection likelihood. Therefore, increasing γ can be strictly profit-enhancing up to the point where the participation constraint becomes binding.

This demonstrates a central insight of the model: *collusion risk reshapes the optimal organizational structure*. When collusion is a concern, the optimal design no longer minimizes redundancy—it leverages it as an anti-collusion device.

Let's assume now that $\bar{V}$ is small enough such that only the collusion constraints are binding but not the participation constraint of the monitor. In this case, using Proposition 1.10, the profit function can be rewritten as:

$$\mathbb{E}\left(\Pi(\mu, \mathbf{W}, \mathbf{S})\right) = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{2}\mu\frac{(1-(1-p(\mu))^2)}{(1-p(\mu))^2}. \tag{1.6}$$

The structure of this profit function is exactly as in the profit function when one monitor is hired, in equation (1.5) but with the incremental probability of two monitors. The first term, is the profits from no monitoring. The second term is a constant times the span of control and the odds of the probability of detection. Notice that, in this case, $(1-(1-p(\mu)^2))$ is the probability of getting at least one informative signal with two monitors. The next proposition finds an implicit expression for the optimal span of control of one monitor.

Proposition 1.11. *Assume that, for any $\gamma \in (0, \mu)$, (PC_M) is not binding. Hence, from Proposition 1.10, $\mu = \gamma$. In that case, the span of control that maximizes profits is unique and given by*

$$\mu = \min\left\{1, \frac{(1-p_0e^{-\lambda\mu}) + (1-p_0e^{-\lambda\mu})^2}{2\lambda}\right\}.$$

When the monitor participation constraint (PC_M) is slack for any $\gamma \in (0, \mu)$, Proposition 1.10 implies full overlap $\gamma = \mu$. The principal's objective in this case collapses to an "enforcement return" in the single span μ, and the first–order condition reduces to the fixed point given by the previous proposition.

The proposition isolates a regime in which the overlap is pinned down by slack PC_M, so the only remaining choice is *how large* the span should be. The fixed-point form makes comparative statics and calibration transparent, and it provides a numerically stable recipe to compute μ^* in structural or quantitative exercises. The next proposition gives those comparative statics.

Proposition 1.12 (Comparative statics of the optimal span)**.** *Suppose $\mu^* \in (0, 1)$ is an interior solution to*

$$\mu = \frac{(1 - p(\mu)) + (1 - p(\mu))^2}{2\lambda}.$$

Then μ^* is continuously differentiable in (p_0, λ) and

$$\frac{\partial \mu^*}{\partial p_0} < 0, \qquad \frac{\partial \mu^*}{\partial \lambda} < 0.$$

A higher p_0 (better baseline detection technology) or a higher λ (faster decay of detection with the span) raise the marginal informativeness of existing coverage relative to the marginal cost of expanding it; the principal optimally *shrinks* the span. The closed-form elasticities above are directly usable for (i) local policy counterfactuals (e.g., how much to adjust μ after a tech upgrade that lifts p_0), (ii) identification (sign and magnitude restrictions on estimated λ, p_0 imply testable shifts in μ^*), and (iii) robustness (as $\lambda \to \infty$ or $p_0 \to 1$, the derivative tends to 0, so μ^* tapers smoothly to smaller spans without knife-edge jumps). Together with Proposition 1.11, this delivers a complete and operational characterization of the optimal span in the two–monitor setting when overlap is endogenously maximal.

These results together with the results in Appendices A.1 and A.2 allow me to illustrate numerically the incentives to hire for the principal that are shown in figure 1.3. The figure maps the optimal monitoring regime over the parameter space $(\lambda, \bar{V})$. The parameter λ summarizes how quickly detection decays with span of control (higher λ means a given supervisor is less effective as her span widens), while $\bar{V}$ collects the stakes of a deviation (and thus the attractiveness

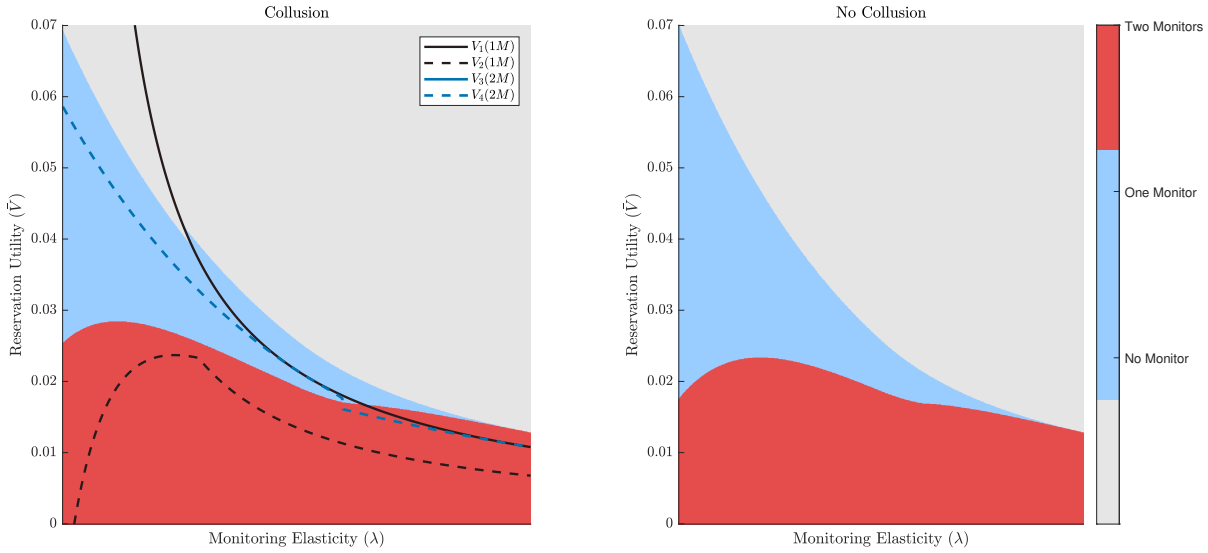


Figure 1.3: Regions for hiring none, one or two monitors in equilibrium. On the left, when collusion is possible. The lines V_i represent the boundaries derived. On the left, where collusion is not possible by assumption. Parameters: $\Delta\theta = 0.75$, $p_0 = 0.75$, $\lambda \in [0, 2]$.

of bribery and the required honesty premia). The dashed boundaries show the no–collusion benchmark; the solid boundaries incorporate collusion. Allowing for collusion *shrinks* the region in which a single supervisor is optimal and *expands* the region in which assigning two independent supervisors to the same worker is optimal. In other words, once bribery is feasible, thresholds shift so that the one–monitor regime is chosen less often, and redundancy becomes optimal across a wider set of $(\lambda, \bar{V})$.

Moving rightward in the figure (larger λ), supervision becomes intrinsically weak. In this range the return to redundancy vanishes and the relevant trade–off collapses to "no monitor" versus "one monitor"; the two–monitor region disappears. For moderate λ and sufficiently high $\bar{V}$, however, the two–monitor regime dominates: overlap is selected precisely where the wage savings from keeping each supervisor honest with a partner exceed the extra headcount cost. The regime boundaries thus encode a simple rule of thumb: collusion risk shifts mass from one– to two–monitor structures, while poor monitoring technology (high λ) pushes the firm toward wider spans with at most a single supervisor per worker.

1.4 Profit losses under naïve firm design

Throughout this section, assume $\bar{V}$ is low enough that only the family of collusion constraints binds. Expected profit under collusion, for a given μ, is

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2 p(\mu)}{2[1-p(\mu)]}\left[\mu + \gamma \cdot \frac{1}{1-p(\mu)}\right], \qquad p(\mu) = p_0 e^{-\lambda\mu}.$$

As stated previously, at the optimal (full overlap, $\gamma = \mu$) this simplifies to

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2}{2}\mu\frac{1-(1-p(\mu))^2}{(1-p(\mu))^2}.$$

Moreover, μ is chosen by the fixed point

$$\mu = \min\left\{1, \frac{(1-p_0e^{-\lambda\mu}) + (1-p_0e^{-\lambda\mu})^2}{2\lambda}\right\}.$$

We analyze two forms of naïveté. In the first, the principal introduces a two-monitor structure but fails to adjust contracts—she reuses the single-monitor, collusion-free contract. In the second, the principal sets the correct (collusion-free) contract but misuses organizational design, treating overlap as wasteful and minimizing it. We compare profits under both behaviors to highlight how contract design and overlap interact.

1.4.1 Profits under naïve contracts

When two monitors are hired, expected profit decomposes by coverage:

$$\mathbb{E}[\Pi(\Omega, \mathbf{W}, \mathbf{S})] = (1 - 2\mu + \gamma)\Pi_{\varnothing M} + 2(\mu - \gamma)\Pi_{1M} + \gamma\Pi_{2M},$$

where $\Pi_{\varnothing M}$, Π_{1M}, and Π_{2M} denote profits in unmonitored, singly monitored, and doubly monitored regions, respectively. The principal offers the same contract everywhere—the one that solves the one-monitor problem when only the collusion constraint binds—so we compute Π_{1M} and Π_{2M} for that contract.

Singly covered region.

$$\Pi_{1M} = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{4}\frac{p(\mu)}{1-p(\mu)}, \qquad p(\mu) = p_0 e^{-\lambda\mu}. \tag{1.7}$$

Doubly covered region. Detection succeeds with probability $1-(1-p(\mu))^2$. Reusing the one-monitor contract yields

$$\Pi_{2M} = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{4}\frac{p^{(2)}(\mu)}{1-p^{(2)}(\mu)}, \qquad p^{(2)}(\mu) = 1-(1-p_0 e^{-\lambda\mu})^2. \tag{1.8}$$

Substituting (1.7)–(1.8) into the decomposition gives

$$\mathbb{E}[\Pi(\Omega,\mathbf{W},\mathbf{S})] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{2}(\mu-\gamma)\frac{p(\mu)}{1-p(\mu)} + \frac{\Delta\theta^2}{4}\gamma\frac{1-(1-p(\mu))^2}{(1-p(\mu))^2}.$$

Regrouping terms,

$$\mathbb{E}[\Pi(\Omega,\mathbf{W},\mathbf{S})] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{2}\frac{p(\mu)}{1-p(\mu)}\left(\mu + \frac{p(\mu)}{2[1-p(\mu)]}\gamma\right).$$

Relative to the benchmark, the coefficient on μ is unchanged, while the overlap term is scaled by $\frac{p(\mu)}{2}$. Thus, using the wrong contract attenuates (but does not reverse) the value of overlap. Since the objective is strictly increasing in γ, the optimal structure under naïve contracts still features full overlap ($\gamma = \mu$), and the profit simplifies to

$$\mathbb{E}[\Pi(\Omega,\mathbf{W},\mathbf{S})] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{4}\mu\frac{1-(1-p(\mu))^2}{(1-p(\mu))^2}. \tag{1.9}$$

This coincides with the benchmark profit up to a factor of $1/2$ on the monitored part, i.e., naïve contracts effectively halve the gains from monitoring.

Proposition 1.13 (Naïve-contracts optimum under single monitoring)**.** *Under naïve contracts with full overlap, any interior optimizer $\mu^* > 0$ satisfies*

$$\mu^* = \frac{\left(1-p_0 e^{-\lambda\mu^*}\right)\left(2-p_0 e^{-\lambda\mu^*}\right)}{2\lambda}. \tag{1.10}$$

Proof. With full overlap, write

$$\mathbb{E}[\Pi(\Omega, \mathbf{W}, \mathbf{S})] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{4}\mu R\Big(p(\mu)\Big), \quad R(p) := \frac{p(2-p)}{(1-p)^2}, \quad p(\mu) = p_0 e^{-\lambda\mu}. \tag{1.11}$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\Big[\mu R\Big(p(\mu)\Big)\Big] = R\Big(p(\mu)\Big) - \lambda\mu p(\mu) R'\Big(p(\mu)\Big), \quad R'(p) = \frac{2}{(1-p)^3}.$$

Setting the derivative to zero and simplifying yields (1.10). □

For sufficiently small λ, the interior solution ceases to exist; for ease of comparison, we focus on parameter values with an interior solution. Since (1.9) coincides with the benchmark objective up to a constant factor, the maximizing μ is the same as in the benchmark. Substituting μ^* into (1.9) gives

$$\mathbb{E}[\Pi(\Omega, \mathbf{W}, \mathbf{S})] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{8\lambda}\frac{p_0 e^{-\lambda\mu^*}\left(2 - p_0 e^{-\lambda\mu^*}\right)^2}{1 - p_0 e^{-\lambda\mu^*}},$$

where μ^* is defined implicitly above.

1.4.2 Profits under naïve firm design

Now consider a different form of naïveté. The principal correctly implements the collusion-free contract but misperceives the role of overlap, viewing it as wasteful; hence she adopts a non-collusive monitoring structure. The resulting expected profit is

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2 p(\mu)}{2[1-p(\mu)]}\left[\mu + \gamma \cdot \frac{1}{1-p(\mu)}\right], \qquad \gamma = \max\{0, 2\mu - 1\}.$$

Proposition 1.14 (Naïve single-monitor design: interior optimum and value)**.** *If $\gamma = 0$, expected profits are*

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2}{2}\frac{\mu p(\mu)}{1-p(\mu)}.$$

An interior optimizer $\mu^* > 0$ (if it exists) satisfies

$$\mu^* = \frac{1 - p(\mu^*)}{\lambda}. \tag{1.12}$$

Proof. Let $f(\mu) := \mu \dfrac{p(\mu)}{1 - p(\mu)}$ with $p'(\mu) = -\lambda p(\mu)$. Then

$$f'(\mu) = \frac{p(\mu)}{1 - p(\mu)} - \mu \frac{\lambda p(\mu)}{(1 - p(\mu))^2}.$$

Setting $f'(\mu^*) = 0$ yields $\mu^* = \dfrac{1 - p(\mu^*)}{\lambda}$. □

Evaluating at μ^* gives

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2}{2\lambda} \frac{p(\mu^*)}{1 - p(\mu^*)}. \tag{1.13}$$

For sufficiently small λ, the interior solution fails and the optimum occurs at the boundary $\mu^* = \frac{1}{2}$. Below we focus on parameter values for which the interior maximum exists.

Assume now $\gamma = 2\mu - 1$ (i.e., $\mu > \frac{1}{2}$). Then

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2 p(\mu)}{2[1 - p(\mu)]^2} \Big[3\mu - 1 - \mu p(\mu)\Big].$$

Proposition 1.15 (Optimal span μ^*)**.** *Any interior maximizer $\mu^* \in \left(\frac{1}{2}, 1\right)$ satisfies*

$$\mu^* = \frac{(1 - p(\mu^*))(3 - p(\mu^*)) + \lambda(1 + p(\mu^*))}{\lambda(3 + p(\mu^*))}, \qquad p(\mu^*) = p_0 e^{-\lambda\mu^*}.$$

If the right-hand side lies outside $\left[\frac{1}{2}, 1\right]$, the optimum occurs at the boundary $\mu^* \in \{\frac{1}{2}, 1\}$.

Proof. Differentiating the objective and simplifying (using $p'(\mu) = -\lambda p(\mu)$) yields the stated fixed point; details are as in the text. □

Substituting μ^* into the profit function,

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2}{2}\frac{p_0 e^{-\lambda\mu^*}}{(1-p_0e^{-\lambda\mu^*})^2}\Bigg[3\frac{(1-p_0e^{-\lambda\mu^*})(3-p_0e^{-\lambda\mu^*})+\lambda(1+p_0e^{-\lambda\mu^*})}{\lambda(3+p_0e^{-\lambda\mu^*})} - 1 - \frac{(1-p_0e^{-\lambda\mu^*})(3-p_0e^{-\lambda\mu^*})+\lambda(1+p_0e^{-\lambda\mu^*})}{\lambda(3+p_0e^{-\lambda\mu^*})}p_0e^{-\lambda\mu^*}\Bigg],$$

which simplifies to

$$\mathbb{E}[\Pi(\mu, W, S)] = \Pi^{\varnothing M} + \frac{\Delta\theta^2}{2\lambda}\frac{p_0e^{-\lambda\mu^*}\left[\lambda p_0e^{-\lambda\mu^*} + (p_0e^{-\lambda\mu^*}-3)^2\right]}{(1-p_0e^{-\lambda\mu^*})(3+p_0e^{-\lambda\mu^*})}.$$

Quantifying losses

Benchmark versus naïve design. Under the benchmark, the principal jointly optimizes incentives and organization. She sets full overlap ($\gamma = \mu$), maximizing enforcement depth: the probability at least one monitor detects wrongdoing is $1-(1-p(\mu))^2$, and profits increase in this term. The objective is

$$\Pi_{\varnothing M} + \frac{\Delta\theta^2}{2}\mu\frac{p(\mu)[2-p(\mu)]}{(1-p(\mu))^2},$$

and at an interior optimum μ^*,

$$\Pi_{\varnothing M} + \frac{\Delta\theta^2}{4\lambda}\frac{p(\mu^*)[2-p(\mu^*)]^2}{1-p(\mu^*)}.$$

Naïve design without overlap ($\gamma = 0$). If the principal forbids overlap, she allocates breadth instead of depth:

$$\Pi_{\varnothing M} + \frac{\Delta\theta^2}{2}\mu\frac{p(\mu)}{1-p(\mu)}, \qquad \mu^* = \frac{1-p(\mu^*)}{\lambda}(\mu \le \tfrac{1}{2}),$$

yielding

$$\Pi_{\varnothing M} + \frac{\Delta\theta^2}{2\lambda}\frac{p(\mu^*)}{1-p(\mu^*)}.$$

This discards the quadratic detection gains from redundancy; enforcement is broad but shallow, leaving each relationship vulnerable to collusion.

Allowing overlap only through $\gamma = 2\mu - 1$ improves outcomes but remains

inferior to full overlap:

$$\Pi_{\varnothing M} + \frac{\Delta\theta^2}{2}\frac{p(\mu)}{(1-p(\mu))^2}[3\mu - 1 - \mu p(\mu)],$$

with

$$\mu^* = \frac{(1-p(\mu^*))(3-p(\mu^*)) + \lambda(1+p(\mu^*))}{\lambda(3+p(\mu^*))},$$
$$\mathbb{E}[\Pi] = \Pi_{\varnothing M} + \frac{\Delta\theta^2}{2\lambda}\frac{p(\mu^*)\Big(\lambda p(\mu^*) + (p(\mu^*)-3)^2\Big)}{(1-p(\mu^*))(3+p(\mu^*))}.$$

This regime partially recovers the depth effect but cannot generally attain full overlap at the optimum.

Profits are strictly increasing in overlap: redundancy raises enforcement credibility, the scarce resource under collusion. The benchmark dominates because it tailors both incentives and architecture to maximize redundancy. Forbidding overlap ($\gamma = 0$) reallocates capacity to breadth, which does not cure bribability and sacrifices large detection gains. Constraining overlap ($\gamma = 2\mu - 1$) is better than no overlap but still falls short of full overlap. The loss is organizational: misallocating monitors away from depth forfeits the non-linear increase in detection—and the associated reduction in collusion rents—that full overlap delivers.

Under both forms of naïveté, the principal suffers profit losses that stem from the same underlying mechanism: a failure to fully exploit the disciplinary power of monitoring. When contracts are naïve, monitors are not properly incentivized, so even with full overlap the firm captures only half of the potential enforcement rents. When the organization is naïve, monitoring is misallocated across workers, weakening the depth of supervision and leaving collusion more effective. In both cases, profits fall because the credibility of enforcement—the key determinant of firm discipline—is diluted, either through insufficient incentives or through an inefficient monitoring architecture.

1.5 Conclusion

This paper has analyzed how firms should design monitoring hierarchies when monitors themselves may collude with subordinates. Standard models treat

oversight as a neutral enforcement technology; here, the possibility of collusion makes organizational architecture part of the incentive system. The central result is that redundancy—often viewed as costly duplication—can be an efficient way to strengthen enforcement: by overlapping monitors, principals make concealment harder, reduce the compensation needed to sustain honesty, and in some environments improve discipline at lower overall cost.

The framework delivers three implications. First, scope for collusion reshapes spans of control: narrower spans and selective overlap become optimal even when supervision technologies are otherwise efficient. Second, overlapping supervision creates new contracting instruments, since cross-monitoring contracts deter bribery at lower cost than isolated oversight. Third, redundancy helps explain why firms in high-risk environments sustain costly "four-eyes" rules or dual approvals that appear inefficient in standard models.

These insights connect to broader debates on firm organization. Work on flattening hierarchies shows that leaner structures can improve efficiency in production, but without parallel improvements in oversight they may increase vulnerability to collusion (Guadalupe and Wulf, 2010). Likewise, the model provides a new perspective on the "missing middle" in developing economies (Hsieh and Olken, 2014; Macchiavello, 2012; Tybout, 2000): if expanding scale requires costly, redundant monitoring to sustain honest enforcement, small-sized firms might find it difficult to expand.

Several avenues for research remain open. One direction is to extend the model beyond two monitors, allowing the principal to choose both the number of supervisors and the depth of monitoring layers. This would clarify the classic trade-off between flat and tall organizations: wide structures economize on oversight costs but heighten collusion risk, while deeper hierarchies provide redundancy at the expense of added layers. A second direction is to study sequential rather than simultaneous monitoring. Sequential oversight may raise detection by concentrating scrutiny on unresolved cases, but it can also introduce delays and opportunities for evidence to be concealed, creating a new speed–credibility trade-off. A third direction is empirical: linking the model to data on auditor assignment, procurement oversight, or internal controls would help quantify the costs of collusion and the returns to organizational remedies. More broadly, the analysis highlights that hierarchies are not only about efficiency or knowledge

allocation: they are also instruments of enforcement, designed to remain credible even when the enforcers themselves may be tempted to collude.

Chapter 2

Feed for Good? On the Effects of Personalization Algorithms in Social Media Platforms

Miguel Risco[1] and Manuel Lleonart-Anguix[2]

Abstract

We study how engagement-maximizing personalization algorithms shape exposure and learning on a social media platform. Users observe private signals, post, and consume ranked feeds; utility reflects sincerity, conformity, and accuracy. Users are heterogeneous: rational users choose engagement; naive users continue scrolling with a payoff-dependent probability. Truth-telling is an equilibrium—unique for rationals—so platforms

[1]IGIER, Università Bocconi. miguel.riscobermejo@unibocconi.it. Risco acknowledges financial support from MUR PRIN 2022, Prot. 2022S5RC7R, "Finanziamento dell'Unione Europea—NextGenerationEU", by the German Research Foundation (DFG) through CRC TR 224 (Project B05), and from the European Research Council (ERC) under the EU Horizon 2020 programme (grant 949465).

[2]Universitat Autònoma de Barcelona and Barcelona School of Economics. manuel.lleonart@bse.eu. We thank Sven Rady, Pau Milán, Francesc Dilmé, Alexander Frug, Carl-Christian Groh, Francesco Decarolis, Rafael Jiménez-Durán, Martino Banchio, Marco Ottaviani, Justus Preusser, Muxin Li, David Jiménez-Gómez, Daniel Krähmer, Philipp Strack, Marc Bourreau, Mikhail Drugov, Manuel Mueller-Frank, Robin Ng, Raquel Lorenzo, Malachy Gavan, Francisco Poggi and Fernando Payro and participants at the BSE Summer Forum in Digital Platforms, EAYE 2024, Paris Digital Economics Conference, MaCCI Annual 2024, CRC TR 224 retreat, 6th Workshop on The Economics of Digitalization, EDP Jamboree, 2nd ECONtribute YEP Workshop, CoED, JEI, and seminars at Bonn, Bocconi, UPF, TSE, UAB, UV, UA, EUI and UB for their helpful comments. Lleonart-Anguix acknowledges financial support of the Spanish *Ministerio de Ciencia e Innovación* through grant PRE2021-099026.

influence behavior only through feed ranking. In general, more similar users are prioritized—especially in naive users' feeds—producing echo-chamber exposure. At scale, naives' learning stalls. Alternative algorithms like reverse-chronological and balanced-insertion restore diversity, though they may underperform overall. Finally, personalization can cause network effects to fail for naive users.

Keywords: personalization algorithms, engagement, social learning, network effects.

JEL Codes: D43, D85, L15, L86.

2.1 Introduction

While entertainment may be the primary reason users open TikTok or Instagram, social media has become a pivotal source of news and opinion formation. More than half of U.S. adults now report getting news from social platforms, and the shares have risen steeply in recent years: between 2020 and 2024, the fraction of adults regularly accessing news on TikTok grew from 3% to 17%, and on Instagram from 11% to 20%.[3] This scale and centrality have sharpened concerns that engagement-maximizing algorithms amplify polarization, misinformation, and echo chambers. The public debate, energized since the 2016 U.S. election, initially focused on the spread of false content and platform incentives (Allcott and Gentzkow, 2017; Silverman, 2016) and has since broadened with empirical evidence on the welfare costs of social media use (Allcott et al., 2022; Braghieri et al., 2022; Bursztyn et al., 2023) and on algorithmic forces that trap users in ideologically narrow information environments (Levy, 2021). Internal platform documents further indicate awareness of such harms, especially for vulnerable groups (Horwitz et al., 2021). At the same time, social media affords sizable benefits—from access to timely information and social connection to professional networking—underscoring that policy must navigate genuine trade-offs (Allcott et al., 2020; Armona, 2023).

The *feed* (the ordered list of posts a user reads when scrolling) is the critical design lever. Before roughly 2009, most platforms presented posts in reverse-chronological order.[4] Today, the feed is a personalized, ordered list optimized to maximize *engagement* because platform revenue is advertising-based.[5] Empirical evidence links algorithmic personalization to higher time-on-platform and more addictive usage (Guess et al., 2023). These observations motivate a positive, micro-founded analysis of how an engagement-maximizing platform designs the feed, how users communicate and learn within it, and which regulatory tools improve outcomes.

We build a tractable model of communication and learning through personal-

[3] Pew Research Center, Social Media and News Fact Sheet, data available here.

[4] Facebook began introducing personalized feeds in 2009; Twitter (now 𝕏) and Instagram transitioned in 2015–2016; TikTok launched with a curated feed.

[5] As an example, see here and Kamath et al. (2014) on RealGraph predictor, the base of Twitter's recommendation algorithm.

ized feeds on a social media platform. So the key aspect is not only what the user reads but also when. Users receive private signals about an unknown state, post a message, and consume a feed whose order is strategically designed by the platform. Users' payoffs have two components. The first, within-the-platform utility, includes (i) direct gratification from reading, (ii) a taste for sincerity (reporting a message close to one's private signal), and (iii) disutility from disagreement (conformity costs).[6] The second, action utility, rewards taking a decision close to the true state after reading the feed (e.g., whether to vaccinate when facing contested information about vaccine efficacy).[7] Engagement is the number of posts read. We allow behavioral heterogeneity: *rational* users choose how far to scroll, whereas *naive* users continue according to a smooth, non-decreasing function of their instantaneous within-the-platform utility, consistent with habit formation and present bias in digital consumption (Allcott et al., 2022; Hoong, 2021).

Methodologically, we impose improper priors so that each user treats her own signal as the local anchor. Our first result pins down the channel through which platforms exercise influence: truth-telling is an equilibrium of the messaging game under any feed design. Among rational users the equilibrium is unique, and for naive users it is the only equilibrium robust to any continuous, non-decreasing continuation rule. Hence, the platform cannot manipulate beliefs by inducing misreporting; it operates solely through *who appears and when* in the user's feed. This separability delivers closed-form posteriors and sharp comparative statics for ranking and engagement.

We leverage this benchmark to deliver three main findings. The first one is that engagement-optimal feeds generate echo chambers and, for naive users, unravel the classic *wisdom-of-the-crowds* result (Golub and Jackson, 2010). For this kind of users, the structure of the algorithm that the platform chooses in equilibrium (the platform-optimal algorithm) is *similars first*: it orders others by expected similarity to the focal (naive) user. The resulting feed is a perfect

[6]Conformity is a core force in social influence, defined as matching attitudes and behaviors to group norms (Cialdini and Goldstein, 2004). We model it as a reduced-form preference, consistent with rational accounts of social conformity and herding (Bernheim, 1994; Chamley, 2004). See also Mosleh et al. (2021) for evidence on shared identity and conformity in online diffusion.

[7]For example, Loomba et al. (2021) document that exposure to misinformation reduced COVID-19 vaccine acceptance in the U.S. and U.K. by about six percentage points.

echo chamber (Pariser, 2011). Because the implemented length is finite under any admissible continuation rule, feed positions concentrate on close copies of the focal user as the platform's selection set grows; thus, the posterior variance fails to fall, and learning vanishes as platform size grows large. Echo chambers thus arise from engagement incentives even when messages are truthful. For rational users, we show in a simplified two-type environment that the platform optimally *front-loads* same-type content to reduce early disagreement and implements a longer path that introduces cross-cutting signals later; among feeds that maximize implemented engagement, there is a front-loaded representative. Echo chambers are therefore not just a byproduct of naiveté—though they are more severe for naive users.

The second main contribution consists of studying policy-relevant alternatives to personalization: the *reverse-chronological algorithm* and a minimal corrective, the *breaking-echo-chambers* algorithm. The reverse-chronological algorithm—the non-profiling benchmark, reintroduced by the Digital Services Act, (DSA)—restores exposure diversity by randomizing order. For a fixed length, it converges to a wisdom-of-the-crowds variance benchmark, but it reduces within-the-platform utility when early disagreement is most costly and therefore shortens naive engagement. We derive the naive welfare comparison relative to the platform-optimal algorithm and the rational user's stopping condition under the reverse-chronological algorithm. We then analyze a minimal tweak that inserts a maximally opposite account at the top of the platform-optimal algorithm for naives' feed and leaves the rest intact, the breaking-echo-chambers algorithm. The insertion raises early disagreement (and can shorten engagement) but lowers posterior variance sharply; in large platforms, this modification beats the platform-optimal algorithm whenever users place sufficiently high weight on learning.

The third main contribution is to show that personalization reshapes network effects and hence the scope for competition. Personalization allows finer matching as a platform grows. For rational users, implemented feeds increasingly front-load similarity of low conformity cost while still delivering enough later diversity, so expected utility rises with platform size. For naive users, the same force can reduce welfare: larger selection sets let the platform fill the top of the feed with close copies, eroding the informational value of additional posts; network effects need not arise. These forces bear directly on

policy. Non-profiling defaults (reverse-chronological) temper echo chambers but rarely dominate engagement-optimal designs in overall welfare. A complementary market-design lever is *horizontal interoperability*: by sharing network effects across platforms, rivalry would shift toward the non-interoperable dimension—the ranking algorithm itself—disciplining engagement-only designs and moving equilibrium feeds toward the utilitarian benchmark.

2.1.1 Related literature

Our paper connects the economics of recommendation and curation (e.g., Aridor et al., 2024) with social learning under correlated signals (Golub and Jackson, 2010) and conformity motives (Bernheim, 1994; Cialdini and Goldstein, 2004), but shifts the focus to the *timing* of exposure in personalized feeds. We also relate to theory on information sharing and learning in networks and misspecification (Bohren and Hauser, 2021; Galeotti et al., 2013; Levy and Razin, 2012; Mossel et al., 2015). Unlike models in which platforms amplify misinformation to stimulate sharing, our users truthfully communicate and value accuracy; echo chambers arise nonetheless because engagement incentives lead the platform to sequence like-minded content early, raising continuation among naive users and, in a two-type rational benchmark, lowering early disagreement at the cost of early diversity. This temporal mechanism complements media-economics accounts of distortion through content selection or slant (Abreu and Jeon, 2019; Ellman and Germano, 2009; Gentzkow and Shapiro, 2010a; Kranton and McAdams, 2022; Reuter and Zitzewitz, 2006): here the distortion is primarily in *when* rather than *what* users see. Closest to us, Acemoglu et al. (2024) study sequential sharing with reputational concerns and show that platforms benefit from homophily because it lowers scrutiny of misinformation; by contrast, we analyze simultaneous truthful posting, model learning explicitly via posterior precision, and derive echo chambers without misinformation or reputational motives. We are also related to models where profit-motivated platforms harm user welfare by complementing time spent or manipulating information flow (e.g., Beknazar-Yuzbashev et al., 2024; Mueller-Frank et al., 2022), to coordination and misinformation dynamics (Chang and Vong, 2025), and to rumor propagation and debunking in networks (Merlino et al., 2023); we model continuation through a bounded-hazard, addiction-driven

process and separate within-the-platform utility from action utility to quantify learning losses.

Our predictions match empirical work showing reverse-chronological algorithms reduces usage while broadening exposure and weakening ideological clustering (Guess et al., 2023), consistent with evidence on personalized feeds and engagement (Gauthier et al., 2025; Kitchens et al., 2020) and with debates on "filter bubbles" (Holtz et al., 2020; Levy, 2021; Pariser, 2011; Sunstein, 2017). Finally, our policy counterfactuals relate to proposals that curb echo chambers through algorithmic changes instead of content moderation (as studied in Jackson et al. (2022) or Guriev et al. (2023)) and to market-level discussions of interoperability and contestability under the EU's Digital Markets Act (DMA) and DSA (Bourreau and Kraemer, 2023; Bourreau et al., 2023; Kades and Scott Morton, 2020).

The remainder of the paper proceeds as follows. Section 2.2 introduces the environment. Section 2.3 establishes truth-telling and characterizes platform-optimal personalization, showing when it generates echo-chamber exposure for naive and rational users. Section 2.4 quantifies the effects of personalization on learning and Section 2.5 evaluates two policy-relevant counterfactuals. Section 2.6 analyzes network effects and discusses the policy implications. Section 2.7 concludes.

2.2 A model of communication and learning through personalized feeds

We study a platform that curates personalized feeds and users who communicate and learn from them. Here we lay out the model assumptions.

Players. There is a social media platform p and a set, $\mathcal{U}$, of n users indexed i that visit it. We assume a complete network. Expectations with respect to the platform's information are denoted $\mathbb{E}_p[.]$; user i's expectations are $\mathbb{E}_i[.]$.

Information structure. There is a state of the world, θ, for which we assume improper priors.[8] Each user i observes a private signal

$$\theta_i = \theta + \varepsilon_i, \qquad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)' \sim \mathcal{N}(0, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma} = (\sigma_{ij})$ is symmetric positive definite and common knowledge. Under improper priors, $\theta|\theta_i \sim \mathcal{N}(\theta_i, \sigma_{ii})$, and $\theta_j|\theta_i \sim \mathcal{N}(\theta_i, \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij})$ for all $j \neq i$ according to Lemma B.1.

User choices. Each user i (i) posts a scalar message $m_i \in \mathbb{R}$, (ii) observes an endogenously determined number of feed posts $e_i \in \{1, \dots, n\}$, and (iii) after reading such posts chooses an action $a_i \in \mathbb{R}$.

Platform choices. The platform designs an algorithm consisting of an assignment that, given a pair of users i, j, tells which position user j's message occupies in user i's feed. Hence, an *algorithm* $\mathscr{F} = (\mathscr{F}_i)_{i \in \mathcal{U}}$ is a collection of bijections $\mathscr{F}_i \in \text{Bij}\left(\{1, \dots, n\}, \mathcal{U}\right)$. We interpret $\mathscr{F}_i(r)$ as the r-th position in i's feed. Given engagement e_i, the realized feed is

$$\mathscr{F}_i^{e_i} = \{\mathscr{F}_i(1), \dots, \mathscr{F}_i(e_i)\}.$$

As user i knows her own signal, we will denote $\mathscr{F}_i(1) = i$ for all user i.

Users preferences. Users derive utility from two streams: *within-the-platform utility* and *action utility*. Within-the-platform (instant) utility has three components: (i) a positive linear payoff coming from reading messages; (ii) *sincerity*: users dislike deviating from their own signals,[9] and (iii) *conformity*: disagreeing with others' opinions is taxing. Formally,

$$u_i(e_i, m_i, m_{-i}, \mathscr{F}_i, \theta_i) = \alpha e_i - \beta \overbrace{(\theta_i - m_i)^2}^{\text{Sincerity}} - (1-\beta) \overbrace{\sum_{j \in \mathscr{F}_i^{e_i}} (m_i - m_j)^2}^{\text{Conformity}}, \tag{2.1}$$

where $\alpha > 0$ and $\beta \in (0, 1)$. Total realized utility is the weighted sum of within-the-platform utility and action utility, which we define as the squared difference of the action a_i to the state of the world:

[8] Equivalently, one can take a normal (proper) prior with variance τ^2 and let $\tau^2 \to \infty$.

[9] Due to improper priors, $\mathbb{E}\left[(m_i - \theta)^2 | \theta_i\right] = (m_i - \theta_i)^2 + \sigma_{ii}$, so penalizing deviations from θ or from θ_i is equivalent up to an additive constant.

$$U_i(e_i, m_i, m_{-i}, a_i, \mathscr{F}_i, \theta_i, \theta) = \lambda u_i(e_i, m_i, m_{-i}, \mathscr{F}_i, \theta_i) - (1-\lambda) \overbrace{(a_i - \theta)^2}^{\text{Action utility}} . \quad (2.2)$$

User heterogeneity. Users differ in how they choose e_i.

- A fraction of users are *rational*, and choose (m_i, e_i, a_i) to maximize $\mathbb{E}_i[U_i]$.
- The remaining fraction of users are *naive*, and follow a myopic, addiction-driven sequential stopping rule: after reading k messages, a naive user i reads the next message with probability $g(u_i(k, m_i, m_{-i}, \mathscr{F}_i, \theta_i))$, where $g : \mathbb{R} \to [0,1]$ is some $\mathcal{C}^1$ and non-decreasing function such that $0 < \underline{g} < g(x) < \overline{g} < 1$ for some $\underline{g}, \overline{g}$ and all $x \in \mathbb{R}$, and exits otherwise.

Platform's revenues and information. The platform monetizes engagement.[10] Given an increasing and positive function $\pi(.)$, the platform chooses $\mathscr{F}$ to maximize expected revenue

$$\mathbb{E}_p \left[\sum_{i=1}^{n} \pi(e_i) \right].$$

For simplicity, we assume π to be linear with slope 1, so effectively the platform maximizes $\mathbb{E}_p \left[\sum_{i=1}^{n} e_i \right]$. The platform observes user types (perfect personalization), knows $g(.)$ and $\mathbf{\Sigma}$, but not θ nor $\{\theta_i\}_{i=1}^n$.

Timing. The game of communication and learning through personalized feeds described above consists of the following sequence of events:

1. The platform publicly commits to an algorithm $\mathscr{F}$.
2. Signals are realized; each user observes θ_i.
3. Each user i posts a message $m_i \in \mathbb{R}$.

[10] We assume a monopolist platform with all users aboard: arguably, most social media platforms in today's landscape are monopolists of their fields. For example, the Bundeskartellamt (the German competition protection authority) states in its case against Facebook (B6-22/16, "Facebook", p. 6): "The facts that competitors are exiting the market and there is a downward trend in the user-based market shares of remaining competitors indicate a market tipping process that will result in Facebook becoming a monopolist." (Franck and Peitz, 2023). See Garcia and Li (2024) for a discussion of monopoly platform strategy after market expansion.

4. Rational users choose e_i; naive users draw e_i from the sequential process above. Given $\mathscr{F}_i^{e_i}$, all users choose a_i.

5. The state of the world is revealed and payoffs are realized.

Equilibrium. The equilibrium concept is a version of Bayesian-Nash Equilibrium (BNE) that accounts for the behavior of naive users. A profile of strategies for users and the platform is an equilibrium if:

- For each naive user i, (m_i, a_i) maximizes $\mathbb{E}_i[U_i]$ given the mechanical process for e_i and correct beliefs about others.
- For each rational user j, (m_j, e_j, a_j) maximizes $\mathbb{E}_j[U_j]$ given correct beliefs about others.
- The platform's algorithm $\mathscr{F}$ maximizes $\mathbb{E}_p[\sum_i \pi(e_i)]$ given users' induced strategies and the naive users' engagement process.

2.2.1 Interpretation of the model

We conclude this section by discussing the model's assumptions.

Improper priors. Users' prior distribution is uniform along $\mathbb{R}$. This is a tractable way to capture that users treat their own signal as locally "central".[11] Formally, $\mathbb{E}[\theta|\theta_i] = \theta_i$, so a user cannot diagnose whether her realization is extreme (Greene, 2004; Ross et al., 1977). We adopt this assumption for tractability. Under normal priors, we can only determine the users' optimal linear messaging strategies for naive users, but we cannot derive an explicit expression for the platform-optimal algorithm.

User heterogeneity and engagement. The empirical evidence (Allcott et al., 2022; Hoong, 2021) shows that a non-trivial share of users display self-control failures (present-biased, time inconsistent "keep-scrolling") on social media, deviating from the standard forward-looking benchmark. The engagement process we specify for naive users is consistent with these findings: continuation is addiction-driven and myopic; it increases with the instantaneous utility the user derives

[11] For a general discussion of improper priors, see Hartigan (1983).

from the last post read, resembling the effect of a dopamine kick. The fact that naives' engagement depends only on instantaneous utility formalizes an extreme present bias: while scrolling, the user heavily discounts the longer-run benefits of improved beliefs and action quality, placing near-zero weight on learning relative to immediate reward (Guriev et al., 2023). The boundedness $0 < \underline{g} \leq g(\cdot) \leq \overline{g} < 1$ matches the empirical fact that even very engaging (or very poor) content does not lead to infinite (or zero) scrolling time.

Platform conditioning on similarities and not on messages. This aligns with practice: 𝕏's RealGraph predicts future interactions from historical affinities, TikTok reports that user interactions are the strongest *For You* signals, and Meta's Feed predicts per-story engagement using prior interaction features under tight latency. Fresh posts, by contrast, face a cold-start problem, making real-time utility estimation noisy.

Perfect personalization. Platforms observe extremely rich traces of user behavior and can rapidly learn engagement propensities; we assume for this paper that personalization is perfect and the platform knows which type the user is.

Platform's profits as a function of total engagement. Ads monetize exposure; more engagement yields more impressions. Modeling revenue as $\sum_i \pi(e_i)$ captures this directly. Allowing π to vary by type (e.g., naive users monetizing differently) leaves the analysis intact and only rescales the platform's objective.

2.3 Truth-telling and the existence of echo chambers

This section establishes our first main result. For any algorithm $\mathscr{F}$ chosen by the platform, reporting the private signal—truth-telling—is an equilibrium. For rational users, this equilibrium is unique. For naives, in turn, it is the only equilibrium that arises for any specification of the engagement function g. We select truth-telling as the equilibrium of interest and stick to it from now on. As a consequence, once the platform commits to $\mathscr{F}$, every user i—naive or rational—behaves identically in equilibrium with respect to messaging: $m_i = \theta_i$. The platform can therefore influence behavior only through the composition and

order of each user's feed, not through others' messages.

Proposition 2.1 (Truth-telling). *For any algorithm $\mathscr{F}$, the profile $m_i^* = \theta_i$ for all $i \in \mathcal{U}$ is a BNE of the messaging game. Moreover, if the population only contains rational users, truth-telling is the unique equilibrium.*

Proof. See Appendix B.1. □

Since users have improper priors, they lack an external anchor for their beliefs. Conditional on their own signal, they expect every other user's signal to coincide with theirs, $\mathbb{E}[\theta_j|\theta_i, \mathscr{F}] = \theta_i$. This implies that, from user i's perspective, the platform's algorithm cannot shift first-order beliefs: when others report their signals truthfully, the expected content aligns with θ_i. In equilibrium, therefore, the best reply to everyone else reporting truthfully is to report truthfully as well.

Lemma B.2 shows that, while some specifications of the continuation rule g can sustain non-truthful equilibria, these hinge on particular functional forms. By contrast, truthful reporting is the only equilibrium that obtains for *any* g satisfying our assumptions (continuous and non-decreasing). In this sense, it is robust to how engagement responds to within-the-platform utility. We therefore restrict attention to truthful reporting throughout the analysis.

For rational users, the equilibrium is unique. The first-order condition defining optimal messages generates a contraction mapping: each user's best response is a convex combination of her own signal and the expected reports of others. Iterating these best-response equations converges to a single fixed point, which coincides with $m_i = \theta_i$ for all i. Any deviation would require shifting expectations that are, by construction, pinned down by the (rational) user's own signal.

As messages are truthful in equilibrium, the platform affects user i's payoffs only through the ordering of posts shown to i, i.e., through her feed $\mathscr{F}_i$. Under perfect personalization, this implies a separability property.

Corollary 2.1 (Separability property of the revenue function). *For the platform, maximizing aggregate expected engagement $\sum_{i=1}^{n} \mathbb{E}_p[e_i]$ is equivalent to maximizing $\mathbb{E}_p[e_i]$ for each user i separately.*

The platform can thus design, for each user type, an algorithm that maximizes that user's expected engagement. We denote by $\mathscr{P}$ the platform-optimal algorithm for naive users and by $\mathcal{P}$ the platform-optimal algorithm for rational users.

Although the feed does not affect messages, it affects actions. Let $\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}$ denote the restriction of $\boldsymbol{\Sigma}$ to the accounts shown to i and $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}$ the corresponding vector of signals.

Proposition 2.2 (Optimal action)**.** *For any algorithm $\mathscr{F}$, naive user i's optimal action after reading e_i messages is*

$$a_i^* = \frac{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}^t_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t} = \mathbb{E}\left[\theta \middle| \boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right],$$

where $\mathbb{1}$ denotes the conformable column vector of ones.

Proof. The action a_i^* minimizes $\mathbb{E}[(a_i-\theta)^2|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}]$. The first-order condition yields the stated conditional expectation; see Lemma B.3. □

2.3.1 The platform-optimal algorithm for naive users

Fix a naive user i. By Proposition 2.1, all messages equal their senders' signals. After $r-1$ messages, user i's within-the-platform utility is

$$\mathbb{E}_p[u_i(r-1,\theta_i,\theta_{-i},\mathscr{F},\theta_i)] = \mathbb{E}_p\left[\alpha(r-1) - (1-\beta)\sum_{j\in\mathscr{F}_i^{r-1}}(\theta_i-\theta_j)^2\right].$$

The expected probability that user i continues to the $(r+1)$-th message is given by $\mathbb{E}_p[g(u_i(r,\theta_i,\theta_{-i},\mathscr{F},\theta_i))]$. Since g is increasing and the conformity term depends only on the identity of the next user in the feed, the platform chooses, for the r-th position and among the not-yet-shown accounts,[12] the user j that minimizes the expected conformity loss. Iterating this choice yields the platform-optimal

[12]We use "user" and "account" interchangeably to denote the entities that post messages appearing in a feed.

algorithm $\mathscr{P}$ for naive user i:

$$\begin{aligned}
\mathscr{P}_i^1 &\in \arg\max_{j\in\mathcal{U}} \left\{-\mathbb{E}_p\left[(\theta_i-\theta_j)^2\right]\right\},\\
\mathscr{P}_i^2 &= \mathscr{P}_i^1 \cup \arg\max_{j\in\mathcal{U}\setminus\mathscr{P}_i^1} \left\{-\mathbb{E}_p\left[(\theta_i-\theta_j)^2\right]\right\},\\
&\vdots\\
\mathscr{P}_i^{e_i} &= \mathscr{P}_i^{e_i-1} \cup \arg\max_{j\in\mathcal{U}\setminus\mathscr{P}_i^{e_i-1}} \left\{-\mathbb{E}_p\left[(\theta_i-\theta_j)^2\right]\right\}. \qquad (2.3)
\end{aligned}$$

Proposition 2.3 (Naive users get echo chambers)**.** *In equilibrium, the platform chooses $\mathscr{P}$ as in Equation* (2.3)*.*

Proof. See Appendix B.1. □

Thus, for each naive user i, the feed ranks others in reverse order of expected conformity loss with i. This is a *perfect* echo chamber in the sense of Pariser (2011). This is a direct consequence of the platform's goal being to minimize the loss on conformity.[13]

2.3.2 The platform-optimal algorithm for rational users

To show that echo chambers are not merely an artifact of the engagement process, we also consider a simplified population of rational users with two types (e.g., *democrats* and *republicans*) whose engagement is strategically chosen; even in this case, the platform may prefer to place same-type users first in a user's feed.

In equilibrium, a rational user chooses an engagement level e_i^* that maximizes her expected utility given the algorithm $\mathscr{F}$,

$$e_i^* \in \arg\max_{e_i\geq 1} \ \mathbb{E}\left[U_i(e_i,\theta_i,\theta_{-i},\mathscr{F})\middle|\theta_i,\mathscr{F}\right].$$

Formally, the platform's problem is to select a feed that implements the highest possible engagement. Let $e_j^*(\mathscr{F}_j) \in \arg\max_{m\in\mathbb{N}} \mathbb{E}[U_j(m;\mathscr{F}_j)]$ be rational user j's best-response engagement under feed $\mathscr{F}_j$, and define the set of implementable

[13] Note that a naive user could, in principle, alter her message to influence engagement and thereby affect how much she learns. However, truthful reporting remains robust to such behavior: any deviation would distort the user's perceived similarity with others without improving expected utility.

engagement levels

$$K_j := \Big\{k \in \mathbb{N} : \exists \mathscr{F}_j \text{ such that } e_j^*(\mathscr{F}_j) = k\Big\}.$$

Proposition 2.4 (Platform-optimal algorithm for rationals)**.** *For each rational user j, the platform chooses a feed $\mathcal{P}_j^{k^*}$ that implements*

$$k^* \in \arg\max_{k \in K_j} k,$$

i.e., the largest engagement level that can be induced by some algorithm.

Proof. Fix any $k \in K_j$ and one implementing feed $\mathscr{F}_j^k$. Since the platform's expected payoff $\mathbb{E}_p[\pi(k)]$ is strictly increasing in k, it selects an algorithm attaining the largest element of K_j. By best response, $\mathbb{E}[U_j(k^*; \mathcal{P}_j^{k^*})] \geq \mathbb{E}[U_j(k; \mathscr{F}_j^k)]$ for all $k \in \mathbb{N}$. □

Characterizing the platform-optimal algorithm for rational users, $\mathcal{P}$, in full generality is analytically infeasible, as it would require closed-form posterior variances under arbitrary correlation matrices. To isolate the core mechanism, we consider a *two-block environment* with two user types—*Democrats (D)* and *Republicans (R)*—that captures the same trade-off.

Assume an *equicorrelation structure* across all users. Consider $\sigma_{ii} = \sigma$. For two users k and k' that are the same type, their correlation is given by: $\rho_{kk'} = x \in (0, 1)$, and $\rho_{kk'} = -x$ if k and k' are of opposite types. Finally, assume that α is small enough that expected marginal conformity costs dominate the direct payoff from consuming additional, even ideologically aligned, content. [14]

A (finite) feed is an ordered list of accounts $\mathscr{F} = (j_1, \ldots, j_n)$. Its prefix of length k is denoted $\mathscr{F}^k = (j_1, \ldots, j_k)$, where $k \leq n$. Let $d(\mathscr{F}, k)$ and $r(\mathscr{F}, k)$ be, respectively, the *counts* of same-type (D) and opposite-type (R) accounts among the first k messages in the feed, so that $d(\mathscr{F}, k) + r(\mathscr{F}, k) = k$. The first position always corresponds to the focal user herself, hence $d(\mathscr{F}, 1) = 1$. The user's expected utility after reading k messages depends only on the *composition* of the feed—through the counts d and r—and not on the specific ordering of the

[14] This assumption simplifies exposition and does not affect the comparative statics discussed below.

accounts.

$$U(\mathscr{F},k)=U(d,r)=\lambda\Big[\alpha(d{+}r){-}(1{-}\beta)2\sigma^2\Big(d(1{-}x){+}r(1{+}x)\Big)\Big]{-}(1{-}\lambda)\,\mathrm{Var}[\theta|d,r]. \tag{2.4}$$

The posterior variance under the two-block structure is order-irrelevant and has the closed form

$$\mathrm{Var}[\theta|d,r]=\sigma^2\frac{(1-x)\Big(1+x(d+r-1)\Big)}{(1-x)(d+r)+4xdr}. \tag{2.5}$$

The platform evaluates an algorithm $\mathscr{F}$ by the engagement it implements, denoted by $e^*(\mathscr{F})=\arg\max_k U(\mathscr{F},k)$ (largest maximizer). Among all algorithms, the platform chooses $\mathcal{P}\in\arg\max_{\mathscr{F}} e^*(\mathscr{F})$, and for reference, the user-optimal algorithm is denoted by $\mathscr{F}^U\in\arg\max_{\mathscr{F}} U(\mathscr{F},e^*(\mathscr{F}))$. Focus on a representative democrat user j. Given a pair (d,r), we can compare the utility induced by adding a democrat to the pool (abusing notation, $U(d+1,r)$) to the utility induced by adding a republican ($U(d,r+1)$):

$$U(d{+}1,r)-U(d,r{+}1)=2\lambda(1-\beta)x-(1-\lambda)\Big(\mathrm{Var}[\theta|d,r{+}1]-\mathrm{Var}[\theta|d{+}1,r]\Big).$$

Moreover,

$$\mathrm{Var}[\theta|d,r{+}1]-\mathrm{Var}[\theta|d{+}1,r]= -\frac{4\sigma^2x(1-x)(d-r)\Big(1+x(d+r)\Big)}{\Big(4xdr-dx+d+3rx+r-x+1\Big)\Big(4xdr+3dx+d-rx+r-x+1\Big)}. \tag{2.6}$$

In particular, the learning gain from the marginal opposite-type signal is (weakly) larger when the user currently observes more same-type signals: if $d>r$ then $\mathrm{Var}[\theta|(d,r+1)]-\mathrm{Var}[\theta|(d+1,r)]>0$, while it is zero at $d=r$ and negative if $d<r$.

The user–optimal algorithm, denoted by $\mathscr{F}^U$, satisfies two key properties. First, it is order independent: given $\mathscr{F}^U$, the engagement level $e_i^*(\mathscr{F}^U)$ maximizes the user's expected utility not only relative to other possible algorithms, but also relative to any reordering of the same feed. In other words, under any ordering

of the feed $(\mathscr{F}_i^{e_i^*})^U$, the user optimally chooses the same level of engagement. This is the main difference with the platform optimal algorithm, which front-loads same-type users to make the initial messages less costly and the final ones more informative. Second, $\mathscr{F}^U$ contains a higher proportion of same–type than opposite–type accounts. Because the optimal learning profile combines exposure to both types, with the proportion of democrat and republican messages satisfying $r \in \{d-1, d, d+1\}$, the algorithm that maximizes learning balances confirmation and diversity. In the limiting case where the user cares only about learning ($\lambda = 0$), this balanced composition is precisely the user–optimal algorithm. As there is an extra cost associated with an opposing view signal, this tilts the balance towards more democrats than republicans in the feed. All the following results are stated for a democrat user; the republican version is similar.

Proposition 2.5 (Echo chambers arise even for rational users)**.** *Given an algorithm $\mathscr{F}$ and a level of engagement $e^*(\mathscr{F})$, let us assume that the number of republicans in a democrat user feed is larger than the number of democrats, i.e., $r(\mathscr{F}, e^*(\mathscr{F})) > d(\mathscr{F}, e^*(\mathscr{F}))$. Then, the platform can weakly increase engagement by choosing an algorithm $\mathscr{F}'$ showing more democrats than republicans in the democrat feed.*

Proof. See Appendix B.1. □

For large λ (low value of learning), the user stops early, consuming only same-type messages. When λ is small, the informational gain from reaching later, opposite-type content outweighs the discomfort cost, leading to a strictly positive interior optimum $e_i^* > 1$. In equilibrium, the platform exploits this forward-looking behavior by clustering similar messages at the top of the feed to keep the user engaged until she reaches the more informative, cross-cutting content that maximizes her overall expected utility.

Even with rational users, personalization delivers less exposure diversity than a random draw from their neighborhood. The platform-optimal design for rationals front-loads same-type exposure. Echo chambers are therefore not confined to naive users.

Proposition 2.6 (Front-loading for rationals)**.** *Fix user j and let $\mathcal{F}^\star$ be the set of feeds that maximize e_j^*. Then there exists $\mathscr{F}_j^\star \in \mathcal{F}_j^\star$ and some $k \in \{1, \ldots, e_j^*\}$ such that the first k positions of $\mathscr{F}^\star$ are democrats and the next $e_j^* - k$ are republicans.*

Proof. See Appendix B.1. □

The platform exploits the user's willingness to learn by front-loading low-conformity-cost content to keep the user engaged until the more informative opposite-type signals arrive. Because the learning benefit of an opposite-type signal is increasing in the stock of same-type signals (Equation (2.6)), a front-loaded feed can implement strictly higher engagement than a balanced or alternating one. When learning matters more (small λ), the platform can tilt the implemented prefix further toward same-type content. Conversely, if the willingness to learn is low (large λ), providing larger marginal learning in farther positions of the feed is not atractive to the user, so engagement level is low.

2.4 The effects of personalization on learning

In this section, we evaluate how the platform-optimal algorithm affects learning for both naive and rational users, i.e., how information acquired on the platform improves decision quality, and how this effect varies with platform size.

First of all, let us assume homogeneous signal variances for tractability: $\sigma_{ii} = \sigma_{jj} =: \sigma^2$ for all $i, j \in \mathcal{U}$. Now, we define learning as the reduction in the expected squared error of the optimal action after reading messages. When user i chooses her optimal action after e_i messages, the expected loss equals the posterior variance:

$$\mathbb{E}\Big[(a_i - \theta)^2 | \boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\Big] = \mathbb{E}\Big[\Big(\mathbb{E}[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}] - \theta\Big)^2 | \boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\Big] = \mathrm{Var}\Big[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\Big].$$

By Lemma B.3, this posterior variance admits the closed form

$$\mathrm{Var}\Big[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\Big] = \frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}, \tag{2.7}$$

which allows us to compute learning for any algorithm and any $(\mathcal{U}, \boldsymbol{\Sigma})$. In particular, $\mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}] \leq \sigma^2$, so reading a feed weakly improves decision quality.

Under the platform-optimal algorithm for naives, the feed is ordered by similarity to minimize conformity losses. This ordering (weakly) raises engagement relative to the user–optimal ordering by front-loading like-minded content. Two

forces determine learning. First, a higher engagement level k delivers more signals, which—holding composition and correlations fixed—reduces the posterior variance. Second, the way the platform raises k is by increasing similarity within the prefix, which lowers signal diversity and might dampen information. The net effect on learning for naives is therefore *a priori* ambiguous. These forces depend critically on the platform's selection set. With a small pool of available accounts (and a given $\mathbf{\Sigma}$), the platform's ability to shape both engagement and the composition of any prefix is limited. As the platform becomes large, the feed can be engineered more flexibly at each depth, and the dependence of learning on the specific covariance structure weakens. In the next subsection we formalize this "large platform" regime, but, before, let us analyze learning for rational users when platform size is fixed.

For rational users, platform and user objectives need not be aligned in terms of learning. By designing a sequence that induces a larger best-response engagement k—for instance, front-loading lower-cost same-type content so that the user keeps scrolling until opposite-type signals arrive—the platform can push the user beyond her stopping rule under $\mathscr{F}^U$. When learning is sufficiently valuable (small λ) and conformity weights are moderate, the additional exposure to opposite signals strictly reduces the posterior variance, potentially yielding higher learning than under $\mathscr{F}^U$. Intuitively, the marginal informational value of an opposite signal is largest when such signals are scarce in the user's history; by increasing k, the platform ensures enough disagreement to improve learning overall.

2.4.1 The effects on learning in large platforms

Motivated by the recent growth of social media usage,[15] we study learning as platform size grows. We show that, for naive users, learning vanishes in the limit. The mechanism is the echo chamber: as the selection set expands, the platform fills the user's finite implemented prefix with close copies of the focal user, so additional messages convey little extra information about θ. This contrasts sharply with classical wisdom-of-the-crowd results, highlighting the platform's strategic role in information design.

[15] See, for example, the number of social media users from 2011 to 2028 (forecasted): https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

We define the expansion protocol as follows. Starting from a given user base $\mathcal{U}$, we add entrants whose covariances with incumbents are drawn *i.i.d.*from a continuous, symmetric, mean–zero distribution with support on $[-\sigma^2, \sigma^2]$ and asymptotically weak in aggregate. We assume a generating process such that the resulting covariance matrix $\mathbf{\Sigma}_n$ is symmetric and positive definite, ensuring a well–defined correlation structure across users.[16]

Learning for naive users in large platforms

As platform size increases, the platform-optimal algorithm $\mathscr{P}$ selects the most similar neighbors to maximize naive engagement. The expected engagement is nevertheless finite: because $g(\cdot) \in (0, 1)$, continuation is never guaranteed and the probability of an infinite reading path is zero. The following lemma formalizes bounded expected engagement.

Lemma 2.1 (Bounded expected engagement)**.** *There exist well-defined $k_{pi}, k_u \in \mathbb{N}$ such that $\mathbb{E}_p[e_i] \leq k_{pi}$ and $\mathbb{E}_i[e_i] \leq k_u$. In particular, there exists k_p with $\mathbb{E}_p[e_i] \leq k_p$ for all $i \in \mathcal{U}$.*

Proof. See Appendix B.1. □

Bounded implemented length implies that, as the platform grows, the naive user's observed prefix under $\mathscr{F}$ contains ever more highly correlated accounts. Consequently, diversity collapses within the prefix and learning disappears asymptotically.

Proposition 2.7 (No learning under $\mathscr{P}$)**.** *Under the platform-optimal algorithm $\mathscr{P}$, user i's learning becomes negligible as $n \to \infty$:*

$$\operatorname*{plim}_{n\to\infty} \operatorname{Var}\left[\theta | \boldsymbol{\theta}_{\mathscr{P}_i^{e_i}}\right] = \sigma^2.$$

Proof. See Appendix B.1. □

This result is in stark contrast with classic environments where the wisdom of the crowd enhances learning as the population grows. Here, the platform's

[16] In particular, a factor-structure data-generating process would suffice to guarantee positive semi-definiteness. The specific construction is standard and omitted for brevity.

strategic feed design undermines diversity, creating the echo chambers and filter bubbles emphasized by Pariser (2011). The policy relevance is immediate. For instance, the DSA's requirement that platforms offer a non-profiling ranking has led to the reinstatement of reverse-chronological feeds. Section 2.5 analyzes such alternatives relative to personalization.

Learning for rational users in large platforms

In contrast to naives, rational users need not experience vanishing learning as the platform expands. Consider the expansion protocol above. For each entrant ℓ, if the correlation $\rho_{j\ell}$ is sufficiently high so that the within-the-platform benefit (scaled by α) outweighs conformity costs, including ℓ at the front of the feed weakly increases user j's expected utility while strictly reducing her posterior variance. Because the platform maximizes engagement, it will front-load all such nonnegative-marginal entrants; the rational user then reads them along the equilibrium path. Hence, as the set of utility-improving entrants grows with platform size, the rational user's posterior variance weakly decreases. In this sense, as platform size increases, rational users benefit from an internal wisdom-of-the-crowd effect: they aggregate more information from correlated peers without incurring conformity costs.

2.5 The reverse-chronological algorithm and other alternatives

This section evaluates algorithms that can replace or modify the engagement-maximizing ones studied above. We consider two families that are especially salient in current debates. The first is the *reverse-chronological* algorithm, the canonical "non-profiling" option contemplated by the DSA. The second is a class of minimal *corrective* modifications to the platform-optimal algorithms that reintroduce diversity in exposure while preserving most of their engagement properties.

To build intuition for the subsequent results, we analyze in more detail how $\text{Var}\left[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right]$ depends on the covariance structure $\boldsymbol{\Sigma}$. Recall that with homogeneous

signal variances ($\sigma_{ii} = \sigma_{jj}$ for all i, j), the posterior variance after reading a prefix $\mathscr{F}_i^{e_i}$ admits the closed form

$$\mathrm{Var}\left[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right] = \frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}, \tag{2.8}$$

(see Equation (2.7)). Let $\boldsymbol{X} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix. The partial correlation between θ_i and θ_j conditional on all other signals equals $-\frac{x_{ij}}{\sqrt{x_{ii}x_{jj}}}$. Hence, the informational value of adding user j to i's feed depends not only on the pairwise covariance σ_{ij}, but also on the pattern of *conditional* relationships among all accounts already in the prefix. Differentiating $\mathrm{Var}\left[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right] = (\mathbb{1}^\top\boldsymbol{\Sigma}^{-1}_{\mathscr{F}^{e_i}}\mathbb{1})^{-1}$ with respect to an off-diagonal covariance σ_{ij} yields

$$\frac{\partial}{\partial\sigma_{ij}}\mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}] = 2\left(\frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}\right)^2\left(\sum_\ell x_{i\ell}\right)\left(\sum_\ell x_{j\ell}\right),$$

so the effect of bringing in a more correlated user j is generally non-monotone: it increases variance when the precision row-sums $s_i = \sum_\ell x_{i\ell}$ and $s_j = \sum_\ell x_{j\ell}$ have the same sign, and decreases it when they have opposite signs. This non-monotonicity explains why, even though random exposure, and thus an increase in diversity, the effect on learning need not always be positive.

The phenomenon is illustrated in the next example. Consider a small network of four individuals ($n = 4$), and assume that $e_i = 3$ and that the distribution of signals, conditional on θ, is

$$\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 & 0.7 & 0.5\\ 0.8 & 1 & 0.3 & 0.6\\ 0.7 & 0.3 & 1 & 0.4\\ 0.5 & 0.6 & 0.4 & 1\end{bmatrix}\right).$$

Define two possible feeds for user 1: $\mathscr{F}_1^{e_i} = \{1, 2, 3\}$, which includes the most correlated users, and $\mathscr{F}_2^{e_i} = \{1, 3, 4\}$, which features less correlated ones. The corresponding posterior variances are $\mathrm{Var}[\theta|\mathscr{F}_1^{e_i}] = 0.58$ and $\mathrm{Var}[\theta|\mathscr{F}_2^{e_i}] = 0.68$. Somewhat counterintuitively, $\mathscr{F}_1^{e_i}$—the more correlated feed—yields better learning. This illustrates that the covariance between user 1 and each of her peers is not the sole determinant of learning: the conditional dependencies among all

users within the feed also shape the amount of information the user can extract.

2.5.1 Reverse-chronological algorithm

Before the adoption of personalized algorithms, most social media platforms displayed posts strictly in the order they were written, with the most recent appearing first. In this model, this corresponds to an algorithm that draws the next message uniformly at random from the remaining accounts, the reverse-chronological algorithm, denoted $\mathscr{R}$. Under $\mathscr{R}$, the next post is drawn uniformly from the remaining accounts. Each neighbor $j \in \mathcal{U} \setminus \{i\}$ is equally likely to appear at any depth; the order is independent of the covariance structure. Unlike $\mathscr{P}$ or $\mathcal{P}$, $\mathscr{R}$ neither amplifies homophily nor minimizes disagreement in early positions. Its salient property is that it restores *exposure diversity*.

Diversity improves learning in large platforms. Because the off–diagonals are centered, bounded, and asymptotically weak in aggregate, and hence

$$\mathbb{E}\Big[\operatorname*{plim}_{n\to\infty} \operatorname{Var}\left[\theta|\mathscr{R}_i^{e_i}\right]\Big] = \sigma^2 \mathbb{E}\Big[\frac{1}{e_i}\Big|\mathscr{R}\Big], \tag{2.9}$$

and, in the large-population limit, $\operatorname{plim}_{n\to\infty} \operatorname{Var}[\theta|\mathscr{R}_i^{e_i}] = \frac{\sigma^2}{e_i}$. Thus, conditional on reading e_i posts, $\mathscr{R}$ eventually delivers the classical wisdom-of-crowds benchmark. The cost is within-platform utility: random early exposure includes dissimilar posts precisely when continuation is most fragile, lowering engagement and platform revenue.

For naive users, $\mathscr{R}$ trades off learning gains from diversity against reduced gratification and higher disagreement costs in the early prefix. The next result compares $\mathscr{P}$ and $\mathscr{R}$ in large platforms (under the expansion protocol introduced above).

Proposition 2.8 (Naive users: $\mathscr{P}$ versus $\mathscr{R}$)**.** *Given the covariance structure* $\mathbf{\Sigma}$ *and for a large platform size, the platform–optimal algorithm* $\mathscr{P}$ *outperforms the reverse–chronological algorithm* $\mathscr{R}$ *in expected utility if and only if*

$$\lambda > \max_{i\in\mathcal{U}} \left\{ \frac{1 - \mathbb{E}\left[\frac{1}{e_i}|\mathscr{R}\right]}{\frac{\alpha}{\sigma^2}(\mathbb{E}[e_i|\mathscr{P}] - \mathbb{E}[e_i|\mathscr{R}]) + 2(1-\beta)\mathbb{E}[e_i - 1|\mathscr{R}] + \left(1 - \mathbb{E}\left[\frac{1}{e_i}|\mathscr{R}\right]\right)} \right\}.$$

Proof. See Appendix B.1. □

Because $e_i \geq 1$ and $\mathscr{P}$ maximizes expected engagement, there is always some $\lambda \in (0,1)$ for which $\mathscr{P}$ dominates $\mathscr{R}$; in many configurations, $\mathscr{P}$ wins for most λ. Intuitively, the learning edge of $\mathscr{R}$ materializes only when users scroll sufficiently long; but random early disagreement lowers continuation, so naives consume fewer posts precisely when diversity would pay off.

Let us now characterize rational users' behavior under the reverse-chronological algorithm. Since the ranking imposed by $\mathscr{R}$ draws the next message uniformly at random from the remaining accounts, the rational user's decision at any depth k reduces to a standard optimal stopping problem: whether to continue reading one more randomly selected post or to stop and act based on the information already acquired. The user weighs the expected instantaneous benefit of continuing against the expected conformity cost generated by disagreement with the next, randomly drawn message. The following proposition formalizes this condition.

Proposition 2.9 (Rational users optimal behavior under $\mathscr{R}$)**.** *Consider a rational user at depth k who has already read the set of messages $\mathscr{R}_i^k$. Then it is optimal to continue to $k{+}1$ if and only if*

$$\lambda\alpha + (1-\lambda)\sigma^2\Big(\operatorname{Var}[\theta|\mathscr{R}_i^k] - \mathbb{E}\Big[\operatorname{Var}[\theta|\mathscr{R}_i^{k+1}]\Big]\Big) > \frac{\lambda(1-\beta)}{n{-}k{-}1}\mathbb{E}\Big[(\theta_j-\theta_i)^2\Big|\mathscr{R}_i^k\Big],$$

where the expectations are taken over the next random user j among the $n{-}k{-}1$ accounts that have not yet appeared in the user's feed.

Proof. See Appendix B.1. □

The stopping condition in Proposition 2.9 formalizes how a rational user balances engagement against information gain when scrolling through a reverse-chronological feed. The left-hand side captures the marginal benefit of reading one more message: a direct engagement payoff $\lambda\alpha$ plus a learning gain proportional to the expected reduction in posterior variance. The right-hand side represents the expected conformity cost from encountering a new message that may differ from the user's current belief. After encountering a sufficiently dissonant post, the expected variance reduction becomes small, so the inequality no longer holds and the rational user optimally disconnects. Conversely, when the next post comes from a highly aligned account, the gain in precision remains positive, sustaining continued scrolling. As correlation differences are stronger in polarized societies,

the informational returns from close users' exposure are steeper, implying that rational users in such environments learn more and disconnect before than those in more ideologically mixed populations.

Overall, the reverse-chronological benchmark restores exposure diversity and attains the wisdom-of-the-crowd asymptotically for a fixed engagement level e_i, but it sacrifices within-the-platform utility and (for naive users) engagement, leading to less learning in equilibrium. This trade-off motivates hybrid designs that preserve diversity while maintaining some of the continuation benefits of targeted similarity.

2.5.2 A minimal corrective: the breaking-echo-chambers algorithm

Several platforms have experimented with tools that surface corrective or contextual content (e.g., community notes, sponsored public-interest messages). We capture this idea with a minimal modification of the platform-optimal algorithm: the breaking-echo-chambers algorithm $\mathscr{B}$ inserts at the very top the account most negatively correlated with the focal user and then follows the platform-optimal order. Formally, for every naive user $i \in \mathcal{U}$,

$$\mathscr{B}_i(2) \in \arg\min_{j \in \mathcal{U}\setminus\{i\}} \rho_{ij}, \qquad \mathscr{B}_i(k) = \mathscr{P}_i(k-1) \ \text{ for } k > 2,$$

and for every rational user $j \in \mathcal{U}$,

$$\mathcal{B}_i(2) \in \arg\min_{j \in \mathcal{U}\setminus\{i\}} \rho_{ij}, \qquad \mathcal{B}_i(k) = \mathcal{P}_i(k-1) \ \text{ for } k > 2.$$

In large platforms, the top insertion can be made nearly maximally opposite at the expense of a small conformity cost, while it sharply reduces posterior variance. The engagement effect is a small reduction in the probability of continuing from the first position. The net welfare comparison for naives is summarized next.

Proposition 2.10 (Naives' welfare comparison $\mathscr{P}$ vs $\mathscr{B}$)**.** *When platform size grows large, the above defined breaking-echo-chambers algorithm outperforms, in*

terms of user's welfare, the platform-optimal algorithm $\mathscr{P}$ for user i if and only if

$$\lambda \leq \frac{1}{5 + \frac{\alpha}{\sigma^2}\Big(\mathbb{E}[e_i|\mathscr{P}] - \mathbb{E}[e_i|\mathscr{B}]\Big)}.$$

Proof. See Appendix B.1. □

Thus, when users place sufficient weight on learning (small λ), the informational gain from a single early opposite post more than compensates for the small engagement loss relative to $\mathscr{P}$. In finite platforms the comparison is ambiguous: if a strongly opposite account exists, $\mathscr{B}$ delivers sizable learning improvements at modest conformity and engagement costs; otherwise $\mathscr{P}$ remains preferable.

For rational users, $\mathcal{B}$ typically brings limited benefits. Under $\mathcal{P}$ the platform already implements a balanced path by front-loading low-cost sameness to secure continuation and delivering cross-cutting content later; adding an opposite post at the top mainly lowers early gratification and may shorten engagement without providing commensurate learning gains.

2.5.3 Policy discussion

The DSA requires very large online platforms to offer a non-profiling algorithm. In practice, platforms satisfied this by (re)introducing reverse-chronological algorithms. In our framework, this corresponds to $\mathscr{R}$. Its effects are clear. As we have shown, on the one hand, $\mathscr{R}$ restores exposure diversity and, for any fixed implemented length, converges to the classical learning benchmark. On the other, early random disagreement lowers within-the-platform utility, shortens engagement, and typically yields lower welfare than engagement-optimal algorithms. This mirrors the evidence that personalization was built to maximize engagement—and has succeeded (Guess et al., 2023): the platform-optimal $\mathscr{P}$ is tuned to deliver immediate gratification.

Hence $\mathscr{R}$ is unlikely to outperform $\mathscr{P}$ in practice. Rational users face a sharp trade-off: random exposure does not guarantee higher learning along the realized path and imposes conformity costs when continuation is most fragile. The reverse-chronological algorithm $\mathscr{R}$ thus fulfills the letter of the DSA while falling short as a welfare substitute for personalization.

Platforms have also tried corrective designs. Twitter (now $\mathbb{X}$) introduced Birdwatch (Community Notes) in January 2021 to crowd-source context. In our model, the breaking-echo-chambers tweak $\mathscr{B}$ captures this logic: inserting a maximally opposite post on top of the similar-first feed restores near first-best learning in large platforms at a small conformity cost (coming only from one user), while preserving the rest of $\mathscr{P}$. Implementation is the hurdle: it needs enforcement or platform cooperation, and users may ignore inserted opposite content.

Taken together, $\mathscr{R}$ and $\mathscr{B}$ do not fully resolve the engagement-learning tension. This motivates the market–design approach pursued next: under horizontal interoperability, network effects are shared and platforms compete on ranking quality; under some conditions, the utilitarian-optimal algorithm $\mathscr{U}$ becomes the implemented option by competing platforms.

2.6 Network effects

This section examines how personalization shapes network effects, i.e., whether a larger user base raises user utility for naives and rationals, and the resulting policy implications.

We say that user i experiences network effects under algorithm $\mathscr{F}$ if her expected utility weakly increases with platform size: $\mathbb{E}_i[U_i(n+1) - U_i(n) | \mathscr{F}(n)] \geq 0$. For clarity, we write $U_i(n)$ to the utility of the user when the platform size is n, $\mathscr{F}(n)$ for the algorithm applied on a platform of size n, and $e_i(n)$ for the engagement induced for user i under $\mathscr{F}(n)$.

Proposition 2.11 (Network effects need not hold for naives)**.** *Fix a naive user i. Under the platform-optimal algorithm $\mathscr{P}$ define*

$$\Delta e_i := \mathbb{E}\Big[e_i(n+1) - e_i(n)\Big|\theta_i, \mathscr{P}(n)\Big];$$

$$\Delta C_i := \mathbb{E}\left[\sum_{j \in \mathscr{P}_i^{e_i(n+1)}(n+1)} (\theta_i - \theta_j)^2 - \sum_{j \in \mathscr{P}_i^{e_i(n)}(n)} (\theta_i - \theta_j)^2 \Big| \theta_i, \mathscr{P}(n)\right];$$

$$\Delta V_i := \mathbb{E}\Big[\mathrm{Var}\Big[\theta\Big|\boldsymbol{\theta}_{\mathscr{P}(n+1)}\Big] - \mathrm{Var}\Big[\theta\Big|\boldsymbol{\theta}_{\mathscr{P}(n)}\Big]\Big|\theta_i, \mathscr{P}(n)\Big].$$

Whenever $\alpha\Delta e_i - (1-\beta)\Delta C_i + \Delta V_i \neq 0$, the platform-optimal algorithm $\mathscr{P}$

features network effects for naive user i if and only if

$$\lambda \geq \frac{\Delta V_i}{\alpha \Delta e_i - (1-\beta)\Delta C_i + \Delta V_i}.$$

If either $\alpha \Delta e_i - (1-\beta)\Delta C_i + \Delta V_i = 0$ or learning improves with platform size (i.e., $\Delta V_i \leq 0$), then $\mathscr{P}$ features network effects for naive user i for every $\lambda \in (0,1)$.

Proof. See Appendix B.1. □

Proposition 2.12 (Rational users always enjoy network effects)**.** *Under the platform-optimal algorithm $\mathcal{P}$, rational users always enjoy network effects.*

Proof. See Appendix B.1. □

The propositions formalize a sharp contrast. Personalization enlarges the platform's selection set and permits finer matching in the feed. For rational users, who internalize both conformity and learning when deciding how far to scroll, this translates into higher expected utility as n grows: the platform can front-load low-cost similar content while still implementing enough exposure to disagreement to sustain learning. By contrast, for naive users the platform over-exploits homophily: as n rises, the implemented feed becomes increasingly populated with close copies of the focal user, which raises conformity and lowers the informational value of each additional message. When the naive user places any weight on learning ($\lambda > 0$), the sign and magnitude of ΔV_i become pivotal; beyond a size threshold the learning loss can dominate, so expected utility can fall with n even if engagement rises.

This aligns with the evidence from Guess et al. (2023), which shows that switching users from Facebook's personalized feed to a chronological (non-personalized) feed significantly reduces engagement but also exposes users to more diverse content, mitigating echo chambers. Analogous patterns are evident on 𝕏 and other social media platforms, where personalization algorithms appear to substantially boost engagement while narrowing exposure diversity (Gauthier et al., 2025; Kitchens et al., 2020). Moreover, these patterns suggest that naive users on highly personalized platforms may not realize the extent of the algorithm's filter bubble, continuing to engage more while actually seeing less variety in viewpoints. Such evidence underscores our model's implication that personalization, when

taken to extremes, can erode the very network benefits that attracted users in the first place.

2.6.1 Network effects, competition, and personalization algorithms

Although our model does not endogenize platform competition, the network–effect results speak directly to current policy debates. Two observations are central. First, the presence of naive users challenges the conventional view that larger platforms unambiguously benefit participants in digital markets. For such users, stronger personalization may raise engagement while eroding learning, so scale can reduce welfare. This suggests that improving contestability in social media markets may require tools beyond the standard repertoire (Banchio et al., 2025). Second, when user behavior does generate platform-specific network effects, the familiar winner-takes-most logic applies: large incumbents enjoy advantages that are difficult for entrants to overturn.

A leading proposal to address these frictions is horizontal interoperability: the capacity of distinct platforms to interconnect so that users from one can interact with users from the other and *vice versa* (Kades and Scott Morton, 2020). Interoperability makes network effects no longer platform-specific, but shared across the whole market. Competition then shifts from for the market to within the market (Belleflamme and Peitz, 2020). This is standard in other communications layers (e.g., telephony and email), where users can communicate across providers without switching accounts.[17]

The EU's DMA is a first step in this direction. Article 7 mandates interoperability for number-independent interpersonal communication (messaging) services provided by gatekeepers. While some authors are skeptical about the magnitude of the resulting competitive gains in messaging (Bourreau and Kraemer, 2023; Bourreau et al., 2023; Dhakar and Yan, 2024), social media differs in a critical respect: personalization algorithms are the key competitive dimension once network effects are shared. In that environment, interoperability weakens lock-in and redirects rivalry toward the quality of feed design. If we had to consider platforms competing and users deciding where to participate, our results would point to hor-

[17]For example, a Yahoo user can seamlessly send an email to a Gmail user, and *vice versa*.

izontal interoperability particularly benefiting naive users: under such a regime, platforms would no longer be able to rely on captive scale and would instead have to compete on algorithms that balance engagement and learning. Naiveté would pertain to within-platform continuation and would not affect platform choice—*ex ante*, naive users would select the platform offering higher expected utility. Thus, rather than prescribing a specific ranking rule, interoperability would harness competitive pressure to discipline engagement-maximizing designs that would otherwise entrench echo chambers.

2.7 Conclusion

This paper develops a model of communication and learning through personalized feeds to study how an engagement-maximizing platform arranges exposure and how that arrangement feeds back into behavior. Truth-telling is the robust equilibrium of the messaging game, so platforms exercise power solely through the composition and order of exposure. Second, the engagement-optimal algorithm fosters homophily: naives receive similar-first feeds (a perfect echo chamber), and rationals receive front-loaded sameness that sustains continuation until cross-cutting content arrives. In large platforms, the naive learning benefit of additional scrolling vanishes, breaking the classical wisdom-of-the-crowds logic; rational users, by contrast, continue to enjoy network effects.

With respect to policy analysis, the non-profiling reverse-chronological algorithm is a natural baseline but not a panacea: absent additional design features, it underperforms engagement-optimal personalization when users heavily value within-platform utility and rarely delivers first-best learning along realized paths. The corrective insertion breaking-echo-chambers algorithm can move outcomes toward the utilitarian benchmark at modest engagement cost, especially on large platforms with rich selection sets. More broadly, competition policy that shares network effects—horizontal interoperability—would redirect rivalry toward algorithmic quality, disciplining pure engagement maximization and aligning platform incentives with user welfare.

Taken together, our results show that when platforms optimize for engagement, echo chambers are not an accident—they are the optimal instrument. Restoring

learning therefore requires either changing the instrument (altering algorithms to time in disagreement, which matters especially for naives) or changing the game (making platforms compete on algorithms rather than on captive network scale).

Two modeling choices invite further work. First, improper priors yield the clean truth-telling benchmark and closed-form posteriors. However, with normal (proper) priors, messaging and ordering would interact, potentially strengthening echo-chamber forces through both content *and* speech. Second, our within-the-platform utility formulation captures simple conformity motives. However, incorporating richer interaction dynamics (e.g., u-shaped engagement in similarity) would allow algorithms to exploit both outrage and affinity.

Chapter 3

The Mechanisms of Media Slant

Manuel Lleonart-Anguix[1] and Luis Menendez[2]

Abstract

We study the mechanisms through which news outlets generate ideological slant. We develop a model of supply and demand for news in which outlets choose whether to cover a story (omission), how to frame it (framing), and how much airtime to assign to it (intensity). The model characterizes equilibrium tone and coverage in closed form and delivers a threshold rule for omission. We then bring the framework to data from Spanish prime-time television news, using a Shapley value decomposition to estimate each mechanism's contribution to observed slant. Framing is the dominant mechanism across all outlets, accounting for at least half and up to nearly 90% of observed slant. Omission plays a secondary but non-negligible role for some channels, while intensity contributes little. These results suggest that ideological distortion in news markets operates primarily through how stories are told, not which stories are selected.

[1]Universitat Autònoma de Barcelona and Barcelona School of Economics.
[2]Universitat Autònoma de Barcelona and Barcelona School of Economics.

3.1 Introduction

Traditional news media are increasingly valued as reliable sources of information, particularly as concerns about online misinformation intensify (Nieman Lab, 2025). In an environment where digital platforms are often criticized for amplifying noise and falsehoods, established outlets continue to play a central role in shaping public understanding of events. Yet these same outlets exhibit substantial dispersion in the ideological content they provide.[3] Coverage of identical events can differ markedly across outlets, not only in tone but also in emphasis and selection, giving rise to distinct ideological profiles. This coexistence of credibility and ideological variation poses a fundamental question. If audiences value accuracy and reliability, how can editorial teams sustain such differences without undermining trust?

In this paper, we study how outlets generate ideological slant. We define slant as a descriptive measure of where an outlet's content falls on the left-right spectrum. We distinguish three mechanisms through which outlets shift their position on that spectrum: *omission* (which events are covered), *framing* (the tone conditional on coverage), and *intensity* (how much airtime an event receives). Each margin distorts audiences' information differently. Omission determines which facts viewers observe; framing shapes how those facts are interpreted and intensity governs how salient each fact becomes relative to others. As a result, two outlets with identical measured slant may correspond to very different patterns of informational distortion. Understanding which mechanism drives slant is therefore important both for diagnosing how media markets work and for designing effective responses.

In the first part of the paper, we develop a theoretical model of news supply and demand. There is a set of potential stories, each characterized by an ideological valence and a newsworthiness level. Outlets observe both and choose, for each story, the tone with which to present it and the airtime share to devote to it. On the demand side, viewers trade off ideological alignment against the intrinsic value of newsworthy coverage. On the cost side, outlets face increasing penalties for deviating from a story's true valence (credibility costs), from their own editorial line (editorial costs), and for concentrating airtime on few stories. The model

[3] See, for instance, `adfontesmedia.com`, `allsides.com`, or `mediabiasdetector.com`.

delivers three results. First, equilibrium tone is a weighted average of the story's true valence, the outlet's editorial position, and a competitive distortion that depends on the location of marginal viewers — and is pinned down in closed form. Second, coverage follows a threshold rule: a story is aired only if its net benefit exceeds the shadow value of airtime. Third, comparative statics show that competitive pressure distorts tone uniformly across stories, making framing the primary margin of adjustment, while credibility pressure moves each story's tone toward its true valence and reshuffles coverage in a story-specific way.

In the second part of the paper we test those predictions using a unique dataset from Spanish prime-time television news. Following the methodology in Menéndez (2025), we combine machine learning and large language models to measure how the same story is framed and covered across outlets. We then use a Shapley value approach to decompose observed slant into the contributions of omission, framing, and intensity. We find that framing is the dominant mechanism for every news outlet, accounting for at least two-thirds of observed slant. Omission and intensity play secondary roles, consistent with the observation that all major Spanish channels cover the same headline political stories but narrate them differently.

This work might inform the design of current efforts to fight misinformation in media markets. On the one hand, when distortion operates primarily through omission or coverage intensity, the problem is one of selective exposure to facts, which points to policies focused to foster transparency and pluralism. On the other hand, when framing dominates, the relevant concern is how facts are interpreted rather than whether they are covered, which shifts attention toward tools such as fact-checking and media literacy.

Related literature. This paper relates first to the theoretical literature on media bias and and information provision. We build on Gentzkow et al. (2014), who provide a unified framework for media bias organized around two mechanisms: *distortion* (slanting the presentation of a covered story) and *filtering* (choosing which stories to report). In our model we separate filtering into *omission* (whether a story is covered at all) and *intensity* (how much airtime a covered story receives), reflecting the fact that television outlets face a binding time constraint and must allocate a continuous coverage share across stories. Another difference with their

work is that our demand side does not require Bayesian updating: viewers choose outlets based on ideological proximity and newsworthiness, which lets us build a framework that links to the empirical data typically observed for TV markets.

Second, our model-free decomposition is related to a broader literature that uses statistical decompositions to attribute observed outcomes to distinct underlying mechanisms. In labor economics, decomposition methods have been used, for example, to separate between-firm inequality into components associated with worker sorting, segregation, and within-firm dispersion (Song et al., 2019). More generally, counterfactual reweighting approaches decompose observed gaps into components attributable to differences in characteristics versus differences in returns or structures (DiNardo et al., 1996; Fortin et al., 2011).

This paper also relates to the empirical literature that measures ideological positions and media slant from text and behavior. Groseclose and Milyo (2005) and Gentzkow and Shapiro (2010b) recover outlet ideology from similarity to external political benchmarks, while text-analysis methods such as Laver et al. (2003) and Slapin and Proksch (2008) estimate latent ideological positions from word frequencies in political texts. Other work uses high-dimensional or supervised text methods to recover political sentiment or group differences in language (Gentzkow et al., 2019; Taddy, 2013), while network-based approaches infer ideal points from online following behavior (Barberá, 2015). Our framework is complementary to this literature. The model provides a foundation for future structural estimations that aim to recover both outlet ideological positions and story true valences.

The rest of the paper is organized as follows. Section 3.2 presents the theoretical model. Section 3.3 describes the data, transcript pipeline, and the construction of tone and coverage measures. Section 3.4 defines the statistical decomposition and reports its results. Section 3.5 concludes.

3.2 Model

This section develops the theoretical framework to study how channels shape their content and how these choices translate into ideological slant. We first describe the primitives of the environment and characterize the equilibrium choices of tone, omission, and coverage. All proofs are relegated to the appendix.

Channels and Stories There are $\mathcal{S} \geq 1$ stories indexed by $s = 1, \ldots, \mathcal{S}$. Each story s has two dimensions: a horizontal property $\theta_s \in \mathbb{R}$ describing its position on the ideological line and a vertical property $n_s \in \mathbb{R}^+$ describing its newsworthiness [4]. Both parts are common knowledge for the channels. There are $N \geq 2$ channels indexed by $j = 1, \ldots, N$. Each channel j has an *editorial point* $\psi_j \in \mathbb{R}$, the position it finds natural to broadcast absent competitive considerations.

Instruments Each channel j chooses a level of slant $\mathfrak{s}_j$ describing its position on the ideological line. To build its slant index, channel j uses two different instruments. For each story s, channel j chooses:

- *Coverage share* $c_{js} \geq 0$: the fraction of airtime devoted to story s. Shares satisfy $\sum_s c_{js} = 1$. In particular, if a channel chooses $c_{js} = 0$ we say that the channel *omits* story s.
- *Tone* $\tau_{js} \in \mathbb{R}$: the ideological position of the broadcast on story s.

The overall ideological slant of channel j is the coverage-weighted average tone

$$\mathfrak{s}_j \equiv \sum_{s=1}^{\mathcal{S}} c_{js}\tau_{js}, \tag{3.1}$$

as developed in Menéndez (2025). This metric takes higher values if the content of an oulets is more pro-right and viceversa.

Viewers There is a continuum of viewers indexed by ideology $x \in [-1, 1]$, distributed according to a density $f(x)$, with $\int_{-1}^{1} f(x)dx = 1$. Each viewer chooses the channel that maximizes their utility or selects the outside option. Viewers value both ideological alignment and the intrinsic newsworthiness of covered stories. A viewer i with ideology x derives utility from channel j:

$$u_{ij}(x) = \sum_{s=1}^{\mathcal{S}} n_s c_{js} - \lambda(x - \mathfrak{s}_j)^2 + \varepsilon_{ij},$$

where ε_{ij} are i.i.d. Type-I extreme value shocks. The outside option's utility is normalized to ε_{i0}. The parameter λ measures the distaste for slant relative to

[4] Notice that, unlike **?** who use newsworthiness to refer to the informational value of reporting, we use the term as an exogenous story-level vertical attribute capturing how valuable or salient coverage of a story is to viewers.

the value of newsworthiness.

The probability that viewer x watches channel j is

$$P_j(x) = \frac{\exp\left(\sum_{s=1}^{\mathcal{S}} n_s c_{js} - \lambda(x - \mathfrak{s}_j)^2\right)}{1 + \sum_{k=1}^{N} \exp\left(\sum_{s=1}^{\mathcal{S}} n_s c_{ks} - \lambda(x - \mathfrak{s}_k)^2\right)}, \qquad \lambda > 0.$$

Consequently, the aggregate demand is

$$D_j = \int_{-1}^{1} P_j(x) f(x) dx.$$

Channel Problem Each channel earns advertising revenue $\rho > 0$ per viewer. The cost of broadcasting story s has two components. The first is proportional to coverage and captures the credibility cost of deviating from the truth and the editorial cost of deviating from the channel's editorial position:

$$c_{js}\left[\frac{\kappa}{2}(\tau_{js} - \theta_s)^2 + \frac{\gamma}{2}(\tau_{js} - \psi_j)^2\right], \qquad \kappa, \gamma > 0.$$

The second is a convex cost of coverage intensity:

$$\frac{\zeta}{2} c_{js}^2, \qquad \zeta > 0,$$

reflecting the fact that allocating more airtime to any single story is increasingly costly, independently of its content.[5] Total cost is

$$C_j = \sum_{s=1}^{\mathcal{S}} \left\{ c_{js}\left[\frac{\kappa}{2}(\tau_{js} - \theta_s)^2 + \frac{\gamma}{2}(\tau_{js} - \psi_j)^2\right] + \frac{\zeta}{2} c_{js}^2 \right\}.$$

Therefore, channel j's profit is

$$\pi_j = \rho D_j - C_j.$$

Remark 3.1 (Notation)**.** *Define* $\Gamma \equiv \kappa + \gamma$, $\alpha \equiv \frac{\kappa}{\Gamma}$, $\eta \equiv \frac{2\rho\lambda}{\Gamma}$, *the cost-minimizing*

[5]The baseline specification uses a quadratic cost $\frac{\zeta}{2} c_{js}^2$, which delivers closed-form characterizations of tone, coverage, and slant. A generalization to $\frac{\zeta}{\alpha+1} c_{js}^{\alpha+1}$ with $\alpha > 1$ preserves the qualitative structure of the omission threshold and the direction of competitive distortions, but precludes closed-form solutions for the Shapley decomposition. We therefore maintain the quadratic specification throughout.

tone

$$\underline{\tau}_{js} \equiv \frac{\kappa\theta_s + \gamma\psi_j}{\Gamma} = \alpha\theta_s + (1-\alpha)\psi_j,$$

and the editorial distortion

$$\delta_{js} \equiv \underline{\tau}_{js} - \theta_s = \frac{\gamma(\psi_j - \theta_s)}{\Gamma} = (1-\alpha)(\psi_j - \theta_s).$$

Also define

$$A_j \equiv \int_{-1}^{1} (x - \mathfrak{s}_j) P_j(x)(1 - P_j(x)) f(x)dx, \qquad B_j \equiv \int_{-1}^{1} P_j(x)(1 - P_j(x)) f(x)dx.$$

A_j is the marginal audience gain from shifting slant; B_j is the marginal audience gain from raising the direct utility of all viewers by one unit.

The next pair of propositions characterizes the optimal usage of tone (Proposition 3.1) and coverage (Proposition 3.2) that a channel does to build its slant.

Proposition 3.1 (Optimal tone)**.** *The equilibrium tone on story s by channel j is*

$$\tau^*_{js} = \alpha\theta_s + (1-\alpha)\psi_j + \eta A_j. \tag{3.2}$$

Proposition 3.1 clarifies the role of competitive incentives in shaping tone. In the absence of audience considerations, the optimal tone coincides with the cost-minimizing message $\underline{\tau}_{js}$. The term ηA_j captures the distortion induced by competition for viewers. In particular, since

$$\tau^*_{js} = \underline{\tau}_{js} + \eta A_j,$$

the sign of A_j determines the direction of the distortion: if $A_j > 0$, the channel tilts its tone upward relative to the cost-minimizing benchmark, while if $A_j < 0$, it tilts it downward. Intuitively, A_j reflects the location of marginal viewers: when these viewers are, on average, to the right of the channel's slant, shifting tone to the right increases demand, and conversely when they are to the left.

The previous proposition allows us to rewrite the coverage problem in terms

of a story-specific net benefit. Substituting the optimal tone into the first-order condition for coverage, define

$$N_{js} := \rho B_j n_s + \Gamma \eta A_j \underline{\tau}_{js} + \frac{\Gamma \eta^2}{2} A_j^2 - \Gamma \frac{\alpha}{2(1-\alpha)} \delta_{js}^2. \tag{3.3}$$

The object N_{js} captures the marginal benefit of increasing coverage of story s for channel j, after optimally adjusting tone. It is composed of four terms. The first, $\rho B_j n_s$, reflects the direct audience value of the story through its newsworthiness. The second, $\Gamma \eta A_j \underline{\tau}_{js}$, captures how the story contributes to shifting the channel's slant toward the viewers that are most valuable at the margin. The third, $\frac{\Gamma \eta^2}{2} A_j^2$, is common across stories and reflects the overall value of being able to adjust slant in response to competitive incentives. The last term, $\Gamma \frac{\alpha}{2(1-\alpha)} \delta_{js}^2$, captures the cost of distorting the message away from the truth and the channel's editorial position. Define $\bar{N}_j$ as the average net benefit over covered stories. The next proposition characterizes the optimal coverage of stories.

Proposition 3.2 (Optimal coverage)**.** *The equilibrium coverage satisfies*

$$c_{js}^* = \frac{1}{|\mathcal{S}_j|} + \frac{N_{js} - \bar{N}_j}{\zeta}, \qquad s \in \mathcal{S}_j,$$

where $\mathcal{S}_j = \{s \in \mathcal{S} : c_{js}^* > 0\}$.

As a consequence of Proposition 3.2, omission is governed by a simple threshold rule: a story is aired only if its net benefit is high enough relative to the shadow value of airtime. The next corollary states this result formally.

Corollary 3.1 (Omission threshold)**.** *A story s is covered by channel j if and only if its net benefit exceeds an endogenous channel-specific threshold:*

$$c_{js}^* > 0 \quad \Longleftrightarrow \quad N_{js} > \mu_j,$$

where μ_j is the Lagrange multiplier associated with the airtime constraint.

Corollary 3.1 shows that omission is entirely summarized by the net benefit N_{js}. Once optimal tone is substituted into the coverage problem, all relevant story characteristics enter through this object: stories with sufficiently high net benefit are aired, whereas stories whose net benefit falls below the shadow value

of airtime are omitted. This threshold structure allows us to build comparative statics with respect to newsworthiness and ideology.

Corollary 3.2 (Omission region)**.** *A story s is covered by outlet j if and only if*

$$n_s \geq \underline{n}_j + \kappa_j(\theta_s - \theta_j^{\max})^2,$$

for some outlet-specific constants $\underline{n}_j$ and $\kappa_j > 0$.

The corollary implies that the omission region takes the form of an inverted parabola in story space. Stories that are either insufficiently newsworthy or too far from the outlet's preferred position $\theta_j^{\max}$ are not covered. The location and curvature of the parabola vary across outlets, generating systematic differences in the set of covered stories. Intuitively, channels face a double selection pressure: stories must be worth the airtime (newsworthiness) and compatible with the outlet's narrative at reasonable distortion cost (ideological proximity). Stories that fail on either dimension are omitted.

Changes in the newsworthiness of one story also affect others through the binding airtime constraint. When a story favorable to the outlet gains in newsworthiness, airtime is reallocated toward it, which may push sufficiently marginal stories below the omission threshold. Conversely, when a previously omitted story's newsworthiness rises enough to clear the threshold, it enters the covered set and incumbent stories lose airtime. Whether incumbent stories survive depends on how much slack they have above their own threshold — those comfortably above it are unaffected, while those near the boundary may be crowded out. Increases in newsworthiness thus either intensify competition within the covered set or induce entry of new stories, depending on whether the affected story was previously above or below the omission threshold.

Finally, the next proposition characterizes the slant in equilibrium of each channel.

Proposition 3.3 (Optimal slant)**.** *The equilibrium ideological slant of channel j implicitly given by*

$$\mathfrak{s}_j^* = \alpha\bar{\theta}_j^* + (1-\alpha)\psi_j + \eta A_j,$$

where

$$\bar{\theta}_j^* \equiv \sum_{s=1}^{\mathcal{S}} c_{js}^* \theta_s = \frac{1}{|\mathcal{S}_j|} \sum_{s \in \mathcal{S}_j} \theta_s + \frac{1}{\zeta} \sum_{s \in \mathcal{S}_j} (N_{js} - \bar{N}_j)\theta_s$$

is the coverage-weighted average true valence of the stories broadcast by channel j*.*

Proposition 3.3 shows that equilibrium slant is driven by three forces. The first term reflects content selection: channels tilt their slant by choosing which stories to cover, as captured by the average ideological position $\bar{\theta}_j^*$. The second term captures the channel's intrinsic editorial position. The third term reflects competitive incentives, which distort slant in the direction of marginal viewers. Through $\bar{\theta}_j^*$, coverage choices endogenously shape slant, linking the extensive and intensive margins of content provision.[6]

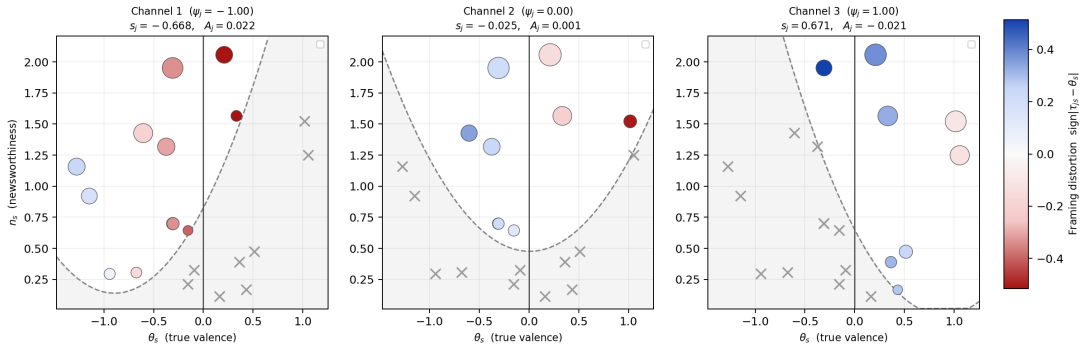
Numerical Example

An example of an equilibrium is represented in Figure 1. Each panel, from left to right, corresponds to the left, centrist, and right outlet, respectively. The horizontal axis reports the true valence of the story, θ_s, while the vertical axis reports its newsworthiness, n_s. The shaded gray region represents the omission region (Corollary 2): a story is covered only if its newsworthiness is sufficiently high relative to its ideological distance from the outlet's coverage-maximizing position. Intuitively, stories that are either not sufficiently newsworthy or too ideologically far from the outlet are not worth covering (gray crosses). Omission is concentrated among stories that go against the outlet's ideological orientation: the left outlet omits relatively more right-leaning stories, while the right outlet does the opposite.

Covered stories are represented as circles, with circle size proportional to equilibrium coverage c_{js}^*. As Proposition 2 shows, conditional on being covered,

[6]The equilibrium characterized in Propositions 3.1–3.3 involves a fixed point: the demand objects A_j and B_j depend on equilibrium slant $\mathfrak{s}_j$, which in turn depends on A_j and B_j through optimal tone and coverage. Existence follows from a standard fixed-point argument: the logit demand structure ensures that A_j and B_j are continuous functions of the slant profile $(\mathfrak{s}_1, \ldots, \mathfrak{s}_N)$, and the best-response map is continuous on a compact domain, so Brouwer's theorem applies. Uniqueness need not hold in general: multiple slant configurations can be mutually consistent with the implied demand objects. The characterizations in Propositions 3.1–3.3 describe any interior equilibrium, and the Shapley decomposition applies to each such equilibrium separately.

Figure 3.1: Equilibrium coverage and framing in story space



Note: Each panel shows one outlet's coverage and framing decisions in story space (θ_s on the horizontal axis, n_s on the vertical axis). Covered stories are represented as bubbles, where bubble size is proportional to the equilibrium coverage share c^*_{js}. Bubble color encodes the signed framing distortion $\tau^*_{js} - \theta_s$: blue (red) indicates rightward (leftward) distortion, with color intensity proportional to magnitude. Omitted stories appear as gray crosses. The vertical dashed line marks $\theta_s = 0$ (ideological center). Outlets are ordered left to right by editorial position $\psi_j \in \{-1, 0, 1\}$. Viewers' ideology follows a symmetric bimodal mixture, $f(x) = 0.5\mathcal{N}(x; -1, 0.3^2) + 0.5\mathcal{N}(x; 1, 0.3^2)$ over $x \in [-1, 1]$, capturing a polarized electorate.

airtime is increasing in the story's net benefit. Since net benefit is increasing in newsworthiness and concave in ideology around the outlet-specific maximizing position, larger bubbles tend to appear for stories that are both more newsworthy and ideologically closer to the outlet's preferred region. Finally, colors represent framing distortion, $\tau^*_{js} - \theta_s$. Darker blue (red) denotes stronger rightward (leftward) distortion. From Eq. 3.2 we get that $\tau^*_{js} - \theta_s = (1 - \alpha)(\psi_j - \theta_s) + \eta A_j$, therefore distortion is smaller for stories whose true valence lies closer to the outlet's preferred position and larger for stories that must be pushed further away from their true valence.

3.2.1 Comparative Statics

We derive comparative statics for the main equilibrium objects with respect to two parameters: competitive pressure A_j and credibility pressure α. The two parameters operate through fundamentally different channels. An increase in A_j shifts all tones uniformly, so that competition distorts primarily through framing. By contrast, an increase in α makes tone track story content more closely, reshuffling coverage in a story-specific way.

Proposition 3.4 (Comparative statics with respect to A_j)**.** *Fix outlet j, and consider a marginal increase in A_j—the marginal audience return to rightward slant—such that the active coverage set $\mathcal{S}_j \equiv \{s : c^*_{js} > 0\}$ is locally unchanged. Then:*

(i) For every story $s \in \mathcal{S}_j$,

$$\frac{\partial \tau^*_{js}}{\partial A_j} = \eta > 0.$$

The shift is identical across all stories: competitive pressure distorts tone by a story-invariant constant.

(ii) For every $s \in \mathcal{S}_j$,

$$\frac{\partial c^*_{js}}{\partial A_j} = \frac{\Gamma\eta}{\zeta}\Big[\tau^*_{js} - \bar{\tau}^*_j\Big], \qquad \bar{\tau}^*_j \equiv \frac{1}{|\mathcal{S}_j|}\sum_{r\in\mathcal{S}_j} \tau^*_{jr}.$$

Stories whose equilibrium tone exceeds the active-set average gain airtime; those below it lose airtime. The reallocation preserves the resource constraint $\sum_s c^*_{js} = 1$ exactly.

(iii) For every story s,

$$\frac{\partial N_{js}}{\partial A_j} = \Gamma\eta\tau^*_{js}.$$

Part (i) is the key result. Because the competitive premium ηA_j enters equilibrium tone additively and independently of the story's true valence θ_s, an increase in competitive pressure shifts every covered story's tone by the same constant η. The outlet becomes uniformly more right-leaning in its narration without selectively emphasizing ideologically congenial stories. Competition therefore operates primarily through *framing* rather than through omission or intensity reallocation.

Part (ii) shows that competitive pressure does nonetheless affect the intensive margin of coverage. The reallocation is governed by deviations of story tone from the active-set average: stories that are already framed more to the right receive more airtime, reinforcing their contribution to slant. Crucially, this reallocation is a *second-order* effect relative to the uniform tone shift of part (i): it affects how airtime is distributed, not the direction of every story's message.

Part (iii) connects the intensive and extensive margins. Because $\partial N_{js}/\partial A_j = \Gamma\eta\tau^*_{js}$, the net benefit of coverage rises faster for stories with higher equilibrium tone. This determines which stories are most likely to cross the omission threshold, characterized in the following corollary.

While framing is the primary margin of adjustment to competitive pressure, omission can reinforce it at the boundary of the coverage set. Consider a discrete increase in A_j large enough to alter $\mathcal{S}_j$. From part (iii), $\partial N_{js}/\partial A_j = \Gamma\eta\tau^*_{js}$, so net benefit rises fastest for stories with high equilibrium tone. Stories with $\tau^*_{js} > 0$ are therefore the first to cross the omission threshold μ_j from below and enter coverage, while stories with $\tau^*_{js} < 0$ are the first to be dropped. An outlet facing a rightward shift in its marginal audience therefore simultaneously covers more right-leaning stories and drops left-leaning ones, amplifying omission-based slant in the same direction as the framing distortion. The two mechanisms are complements: the uniform tone shift of part (i) raises the net benefit of right-leaning stories and lowers that of left-leaning ones, making coverage adjustments more likely precisely where they would reinforce framing-based slant.

Proposition 3.5 (Comparative statics with respect to α)**.** *Fix outlet j, and consider a marginal increase in α such that $\mathcal{S}_j$ is locally unchanged. Then:*

(i) For every story s,

$$\frac{\partial \tau^*_{js}}{\partial \alpha} = \theta_s - \psi_j.$$

Stories with $\theta_s > \psi_j$ shift rightward; stories with $\theta_s < \psi_j$ shift leftward. Unlike competitive pressure, credibility pressure has a heterogeneous effect on tone that depends on the ideological content of each story.

(ii) For every $s \in \mathcal{S}_j$,

$$\frac{\partial c^*_{js}}{\partial \alpha} = \frac{1}{\zeta}\left[\frac{\partial N_{js}}{\partial \alpha} - \frac{1}{|\mathcal{S}_j|}\sum_{r\in\mathcal{S}_j}\frac{\partial N_{jr}}{\partial \alpha}\right].$$

Coverage shifts toward stories whose net benefit rises by more than the active-set average.

(iii) For every story s,

$$\frac{\partial N_{js}}{\partial \alpha} = \underbrace{\Gamma \eta A_j(\theta_s - \psi_j)}_{\text{audience alignment gain}} - \underbrace{\Gamma \frac{\delta_{js}^2}{2(1-\alpha)^2}}_{\text{distortion penalty}} .$$

The first term is positive for stories ideologically close to the outlet's audience and negative for distant ones. The second term is always negative and grows with the squared editorial distortion $\delta_{js}^2 = (1-\alpha)^2(\psi_j - \theta_s)^2$, so that stories already far from the outlet's editorial line suffer a disproportionately larger penalty as α rises.

The contrast with Proposition 3.4 is instructive. Competitive pressure shifts all tones by the same constant, so its primary effect is on framing; coverage reallocations are a second-order consequence. Credibility pressure, by contrast, is story-specific: it pushes each tone toward its true valence by an amount that depends on the story's ideological distance from the editorial line. This generates a different pattern of coverage reallocation — and a different pattern of omission. The two parameters therefore operate through different combinations of the three mechanisms, which motivates using the Shapley decomposition empirically to separate their contributions.

Part (iii) reveals a selection effect: a rise in α favors coverage of stories for which better alignment with truth is also valuable to the audience (high $\theta_s - \psi_j$ when $A_j > 0$), while penalizing stories that require large editorial distortions to fit the outlet's line. As $\alpha \to 1$, coverage concentrates on stories whose true valence is close to the outlet's preferred position, because the cost of framing ideologically distant stories becomes prohibitive. This generates a form of *credibility-driven omission*: outlets under strong fact-checking pressure do not simply report the truth about all stories, but selectively cover stories they can report truthfully without large ideological costs.

Note that whereas changes in A_j can generate entry or exit from coverage based solely on the sign of equilibrium tone, changes in α induce omission that depends on the full ideological distance $|\theta_s - \psi_j|$. Credibility pressure therefore produces a different omission pattern: outlets drop the stories furthest from their editorial line first, regardless of directional slant.

Two Extreme Cases

We study two extreme cases where $\alpha \in \{0, 1\}$, that isolate the role of credibility costs and clarify the mechanisms through which slant is generated.

No credibility costs: $\alpha = 0$. When $\alpha = 0$, outlets face no penalty for deviating from a story's true valence. From Proposition 3.1, equilibrium tone simplifies to

$$\tau_{js}^* = \psi_j + \eta A_j,$$

which is independent of the story's true valence θ_s. Every story is narrated at the same ideological position, determined entirely by the outlet's editorial line and its competitive incentive. This is the most extreme form of framing dominance: the outlet imposes a uniform ideological lens on all content regardless of what the stories are actually about.

Coverage in this case is governed by net benefit (Eq. 3.3), which simplifies to

$$N_{js} = \rho B_j n_s + \Gamma \eta A_j \psi_j + \frac{\Gamma \eta^2}{2} A_j^2.$$

The last two terms are common across stories, so coverage shares satisfy

$$c_{js}^* = \frac{1}{|\mathcal{S}_j|} + \frac{\rho B_j}{\zeta}(n_s - \bar{n}_j),$$

where $\bar{n}_j$ is the average newsworthiness of covered stories. Coverage is therefore determined solely by newsworthiness: ideologically distant stories are not penalized in coverage decisions because there is no cost to framing them at the outlet's preferred position. The omission threshold likewise depends only on newsworthiness: a story is covered if and only if n_s exceeds a channel-specific cutoff. In this polar case, omission and intensity are newsworthiness-driven, while all ideological variation in slant is generated by framing alone.

Full credibility: $\alpha = 1$. When $\alpha = 1$, the editorial cost vanishes and outlets bear only the credibility cost of deviating from the truth. From Proposition 3.1, equilibrium tone becomes

$$\tau_{js}^* = \theta_s + \eta A_j,$$

so tone tracks the story's true valence one-for-one, up to the outlet-specific competitive shift ηA_j. The editorial position ψ_j disappears entirely from tone: outlets cannot impose their ideological line on how stories are narrated. The only framing variation across outlets is the common competitive premium ηA_j, which shifts all tones by the same constant and is therefore story-invariant by Proposition 3.4.

Net benefit in this case is

$$N_{js} = \rho B_j n_s + 2\rho\lambda A_j \theta_s + \frac{2\rho^2\lambda^2}{\kappa} A_j^2,$$

which is linear in ideological position θ_s. If $A_j > 0$, net benefit is strictly increasing in θ_s: the outlet covers more right-leaning stories and progressively drops left-leaning ones as their ideological distance grows. Coverage is therefore *monotone* in story ideology rather than concave, unlike the interior case $0 < \alpha < 1$ where the omission region is symmetric around an outlet-specific preferred position (Corollary 3.2). In this polar case, all cross-outlet differences in slant are generated by omission and intensity, since framing differences are limited to the story-invariant competitive shift $\eta(A_j - A_{j'})$.

3.3 Context, Data and Descriptive Statistics

3.3.1 Context

The product we study is the evening edition of the main TV newscast. During the incumbent period, four outlets compete in the same slot starting at 21h: the public broadcaster TVE and two private channels (Antena 3 and Telecinco). La Sexta broadcasts its show one hour before, at 20h.[7] The period considered in this analysis spans September 2023 until January 2024.

[7]We abstain from modelling competition through different time slots; that is, as considered in the theory model, viewers are available from 20h on and decide which outlet they want to consume news on.

3.3.2 Data

We build story-level transcript measures following the pipeline in Menéndez (2025). The key output is a set of *stories*—contiguous segments of a broadcast that discuss the same event.

On each day d, channel j produces a set of stories $\mathcal{S}_{jd}$, indexed by s.[8] Empirically, stories are obtained by clustering caption transcripts using BERTopic. We feed the full text of each story to a large language model (ChatGPT) to ensure that classification is based on sufficient context. This procedure yields 20,674 distinct stories at the channel day level in the sample.

Each broadcast is transcribed using automated speech recognition and segmented into contiguous story-level units by clustering caption text with BERTopic, a topic model that assigns segments to coherent thematic clusters. Stories spanning multiple segments on the same event are merged, and segments not attributable to a political story are dropped. The tone and political valence of each story are then labeled by querying GPT-4 with the full story text, prompting it to classify tone as positive, negative, or neutral, and — for non-neutral stories — to identify whether the tone applies to the left or right political bloc. Inter-rater reliability checks are reported in Menéndez (2025).

For each $s \in \mathcal{S}_{jd}$, there is a tone $\tau(s) \in \{-1, 0, 1\}$,, where $\tau(s) = 1$ ($\tau(s) = -1$) denotes positive (negative) tone and $\tau(s) = 0$ denotes neutral tone. Stories with a non-neutral tone also get assigned a label $p(s) \in \{L, R\}$ indicating whether the tone applies to the left or right bloc.

We construct its empirical counterpart of Eq. 3.2 as:

$$\begin{aligned}\hat{\tau}_{jd}(s) \equiv &\mathbb{1}\{\tau(s) = +1, p(s) = R\} + \mathbb{1}\{\tau(s) = -1, p(s) = L\} - \\ &-\mathbb{1}\{\tau(s) = +1, p(s) = L\} - \mathbb{1}\{\tau(s) = -1, p(s) = R\},\end{aligned} \tag{3.4}$$

which equals $+1$ when the story is favorable to the right bloc or unfavorable to the left, -1 in the opposite case, and 0 for neutral stories. The empirical coverage share is simple the proportion of minutes devoted to each story in the bulletin:

[8] We restrict attention only to the subset of political stories. We flag a story as political if it mentions national parties or prominent politicians, or contains general political keywords. This leaves unmodeled the intensity margin of producing other type of content.

$$\hat{c}_{jd}(s) \equiv \frac{\ell_j(s)}{L_{jd}}, \qquad L_{jd} \equiv \sum_{s \in \mathcal{S}_{jd}} \ell_j(s), \tag{3.5}$$

which satisfies $\sum_{s \in \mathcal{S}_{jd}} c_{jd}(s) = 1$ by construction. Then, the estimate of the slant index for outlet in Eq. 3.1 is:

$$\hat{\mathfrak{s}}_j = \frac{1}{D} \sum_{d=1}^{D} \sum_{s \in \mathcal{S}_{jd}} \hat{c}_{jd}(s) \hat{\tau}_{jd}(s), \tag{3.6}$$

where D refers to the total number of days in the sample.

3.3.3 Descriptive Statistics

Table 3.1 presents descriptive statistics for each channel. All four outlets exhibit negative slant indices, indicating a net left-leaning tone on average. There is, however, substantial heterogeneity across channels in both the extent of slant and the way political coverage is organized. TVE and La Sexta display the most negative slant index, while Telecinco and Antena 3 are closer to the center. Channels also differ markedly in the number of distinct political stories they cover and in the intensity with which they cover them. La Sexta and TVE cover the largest number of stories, whereas Telecinco covers fewer stories but assigns a higher average coverage share per story-day observation, consistent with a more concentrated allocation of airtime. By contrast, Antena 3 and La Sexta devote more total minutes per day to political news, suggesting broader or more sustained political coverage.

Table 3.1: Descriptive statistics by channel

Channel	Stories	$\hat{\mathfrak{s}}_j$	Avg. c_{jd}	Avg. L_{jd}
TVE	167	-0.1975	0.1656	8.6
Antena 3	126	-0.0307	0.1467	10.7
Telecinco	113	-0.0588	0.2271	5.5
La Sexta	163	-0.1296	0.1431	10.3
$D = 104$ days				

Notes: Days are restricted to those where all outlets are present. Stories is the total number of distinct political stories covered by the channel in the period. $\hat{\mathfrak{s}}_j$ is the slant index (Eq. 3.6). Avg. c_{jd} is the mean coverage share per story-day observation. Avg. L_{jd} is the mean daily political airtime in minutes.

From Eq. 3.2, tone is divided in two components: a story effect θ_s and

outlet effect. Assuming there is measurement error to incorporate a residual component, we can do a two-way fixed-effects estimation to see how each of them contribute to the tone before pooling days. Table 3.2 presents the implied variance decomposition.

Table 3.2: Variance decomposition of tone in the two-way fixed-effects regression

Component	Variance	Share of total variance
Story fixed effects,	0.1302	52.96%
Outlet fixed effects,	0.0029	1.16%
Twice covariance	-0.0008	-0.33%
Fitted component	0.1322	53.79%
Residual, ε_{js}	0.1136	46.21%
Total variance of τ_{js}	0.2458	100.00%

Notes: The table reports the variance decomposition from a TWFE estimation on the empirical tone in Eq. 3.4.

Story fixed effects account for about 53% of total variation in tone, indicating that differences across stories are an important source of heterogeneity. Outlet fixed effects, by contrast, explain only about 1% of the variance, suggesting that unconditional differences in tone across channels are relatively small. At the same time, the residual remains sizable, capturing close to 46% of total variance. This residual component may reflect measurement error, but also outlet-specific reactions to particular stories that are not captured by the additive specification. Taken together, these tables suggest that, while story identity is the main source of tone variation, there is substantial outlet-specific narration of common stories — precisely what the Shapley decomposition in Section 3.4 aims to quantify.

3.4 Statistical Decomposition of Slant

Before turning to structural estimation, we present a model-free decomposition of each outlet's slant into the three mechanisms identified in Section 3.2: omission, framing, and intensity.

Because these margins interact mechanically, the contribution of any one mechanism depends on the order in which the others are shut down. We therefore use Shapley values, following Shorrocks (2013), and treat omission, framing, and intensity as players in a cooperative game whose total surplus is the observed slant

gap. This yields an exact additive decomposition that averages each mechanism's marginal contribution over all possible elimination orders, thereby avoiding the path-dependence problems that arise in sequential decompositions.

3.4.1 Setup

Let j index outlets and s index stories. Define c_{js} as the coverage share outlet j devotes to story s over the sample period, and τ_{js} as its average tone. Let $\mathcal{S}_j$ be the set of political stories covered by outlet j, $\mathcal{C} = \bigcap_j \mathcal{P}_j$ the *common set* covered by all outlets, and $\mathcal{S} = \bigcup_j \mathcal{S}_j$ the full set covered by at least one. The cross-outlet consensus tone on story s is

$$\bar{\tau}_s = \frac{1}{J} \sum_j \tau_{js},$$

and the consensus coverage share is

$$\bar{c}_s = \frac{1}{J} \sum_j c_{js}.$$

For each outlet j, define the counterfactual slant $\mathfrak{s}_j^{ofi}$, where each index equals 1 if the corresponding mechanism is outlet-specific and 0 if it is replaced by the market consensus: o for omission, f for framing, and i for intensity. The eight counterfactual slant objects are

- $\mathfrak{s}^{000} = \sum_{s \in \mathcal{C}} \bar{c}_s \bar{\tau}_s$ (global benchmark: consensus coverage and consensus tone)
- $\mathfrak{s}_j^{010} = \sum_{s \in \mathcal{C}} \bar{c}_s \tau_{js}$ (outlet framing, consensus coverage on common stories)
- $\mathfrak{s}_j^{001} = \sum_{s \in \mathcal{C}} c_{js} \bar{\tau}_s$ (outlet intensity on common stories, consensus tone)
- $\mathfrak{s}_j^{011} = \sum_{s \in \mathcal{C}} c_{js} \tau_{js}$ (outlet slant on the common set)
- $\mathfrak{s}_j^{100} = \sum_{s \in \mathcal{S}} c_s, \bar{\tau}_s$ (market story selection, consensus tone and consensus intensity)
- $\mathfrak{s}_j^{110} = \sum_{s \in \mathcal{S}} c_s, \tau_{js}$ (selection and framing)
- $\mathfrak{s}_j^{101} = \sum_{s \in \mathcal{S}} c_{js} \bar{\tau}_s$ (selection and intensity)

- $\mathfrak{s}_j^{111} = \sum_{s \in \mathcal{S}} c_{js}\tau_{js} \equiv \mathfrak{s}_j$ (observed full slant)

The Shapley value of each mechanism is its average marginal contribution across all six orderings of the three players:

$$\begin{aligned}
\phi_j^O &= \frac{1}{3}\left(\mathfrak{s}_j^{111} - \mathfrak{s}_j^{011}\right) + \frac{1}{6}\left(\mathfrak{s}_j^{110} - \mathfrak{s}_j^{010}\right) + \frac{1}{6}\left(\mathfrak{s}_j^{101} - \mathfrak{s}_j^{001}\right) + \frac{1}{3}\left(\mathfrak{s}_j^{100} - \mathfrak{s}^{000}\right), \\
\phi_j^F &= \frac{1}{3}\left(\mathfrak{s}_j^{111} - \mathfrak{s}_j^{101}\right) + \frac{1}{6}\left(\mathfrak{s}_j^{110} - \mathfrak{s}_j^{100}\right) + \frac{1}{6}\left(\mathfrak{s}_j^{011} - \mathfrak{s}_j^{001}\right) + \frac{1}{3}\left(\mathfrak{s}_j^{010} - \mathfrak{s}^{000}\right), \\
\phi_j^I &= \frac{1}{3}\left(\mathfrak{s}_j^{111} - \mathfrak{s}_j^{110}\right) + \frac{1}{6}\left(\mathfrak{s}_j^{101} - \mathfrak{s}_j^{100}\right) + \frac{1}{6}\left(\mathfrak{s}_j^{011} - \mathfrak{s}_j^{010}\right) + \frac{1}{3}\left(\mathfrak{s}_j^{001} - \mathfrak{s}^{000}\right).
\end{aligned} \tag{3.7}$$

By construction,

$$\phi_j^O + \phi_j^F + \phi_j^I = \mathfrak{s}_j^{111} - \mathfrak{s}^{000}.$$

Since the tone of the stories that are omitted is not observed empirically, the Shapley value of omission is calculated among the set of common stories $\mathcal{C}$. This means that we decompose omission with the degree of specialization of an outlet.

3.4.2 Results

Table 3.3 reports the Shapley decomposition of each outlet's slant into omission, framing, and intensity.

Table 3.3: Shapley decomposition of slant by channel

	ϕ^O	ϕ^F	ϕ^I	$Share^O$	$Share^F$	$Share^I$
Antena 3	−0.007	+0.058	−0.001	10.0%	88.1%	2.0%
La Sexta	−0.015	−0.021	−0.006	35.8%	49.8%	14.3%
Telecinco	−0.008	+0.040	+0.001	16.6%	81.3%	2.0%
TVE	−0.025	−0.079	−0.008	22.1%	70.8%	7.0%

Notes: The table shows the estimates of the Shapley factor decomposition. The share of each mechanism m equals $|\phi^m|/(|\phi^O| + |\phi^F| + |\phi^I|)$.

A clear pattern emerges: slant is primarily driven by framing across all outlets. For Antena 3 and Telecinco, framing accounts for more than 80% of total slant, indicating that differences arise mainly from how common political stories are narrated rather than from differences in coverage. TVE displays a similar pattern, with framing explaining roughly 70% of its slant.

Omission nonetheless plays a quantitatively important role for some outlets. In particular, it accounts for more than one third of slant at La Sexta and about one fifth at TVE, suggesting that selective coverage is a key margin through which these outlets shape their ideological profile. By contrast, coverage intensity contributes relatively little across all channels, never exceeding 15%.

Taken together, these results show that while framing is the dominant mechanism overall, outlets differ in how slant is generated. In particular, La Sexta stands out as relatively more omission-driven, highlighting that similar levels of slant can arise from different editorial strategies. A limitation of the empirical decomposition is that, with archival data, common cross-outlet responses to the same news event are not separately identified. When all channels react similarly to a given story, that common variation is absorbed by the story component rather than loaded onto framing or intensity. The decomposition therefore captures outlet-specific deviations in the treatment of common stories, not the full extent of editorial adjustment. Under this interpretation, the estimated shares of framing and intensity are best viewed as conservative, and may constitute a lower bound on the underlying importance of these mechanisms when outlets move together.

3.5 Conclusion

This paper studies how news outlets generate ideological slant and, crucially, through which editorial margins that slant is produced. We develop a model in which outlets choose story selection, narrative tone, and airtime allocation, and use it to organize slant into three distinct mechanisms: omission, framing, and coverage intensity. We then bring this framework to transcript-level data from Spanish prime-time television news.

Our findings show that framing is the dominant driver of slant for most outlets. Differences across channels arise primarily from how common political stories are narrated, rather than from large differences in story selection or airtime allocation. Omission is nonetheless quantitatively important for a subset of outlets, indicating that selective coverage can play a central role even when overall slant is moderate. By contrast, coverage intensity contributes relatively little on average. Taken together, these results highlight that similar levels of

observed slant can be generated through different editorial strategies, and that distinguishing between margins is key to understanding how ideological distortion emerges in news markets.

More broadly, the analysis underscores that media bias is inherently multi-dimensional. Focusing solely on aggregate measures of slant may obscure the mechanisms through which outlets shape information environments, and therefore the ways in which audiences are exposed to ideological content.

A natural next step is to embed the theoretical framework in a structural estimation that recovers the primitives governing audience demand and editorial costs. This would allow one to move beyond the reduced-form decomposition and quantify the latent ideological positions of both stories and outlets, as well as to evaluate counterfactual changes in the media environment.

Bibliography

Luis Abreu and Doh-Shin Jeon. Homophily in social media and news polarization. working paper, available at *https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3468416*, 2019.

Daron Acemoglu and Andrew F. Newman. The labor market and corporate structure. *European Economic Review*, 46(10):1733–1756, 2002. ISSN 0014-2921. doi: 10.1016/S0014-2921(01)00199-4. URL https://doi.org/10.1016/S0014-2921(01)00199-4.

Daron Acemoglu, Asuman Ozdaglar, and James Siderius. A model of online misinformation. *The Review of Economic Studies*, 91(6):3117–3150, 2024. doi: 10.1093/restud/rdad111.

ACFE. Occupational fraud 2024: Report to the nations. https://www.acfe.com/report-to-the-nations/2024/, 2024. Accessed: 2025-08-11.

Armen A. Alchian and Harold Demsetz. Production, information costs, and economic organization. *The American Economic Review*, 62(5):777–795, 1972.

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–76, 2020.

Hunt Allcott, Matthew Gentzkow, and Lena Song. Digital addiction. *American Economic Review*, 112(7):2424–2463, 2022.

Guy Aridor, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. The economics of social media. 2024.

Luis Armona. Online social network effects in labor markets: Evidence from facebook's entry to college campuses. *Review of Economics and Statistics*, pages 1–47, 2023.

Martino Banchio, Francesco Decarolis, Rafael Jiménez-Durán, Carl-Christian Groh, and Miguel Risco. Contestability and the optimal regulation of social media platforms. working paper, 2025.

Oriana Bandiera, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. The allocation of authority in organizations: A field experiment with bureaucrats. *Quarterly Journal of Economics*, 136(4):2195–2242, 2021. doi: 10.1093/qje/qjab029. URL https://doi.org/10.1093/qje/qjab029.

Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.

Basel Committee on Banking Supervision. Principles for the sound management of operational risk. Technical Report BCBS 195, Bank for International Settlements, Basel, jun 2011. URL https://www.bis.org/publ/bcbsc102.pdf. Includes guidance on dual control and segregation of duties (maker–checker principle).

George Beknazar-Yuzbashev, Rafael Jiménez-Durán, and Mateusz Stalinski. A model of harmful yet engaging content on social media. In *AEA Papers and Proceedings*, volume 114, pages 678–683. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2024.

Paul Belleflamme and Martin Peitz. The competitive impacts of exclusivity and price transparency in markets with digital platforms. *Concurrences*, (1):2–12, 2020. URL https://EconPapers.repec.org/RePEc:cor:louvco:2019019.

B. Douglas Bernheim. A theory of conformity. *Journal of Political Economy*, 102 (5):841–877, 1994.

B. Douglas Bernheim and Michael D. Whinston. Common agency. *Econometrica*, 54(4):923–942, jul 1986. doi: 10.2307/1912827.

J. Aislinn Bohren and Daniel N. Hauser. Learning with heterogeneous misspecified models: Characterization and robustness. *Econometrica*, 89(6):3025–3077, 2021. doi: 10.3982/ECTA15318.

Marc Bourreau and Jan Kraemer. Interoperability in digital markets: Boon or bane for market contestability? working paper, available at https://ssrn.com/abstract=4172255, 2023.

Marc Bourreau, Adrien Raizonville, and Guillaume Thébaudin. Interoperability between ad-financed platforms with endogenous multi-homing. 2023.

Luca Braghieri, Ro'ee Levy, and Alexey Makarin. Social media and mental health. *American Economic Review*, 112(11):3660–3693, 2022.

Leonardo Bursztyn, Benjamin Handel, Rafael Jiménez-Durán, and Christopher Roth. When product markets become collective traps: the case of social media. working paper, available at https://ssrn.com/abstract=4596071, 2023.

Guillermo A. Calvo and Stanislaw Wellisz. Supervision, loss of control, and the optimum size of the firm. *Journal of Political Economy*, 86(5):943–952, 1978. URL https://www.jstor.org/stable/1828417.

Guillermo A. Calvo and Stanislaw Wellisz. Hierarchy, ability, and income distribution. *Journal of Political Economy*, 87(5):991–1010, 1979. URL https://www.jstor.org/stable/1833180.

Christophe Chamley. *Rational herds: Economic models of social learning*. Cambridge University Press, 2004.

Dongkyu Chang and Allen Vong. Perverse ethical concerns: Misinformation and coordination. *Journal of Economic Theory*, 226:106011, 2025. doi: 10.1016/j.jet.2025.106011.

Cheng Chen. Management quality and firm hierarchy in industry equilibrium. *American Economic Journal: Microeconomics*, 9(4):203–244, 2017. doi: 10.1257/mic.20160305. URL https://www.aeaweb.org/articles?id=10.1257/mic.20160305.

Cheng Chen and Wing Suen. The comparative statics of optimal hierarchies. *American Economic Journal: Microeconomics*, 11(2):1–25, 2019. URL https://www.jstor.org/stable/26641414.

Robert B. Cialdini and Noah J. Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55(1):591–621, 2004.

Olivier Compte. Redundancy and conformity in organizations. *European Economic Review*, 39(2):275–292, 1995. doi: 10.1016/0014-2921(94)00052-O.

Mudit Dhakar and Jun Yan. Interoperability & privacy: A case of messaging apps. 2024.

John DiNardo, Nicole M. Fortin, and Thomas Lemieux. Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044, 1996.

District of Columbia Public Schools. Impact: The dcps evaluation and feedback system for school-based personnel, 2023. URL https://dcps.dc.gov/page/impact-dcps-evaluation-and-feedback-system-school-based-personnel. Accessed: 2025-08-22.

Mikhail Drugov. Competition in bureaucracy and corruption. *Journal of Development Economics*, 92(2):107–114, 2010. doi: 10.1016/j.jdeveco.2009.02.004. URL https://doi.org/10.1016/j.jdeveco.2009.02.004.

Esther Duflo, Michael Greenstone, Rohini Pande, and Nicholas Ryan. Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from india. *Quarterly Journal of Economics*, 128(4):1499–1545, 2013. doi: 10.1093/qje/qjt024. URL https://doi.org/10.1093/qje/qjt024.

Matthew Ellman and Fabrizio Germano. What do the papers sell? a model of advertising and media bias. *Economic Journal*, 119(537):680–704, 2009.

Jingyu Fan. Corruption networks. *International Economic Review*, 2022. URL https://jingyufan.net/talks/. Revise & resubmit.

Federal Deposit Insurance Corporation. Examination policies manual: Section 4.2 – internal routine and controls. Federal Deposit Insurance Corporation, 2014. URL https://www.fdic.gov/resources/supervision-and-examinations/examination-policies-manual/section4-2.pdf. Accessed: 2025-08-21.

Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4A, pages 1–102. Elsevier, 2011.

Jens-Uwe Franck and Martin Peitz. Market power of digital platforms. *Oxford Review of Economic Policy*, 39(1):34–46, 2023.

Andrea Galeotti, Christian Ghiglino, and Francesco Squintani. Strategic information transmission networks. *Journal of Economic Theory*, 148(5):1751–1769, 2013. doi: 10.1016/j.jet.2013.04.016.

Filomena Garcia and Muxin Li. Post-merger strategies of multiplatform monopolies. *mimeo*, 2024.

Luis Garicano. Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5):874–904, 2000. doi: 10.1086/317710.

Luis Garicano and Esteban Rossi-Hansberg. Inequality and the organization of knowledge. *American Economic Review*, 94(2):197–202, 2004. doi: 10.1257/0002828041302037.

Luis Garicano and Esteban Rossi-Hansberg. Organization and inequality in a knowledge economy. *Quarterly Journal of Economics*, 121(4):1383–1435, 2006. doi: 10.1093/qje/121.4.1383. URL https://doi.org/10.1093/qje/121.4.1383.

Germain Gauthier, Roland Hodler, Ekaterina Zhuravskaya, and Philine Widmer. The political effects of x's recommender algorithm. working paper, 2025.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010a.

Matthew Gentzkow and Jesse M. Shapiro. What drives media slant? evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71, 2010b.

Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. Media bias in the marketplace: Theory. NBER Working Paper 19880, National Bureau of Economic Research, 2014. URL https://www.nber.org/papers/w19880.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340, July 2019. doi: 10.3982/ECTA16566. URL https://doi.org/10.3982/ECTA16566.

Jacob Gersen. Overlapping and underlapping jurisdiction in administrative law. *University of Chicago Public Law & Legal Theory Working Papers*, 2007. Explores how overlapping authority can structure administrative incentives.

Benjamin Golub and Matthew O. Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1): 112–149, 2010.

Steven Greene. Social identity theory and party identification. *Social science quarterly*, 85(1):136–153, 2004.

Tim Groseclose and Jeffrey Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237, 2005.

Maria Guadalupe and Julie Wulf. The flattening firm and product market competition: The effect of trade liberalization on corporate hierarchies. *American Economic Journal: Applied Economics*, 2(4):105–27, oct 2010. doi: 10.1257/app.2.4.105. URL https://www.aeaweb.org/articles?id=10.1257/app.2.4.105.

Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023.

Sergei Guriev, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya. Curtailing false news, amplifying truth. working paper, available at https://shs.hal.science/halshs-04315924/document, 2023.

John A. Hartigan. *Bayes Theory*. Springer Science & Business Media, 1983.

David Holtz, Ben Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral. The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 75–76, 2020.

Ruru Hoong. Self control and smartphone use: An experimental study of soft commitment devices. *European Economic Review*, 140:103924, 2021.

Jeff Horwitz et al. The facebook files. *The Wall Street Journal,* available online at: https://www.wsj.com/articles/the-facebook-files-11631713039, 2021.

Chang-Tai Hsieh and Benjamin A. Olken. The missing "missing middle". *Journal of Economic Perspectives*, 28(3):89–108, 2014.

Yannis M. Ioannides. Complexity and organizational architecture. *Mathematical Social Sciences*, 64(2):193–202, 2012. doi: 10.1016/j.mathsocsci.2012.04.005.

Matthew O. Jackson, Suraj Malladi, and David McAdams. Learning through the grapevine and the impact of the breadth and depth of social networks. *Proceedings of the National Academy of Sciences*, 119(34), 2022.

Michael Kades and Fiona Scott Morton. Interoperability as a competition remedy for digital networks. *Washington Center for Equitable Growth Working Paper Series*, 2020.

Krishna Kamath, Aneesh Sharma, Dong Wang, and Zhijun Yin. Realgraph: User interaction prediction at twitter. In *user engagement optimization workshop@ KDD*, number ii, 2014.

Mark J. Keightley and Mark Jickling. Banking regulation: Overview of regulatory agencies, September 2017. URL https://sgp.fas.org/crs/misc/R44918.pdf. Accessed: 2025-08-22.

Michael Keren and David Levhari. The internal organization of the firm and the shape of average costs. *The Bell Journal of Economics*, 14(2):474–486, 1983. URL http://www.jstor.org/stable/3003648.

Brent Kitchens, Steven L Johnson, and Peter Gray. Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS quarterly*, 44(4):1619–1649, 2020.

Kevin W. Kobelsky. A conceptual model for segregation of duties: Integrating theory and practice for manual and it-supported processes. *International Journal of Accounting Information Systems*, 15(4):304–322, 2014. ISSN 1467-0895. doi: https://doi.org/10.1016/j.accinf.2014.05.003. URL https://www.sciencedirect.com/science/article/pii/S1467089514000293.

Fred Kofman and Jacques Lawarrée. A prisoner's dilemma model of collusion deterrence. *Journal of Public Economics*, 59(1):117–136, 1996. URL https://doi.org/10.1016/0047-2727(95)01404-2.

Rachel Kranton and David McAdams. Social connectedness and information markets. working paper, available at https://sites.duke.edu/rachelkranton/

files/2022/12/Social_Connectedness___Information_Markets-Dec-17_2022-final.pdf, 2022.

Jean-Jacques Laffont and David Martimort. Collusion and delegation. *RAND Journal of Economics*, 31(2):280–305, 2000. doi: 10.2307/2601046.

Jean-Jacques Laffont and Jean Tirole. The politics of government decision-making: A theory of regulatory capture. *Quarterly Journal of Economics*, 106(4):1089–1127, 1991. doi: 10.2307/2937958. URL https://doi.org/10.2307/2937958.

Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

Gilat Levy and Ronny Razin. Religious beliefs, religious participation, and cooperation. *American Economic Journal: Microeconomics*, 4(3):121–151, 2012. doi: 10.1257/mic.4.3.121.

Ro'ee Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–870, 2021.

Sahil Loomba, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature Human Behaviour*, 5(3):337–348, 2021.

Rocco Macchiavello. Financial development and vertical integration: Theory and evidence. *Journal of the European Economic Association*, 10(2):255–289, 2012. doi: 10.1111/j.1542-4774.2011.01042.x. URL https://doi.org/10.1111/j.1542-4774.2011.01042.x.

John Paul MacDuffie. Human resource bundles and manufacturing performance: Organizational logic and flexible production systems in the world auto industry. *Industrial and Labor Relations Review*, 48(2):197–221, 1995. doi: 10.1177/001979399504800201. URL https://doi.org/10.1177/001979399504800201.

Luis Menéndez. The impact of political campaigns on demand for partisan news. Job Market Paper, Universitat Autònoma de Barcelona, CSIC and Barcelona

School of Economics, December 2025. URL https://luismenendezgarcia.com.

Luca P. Merlino, Paolo Pin, and Nicole Tabasso. Debunking rumors in networks. *American Economic Journal: Microeconomics*, 15(1):467–496, 2023. doi: 10.1257/mic.20200403.

Mohsen Mosleh, Cameron Martel, Dean Eckles, and David G. Rand. Shared partisanship dramatically increases social tie formation in a twitter field experiment. *Proceedings of the National Academy of Sciences*, 118(7):e2022761118, 2021.

Elchanan Mossel, Allan Sly, and Omer Tamuz. Strategic learning and the topology of social networks. *Econometrica*, 83(5):1755–1794, 2015. doi: 10.3982/ECTA12058.

Manuel Mueller-Frank, Mallesh M. Pai, Carlo Reggini, Alejandro Saporiti, and Luis Simantujak. Strategic management of social information. working paper, available at *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4434685*, 2022.

Nieman Lab. Trusted news sites may benefit in an internet full of ai-generated fakes, a new study finds. *Nieman Journalism Lab*, 2025. URL https://www.niemanlab.org/2025/08/trusted-news-sites-may-benefit-in-an-internet-full-of-ai-generated-fakes-a-new Accessed: 2026-03-17.

OECD. *State-Owned Enterprises and Corruption: What Are the Risks and What Can Be Done?* OECD Publishing, Paris, 2018. doi: 10.1787/9789264303058-en. URL https://doi.org/10.1787/9789264303058-en.

Office of the Comptroller of the Currency. Comptroller's handbook: Custody services. Technical report, U.S. Department of the Treasury, 2012. States that "procedures should require dual control in processing of all custody assets, including securities, cash, income payments, and corporate actions.".

Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

Yingyi Qian. Incentives and loss of control in an optimal hierarchy. *The Review of Economic Studies*, 61(3):527–544, 1994. ISSN 0034-6527. doi: 10.2307/2297902. URL https://doi.org/10.2307/2297902.

David Rahman. But who will monitor the monitor? *American Economic Review*, 102(6):2767–2797, 2012. doi: 10.1257/aer.102.6.2767. URL https://www.aeaweb.org/articles?id=10.1257/aer.102.6.2767.

Reserve Bank of India. Implementation of recommendations of the committee on customer service in banks. Notification, Department of Banking Operations and Development, DBOD.No.Leg.BC.104/09.07.005/2010-11, may 2011. URL https://www.rbi.org.in/commonman/english/scripts/Notification.aspx?Id=940. Mandates the adoption of maker–checker (dual control) procedures in bank branches.

Jonathan Reuter and Eric Zitzewitz. Do ads influence editors? advertising and bias in the financial media. *Quarterly Journal of Economics*, 121(1):197–227, 2006.

Susan Rose-Ackerman. *Corruption: A Study in Political Economy*. Academic Press, 1978.

Lee Ross, David Greene, and Pamela House. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301, 1977. ISSN 0022-1031. doi: https://doi.org/10.1016/0022-1031(77)90049-X. URL https://www.sciencedirect.com/science/article/pii/002210317790049X.

Raaj Kumar Sah and Joseph E. Stiglitz. The architecture of economic systems: Hierarchies and polyarchies. *The American Economic Review*, 76(4):716–727, 1986. ISSN 00028282. URL http://www.jstor.org/stable/1806069.

Carl Shapiro and Joseph E. Stiglitz. Equilibrium unemployment as a worker discipline device. *American Economic Review*, 74(3):433–444, 1984. URL https://www.jstor.org/stable/1804018.

Andrei Shleifer and Robert W. Vishny. Corruption. *Quarterly Journal of Economics*, 108(3):599–617, 1993. doi: 10.2307/2118402.

Anthony F. Shorrocks. Decomposition procedures for distributional analysis: A unified framework based on the shapley value. *Journal of Economic Inequality*, 11(1):99–126, 2013. doi: 10.1007/s10888-011-9214-z.

Craig Silverman. This analysis shows how viral fake election news stories outperformed real news on facebook. *BuzzFeed*, available at https://newsmediauk.org/wp-content/uploads/2022/10/Buzzfeed-_This_Analysis_Shows_How_Viral_Fake_Election_News_Stories_Outperformed_Real_News_On_Facebook_-_BuzzFeed_News.pdf, November 16, 2016.

Jonathan B. Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008.

Jae Song, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till von Wachter. Firming up inequality. *Quarterly Journal of Economics*, 134(1):1–50, 2019.

Cass R. Sunstein. A prison of our own design: Divided democracy in the age of social media. *Democratic Audit UK*, 2017.

Matt Taddy. Measuring political sentiment on Twitter: Factor-optimal design for multinomial inverse regression. *Technometrics*, 55(4):415–425, 2013.

Jean Tirole. Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, and Organization*, 2(2):181–214, 1986. doi: 10.1093/jleo/2.2.181. URL https://doi.org/10.1093/oxfordjournals.jleo.a036907.

James R. Tybout. Manufacturing firms in developing countries: How well do they do, and why? *Journal of Economic Literature*, 38(1):11–44, 2000. doi: 10.1257/jel.38.1.11. URL https://www.aeaweb.org/articles?id=10.1257/jel.38.1.11.

Oliver E. Williamson. Hierarchical control and optimum firm size. *Journal of Political Economy*, 75(2):123–138, 1967.

World Bank. Enhancing government effectiveness and transparency: The fight against corruption. https://

documents1.worldbank.org/curated/en/646781468330980665/pdf/Governance-and-anti-corruption-ways-to-enhance-the-World-Banks-impact.pdf, 2006. Accessed: 2025-08-11.

Appendix A

Appendix to Chapter 1

A.1 One monitor, CC_{HH} and PC_M binding

Proposition A.1 (Shape of the optimal span in the CC + PCM regime)**.** *Fix* $\Delta\theta \in (0, \frac{1}{2})$, $p_0 \in (0,1)$ *and* $\bar{V} > 0$. *In the intermediate regime where the collusion constraint and the monitor participation constraint bind, the principal's profit (relative to the no–monitor baseline* $\Pi_{\varnothing M} = \frac{1}{2} + \theta_L + \frac{\Delta\theta^2}{2}$*) as a function of* $\mu \in (0,1]$ *is*

$$\begin{aligned}\Delta\theta\Pi(\mu) =& \tfrac{1}{2}\mu\left[\Delta\theta(1-\Delta\theta) - \tfrac{1}{2}\Big(1-p(\mu)\Big)\left(1-\tfrac{\Delta\theta}{2}\right)^2\right] \\ &+ \left(\tfrac{1}{\Delta\theta p(\mu)} - \tfrac{3}{2p(\mu)} + \tfrac{1}{2} - \tfrac{1}{\Delta}\theta\right)\bar{V} + \left(\tfrac{1}{\Delta\theta^2\mu p(\mu)} - \tfrac{1}{\Delta\theta^2\mu p(\mu)^2}\right)\bar{V}^2,\end{aligned}$$

with $p(\mu) = p_0 e^{-\lambda\mu}$. *Let* $\mu^*(\lambda) \in (0,1]$ *maximize* $\Delta\theta\Pi(\mu)$ *(assumed interior when stated). Then:*

*(i) (*Eventual collapse of span*) As* $\lambda \to \infty$,

$$\mu^*(\lambda) \longrightarrow 0.$$

More precisely, there exists $x^* \in (0,\infty)$ *(depending on parameters) such that* $\mu^*(\lambda) = x^*/\lambda + o(1/\lambda)$.

*(ii) (*Non-monotonicity in λ*) There exist parameter values and an interval of* λ *contained in the CC + PCM regime for which* $\mu^*(\lambda)$ *is non-monotonic in* λ *(it increases for some range of* λ *and decreases for larger* λ*). In particular, for* $\Delta\theta < \frac{2}{3}$ *and for moderate* $\bar{V}$ *one has* $d\mu^*/d\lambda > 0$ *at some interior optimum, while for all sufficiently large* λ *one has* $d\mu^*/d\lambda < 0$*; by*

continuity, $\mu^*(\lambda)$ must turn at least once.

A.2 Two monitors, all constraints binding.

We are in the two–monitor environment with spans $\mu > \mu_2 \in (0, 1]$. A mass $1 - 2\mu + \mu_2$ is unmonitored, a mass $\mu - \mu_2$ is covered by exactly one monitor, and a mass μ_2 is covered by both (overlap). Detection is $p := p(\mu) = p_0 e^{-\lambda\mu}$ and reports are independent across monitors in the overlap block. In this section, I focus on the *intermediate* region for the monitoring cost $\bar{V}$, namely $\bar{V} \in [\bar{V}_3, \bar{V}_4]$. In this region the monitor Participation Constraint (PC_M) *binds* and at least one Collusion Constraint (CC) *binds*, while all low–type incentive constraints bind and the remaining constraints are slack. Intuitively, PCM pinning down the expected transfer to monitors forces the principal to use the CCs to steer the high type's behavior via truthful–report transfers, and this shows up entirely in the uninformative–state efforts $e_{2,1}$ and $e_{2,2}$ that the principal sets in the one–monitor and two–monitor blocks. The proposition below states the optimal contract in this regime and gives the closed forms of the only two endogenous efforts, $e_{2,1}$ and $e_{2,2}$, as functions of primitives and (μ, μ_2).

Proposition A.2 (Optimal two–monitor contract when CC & PCM bind)**.** *Fix $\mu > \mu_2 \in (0, 1]$ and let $p := p(\mu) = p_0 e^{-\lambda\mu}$ and $\Delta\theta \in (0, \frac{1}{2})$. Suppose $\bar{V} \in [\bar{V}_3, \bar{V}_4]$, so that the monitor Participation Constraint (PC_M) and at least one Collusion Constraint is binding, while the low–type ICs bind and the other constraints are slack. Then, an optimal contract exists and is given by:*

Efforts.

$$e_{1,1} = e_{3,1} = e_{4,1} = 1, \qquad e_{1,2} = e_{3,2} = e_{4,2} = 1,$$

and the only endogenous efforts are $e_{2,1}, e_{2,2}$ determined below.

Wages to agents.

$$W^1(x_{\theta_L,\theta_L}, \theta_L) = W^2(x_{\theta_L,\theta_L}, \theta_L) = \tfrac{1}{2}, \qquad W^1(x_{\theta_L,\varnothing}, \varnothing) = \tfrac{1}{2}e_{2,1}^2, \quad W^2(x_{\theta_L,\varnothing}, \varnothing) = \tfrac{1}{2}e_{2,2}^2,$$
$$W^1(x_{\theta_H,\theta_H}, \theta_H) = W^2(x_{\theta_H,\theta_H}, \theta_H) = \tfrac{1}{2},$$
$$W^1(x_{\theta_H,\varnothing}, \varnothing) = \tfrac{1}{2} + \Delta\theta e_{2,1} - \tfrac{\Delta\theta^2}{2}, \quad W^2(x_{\theta_H,\varnothing}, \varnothing) = \tfrac{1}{2} + \Delta\theta e_{2,2} - \tfrac{\Delta\theta^2}{2}.$$

Transfers to monitors and instruments.

$$S^1(x_{\theta_H,\varnothing},\varnothing)=S^2(x_{\theta_H,\varnothing},\varnothing)=0,\qquad S^1(x_{\theta_H,\theta_H},\theta_H)=\Delta\theta\Big(e_{2,1}-\tfrac{\Delta\theta}{2}\Big),$$

$$S^2(x_{\theta_H,\theta_H},\theta_H)=\tfrac{1}{2}\Delta\theta\Big(e_{2,2}-\tfrac{\Delta\theta}{2}\Big),\qquad \beta=0,\qquad \delta=S^2(x_{\theta_H,\theta_H},\theta_H).$$

Binding constraints.

- *Low–type ICs:* $W^j(x_{\theta_L,\varnothing},\varnothing)=\frac{1}{2}e_{2,j}^2$ *for* $j=1,2$.
- CC_{HH}, CC_{H2H}, CC_{H1H}, CC_M *bind; with the above* (S^1,S^2,δ,β) *this is equivalent to*

$$S^1(x_{\theta_H,\theta_H},\theta_H)=\Delta\theta e_{2,1}-\tfrac{\Delta\theta^2}{2},\qquad 2S^2(x_{\theta_H,\theta_H},\theta_H)=\Delta\theta e_{2,2}-\tfrac{\Delta\theta^2}{2},\qquad \delta=S^2(x_{\theta_H,\theta_H},\theta_H).$$

- *PCM binds:*

$$(\mu-\mu_2)pS^1(x_{\theta_H,\theta_H},\theta_H)+\mu_2 p\Big(S^2(x_{\theta_H,\theta_H},\theta_H)+(1-p)\delta\Big)=2\bar{V}.$$

Closed forms for $e_{2,1}$ ***and*** $e_{2,2}$***.*** *The first–order conditions together with PCM yield*

$$e_{2,1}=\frac{\dfrac{2\bar{V}}{\Delta\theta p}+(\mu-\mu_2)\dfrac{\Delta\theta}{2}+\mu_2\Big(1-\dfrac{p}{2}\Big)\Big(\dfrac{1-\Delta\theta}{1-p}+\dfrac{\Delta\theta}{2}\Big)}{(\mu-\mu_2)+\mu_2\Big(1-\dfrac{p}{2}\Big)\dfrac{2-p}{1-p}},$$

and,

$$e_{2,2}=\frac{\dfrac{2-p}{1-p}\left(\dfrac{2\bar{V}}{\Delta\theta p}+\mu_2\Big(1-\dfrac{p}{2}\Big)\dfrac{\Delta\theta}{2}\right)+\dfrac{\mu-\mu_2}{1-p}\left(\dfrac{\Delta\theta}{2}(2-p)-(1-\Delta\theta)\right)}{(\mu-\mu_2)+\mu_2\Big(1-\dfrac{p}{2}\Big)\dfrac{2-p}{1-p}}.$$

We are in the two–monitor environment with spans $\mu>\mu_2\in(0,1]$. A mass $1-2\mu+\mu_2$ is unmonitored, a mass $\mu-\mu_2$ is covered by exactly one monitor, and a mass μ_2 is covered by both (overlap). Detection is $p:=p(\mu)=p_0e^{-\lambda\mu}$ and reports are independent across monitors in the overlap block. We focus

on the *intermediate* region for the monitoring cost $\bar{V}$, namely $\bar{V} \in [\bar{V}_3, \bar{V}_4]$. In this region the monitor Participation Constraint (PC_M) *binds* and at least one Collusion Constraint (CC) *binds*, while all low–type incentive constraints bind and the remaining constraints are slack. Intuitively, PCM pinning down the expected transfer to monitors forces the principal to use the CCs to steer the high type's behavior via truthful–report transfers, and this shows up entirely in the uninformative–state efforts $e_{2,1}$ and $e_{2,2}$ that the principal sets in the one–monitor and two–monitor blocks. The proposition below states the optimal contract in this regime and gives the closed forms of the only two endogenous efforts, $e_{2,1}$ and $e_{2,2}$, as functions of primitives and (μ, μ_2).

Let $\Delta \in (0, \frac{1}{2})$ and define

$$\Phi(e) := (1-\Delta)e + \tfrac{1}{2} - \tfrac{1}{2}e^2 + \tfrac{\Delta^2}{2}.$$

In the two-monitor intermediate regime (CC and PCM binding), with $p \equiv p(\mu) = p_0 e^{-\lambda\mu}$ and $q := 1-p$, the profit can be written as

$$\Pi(\bar{V}; \mu, \mu_2) = (1-2\mu+\mu_2)\Pi_{\varnothing M} + (\mu-\mu_2)\Big[p+q\Phi(e_{2,1})\Big] + \frac{\mu_2}{2}\Big[p(2-p)+q^2\Phi(e_{2,2})\Big] - 2\bar{V}, \tag{A.1}$$

where $e_{2,1}$ and $e_{2,2}$ are the optimal uninformative-state efforts in the 1-monitor and 2-monitor blocks given by Proposition A.2, respectively. Note that $p = p(\mu)$ does not depend on μ_2. Differentiating (A.1) with respect to μ_2 gives

$$\begin{aligned}\frac{\partial \Pi}{\partial \mu_2} &= \Pi_{\varnothing M} - \Big[p + q\Phi(e_{2,1})\Big] + \frac{1}{2}\Big[p(2-p) + q^2\Phi(e_{2,2})\Big] \\ &\quad + (\mu-\mu_2)q\Phi'(e_{2,1})\frac{\partial e_{2,1}}{\partial \mu_2} + \frac{\mu_2}{2}q^2\Phi'(e_{2,2})\frac{\partial e_{2,2}}{\partial \mu_2}.\end{aligned} \tag{A.2}$$

Here $\Phi'(e) = (1-\Delta) - e$.

Because $e_{2,1}$ and $e_{2,2}$ are chosen optimally given the binding constraints in this regime, the envelope theorem (for constrained problems) implies that the terms multiplying $\partial e_{2,1}/\partial \mu_2$ and $\partial e_{2,2}/\partial \mu_2$ are zero at the optimum. Therefore, at the optimal $(e^*_{2,1}, e^*_{2,2})$ we can drop the last line in (A.2) and write the clean envelope derivative

$$\frac{\partial \Pi}{\partial \mu_2} = \Pi_{\varnothing M} - \Big[p + q\Phi(e^*_{2,1})\Big] + \frac{1}{2}\Big[p(2-p) + q^2\Phi(e^*_{2,2})\Big]. \tag{A.3}$$

Sign and the role of λ. The sign of $\partial\Pi/\partial\mu_2$ in (A.3) depends on $p = p(\mu)$ and on the optimal efforts $e^*_{2,1}, e^*_{2,2}$, which themselves depend on $(\mu, \bar{V}, \Delta, p_0, \lambda)$ through the binding CC and PCM. Crucially, λ shapes p via $p(\mu) = p_0 e^{-\lambda\mu}$ and also affects the optimal $\mu^*(\lambda)$ (interior μ^* scales roughly like $1/\lambda$ when the interior solution is feasible).

Intuitively:

- For small λ (slow decay), the optimal μ^* is large and p is high. Then the 2-monitor term $\frac{1}{2}\Big[p(2-p) + q^2\Phi(e^*_{2,2})\Big]$ dominates the 1-monitor term $p + q\Phi(e^*_{2,1})$, making $\partial\Pi/\partial\mu_2 > 0$ and favoring full overlap ($\mu_2 = \mu$).
- For large λ (fast decay), the optimal μ^* is small and p is low. The overlap advantage shrinks and the balance can flip to $\partial\Pi/\partial\mu_2 < 0$, favoring zero overlap ($\mu_2 = 0$).
- By continuity in λ and in the primitives, an interior $\mu_2^* \in (0, \mu)$ solving $\partial\Pi/\partial\mu_2 = 0$ can arise for intermediate values of λ.

Hence, the sign of $\partial\Pi/\partial\mu_2$ is not globally fixed; it is governed by the detection level $p(\mu^*(\lambda))$ induced by λ (and by the corresponding optimal $e^*_{2,1}, e^*_{2,2}$), delivering the comparative-statics pattern you conjectured.

Regime thresholds with two monitors: $\bar{V}_3$ and $\bar{V}_4$. Now, I characterize the two thresholds $\bar{V}_3 < \bar{V}_4$ that separate these regimes.

To solve for V_4, I solve the CC-only problem and denote by (μ^C, μ_2^C) an optimal pair and by $p^C = p(\mu^C)$ the associated detection. Let E^C_{tot} be the *minimal* expected transfer to monitors that implements the CC-only optimum; PCM is slack if and only if $E^C_{\text{tot}} > \bar{V}$. Hence the boundary where PCM starts binding is

$$\bar{V}_3 = E^C_{\text{tot}}.$$

Under full overlap $\mu^C = \mu_2^C = \mu^C$ one obtains the closed expression

$$\bar{V}_3(\lambda, p_0, \Delta\theta) = \mu^C p^C (2 - p^C) \cdot \frac{1}{4}, \Delta\theta \left[1 - \Delta\theta, \frac{2 + (1-p^C)^2}{2(1-p^C)^2}\right], \tag{A.4}$$

where p^C is the interior CC-only optimizer (the root of the p-only FOC for that regime).

To solve for V_4, I start from the PCM-only (non-collusion) optimum and denote by (μ^N, μ_2^N) the optimal spans and $p^N = p(\mu^N)$. In the uninformative branch the second-best choice is $e_{L\varnothing} = 1 - \Delta\theta$, so the coalition CC requires a truthful transfer per informative event

$$S_{HH}^{\text{req}} = \Delta\theta e_{L\varnothing} - \tfrac{1}{2}\Delta\theta^2 = \Delta\theta - \tfrac{3}{2}\Delta\theta^2.$$

The cheapest way to implement this at $\mu_2^N = \max\left\{0, 2\mu_1^N - 1\right\}$, yielding a minimal total expected wage

$$\begin{aligned}\mathbb{E}(S) &= \frac{1}{2}p(\mu)(\mu - \mu_2)S_{HH} + \frac{1}{2}p(\mu)\mu_2 S_{HH} \\ &= \frac{1}{2}p(\mu)\mu\Delta\theta\left(1 - \frac{3}{2}\Delta\theta\right)\end{aligned}$$

The PCM-only optimum is collusion-proof, provided $\mathbb{E}(S) = \bar{V}$, so the boundary is

$$\bar{V}_4(\lambda, p_0, \Delta\theta) = \tfrac{1}{2}\mu^N p_0 e^{-\lambda\mu^N}\left(\Delta\theta - \tfrac{3}{2}\Delta\theta^2\right). \tag{A.5}$$

A.3 Proofs

Proof of Proposition 1.1.

Proof. Multiplying the objective function by two, the Lagrangian is,

$$\begin{aligned}\mathcal{L}(e_{\theta_L,\varnothing}, e_{\theta_H,\varnothing}, x_{\theta_L,\varnothing}, x_{\theta_H,\varnothing}) = \theta_H + \theta_L + (1-q)e_{\theta_L,\varnothing} + qe_{\theta_H,\varnothing} - (1-q)W(x_{\theta_L,\varnothing},\varnothing) \\ - qW(x_{\theta_H,\varnothing},\varnothing) + \lambda_1\left(W(x_{\theta_L,\varnothing},\varnothing) - \frac{e^2_{\theta_L,\varnothing}}{2}\right) \\ + \lambda_2\left(W(x_{\theta_H,\varnothing},\varnothing) - \frac{e^2_{\theta_H,\varnothing}}{2} - W(x_{\theta_L,\varnothing},\varnothing) + \frac{(e_{\theta_L,\varnothing}-\Delta\theta)^2}{2}\right).\end{aligned}$$

The fist order conditions,

$$\frac{\partial\mathcal{L}}{\partial e_{\theta_L,\varnothing}} = (1-q) - \lambda_1 e_{\theta_L,\varnothing} + \lambda_2(e_{\theta_L,\varnothing} - \Delta\theta) = 0 \tag{A.6}$$

$$\frac{\partial\mathcal{L}}{\partial e_{\theta_H,\varnothing}} = q - \lambda_2 e_{\theta_H,\varnothing} = 0 \tag{A.7}$$

$$\frac{\partial\mathcal{L}}{\partial W(x_{\theta_L,\varnothing},\varnothing)} = -1 + q + \lambda_1 - \lambda_2 = 0 \tag{A.8}$$

$$\frac{\partial\mathcal{L}}{\partial W(x_{\theta_H,\varnothing},\varnothing)} = -q + \lambda_2 = 0 \tag{A.9}$$

From (A.9), $\lambda_2 = q$. Using the first order condition of $e_{\theta_H,\varnothing}$ we get that $e_{\theta_H,\varnothing} = 1$. From (A.8), $\lambda_1 = 1$, which implies that $W(x_{\theta_L,\varnothing},\varnothing) = \frac{e^2_{\theta_L,\varnothing}}{2}$. At the same time, from (A.6),

$$(1-q) - (1-q)e_{\theta_L,\varnothing} - q\Delta\theta = 0$$

which means that

$$1 - \frac{q\Delta\theta}{1-q} = e_{\theta_L,\varnothing}$$

which is precisely $e_{\theta_H,\varnothing} - \frac{q\Delta\theta}{1-q}$. Finally, as $\lambda_2 > 0$, ($IC_{H\varnothing}$) is binding, which means that

$$\begin{aligned}W(x_{\theta_H,\varnothing},\varnothing) &= \frac{1}{2} + \frac{e^2_{\theta_L,\varnothing}}{2} - \frac{(e_{\theta_L,\varnothing}-\Delta\theta)^2}{2} = \frac{1}{2} + \frac{\left(1-\frac{q\Delta\theta}{1-q}\right)^2}{2} - \frac{\left(\left(1-\frac{q\Delta\theta}{1-q}\right)-\Delta\theta\right)^2}{2} \\ &= \frac{1}{2} - \frac{\Delta\theta^2 - 2\left(1-\frac{q\Delta\theta}{1-q}\right)\Delta\theta}{2}\end{aligned}$$

□

Proof of Proposition 1.2.

Proof. It is useful to note that the problem of the monitor and the problem of the agents are not related by any inequality. Therefore, both can be solved separately. If $S(x_{\theta_L,\theta_L},\theta_L) = S(x_{\theta_L,\varnothing},\varnothing) = S(x_{\theta_H,\varnothing},\varnothing) = S(x_{\theta_H,\theta_H},\theta_H) = \frac{\bar{V}}{\mu}$ the participation constraint of the monitor holds. As in the cases where there is a report there is full information, setting $W(x_{\theta_L,\theta_L}) = W(x_{\theta_H,\theta_H}) = \frac{1}{2}$, $e_{\theta_L,\theta_L} = e_{\theta_H,\theta_H} = 1$ and the other variables as before gives the best contract.

□

Proof of Proposition 1.3

Proof. The expected profit function is

$$\mathbb{E}\left[\Pi(\mu,\mathbf{W},\mathbf{S})\right] = \Pi_{\varnothing M} - \bar{V} + \mu p(\mu)\frac{\Delta\theta(1-\Delta\theta)}{2}$$

whose first order condition with respect to μ is,

$$\frac{\partial\mathbb{E}\left[\Pi(\mu,\mathbf{W},\mathbf{S})\right]}{\partial\mu} = 0 \Rightarrow p(\mu) + \mu p'(\mu) = 0$$

Clearly, for $\mu = 0$ the profit function is increasing, which means that the optimal span of control is always strictly positive. If there exist $\mu \leq 1$ such that the first order condition is satisfied,

$$\mu = -\frac{p(\mu)}{p'(\mu)}$$

else, $\mu = 1$, as the first derivative of profits is always positive. A sufficient condition for μ to be a global maximum is that the function $\mu p(\mu)$ is convex everywhere, or equivalently that

$$2p'(\mu) + \mu p''(\mu) < 0, \quad \mu \in [0,1]$$

In the particular case of $p(\mu) = p_0 e^{-\lambda\mu}$, the first order condition gives

$$\mu = \frac{1}{\lambda}$$

and the second order condition is satisfied, as $p_0 e^{-\lambda\mu}$. □

Proof of Proposition 1.5

Proof. It is immediate to see from equation (1.4) that the profit function is linear and strictly decreasing on γ. Therefore γ needs be larger than zero and $2\mu - 1$. □

Proof of Proposition 1.6

Proof. By Proposition 1.5, if a profit-maximizing span of control satisfies $\mu \leq \frac{1}{2}$, it must be a global maximum. Assume first that $\gamma = 0$. Then the first-order condition yields $\mu = \frac{1}{\lambda}$, which is only feasible when $\lambda \geq 2$, so that $\mu \leq \frac{1}{2}$.

Assume then that $\lambda < 2$. The optimal span must satisfy $\mu \geq \frac{1}{2}$ and hence $\gamma = 2\mu - 1$. The profit function becomes:

$$\Pi = \mu p_0 e^{-\lambda\mu} - \frac{1}{2}(2\mu - 1)p_0^2 e^{-2\lambda\mu} + \Pi_{\varnothing M} - 2\bar{V}.$$

The first-order condition is:

$$\frac{\partial \Pi}{\partial \mu} = p_0 e^{-\lambda\mu}(1 - \lambda\mu) - p_0^2 e^{-2\lambda\mu}\left(1 - \lambda(2\mu - 1)\right).$$

Dividing by $p(\mu)$ and defining:

$$f(\mu; \lambda) = 1 - \lambda\mu - p_0 e^{-\lambda\mu}\left(1 - \lambda(2\mu - 1)\right),$$

we study when $f(\mu; \lambda) = 0$.

Assume that $\lambda = 1$. The first-order condition becomes:

$$1 - \mu - p_0 e^{-\mu}(1 - 2\mu + 1) = 0 \quad \Rightarrow \quad (1 - 2p_0 e^{-\mu})(1 - \mu) = 0.$$

This implies either $\mu = 1$ or $e^{\mu} = 2p_0$. The first term, $(1 - 2p_0 e^{-\mu})$, is increasing, so if it crosses zero, the root is a local minimum. Therefore, $\mu = 1$ is the global maximum.

Assume now that $\lambda \in (0, 1)$. We now show that the derivative is always

positive on $\left[\frac{1}{2}, 1\right]$, so the maximum occurs at $\mu = 1$. Rearranging $f(\mu; \lambda) \geq 0$:

$$p_0 \leq \frac{1 - \lambda\mu}{e^{-\lambda\mu}(1 - \lambda(2\mu - 1))}.$$

Since $1 - \lambda(2\mu - 1) \leq 1 - \lambda\mu$, we obtain:

$$p_0 \leq \frac{1 - \lambda\mu}{e^{-\lambda\mu}(1 - \lambda\mu)} = e^{\lambda\mu} \geq 1,$$

so $f(\mu; \lambda) \geq 0$ holds for all $p_0 \leq 1$, implying the maximum is at $\mu = 1$.

Finally, let's study the optimal span of control when $\lambda \in (1, 2)$. First, identify when $\mu = \frac{1}{2}$ is a local maximum. At $\mu = \frac{1}{2}$, we get:

$$f\left(\tfrac{1}{2}; \lambda\right) = 1 - \frac{\lambda}{2} - p_0 e^{-\frac{\lambda}{2}} \leq 0 \quad \Leftrightarrow \quad p_0 \geq \left(1 - \frac{\lambda}{2}\right) e^{\frac{\lambda}{2}}.$$

Define $x = 1 - \frac{\lambda}{2} \in (0, 1)$ so the inequality becomes:

$$x e^{1-x} \leq p_0.$$

Using the Lambert W function, the condition binds when:

$$x = -W_0(-e^{-1} p_0) \quad \Rightarrow \quad \lambda = 2\left(1 + W_0\left(-\frac{p_0}{e}\right)\right).$$

Hence, if $\lambda \geq 2\left(1 + W_0\left(-\frac{p_0}{e}\right)\right)$, then $f\left(\frac{1}{2}; \lambda\right) \leq 0$, and $\mu = \frac{1}{2}$ is a local maximum. In particular, if $p_0 \geq \frac{1}{2} e^{1/2} \approx 0.8244$, then this holds for any $\lambda \geq 1$. Notice that this does not rule out the possibility of having several local maximum.

In particular, I will show that when p_0 is large enough, there is at least another local maximum. If $f(1; \lambda) < 0$ and $f(\mu^*; \lambda) > 0$ for some $\mu^* \in (\frac{1}{2}, 1)$, then by the Intermediate Value Theorem there is an interior maximum. Indeed, $f(1; \lambda) = (1 - \lambda)(1 - p_0 e^{-\lambda}) < 0$ for $\lambda > 1$. At $\mu = \frac{1}{2} + \frac{1}{2\lambda} \in (\frac{1}{2}, 1)$:

$$f\left(\tfrac{1}{2} + \tfrac{1}{2\lambda}; \lambda\right) = 1 - \frac{\lambda}{4} - \frac{\lambda}{2} p_0 e^{-\frac{\lambda}{2}(1 + \frac{1}{\lambda})},$$

which is positive for any $\lambda \geq 1$ and large p_0. Hence, by the Intermediate Value Theorem, there exists an interior maximum.

Now, if $\lambda < 2\left(1 + W(-p_0 e^{-1})\right)$, we can use again the Intermediate Value Theorem to show there exists an interior solution, as

$$f\left(\frac{1}{2};\lambda\right) > 0 > f\left(1;\lambda\right).$$

I will prove now that this solution is below $\frac{1}{\lambda}$. At $\mu = \frac{1}{\lambda}$, we get:

$$f\left(\tfrac{1}{\lambda};\lambda\right) = -p_0 e^{-1}(\lambda - 1) < 0.$$

For $\mu > \frac{1}{\lambda}$,

$$f(\mu;\lambda) = (1 - \lambda\mu)(1 - p_0 e^{-\lambda\mu}) - \lambda p_0 e^{-\lambda\mu}(1 - \mu) < 0,$$

as both terms are negative. So any interior solution satisfies $\mu < \frac{1}{\lambda}$. Hence, if $f(\mu;\lambda) = 0$ it must be when $\mu \in \left(\frac{1}{2}, \frac{1}{\lambda}\right)$.

To finish the proof, I will show the comparative statics using the Implicit Function Theorem. Let

$$F(\mu, p_0, \lambda) = 1 - \lambda\mu - p_0 e^{-\lambda\mu}(1 - \lambda(2\mu - 1)).$$

Then at the optimum $F = 0$. Its partial derivatives are, using the first order condition:

$$\begin{aligned}
\frac{\partial F}{\partial p_0} &= -\frac{1 - \lambda\mu}{p_0} < 0, \\
\frac{\partial F}{\partial \lambda} &= -\lambda\gamma + p_0 e^{-\lambda\mu}(2\mu - 1), \\
\frac{\partial F}{\partial \mu} &= -\lambda\left[1 - 2p_0 e^{-\lambda\mu} + (1 - \lambda\mu)\right].
\end{aligned}$$

I will show that $F_\lambda < 0$. This is a decreasing function of λ. Hence, it reaches its maximum when $\lambda = 1$.

$$F_\lambda < 0 \quad \text{if and only if} \quad \frac{\gamma}{2\mu - 1} > p_0 e^{-\mu}$$

The left hand side is a decreasing function of μ. Hence, it reaches its minimum when $\mu = 1$. As the right hand side is a probability, $1 > p_0 e^{-\mu}$ and the statement is always true.

Finally, let's show that $F_\mu < 0$. The condition can be rewritten as

$$2(1 - p_0 e^{-\lambda\mu}) > \lambda\mu$$

The left hand side reaches a minimum when $p_0 = 1$, and therefore, it is enough to check that

$$2 > \lambda\mu + 2e^{-\lambda\mu}.$$

As $\mu < \frac{1}{\lambda}$, define $x = \lambda\mu \in \left(\frac{1}{2}, 1\right)$. The function $g(x) = x + 2e^{-x}$ has a derivative equal to

$$g'(x) = 1 - 2e^{-x}$$

which is zero at $x = \log(2)$. As the derivative has a unique solution, there is a unique critical point, that coincides with a minimum as the derivative is negative for small values of x. Hence, it is enough to compare the values of $g(x)$ at the boundaries of x.

$$g\left(\frac{1}{2}\right) = \frac{1}{2} + 2e^{-\frac{1}{2}} < g(1) = 1 + 2e^{-1} < 2$$

This proves that the partial derivative of λ is negative too. Using the Implicit Function Theorem, the partial derivatives of μ are given by

$$\frac{\partial\mu}{\partial p_0} = -\frac{F_{p_0}}{F_\mu} = -\frac{e^{-\lambda\mu}\left(1 - \lambda(2\mu - 1)\right)}{\lambda\left[1 - 2p_0 e^{-\lambda\mu} - p_0 e^{-\lambda\mu}(1 - \lambda(2\mu - 1))\right]}$$
$$\frac{\partial\mu}{\partial\lambda} = -\frac{F_\lambda}{F_\mu} = -\frac{\mu + p_0 e^{-\lambda\mu}\left[-\mu(1 - \lambda(2\mu - 1)) - (2\mu - 1)\right]}{\lambda\left[1 + p_0 e^{-\lambda\mu}(-3 + \lambda(2\mu - 1))\right]}$$

To study $\frac{\partial\mu}{\partial p_0}$, and $\frac{\partial\mu}{\partial\lambda}$ we just need to compare the signs of the two partial derivatives of F. As all three are negative, this shows that the partial derivatives are also negative when μ is an interior solution. □

Proof of Corollary 1.2

Proof. Assume $\lambda \geq 2$. Then the optimal structure is given by $\Omega = \left(\frac{1}{\lambda}, 0\right)$.

Comparing expected profits under this structure with $\Pi_{\varnothing M}$ yields:

$$\frac{p_0 e^{-1}}{\lambda} \cdot \frac{\Delta\theta(1-\Delta\theta)}{2} \geq \bar{V}$$

which is exactly the same condition as for hiring a single monitor under $\lambda > 2$, given in Corollary 1.1. □

Proof of Corollary 1.3

Proof. Assume now that $\lambda \leq 1$. Comparing expected profits under two-monitor and no-monitor regimes yields:

$$p_0 e^{-\lambda}\left(1 - \frac{1}{2}p_0 e^{-\lambda}\right) \cdot \frac{\Delta\theta(1-\Delta\theta)}{2} \geq \bar{V}$$

Now compare expected profits under two monitors vs. a single monitor:

$$p_0 e^{-\lambda}\left(1 - p_0 e^{-\lambda}\right) \cdot \frac{\Delta\theta(1-\Delta\theta)}{2} \geq \bar{V}$$

By comparing the two right-hand sides, we observe:

$$p_0 e^{-\lambda}\left(1 - \frac{1}{2}p_0 e^{-\lambda}\right) \geq p_0 e^{-\lambda}\left(1 - p_0 e^{-\lambda}\right)$$

Therefore, if the principal prefers two monitors over one, she also prefers two over none. However, the reverse is not necessarily true: hiring one monitor may be optimal even if two is not. This occurs when $\bar{V}$ lies between the two thresholds:

$$p_0 e^{-\lambda}\left(1 - \frac{1}{2}p_0 e^{-\lambda}\right) \cdot \frac{\Delta\theta(1-\Delta\theta)}{2} \geq \bar{V} \geq p_0 e^{-\lambda}\left(1 - p_0 e^{-\lambda}\right) \cdot \frac{\Delta\theta(1-\Delta\theta)}{2}$$

Thus, the only case in which two monitors are strictly preferred over both one and zero is when the first (larger) condition holds. □

Proof of Corollary 1.4

Proof. Assume $\lambda \in (1,2)$ and let $\Delta \equiv \frac{\Delta\theta(1-\Delta\theta)}{2}$ and write $p(\mu) = p_0 e^{-\lambda\mu}$. For $\lambda \in (1,2)$ and the optimal span $\mu_1^*(p_0,\lambda)$ in the two-monitor regime (with

$\mu^2 = 2\mu_1^* - 1$), define

$$A(\mu_1^*; p_0, \lambda) \equiv 2\mu_1^* p(\mu_1^*) - \mu^2 p(\mu_1^*)^2.$$

From Corollary 1.1, the principal hires *one* monitor rather than none iff

$$\bar{V} < V_1(p_0, \lambda) \equiv \frac{p_0}{\lambda} e^{-1} \Delta.$$

Also, the principal hires *two* monitors rather than one iff

$$\bar{V} \leq V_2(p_0, \lambda) \equiv \left(A(\mu_1^*; p_0, \lambda) - \tfrac{p_0}{\lambda} e^{-1} \right) \Delta.$$

Because $A(\mu_1^*; p_0, \lambda) \geq 0$, it follows that $V_2(p_0, \lambda) \leq V_1(p_0, \lambda)$. Thus this gives a monotonic ordering for the three hiring conditions.

$$\begin{cases} \bar{V} \leq V_2(p_0, \lambda) & \Rightarrow \text{two monitors}, \\ V_2(p_0, \lambda) < \bar{V} < V_1(p_0, \lambda) & \Rightarrow \text{one monitor}, \\ \bar{V} \geq V_1(p_0, \lambda) & \Rightarrow \text{no monitor}. \end{cases}$$

Now, let's explain the comparative statics of the incentives to hire as a function fo p_0 and λ. The one-monitor cutoff is explicit:

$$\frac{\partial V_1}{\partial p_0} = \frac{e^{-1}}{\lambda} \Delta > 0, \qquad \frac{\partial V_1}{\partial \lambda} = -\frac{p_0 e^{-1}}{\lambda^2} \Delta < 0.$$

Hence better verifiability ($p_0 \uparrow$) makes hiring one monitor attractive at higher $\bar{V}$, while faster span decay ($\lambda \uparrow$) makes it less attractive.

For V_2, note that μ_1^* satisfies the first-order condition $F(\mu_1, p_0, \lambda) = 0$ from Proposition 1.5. By the Implicit Function Theorem,

$$\frac{\partial \mu_1^*}{\partial p_0} = -\frac{F_{p_0}}{F_{\mu_1}} < 0, \qquad \frac{\partial \mu_1^*}{\partial \lambda} = -\frac{F_\lambda}{F_{\mu_1}} < 0,$$

so the optimal span shrinks as either p_0 or λ rises. Differentiating V_2 shows

$$\frac{\partial V_2}{\partial p_0} > 0, \qquad \frac{\partial V_2}{\partial \lambda} < 0.$$

Thus higher p_0 expands the two-monitor region, while higher λ contracts it.

□

Proof of Proposition 1.7

Proof. First, I write the Lagrangian of the problem and its first order conditions.

Lagrangian. Let e_i, $W(x_i, r_i)$, and $S(x_i, r_i)$ denote efforts, wages to agents, and wages to the monitor respectively. The Lagrangian is:

$$\begin{aligned}
\mathcal{L} = {} & \mathbb{E}\left[\theta_i + e_i - W(x_i, r_i) - S(x_i, r_i)\right] \\
& + \lambda_1 S(x_{\theta_L,\theta_L}, \theta_L) + \lambda_2 S(x_{\theta_L,\varnothing}, \varnothing) + \lambda_3 S(x_{\theta_H,\varnothing}, \varnothing) + \lambda_4 S(x_{\theta_H,\theta_H}, \theta_H) \\
& + \psi_1 \left(W(x_{\theta_L,\theta_L}, \theta_L) - \frac{e^2_{\theta_L,\theta_L}}{2} \right) + \psi_2 \left(W(x_{\theta_L,\varnothing}, \varnothing) - \frac{e^2_{\theta_L,\varnothing}}{2} \right) + \psi_4 \left(W(x_{\theta_H,\theta_H}, \theta_H) - \frac{e^2_{\theta_H,\theta_H}}{2} \right) \\
& + \psi_3 \left(W(x_{\theta_H,\varnothing}, \varnothing) - \frac{e^2_{\theta_H,\varnothing}}{2} - W(x_{\theta_L,\varnothing}, \varnothing) + \frac{(e_{\theta_L,\varnothing} - \Delta\theta)^2}{2} \right) \\
& + \varphi \left(p(\mu) \left(S(x_{\theta_L,\theta_L}, \theta_L) + S(x_{\theta_H,\theta_H}, \theta_H) \right) + (1 - p(\mu)) \left(S(x_{\theta_L,\varnothing}, \varnothing) + S(x_{\theta_H,\varnothing}, \varnothing) \right) - \frac{2\bar{V}}{\mu} \right) \\
& + \Pi \left(S(x_{\theta_H,\theta_H}, \theta_H) + W(x_{\theta_H,\theta_H}, \theta_H) - \frac{e^2_{\theta_H,\theta_H}}{2} - S(x_{\theta_H,\varnothing}, \varnothing) - W(x_{\theta_H,\varnothing}, \varnothing) + \frac{e^2_{\theta_H,\varnothing}}{2} \right).
\end{aligned}$$

First-order Conditions.

$$\frac{\partial \mathcal{L}}{\partial e_{\theta_L,\theta_L}} = 0 \Rightarrow \frac{p(\mu)}{2} - \psi_1 e_{\theta_L,\theta_L} = 0 \tag{A.10}$$

$$\frac{\partial \mathcal{L}}{\partial e_{\theta_L,\varnothing}} = 0 \Rightarrow \frac{1}{2}(1-p(\mu)) - \psi_2 e_{\theta_L,\varnothing} + \psi_3(e_{\theta_L,\varnothing} - \Delta\theta) = 0 \tag{A.11}$$

$$\frac{\partial \mathcal{L}}{\partial e_{\theta_H,\varnothing}} = 0 \Rightarrow \frac{1}{2}(1-p(\mu)) - \psi_3 e_{\theta_H,\varnothing} + \Pi e_{\theta_H,\varnothing} = 0 \tag{A.12}$$

$$\frac{\partial \mathcal{L}}{\partial e_{\theta_H,\theta_H}} = 0 \Rightarrow \frac{p(\mu)}{2} - \psi_4 e_{\theta_H,\theta_H} - \Pi e_{\theta_H,\theta_H} = 0 \tag{A.13}$$

$$\frac{\partial \mathcal{L}}{\partial W(x_{\theta_L,\theta_L},\theta_L)} = 0 \Rightarrow -\frac{p(\mu)}{2} + \psi_1 = 0 \tag{A.14}$$

$$\frac{\partial \mathcal{L}}{\partial W(x_{\theta_L,\varnothing},\varnothing)} = 0 \Rightarrow -\frac{1}{2}(1-p(\mu)) + \psi_2 - \psi_3 = 0 \tag{A.15}$$

$$\frac{\partial \mathcal{L}}{\partial W(x_{\theta_H,\varnothing},\varnothing)} = 0 \Rightarrow -\frac{1}{2}(1-p(\mu)) + \psi_3 - \Pi = 0 \tag{A.16}$$

$$\frac{\partial \mathcal{L}}{\partial W(x_{\theta_H,\theta_H},\theta_H)} = 0 \Rightarrow -\frac{p(\mu)}{2} + \psi_4 + \Pi = 0 \tag{A.17}$$

$$\frac{\partial \mathcal{L}}{\partial S(x_{\theta_L,\theta_L},\theta_L)} = 0 \Rightarrow -\frac{p(\mu)}{2} + \lambda_1 + \varphi p(\mu) = 0 \tag{A.18}$$

$$\frac{\partial \mathcal{L}}{\partial S(x_{\theta_L,\varnothing},\varnothing)} = 0 \Rightarrow -\frac{1}{2}(1-p(\mu)) + \lambda_2 + \varphi(1-p(\mu)) = 0 \tag{A.19}$$

$$\frac{\partial \mathcal{L}}{\partial S(x_{\theta_H,\varnothing},\varnothing)} = 0 \Rightarrow -\frac{1}{2}(1-p(\mu)) + \lambda_3 - \Pi + \varphi(1-p(\mu)) = 0 \tag{A.20}$$

$$\frac{\partial \mathcal{L}}{\partial S(x_{\theta_H,\theta_H},\theta_H)} = 0 \Rightarrow -\frac{p(\mu)}{2} + \lambda_4 + \Pi + \varphi p(\mu) = 0 \tag{A.21}$$

From the first order condition of the efforts and the wages, it is immediate that:

$$e_{\theta_L,\theta_L} = e_{\theta_H,\theta_H} = e_{\theta_H,\varnothing} = 1.$$

From equation (A.14), $\psi_1 > 0$, and from (A.15) $\psi_2 > 0$ this implies that the constraints for low-productivity types ($IC_{L\varnothing}$) are binding. Hence,

$$W(x_{\theta_L,\theta_L},\theta_L) = \frac{1}{2}, \quad W(x_{\theta_L,\varnothing},\varnothing) = \frac{e^2_{\theta_L,\varnothing}}{2}.$$

To show that ($IC_{H\varnothing}$) binds, it is enough to notice that $\psi_3 > 0$ in equation (A.16). Then,

$$W(x_{\theta_H,\varnothing},\varnothing) = \frac{1}{2} + W(x_{\theta_L,\varnothing},\varnothing) - \frac{(e_{\theta_L,\varnothing} - \Delta\theta)^2}{2}.$$

And, substituting the value of $W(x_{\theta_L,\varnothing},\varnothing)$,

$$W(x_{\theta_H,\varnothing},\varnothing)=\frac{1}{2}+\Delta\theta e_{\theta_L,\varnothing}-\frac{\Delta\theta^2}{2}.$$

We already know that if the Participation Constraint of the Monitor is binding but the Collusion Constraint is not, we are under the non-collusion contract. Now consider the wage contract W as in Proposition 1.3, and let the monitor's wage for the truthful report be

$$S(x_{\theta_H,\theta_H},\theta_H)=\frac{2\bar{V}}{\mu p(\mu)}.$$

This contract is incentive compatible for both agents and monitor and satisfies ($\mathrm{CC}_{H\varnothing}$). If it also satisfies the first collusion constraint (CC_{HH}),

$$\bar{V}\geq\frac{1}{\lambda}p_0e^{-1}\cdot\frac{\Delta\theta}{2}\left(1-\frac{3}{2}\Delta\theta\right),$$

then the contract is collusion-proof. This means that the contract developed in Section 1.2 is also collusion proof.

Assume now that this inequality does not hold:

$$\bar{V}<\frac{1}{\lambda}p_0e^{-1}\cdot\frac{\Delta\theta}{2}\left(1-\frac{3}{2}\Delta\theta\right):=\bar{V}_1,$$

I explore the optimal contract in this case. Begin by assuming that the monitor's participation constraint is not binding ($\varphi=0$); we will later examine when this assumption holds and its implications. From equations (A.19) and (A.20), $S(x_{\theta_L,\varnothing},\varnothing)=S(x_{\theta_H,\varnothing},\varnothing)=0$. Hence, ($\mathrm{CC}_{H\varnothing}$) reduces to ($\mathrm{IC}_{H\varnothing}$) as stated before. Then, the Collusion Constraint (CC_{HH}) must bind. This implies:

$$W(x_{\theta_H,\theta_H},\theta_H)+S(x_{\theta_H,\theta_H},\theta_H)=W(x_{\theta_H,\varnothing},\varnothing).$$

Solving for $e_{\theta_L,\varnothing}$:

Recall from earlier that:

$$\psi_2=1-\frac{p(\mu)}{2},\quad \Pi=\frac{p(\mu)}{2},\quad \psi_3+\pi_2=\Pi+\frac{1}{2}(1-p(\mu)).$$

The FOC for $e_{\theta_L,\varnothing}$ becomes:

$$\frac{1}{2}(1-p(\mu)) - \psi_2 e_{\theta_L,\varnothing} + (\psi_3+\pi_2)(e_{\theta_L,\varnothing} - \Delta\theta) = 0.$$

Substitute in and simplify:

$$\frac{1}{2}(1-p(\mu)) - \left(1-\frac{p(\mu)}{2}\right) e_{\theta_L,\varnothing} + \left(\frac{p(\mu)}{2} + \frac{1}{2}(1-p(\mu))\right)(e_{\theta_L,\varnothing} - \Delta\theta) = 0.$$

Solving yields:

$$e_{\theta_L,\varnothing} = 1 - \frac{\Delta\theta}{1-p(\mu)}.$$

This completes the derivation of the optimal contract when only the Collusion Constraint (CC_{HH}) is binding. Substituting the values into the participation constraint of the monitor, we require to have

$$\bar{V} \leq \mu p(\mu)\frac{\Delta\theta}{2}\left(1 - \Delta\theta\frac{3-p(\mu)}{2(1-p(\mu))}\right) := \bar{V}_2(\mu)$$

Now, suppose instead that the participation constraint of the monitor is also binding. If either $\lambda_1 > 0$ or $\lambda_2 > 0$ we can show that $\lambda_3,\ \lambda_4 > 0$ and the optimal contract is the non-collusion one. But, by assumption, we know that, in that case, (CC_{HH}) is slack. Hence, $S(x_{\theta_L,\theta_L},\theta_L) = S(x_{\theta_L,\varnothing},\varnothing) = 0$. Adding equations (A.19) and (A.20), we get that

$$\lambda_3 = \lambda_2 + \Pi > 0,$$

and, hence, $S(x_{\theta_H,\varnothing},\varnothing) = 0$. Then, we can obtain the contract from the monitor by substituting $S(x_{\theta_L,\varnothing},\varnothing) = S(x_{\theta_H,\varnothing},\varnothing) = 0$:

$$S(x_{\theta_H,\theta_H},\theta_H) = \frac{2\bar{V}}{\mu p(\mu)}.$$

At the same time, as (CC_{HH}) and ($IC_{H\varnothing}$) are binding,

$$W(x_{\theta_H,\theta_H},\theta_H) + S(x_{\theta_H,\theta_H},\theta_H) = \frac{1}{2} + \Delta\theta e_{\theta_L,\varnothing} - \frac{\Delta\theta^2}{2}.$$

I will distinguish two cases:

Case A. $W(x_{\theta_H,\theta_H},\theta_H) > \frac{1}{2}$ This immediately implies that $\varphi_4 = 0$. Using a similar reasoning to before, we get $e_{\theta_L,\varnothing} = 1 - \frac{\Delta\theta}{1-p(\mu)}$. Finally, from the ($IC_{L\varnothing}$),

$$W(x_{\theta_H,\theta_H},\theta_H) = \frac{1}{2} + \Delta\theta - \Delta\theta^2\frac{3-p(\mu)}{2(1-p(\mu))} - \frac{2\bar{V}}{\mu p(\mu)}$$

Notice that we require that $W(x_{\theta_H,\theta_H},\theta_H) > \frac{1}{2}$. Or equivalently,

$$\Delta\theta - \Delta\theta^2\frac{3-p(\mu)}{2(1-p(\mu))} - \frac{2\bar{V}}{\mu p(\mu)} > 0$$

Which is a contradiction with the fact that $\bar{V} \geq \bar{V}_2$.

Case B. $W(x_{\theta_H,\theta_H},\theta_H) = \frac{1}{2}$ So, from the previous two equations,

$$S(x_{\theta_H,\theta_H},\theta_H) = \Delta\theta e_{\theta_L,\varnothing} - \frac{\Delta\theta^2}{2},$$

and

$$S(x_{\theta_L,\theta_L},\theta_L) = \frac{2\bar{V}}{\mu p(\mu)} - \Delta\theta e_{\theta_L,\varnothing} + \frac{\Delta\theta^2}{2}.$$

As $S(x_{\theta_L,\theta_L},\theta_L) = 0$,

$$e_2 = \frac{2\bar{V}}{\Delta\theta\mu p(\mu)} + \frac{\Delta\theta}{2}.$$

□

Proof of Proposition 1.8

Proof. The first order condition of the Lagrangian of the proof of Proposition 1.7 when μ is an interior solution is given by,

$$\begin{aligned}\frac{1}{2}p(\mu)((e_{\theta_L,\theta_L} + e_{\theta_H,\theta_H} - W(x_{\theta_L,\theta_L},\theta_L) - W(x_{\theta_H,\theta_H},\theta_H) - S(x_{\theta_H,\theta_H},\theta_H)\\ -(e_{\theta_H,\varnothing} + e_{\theta_L,\varnothing} - W(x_{\theta_L,\varnothing},\varnothing) - W(x_{\theta_H,\varnothing},\varnothing))) + \frac{1}{2}(e_{\theta_H,\varnothing} + e_{\theta_L,\varnothing} - W(x_{\theta_L,\varnothing},\varnothing)\\ -W(x_{\theta_H,\varnothing},\varnothing)) - \Pi_{\varnothing M} + \mu\frac{1}{2}p'(\mu)((e_{\theta_L,\theta_L} + e_{\theta_H,\theta_H} - W(x_{\theta_L,\theta_L},\theta_L)\\ -W(x_{\theta_H,\theta_H},\theta_H) - S(x_{\theta_H,\theta_H},\theta_H) - (e_{\theta_H,\varnothing} + e_{\theta_L,\varnothing} - W(x_{\theta_L,\varnothing},\varnothing) - W(x_{\theta_H,\varnothing},\varnothing))) + \varphi_2 = 0,\end{aligned}$$

where φ_2 is the Lagrange multiplier of (PC_M). Assume that only the (CC_{HH}) binds while the Participation Constraint of the monitor is slack. Then $\varphi_2 = 0$. Substituting by the values obtained in Proposition 1.7, we can rewrite the first order condition as,

$$\frac{1}{2}p(\mu)\frac{\Delta\theta^2}{2(1-p(\mu))^2} - \frac{1}{4}\frac{\Delta\theta^2}{2(1-p(\mu))^2}p(\mu)^2 + \mu\frac{1}{2}p'(\mu)\frac{\Delta\theta^2}{2(1-p(\mu))^2} = 0. \quad \text{(A.22)}$$

Which can be further simplified to

$$\frac{\Delta\theta^2}{4(1-p(\mu))^2}p(\mu)\left(1 - p_0e^{-\lambda\mu} - \lambda\mu\right) = 0$$

Notice that if $\mu = 0$ that equation is positive, hence $\mu = 0$ is never a solution to the firm design. In particular, the solution is implicitly given by

$$\mu = \frac{1 - p_0e^{-\lambda\mu}}{\lambda}$$

Rearraging and defining $u = 1 - \lambda\mu$,

$$-ue^{-u} = -p_0e^{-1},$$

which is the standard form of the Lambert W, which means that $u = W(-p_0e^{-1})$. Substituting again u, we get

$$\mu = \frac{1 + W_0\left(-p_0e^{-1}\right)}{\lambda}$$

Going back to the Participation Constraint of the Monitor, we can substitute the values of the contract in (PC_M) and define $\bar{V}_2$ as,

$$\bar{V}_2 := -W_0\left(-p_0e^{-1}\right) \cdot \frac{1 + W_0\left(-p_0e^{-1}\right)}{\lambda} \cdot \frac{\Delta\theta}{2} \cdot \left(1 - \Delta\theta\frac{3 - p(\mu)}{2(1-p(\mu))}\right)$$

□

Proof of Corollary 1.5

Proof. First, it is important to know that the main branch of the Lambert W function is an increasing function with

$$W_0\left(-\frac{1}{e}\right) = -1,\ W_0(0) = 0$$

This guarantees that $1 + W_0\left(-p_0 e^{-1}\right)$ is always positive. As $W_0\left(-p_0 e^{-1}\right) < 0$, we have that

$$\frac{1 + W_0\left(-\frac{p_0}{e}\right)}{\lambda} < \frac{1}{\lambda}$$

which guarantees that the optimal span of control under collusion is always smaller than in the non collusion case. At the same time, if $\mu \in (0, 1)$, applying the implicit function theorem and the chain rule, we can see that

$$\frac{\partial \mu}{\partial \lambda} = -\frac{1 + W_0\left(-p_0 e^{-1}\right)}{\lambda^2} < 0$$
$$\frac{\partial \mu}{\partial p_0} = -\frac{W_0'\left(-p_0 e^{-1}\right)}{\lambda e} < 0$$

where the second inequality comes from the fact that W_0 is an increasing function. Finally, the interior optimal condition is not satisfied if

$$\frac{1 + W_0\left(-p_0 e^{-1}\right)}{\lambda} \geq 1$$

or equivalently,

$$\lambda \leq 1 + W_0\left(-p_0 e^{-1}\right)$$

In this case, the optimal $\mu = 1$. □

Proof of Proposition 1.9 If only the participation constraint is binding, the problem of the firm can be solved as in the non-collusion case, and therefore, the expected cost of a monitor is $\bar{V}$. Else, if the participation constraint is not binding, the problem under one monitor and under two monitors are unrelated, as they do not share any restriction in common. Hence, both problems can be solved separately. Notice that the solution to the problem of one monitor is given by Proposition 1.7. I will rewrite the problem taking into account only the problem of two monitors and assuming that the participation constraint is not binding.

$$\max_{\{S,W,e\}} \left(1-(1-p(\mu))^2\right)\left(e_{1,2}+e_{4,2}-W(x_{\theta_L,\theta_L},\theta_L)-W(x_{\theta_H,\theta_H},\theta_H)\right)$$
$$+\left(1-p(\mu)\right)^2\left(e_{2,2}+e_{3,2}-W(x_{\theta_L,\varnothing},\varnothing)-W(x_{\theta_H,\varnothing},\varnothing)\right)$$
$$-2p(\mu)(S(x_{\theta_L,\theta_L},\theta_L)+S(x_{\theta_H,\theta_H},\theta_H))$$
$$-2(1-p(\mu))(S(x_{\theta_L,\varnothing},\varnothing)+S(x_{\theta_H,\varnothing},\varnothing))-2p(\mu)(1-p(\mu))(\delta-\beta)$$

s.t.

$$S(x_i,r_i)\geq 0,\ i=1,2,4 \tag{LLM}$$
$$S^2(x_i,r_i)-\beta\geq 0,\ i=3 \tag{LLM'}$$
$$\beta\geq 0, \tag{PenC}$$
$$W^2(x_i,r_i)-\frac{e_i^2}{2}\geq 0,\ i=1,2,3,4 \tag{IC1'}$$
$$W(x_{\theta_H,\varnothing},\varnothing)-\frac{e_{3,2}^2}{2}\geq W(x_{\theta_L,\varnothing},\varnothing)-\frac{(e_{2,2}-\Delta\theta)^2}{2} \tag{IC3'}$$
$$2S(x_{\theta_H,\theta_H},\theta_H)+W(x_{\theta_H,\theta_H},\theta_H)-c(e_{4,1})\geq$$
$$2S(x_{\theta_H,\varnothing},\varnothing)+W(x_{\theta_H,\varnothing},\varnothing)-c(e_{3,1}) \tag{CC$_{H2H}$}$$
$$S(x_{\theta_H,\theta_H},\theta_H)+\delta-\beta+W(x_{\theta_H,\theta_H},\theta_H)-c(e_{4,1})\geq$$
$$S(x_{\theta_H,\varnothing},\varnothing)+W(x_{\theta_H,\varnothing},\varnothing)-c(e_{3,1}) \tag{CC$_{H1H}$}$$
$$S(x_{\theta_H,\theta_H},\theta_H)\geq S(x_{\theta_H,\varnothing},\varnothing)+\delta-\beta \tag{CC$_M$}$$

The Lagrangian is

$$\mathcal{L}() = \left(1-(1-p(\mu))^2\right)\left(e_{1,2}+e_{4,2}-W(x_{\theta_L,\theta_L},\theta_L)-W(x_{\theta_H,\theta_H},\theta_H)\right)$$
$$+\left(1-p(\mu)\right)^2\left(e_{2,2}+e_{3,2}-W(x_{\theta_L,\varnothing},\varnothing)-W(x_{\theta_H,\varnothing},\varnothing)\right)$$
$$-2p(\mu)(S(x_{\theta_L,\theta_L},\theta_L)+S(x_{\theta_H,\theta_H},\theta_H))$$
$$-2(1-p(\mu))(S(x_{\theta_L,\varnothing},\varnothing)+S(x_{\theta_H,\varnothing},\varnothing))-p(\mu)(1-p(\mu))(\delta-\beta)$$
$$+\sum_{i=1,2,4}\lambda_i\left(S(x_i,r_i)\right)+\lambda_3(S_3-\beta)+\gamma\beta+\sum_{i=1,2,4}\varphi_i\left(W_i-\frac{e_i^2}{2}\right)$$
$$+\varphi_3\left(W_3-\frac{e_3^2}{2}-W_2+\frac{(e_2-\Delta\theta)^2}{2}\right)+\Pi_3\left(2S_4+W_4-\frac{e_4^2}{2}-2S_3-W_3+\frac{e_3^2}{2}\right)$$
$$\Pi_4\left(S_4+\delta-\beta+W_4-\frac{e_4^2}{2}-S_3-W_3+\frac{e_3^2}{2}\right)+\Pi_5\left(S_4-S_3-\delta+\beta\right)$$

And the first order conditions of the problem are:

$$\frac{\partial \mathcal{L}}{\partial e_1} = 0 \Rightarrow (1-(1-p(\mu))^2) - \varphi_1 e_1 = 0 \tag{A.23}$$
$$\frac{\partial \mathcal{L}}{\partial e_2} = 0 \Rightarrow (1-p(\mu))^2 - \varphi_2 e_2 + \varphi_3(e_2 - \Delta\theta) = 0 \tag{A.24}$$
$$\frac{\partial \mathcal{L}}{\partial e_3} = 0 \Rightarrow (1-p(\mu))^2 - \varphi_3 e_3 + \Pi_3 e_3 + \Pi_4 e_3 = 0 \tag{A.25}$$
$$\frac{\partial \mathcal{L}}{\partial e_4} = 0 \Rightarrow (1-(1-p(\mu))^2) - \varphi_4 e_4 - \Pi_3 e_4 - \Pi_4 e_4 = 0 \tag{A.26}$$
$$\frac{\partial \mathcal{L}}{\partial W_1} = 0 \Rightarrow (1-(1-p(\mu))^2) - \varphi_1 = 0 \tag{A.27}$$
$$\frac{\partial \mathcal{L}}{\partial W_2} = 0 \Rightarrow (1-p(\mu))^2 - \varphi_2 + \varphi_3 = 0 \tag{A.28}$$
$$\frac{\partial \mathcal{L}}{\partial W_3} = 0 \Rightarrow (1-p(\mu))^2 - \varphi_3 + \Pi_3 + \Pi_4 = 0 \tag{A.29}$$
$$\frac{\partial \mathcal{L}}{\partial W_4} = 0 \Rightarrow (1-(1-p(\mu))^2) - \varphi_4 - \Pi_3 - \Pi_4 = 0 \tag{A.30}$$
$$\frac{\partial \mathcal{L}}{\partial S_1} = 0 \Rightarrow -2p(\mu) + \lambda_1 = 0 \tag{A.31}$$
$$\frac{\partial \mathcal{L}}{\partial S_2} = 0 \Rightarrow -2(1-p(\mu)) + \lambda_2 = 0 \tag{A.32}$$
$$\frac{\partial \mathcal{L}}{\partial S_3} = 0 \Rightarrow -2(1-p(\mu)) + \lambda_3 - 2\Pi_3 - \Pi_4 - \Pi_5 = 0 \tag{A.33}$$
$$\frac{\partial \mathcal{L}}{\partial S_4} = 0 \Rightarrow -2p(\mu) + \lambda_4 + 2\Pi_3 + \Pi_4 + \Pi_5 = 0 \tag{A.34}$$
$$\frac{\partial \mathcal{L}}{\partial \delta} = 0 \Rightarrow -p(\mu)(1-p(\mu)) + \Pi_4 - \Pi_5 = 0 \tag{A.35}$$
$$\frac{\partial \mathcal{L}}{\partial \beta} = 0 \Rightarrow p(\mu)(1-p(\mu)) + \gamma - \lambda_3 - \Pi_4 + \Pi_5 = 0 \tag{A.36}$$

As in previous proofs, it is immediate to see that $e_{1,2} = e_{3,2} = e_{4,2} = 1$. From the first order condition of $W(x_{\theta_L,\theta_L}, \theta_L)$, we can see that $\varphi_1 > 0$, which means that $W_1 = \frac{1}{2}$. At the same time, form (A.28), $\varphi_2 > 0$, which also means that $W_2 = \frac{e_2^2}{2}$. So the low type agents do not receive any rents under the optimal contract. From equation (A.29), $\varphi_3 > 0$. This means that the ($IC_{H\varnothing}$) is binding and,

$$W_3 = \frac{1}{2} + \frac{e_2^2}{2} - \frac{(e_2 - \Delta\theta)^2}{2} = \frac{1-\Delta\theta^2}{2} + e_2\Delta\theta$$

From equations (A.31) and (A.32) it is immediate that $\lambda_1, \lambda_2 > 0$, which implies that $S_1 = S_2 = 0$. From equation (A.33),

$$\lambda_3 = 2(1-p(\mu)) + 2\Pi_3 + \Pi_4 + \Pi_5 > 0,$$

which implies that $S_3 = 0$. This implies that, when the *Participation Constraint of the Monitor* is not binding, she only receives wages from reporting high types.

From the first order condition of δ, we have that $\Pi_4 > 0$, which means that (CC_{H1H}) is binding. Adding (A.35) and (A.36), we get $\gamma = \lambda_3$, which implies that $\gamma > 0$ and therefore $\beta = 0$.

If $S_4 = \delta$, equation (CC_{H2H}) becomes (CC_{H1H}), which we already know is binding. If, instead $S_4 > \delta$, $\Pi_5 = 0$ and (CC_{H2H}) is not binding, as

$$2S_4 + W_4 > S_4 + \delta + W_4 = W_3.$$

This implies that $\Pi_3 = 0$. From (A.34) we have that $\Pi_4 + \lambda_4 = 2p(\mu)$, and from (A.36) $\Pi_4 = p(\mu)(1 - p(\mu))$. If $S_4 > 0$, this is a contradiction. Now, if $S_4 = 0$, then $S_4 = \delta = 0$, which makes (CC_M) binding and thus, we arrive to a contradiction.

To finish the proof, we just need to compute e_2 and W_4. As (CC_{H1H}) is binding, we know that

$$W_4 + 2S_4 = W_3.$$

Notice that any combination of S_4 and W_4 that satisfies the restrictions and the previous equation would give the same profits for the monitor. This implies that $\varphi_4 = 0$. As in previous proofs, I will assume that $W_4 = \frac{1}{2}$, satisfying the incentive compatibility constraint, and $2S_4 = W_3 - \frac{1}{2}$.

Adding equations (A.29) and (A.30) we get that $\varphi_3 = 1$. And from equation (A.28) $\varphi_3 - \varphi_2 = (1 - p(\mu))^2$. Substituting into the first order condition of e_2,

$$e_2 = 1 - \frac{\Delta\theta}{(1 - p(\mu))^2},$$

finishing the proof.

Proof of Proposition 1.11.

Proof. As (PC_M) is never binding, Proposition 1.10 guarantees that $\mu = \gamma$. Hence, the profit function can be written as,

$$\mathbb{E}\left[\Pi(\Omega, \mathbf{W}, \mathbf{S})\right] = \Pi_{\varnothing M} + \frac{\Delta\theta^2(1-(1-p(\mu)^2))}{4(1-p(\mu))^2}\mu$$

In this case,

$$\begin{aligned}\mu =& \frac{(1-p_0e^{-\lambda\mu}) + (1-p_0e^{-\lambda\mu})^2}{2\lambda} \\ =& \frac{(1-p_0e^{-\lambda\mu})(2-p_0e^{-\lambda\mu})}{2\lambda}\end{aligned}$$

To show the uniqueness, assume that μ is interior (as otherwise is trivial) and define $\phi(\mu) := \frac{(1-p(\mu))+(1-p(\mu))^2}{2\lambda}$. One checks

$$\phi'(\mu) = \frac{p(\mu)\left[3-2p(\mu)\right]}{2} \in (0, \tfrac{9}{16}) \qquad \text{for } p(\mu) \in (0,1),$$

so ϕ is a contraction. Then, Banach theorem ensures that there is a unique interior fixed point $\mu = \phi(\mu)$. If that fixed point exceeds 1 (small λ or small p_0), the optimum hits the boundary $\mu = 1$.

□

Proof of Proposition 1.12

Proof. Define

$$F(\mu, p_0, \lambda) \;:=\; \mu - \frac{g(p(\mu))}{2\lambda}, \qquad g(p) := (1-p)(2-p) = 2-3p+p^2, \qquad p(\mu) = p_0e^{-\lambda\mu}.$$

An interior optimum μ^* satisfies $F(\mu^*, p_0, \lambda) = 0$. Compute the needed partial derivatives. First,

$$\frac{\partial F}{\partial \mu} = 1 - \frac{g'(p)}{2\lambda}\frac{\partial p}{\partial \mu} = 1 - \frac{(-3+2p)(-\lambda p)}{2\lambda} = 1 - \frac{p(3-2p)}{2} = \frac{2-3p+2p^2}{2}.$$

For $p \in (0,1)$ the quadratic $2-3p+2p^2 > 0$, hence $\partial F/\partial\mu > 0$. By the Implicit

Function Theorem, there is a C^1 function $\mu^*(p_0, \lambda)$ solving $F = 0$, and

$$\frac{\partial \mu^*}{\partial x} = -\left.\frac{\partial F/\partial x}{\partial F/\partial \mu}\right|_{F=0}, \qquad x \in \{p_0, \lambda\}.$$

Derivative with respect to p_0. Holding μ fixed,

$$\frac{\partial F}{\partial p_0} = -\frac{g'(p)}{2\lambda}\frac{\partial p}{\partial p_0} = -\frac{(-3+2p)}{2\lambda}e^{-\lambda\mu} = \frac{(3-2p)p}{2\lambda p_0}.$$

Therefore

$$\frac{\partial \mu^*}{\partial p_0} = -\frac{\dfrac{(3-2p)p}{2\lambda p_0}}{\dfrac{2-3p+2p^2}{2}} = -\frac{p(3-2p)}{\lambda p_0(2-3p+2p^2)} < 0,$$

since $p > 0$, $3 - 2p > 0$ and the denominator is positive.

Derivative with respect to λ. Holding μ fixed,

$$\frac{\partial F}{\partial \lambda} = -\frac{\partial}{\partial \lambda}\left(\frac{g(p)}{2\lambda}\right) = \frac{g(p)}{2\lambda^2} - \frac{g'(p)}{2\lambda}\frac{\partial p}{\partial \lambda} = \frac{g(p)}{2\lambda^2} - \frac{(-3+2p)}{2\lambda}(-\mu p).$$

Using the first order condition, $\mu = \frac{g(p)}{2\lambda}$, gives

$$\frac{\partial F}{\partial \lambda} = \frac{g(p)}{2\lambda^2} - \frac{g(p)p(3-2p)}{4\lambda^2} = \frac{g(p)}{4\lambda^2}\big(2 - p(3-2p)\big) = \frac{g(p)}{2\lambda^2}\frac{\partial F}{\partial \mu}.$$

Hence

$$\frac{\partial \mu^*}{\partial \lambda} = -\frac{\dfrac{g(p)}{2\lambda^2}\dfrac{\partial F}{\partial \mu}}{\dfrac{\partial F}{\partial \mu}} = -\frac{g(p)}{2\lambda^2} = -\frac{(1-p)(2-p)}{2\lambda^2} < 0,$$

because $(1-p)(2-p) > 0$ for $p \in (0,1)$. □

Proof of proposition A.1

Proof. Throughout write $K := (1 - \Delta\theta/2)^2$ and $p := p(\mu)$.

Step 1 (FOC and derivatives). For an interior maximizer,

$$F(\mu; \lambda) := \frac{\partial \Delta\theta \Pi(\mu)}{\partial \mu} = 0.$$

A direct differentiation yields

$$F(\mu;\lambda) = \underbrace{\tfrac{1}{2}\Big[\Delta\theta(1-\Delta\theta) - \tfrac{1}{2}(1-p)K\Big] - \tfrac{1}{4}\lambda\mu K p}_{=:F_A} + \underbrace{\lambda\frac{(1/\Delta\theta - 3/2)}{p}\bar{V}}_{=:F_B} + \underbrace{\left(\frac{1-p}{\Delta\theta^2\gamma p^2} - \frac{\lambda}{\Delta\theta^2\mu}\frac{2-p}{p^2}\right)}_{=:F_C} \tag{A.37}$$

The second derivative $F_\mu = \partial^2\Delta\theta\Pi/\partial\gamma$ is negative at the interior optimum (SOC), so by the Implicit Function Theorem

$$\frac{d\mu^*}{d\lambda} = -\frac{F_\lambda}{F_\mu}\Big|_{\mu=\mu^*}, \qquad \text{hence} \quad \operatorname{sign}\left(\tfrac{d\mu^*}{d\lambda}\right) = \operatorname{sign}(F_\lambda) \ \text{ at } \mu^*.$$

Step 2 (Asymptotics $\lambda \to \infty$). Let $x := \lambda\mu \in (0, \lambda]$, so $p(\mu) = p_0 e^{-x}$. Rewrite $\Delta\theta\Pi$ grouping orders in λ:

$$\Delta\theta\Pi(\mu) = \underbrace{\frac{x}{2\lambda}\Big[\Delta\theta(1-\Delta\theta) - \tfrac{1}{2}\big(1-p_0e^{-x}\big)K\Big]}_{O(1/\lambda)} + \underbrace{\Big(\tfrac{1}{\Delta\theta p_0 e^{-x}} - \tfrac{3}{2p_0e^{-x}} + \tfrac{1}{2} - \tfrac{1}{\Delta}\theta\Big)\bar{V}}_{O(1)} - \underbrace{\lambda\frac{e^{2x} - p_0e^x}{\Delta\theta^2 x p_0^2}\bar{V}^2}_{=:\ \lambda g(x) > 0}.$$

As $\lambda \to \infty$, the dominant term is $-\lambda g(x)\bar{V}^2$, with g continuous on $(0,\infty)$ and $g(x) \to \infty$ as $x \downarrow 0$ or $x \to \infty$. Hence g attains a finite minimum at some $x^* \in (0,\infty)$, and the maximizer of $\Delta\theta\Pi$ must choose x near x^*, i.e.

$$\mu^*(\lambda) = \frac{x^*}{\lambda} + o\Big(\frac{1}{\lambda}\Big) \quad \Longrightarrow \quad \mu^*(\lambda) \to 0,$$

which proves part (i).

Step 3 (Sign of $d\mu^*/d\lambda$ for moderate λ). Differentiate F in (A.37) w.r.t. λ at an interior optimum. Using $p_\lambda = -\mu p$ and collecting terms,

$$F_\lambda = \underbrace{-\tfrac{1}{4}\mu^* K p + \tfrac{1}{4}\lambda\mu^{*2}Kp}_{\text{from } F_A} + \underbrace{\lambda\frac{(1/\Delta\theta - 3/2)}{p}\bar{V}}_{\text{from } F_B} + \underbrace{\frac{\bar{V}^2}{\Delta\theta^2}\Big(\frac{1}{\mu^{*2}}\frac{1-p}{p^2} - \frac{\lambda}{\mu^*}\frac{2-p}{p^2}\Big)}_{\text{from } F_C}.$$

For $\Delta\theta < \frac{2}{3}$ one has $(1/\Delta\theta - 3/2) > 0$. On any compact set of (λ,μ) where the CC + PCM solution is interior and p stays away from 0 and 1, the (positive) middle term $\lambda\frac{(1/\Delta\theta - 3/2)}{p}\bar{V}$ dominates the negative contribution from the $\bar{V}^2$–term for all *moderate* λ and $\bar{V}$ small enough. Hence there exists $\lambda_0 > 0$ (within the regime) such that $F_\lambda(\mu^*(\lambda_0);\lambda_0) > 0$, and therefore $d\mu^*/d\lambda > 0$ at λ_0.

Step 4 (Sign of $d\mu^*/d\lambda$ for large λ). From Step 2, $\mu^*(\lambda) = x^*/\lambda + o(1/\lambda)$ and $p(\mu^*) = p_0 e^{-x^*} \in (0, p_0]$ as $\lambda \to \infty$. Substituting this scaling into F_λ and keeping the leading order gives

$$F_\lambda(\mu^*; \lambda) \;=\; -\frac{\lambda}{\Delta\theta^2} \frac{2 - p_0 e^{-x^*}}{p_0^2 e^{-2x^*}} \cdot \frac{\bar{V}^2}{x^*} \;+\; O(1),$$

which is strictly negative for all sufficiently large λ. Since $F_\mu < 0$ at the optimum, we conclude $d\mu^*/d\lambda < 0$ eventually.

Step 5 (Non-monotonicity). By Steps 3–4 there exist λ_0 with $d\mu^*/d\lambda > 0$ and $\Lambda > \lambda_0$ with $d\mu^*/d\lambda < 0$ (both within the CC + PCM regime). Continuity of $\mu^*(\cdot)$ and of F_λ implies at least one turning point $\lambda^* \in (\lambda_0, \Lambda)$ where $d\mu^*/d\lambda = 0$, hence $\mu^*(\lambda)$ is non-monotonic. This proves part (ii). □

Proof of Proposition A.2

Proof. For this proof assume that $\mu > \mu_2$. In the opposite case, the optimal contract can be solved as in Proposition 1.7 with a different probability of detection. For this part, I will assume that both the (PC_M) and at least a collusion constraint are binding, as if all the Collusion Constraints are non-binding, we are exactly in the non-collusion scenario discussed in the previous section. So, at least, one collusion constraint must be binding. Take again the maximization problem of the principal. Now, as (PC_M) is binding, we can not split the two problems of the monitor. For completeness I will write the maximization problem of the monitor, the Lagrangian and its first order conditions.

$$\max_{\{S,W,e\}} (1-2\mu+\mu_2)\Pi_{\varnothing M} + (\mu-\mu_2)p(\mu)(e_{1,1}+e_{4,1}-W^1(x_{\theta_L,\theta_L},\theta_L)-W^1(x_{\theta_H,\theta_H},\theta_H))$$
$$+(\mu-\mu_2)(1-p(\mu))(e_{2,1}+e_{3,1}-W^1(x_{\theta_L,\varnothing},\varnothing)-W^1(x_{\theta_H,\varnothing},\varnothing))$$
$$+\frac{1}{2}\mu_2(1-(1-p(\mu))^2)\left(e_{1,2}+e_{4,2}-W^2(x_{\theta_L,\theta_L},\theta_L)-W^2(x_{\theta_H,\theta_H},\theta_H)\right)$$
$$+\frac{1}{2}\mu_2(1-p(\mu))^2\left(e_{2,2}+e_{3,2}-W^2(x_{\theta_L,\varnothing},\varnothing)-W^2(x_{\theta_H,\varnothing},\varnothing)\right)$$
$$-(\mu-\mu_2)\left(p(\mu)(S^1(x_{\theta_L,\theta_L},\theta_L)+S^1(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S^1(x_{\theta_L,\varnothing},\varnothing)+S^1(x_{\theta_H,\varnothing},\varnothing)\right.$$
$$-\mu_2\left[p(\mu)(S^2(x_{\theta_L,\theta_L},\theta_L)+S^2(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S^2(x_{\theta_L,\varnothing},\varnothing)+S^2(x_{\theta_H,\varnothing},\varnothing))\right.$$
$$+p(\mu)(1-p(\mu))(\delta-\beta)]$$

s.t.

$$S^1(x_i,r_i)\geq 0,\ i=1,2,3,4 \qquad \text{(LLM)}$$
$$S^2(x_i,r_i)\geq 0,\ i=1,2,4 \qquad \text{(LLM)}$$
$$S^2(x_3,r_3)-\beta\geq 0, \qquad \text{(LLM')}$$
$$\beta\geq 0, \qquad \text{(PenC)}$$
$$W^1(x_i,r_i)-\frac{e_{i,1}^2}{2}\geq 0,\ i=1,2,3,4 \qquad \text{(ICL)}$$
$$W^2(x_i,r_i)-\frac{e_{i,2}^2}{2}\geq 0,\ i=1,2,3,4 \qquad \text{(ICL')}$$
$$W^1(x_{\theta_H,\varnothing},\varnothing)-\frac{e_{3,1}^2}{2}\geq W^1(x_{\theta_L,\varnothing},\varnothing)-\frac{(e_{2,1}-\Delta\theta)^2}{2} \qquad \text{(ICH)}$$
$$W^2(x_{\theta_H,\varnothing},\varnothing)-\frac{e_{3,2}^2}{2}\geq W^2(x_{\theta_L,\varnothing},\varnothing)-\frac{(e_{2,2}-\Delta\theta)^2}{2} \qquad \text{(IC3')}$$
$$(\mu-\mu_2)\left(p(\mu)(S^1(x_{\theta_L,\theta_L},\theta_L)+S^1(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S^1(x_{\theta_L,\varnothing},\varnothing)+S^1(x_{\theta_H,\varnothing},\varnothing))\right)$$
$$+\mu_2\left(p(\mu)(S^2(x_{\theta_L,\theta_L},\theta_L)+S^2(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S^2(x_{\theta_L,\varnothing},\varnothing)+S^2(x_{\theta_H,\varnothing},\varnothing))\right.$$
$$+p(\mu)(1-p(\mu))(\delta-\beta))\geq 2\bar{V} \qquad \text{(PCM)}$$
$$S^1(x_{\theta_H,\theta_H},\theta_H)+W^1(x_{\theta_H,\theta_H},\theta_H)-c(e_{4,1})\geq$$
$$S^1(x_{\theta_H,\varnothing},\varnothing)+W^1(x_{\theta_H,\varnothing},\varnothing)-c(e_{3,1}) \qquad (\text{CC}_{HH})$$
$$2S^2(x_{\theta_H,\theta_H},\theta_H)+W^2(x_{\theta_H,\theta_H},\theta_H)-c(e_{4,2})\geq$$
$$2S^2(x_{\theta_H,\varnothing},\varnothing)+W^2(x_{\theta_H,\varnothing},\varnothing)-c(e_{3,2}) \qquad (\text{CC}_{H2H})$$
$$S^2(x_{\theta_H,\theta_H},\theta_H)+\delta-\beta+W^2(x_{\theta_H,\theta_H},\theta_H)-c(e_{4,2})\geq$$
$$S^2(x_{\theta_H,\varnothing},\varnothing)+W^2(x_{\theta_H,\varnothing},\varnothing)-c(e_{3,2}) \qquad (\text{CC}_{H1H})$$
$$S^2(x_{\theta_H,\theta_H},\theta_H)\geq S^2(x_{\theta_H,\varnothing},\varnothing)+\delta-\beta \qquad (\text{CC}_M)$$

The Lagrangian.

$$
\begin{aligned}
\mathcal{L}() = {} & (1-2\mu+\mu_2)\Pi_{\varnothing M} + (\mu-\mu_2)p(\mu)(e_{1,1}+e_{4,1}-W^1(x_{\theta_L,\theta_L},\theta_L)-W^1(x_{\theta_H,\theta_H},\theta_H)) \\
& + (\mu-\mu_2)(1-p(\mu))(e_{2,1}+e_{3,1}-W^1(x_{\theta_L,\varnothing},\varnothing)-W^1(x_{\theta_H,\varnothing},\varnothing)) \\
& + \frac{1}{2}\mu_2(1-(1-p(\mu))^2)\left(e_{1,2}+e_{4,2}-W^2(x_{\theta_L,\theta_L},\theta_L)-W^2(x_{\theta_H,\theta_H},\theta_H)\right) \\
& + \frac{1}{2}\mu_2(1-p(\mu))^2\left(e_{2,2}+e_{3,2}-W^2(x_{\theta_L,\varnothing},\varnothing)-W^2(x_{\theta_H,\varnothing},\varnothing)\right) \\
& - (\mu-\mu_2)\left(p(\mu)(S^1(x_{\theta_L,\theta_L},\theta_L)+S^1(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S^1(x_{\theta_L,\varnothing},\varnothing)+S^1(x_{\theta_H,\varnothing},\varnothing))\right) \\
& - \mu_2\Big[p(\mu)(S^2(x_{\theta_L,\theta_L},\theta_L)+S^2(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S^2(x_{\theta_L,\varnothing},\varnothing)+S^2(x_{\theta_H,\varnothing},\varnothing)) \\
& + p(\mu)(1-p(\mu))(\delta-\beta)] \\
& + \sum_{i=1,2,3,4}\lambda_i^1\left(S^1(x_i,r_i)\right)+\sum_{i=1,2,4}\lambda_i^2\left(S^2(x_i,r_i)\right)+\lambda_3^2(S^2(x_{\theta_H,\varnothing},\varnothing)-\beta)+\gamma\beta \\
& + \sum_{i=1,2,4}\varphi_i^1\left(W_i^1-\frac{e_{i,1}^2}{2}\right)+\sum_{i=1,2,4}\varphi_i^2\left(W_i^2-\frac{e_{i,2}^2}{2}\right) \\
& + \varphi_3^1\left(W^1(x_{\theta_H,\varnothing},\varnothing)-\frac{e_{3,1}^2}{2}-W^1(x_{\theta_L,\varnothing},\varnothing)+\frac{(e_{2,1}-\Delta\theta)^2}{2}\right) \\
& + \varphi_3^2\left(W^2(x_{\theta_H,\varnothing},\varnothing)-\frac{e_{3,2}^2}{2}W^2(x_{\theta_L,\varnothing},\varnothing)+\frac{(e_{2,2}-\Delta\theta)^2}{2}\right) \\
& + \Pi_1\left(S^1(x_{\theta_H,\theta_H},\theta_H)+W^1(x_{\theta_H,\theta_H},\theta_H)-\frac{e_{4,1}^2}{2}-S^1(x_{\theta_H,\varnothing},\varnothing)-W^1(x_{\theta_H,\varnothing},\varnothing)+\frac{e_{3,1}^2}{2}\right) \\
& + \Pi_3\left(2S^2(x_{\theta_H,\theta_H},\theta_H)+W_4^2-\frac{e_{4,2}^2}{2}-2S^2(x_{\theta_H,\varnothing},\varnothing)-W^2(x_{\theta_H,\varnothing},\varnothing)+\frac{e_{3,2}^2}{2}\right) \\
& + \Pi_4\left(S^2(x_{\theta_H,\theta_H},\theta_H)+\delta-\beta+W_4^2-\frac{e_{4,2}^2}{2}-S^2(x_{\theta_H,\varnothing},\varnothing)-W^2(x_{\theta_H,\varnothing},\varnothing)+\frac{e_{3,2}^2}{2}\right)+\Pi_5\left(S^2(x\right. \\
& + \alpha\Big((\mu-\mu_2)\left(p(\mu)(S^1(x_{\theta_L,\theta_L},\theta_L)+S^1(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S(x_{\theta_L,\varnothing},\varnothing)^1+S^1(x_{\theta_H,\varnothing},\varnothing))\right. \\
& + \mu_2\left(p(\mu)(S^2(x_{\theta_L,\theta_L},\theta_L)+S^2(x_{\theta_H,\theta_H},\theta_H))+(1-p(\mu))(S(x_{\theta_L,\varnothing},\varnothing)^2+S^2(x_{\theta_H,\varnothing},\varnothing))\right. \\
& \left.+p(\mu)(1-p(\mu))(\delta-\beta))-2\bar{V}\right)
\end{aligned}
$$

First order conditions.

$$\frac{\partial \mathcal{L}}{\partial e_{1,1}} = 0 \Rightarrow (\mu - \mu_2)p(\mu) - \varphi_1^1 e_{1,1} = 0 \tag{A.38}$$

$$\frac{\partial \mathcal{L}}{\partial e_{2,1}} = 0 \Rightarrow (\mu - \mu_2)(1 - p(\mu)) - \varphi_2^1 e_{2,1} + \varphi_3^1(e_{2,1} - \Delta\theta) = 0 \tag{A.39}$$

$$\frac{\partial \mathcal{L}}{\partial e_{3,1}} = 0 \Rightarrow (\mu - \mu_2)(1 - p(\mu)) - \varphi_3^1 e_{3,1} + \Pi_1 e_{3,1} = 0 \tag{A.40}$$

$$\frac{\partial \mathcal{L}}{\partial e_{4,1}} = 0 \Rightarrow (\mu - \mu_2)p(\mu) - \varphi_4^1 e_{4,1} - \Pi_1 e_{4,1} = 0 \tag{A.41}$$

$$\frac{\partial \mathcal{L}}{\partial e_{1,2}} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - (1 - p(\mu))^2) - \varphi_1^2 e_{1,2} = 0 \tag{A.42}$$

$$\frac{\partial \mathcal{L}}{\partial e_{2,2}} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - p(\mu))^2 - \varphi_2^2 e_{2,2} + \varphi_3^2(e_{2,2} - \Delta\theta) = 0 \tag{A.43}$$

$$\frac{\partial \mathcal{L}}{\partial e_{3,2}} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - p(\mu))^2 - \varphi_3^2 e_{3,2} + \Pi_3 e_{3,2} + \Pi_4 e_{3,2} = 0 \tag{A.44}$$

$$\frac{\partial \mathcal{L}}{\partial e_{4,2}} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - (1 - p(\mu))^2) - \varphi_4^2 e_{4,2} - \Pi_3 e_{4,2} - \Pi_4 e_{4,2} = 0 \tag{A.45}$$

$$\frac{\partial \mathcal{L}}{\partial W^1(x_{\theta_L,\theta_L}, \theta_L)} = 0 \Rightarrow (\mu - \mu_2)p(\mu) - \varphi_1^1 = 0 \tag{A.46}$$

$$\frac{\partial \mathcal{L}}{\partial W^1(x_{\theta_L,\varnothing}, \varnothing)} = 0 \Rightarrow (\mu - \mu_2)(1 - p(\mu)) - \varphi_2^1 + \varphi_3^1 = 0 \tag{A.47}$$

$$\frac{\partial \mathcal{L}}{\partial W^1(x_{\theta_H,\varnothing}, \varnothing)} = 0 \Rightarrow (\mu - \mu_2)(1 - p(\mu)) - \varphi_3^1 + \Pi_1 = 0 \tag{A.48}$$

$$\frac{\partial \mathcal{L}}{\partial W^1(x_{\theta_H,\theta_H}, \theta_H)} = 0 \Rightarrow (\mu - \mu_2)p(\mu) - \varphi_4^1 - \Pi_1 = 0 \tag{A.49}$$

$$\frac{\partial \mathcal{L}}{\partial W^2(x_{\theta_L,\theta_L}, \theta_L)} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - (1 - p(\mu))^2) - \varphi_1^2 = 0 \tag{A.50}$$

$$\frac{\partial \mathcal{L}}{\partial W^2(x_{\theta_L,\varnothing}, \varnothing)} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - p(\mu))^2 - \varphi_2^2 + \varphi_3^2 = 0 \tag{A.51}$$

$$\frac{\partial \mathcal{L}}{\partial W^1(x_{\theta_H,\varnothing}, \varnothing)} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - p(\mu))^2 - \varphi_3^2 + \Pi_3 + \Pi_4 = 0 \tag{A.52}$$

$$\frac{\partial \mathcal{L}}{\partial W^2(x_{\theta_H,\theta_H}, \theta_H)} = 0 \Rightarrow \frac{1}{2}\mu_2(1 - (1 - p(\mu))^2) - \varphi_4^2 - \Pi_3 - \Pi_4 = 0 \tag{A.53}$$

$$\frac{\partial \mathcal{L}}{\partial S^1(x_{\theta_L,\theta_L},\theta_L)} = 0 \Rightarrow -(\mu-\mu_2)p(\mu) + \lambda_1^1 + \alpha(\mu-\mu_2)p(\mu) = 0 \tag{A.54}$$

$$\frac{\partial \mathcal{L}}{\partial S^1(x_{\theta_L,\varnothing},\varnothing)} = 0 \Rightarrow -(\mu-\mu_2)(1-p(\mu)) + \lambda_2^1 + \alpha(\mu-\mu_2)(1-p(\mu)) = 0 \tag{A.55}$$

$$\frac{\partial \mathcal{L}}{\partial S^1(x_{\theta_H,\varnothing},\varnothing)} = 0 \Rightarrow -(\mu-\mu_2)(1-p(\mu)) + \alpha(\mu-\mu_2)(1-p(\mu)) + \lambda_3^1 - \Pi_1 = 0 \tag{A.56}$$

$$\frac{\partial \mathcal{L}}{\partial S^1(x_{\theta_H,\theta_H},\theta_H)} = 0 \Rightarrow -(\mu-\mu_2)p(\mu) + \lambda_4^1 + \alpha(\mu-\mu_2)p(\mu) + \Pi_1 = 0 \tag{A.57}$$

$$\frac{\partial \mathcal{L}}{\partial S^2(x_{\theta_L,\theta_L},\theta_L)} = 0 \Rightarrow -\mu_2 p(\mu) + \lambda_1^2 + \alpha\mu_2 p(\mu) = 0 \tag{A.58}$$

$$\frac{\partial \mathcal{L}}{\partial S^2(x_{\theta_L,\varnothing},\varnothing)} = 0 \Rightarrow -\mu_2(1-p(\mu)) + \lambda_2^2 + \alpha\mu_2(1-p(\mu)) = 0 \tag{A.59}$$

$$\frac{\partial \mathcal{L}}{\partial S^2(x_{\theta_H,\varnothing},\varnothing)} = 0 \Rightarrow -\mu_2(1-p(\mu)) + \alpha\mu_2(1-p(\mu)) + \lambda_3^2 - 2\Pi_3 - \Pi_4 - \Pi_5 = 0 \tag{A.60}$$

$$\frac{\partial \mathcal{L}}{\partial S^2(x_{\theta_H,\theta_H},\theta_H)} = 0 \Rightarrow -\mu_2 p(\mu) + \lambda_4^2 + \alpha\mu_2 p(\mu) + 2\Pi_3 + \Pi_4 + \Pi_5 = 0 \tag{A.61}$$

$$\frac{\partial \mathcal{L}}{\partial \delta} = 0 \Rightarrow -\mu_2 p(\mu)(1-p(\mu)) + \alpha\mu_2 p(\mu)(1-p(\mu)) + \Pi_4 - \Pi_5 = 0 \tag{A.62}$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = 0 \Rightarrow \mu_2 p(\mu)(1-p(\mu)) + \gamma - \alpha\mu_2 p(\mu)(1-p(\mu)) - \lambda_3^2 - \Pi_4 + \Pi_5 = 0 \tag{A.63}$$

As in previous proofs, it is immediate to see that $e_{1,1} = e_{3,1} = e_{4,1} = 1$ and $e_{1,2} = e_{3,2} = e_{4,2} = 1$. At the same time, using a similar reasoning to previous proofs, we can see that $W^1(x_{\theta_L,\theta_L},\theta_L) = W^2(x_{\theta_L,\theta_L},\theta_L) = \frac{1}{2}$, $W^1(x_{\theta_L,\varnothing},\varnothing) = \frac{e_{2,1}^2}{2}$, and $W^2(x_{\theta_L,\varnothing},\varnothing) = \frac{e_{2,2}^2}{2}$. So the low type agents do not receive any rents under the optimal contract. From the FOCs of $W^1(x_{\theta_H,\varnothing},\varnothing)$ and $W^1(x_{\theta_H,\varnothing},\varnothing)$ we have that both ($IC_{H\varnothing}$) and (IC3') are binding.

$$W^1(x_{\theta_H,\varnothing},\varnothing) = \frac{1-\Delta\theta^2}{2} + e_{2,1}\Delta\theta$$

$$W^2(x_{\theta_H,\varnothing},\varnothing) = \frac{1-\Delta\theta^2}{2} + e_{2,2}\Delta\theta$$

Before continuing with the proof, notice that if $\alpha = 1$, all the $\{\lambda_i^j\}_{i=1,2,3,4}^{j=1,2}$ and the $\{\Pi_k\}_{k=1,3,4,5}$, multipliers will be zero. This implies that the contract is the non-collusion contract, which by assumption violates the collusion constraints.

For this reason, assuming that any $\{\lambda_i^j\}_{i=1,2}^{j=1,2}$ is zero makes $\alpha = 1$ and drives to a contradiction. Hence, $\lambda_1^1 > 0$, $\lambda_2^1 > 0$, $\lambda_1^2 > 0$, $\lambda_2^2 > 0$. This implies that the associated wages are equal to zero. Adding equations (18) and (19), we can argue that $\lambda_3^1 > 0$, and then $S^1(x_{\theta_H,\varnothing}, \varnothing) = 0$. The same logic applies to equations (22) and (23) which mean that $S^2(x_{\theta_H,\varnothing}, \varnothing) = 0$.

Adding (25) and (26), we can see that $\gamma > 0$, which implies that $\beta = 0$. Hence, substituting into the Participation Constraint of the Monitor we get,

$$(\mu - \mu_2)p(\mu)S^1(x_{\theta_H,\theta_H}, \theta_H) + \mu_2 p(\mu)(S^2(x_{\theta_H,\theta_H}, \theta_H) + (1 - p(\mu))\delta) = 2\bar{V}$$

If $\Pi_1 = 0$, by (20), $\alpha = 1$, which is a contradiction as already stated. Hence, $\Pi_1 > 0$, and (CC_{HH}) binds. Using the same reasoning, if $\Pi_4 = \Pi_5$, equation (25) shows that $\alpha = 2$, which is a contradiction with equation (24) as it requires that some multiplier is negative. Hence, $\Pi_4 > \Pi_5 \geq 0$, and (CC_{H1H}) binds.

Now assume that $\Pi_5 = 0$. This means that (CC_M) is not binding and therefore (CC_{H2H}) is also slack. Hence, $\Pi_3 = 0$. Adding equations (24) and (25) we can see that $\alpha = 1$ reaching again a contradiction. This implies that all the collusion constrains and the participation constraint is binding.

Assume now that $W^1(x_{\theta_H,\theta_H}, \theta_H) > \frac{1}{2}$ (a similar logic applies to $W^2(x_{\theta_H,\theta_H}, \theta_H)$). Then, $\varphi_4^1 = 0$ and, by (12), $\Pi_1 = (\mu - \mu_2)p(\mu)$. Using a similar procedure to previous proof, we can argue that, under this assumption, $e_{2,1} = 1 - \frac{\Delta\theta}{1-p(\mu)}$. If both $W^1(x_{\theta_H,\theta_H}, \theta_H)$ and $W^2(x_{\theta_H,\theta_H}, \theta_H)$ are larger than $\frac{1}{2}$, the optimal contract coincides with the collusion contract, which by assumption does not bind in the participation constraint of the monitor.

As in previous proofs, assume that, $W^1(x_{\theta_H,\theta_H}, \theta_H) = \frac{1}{2}$, $W^2(x_{\theta_H,\theta_H}, \theta_H) = \frac{1}{2}$. From the collusion constraints,

$$S^1(x_{\theta_H,\theta_H},\theta_H) = \Delta\theta\left(e_{2,1} - \frac{\Delta\theta}{2}\right)$$
$$S^2(x_{\theta_H,\theta_H},\theta_H) = \frac{1}{2}\Delta\theta\left(e_{2,2} - \frac{\Delta\theta}{2}\right)$$

Notice that the whole maximization problem is written in terms of $e_{2,1}$ and $e_{2,2}$ with (PC_M) being binding.

$$(\mu-\mu_2)\left(e_{2,1} - \frac{\Delta\theta}{2}\right) + \mu_2\left(1 - \frac{1}{2}p(\mu)\right)\left(e_{2,2} - \frac{\Delta\theta}{2}\right) = \frac{2\bar{V}}{\Delta\theta p(\mu)}.$$

Hence, I can rewrite the whole program and maximize as a function of only these two variables with λ being the multiplier associated to (PC_M). Call this new lagrangian $\tilde{\mathcal{L}}$. This means that,

$$\frac{\partial\tilde{\mathcal{L}}}{\partial e_{2,1}} = 0 \Rightarrow (\mu-\mu_2)(1-p(\mu)(1-e_{2,1}-\Delta\theta) + (\mu-\mu_2)\lambda = 0$$
$$\frac{\partial\tilde{\mathcal{L}}}{\partial e_{2,1}} = 0, \Rightarrow \frac{1}{2}\mu_2(1-p(\mu))^2(1-e_{2,2}-\Delta\theta) + \mu_2\frac{1}{2}\left(2-p(\mu)\right)\lambda = 0.$$

This allows me to express λ as a function of $e_{2,1}$ and substitute.

$$\lambda = (1-p(\mu)(e_{2,1}-1+\Delta\theta)$$

$$\frac{(2-p(\mu))}{1-p(\mu)}e_{2,1} - \frac{1-\Delta\theta}{1-p(\mu)} = e_{2,2}$$

Substituting back into the (PC_M) we get,

$$e_{2,1} = \frac{\dfrac{2\bar{V}}{\Delta\theta p(\mu)} + (\mu-\mu_2)\cdot\dfrac{\Delta\theta}{2} + \mu_2\left(1-\dfrac{p(\mu)}{2}\right)\left(\dfrac{1-\Delta\theta}{1-p(\mu)} + \dfrac{\Delta\theta}{2}\right)}{(\mu-\mu_2) + \mu_2\left(1-\dfrac{p(\mu)}{2}\right)\cdot\dfrac{2-p(\mu)}{1-p(\mu)}},$$

and,

$$e_{2,2} = \frac{2-p}{1-p} \cdot \frac{\frac{2\bar{V}}{\Delta\theta p} + (\mu - \mu_2) \cdot \frac{\Delta\theta}{2} + \mu_2 \left(1 - \frac{p}{2}\right) \left(\frac{1-\Delta\theta}{1-p} + \frac{\Delta\theta}{2}\right)}{(\mu - \mu_2) + \mu_2 \left(1 - \frac{p}{2}\right) \cdot \frac{2-p}{1-p}} - \frac{1-\Delta\theta}{1-p}.$$

□

Appendix B

Appendix to Chapter 2

B.1 Proofs

Lemma B.1 (Variance of θ_j given θ_i)**.** *Let $\theta_i = \theta + \varepsilon_i$, and $\theta_j = \theta + \varepsilon_j$, where $(\varepsilon_i, \varepsilon_j)$ is bivariate normal with variances σ_{ii}, σ_{jj} and correlation σ_{ij}. Then, if we assume an improper prior for θ,*

$$\theta|\theta_i \sim \mathcal{N}\Big(\theta_i,\ \sigma_{ii}\Big),\ \textit{and} \quad \theta_j|\theta_i \sim \mathcal{N}\Big(\theta_i,\ \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}\Big).$$

Proof. The conditional distribution of θ given θ_i follows from Bayes's rule with an improper prior on θ, which implies $f(\theta|\theta_i) \propto f(\theta_i|\theta)$. Since $\theta_i|\theta \sim \mathcal{N}(\theta, \sigma_{ii})$, the posterior distribution is

$$\theta|\theta_i \sim \mathcal{N}(\theta_i, \sigma_{ii}).$$

The conditional distribution of θ_j given θ_i is derived analogously below. Using the definition of conditional density,

$$f(\theta_j|\theta_i, \theta) = \frac{f(\theta_j, \theta_i|\theta)}{f(\theta_i|\theta)}.$$

Under the joint normality of (θ_i, θ_j) conditional on θ, the conditional distribution is given by the standard bivariate–normal formula

$$\theta_j|\theta_i, \theta \sim \mathcal{N}\left(\theta + \rho_{ij}\sqrt{\frac{\sigma_{jj}}{\sigma_{ii}}}(\theta_i - \theta),\ \sigma_{jj}(1 - \rho_{ij}^2)\right), \tag{2}$$

where $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$. For fixed θ_i, the joint conditional density of (θ, θ_j) is given by

$$f(\theta, \theta_j|\theta_i) = f(\theta_j|\theta, \theta_i) f(\theta|\theta_i).$$

Because both $\theta_j|(\theta_i, \theta)$ and $\theta|\theta_i$ are normal, their product is proportional to the exponential of a quadratic form in (θ, θ_j). Hence the joint density $f(\theta, \theta_j|\theta_i)$ is normal in (θ, θ_j), and up to a normalizing constant, its exponent is

$$-\frac{1}{2}\left[\frac{\left(\theta_j - \theta - \rho_{ij}\frac{\sqrt{\sigma_{jj}}}{\sqrt{\sigma_{ii}}}(\theta_i - \theta)\right)^2}{\sigma_{jj}(1-\rho_{ij}^2)} + \frac{(\theta - \theta_i)^2}{\sigma_{ii}}\right]. \tag{B.1}$$

The numerator of the first fraction in Equation (B.1) is

$$\left(\theta_j - \theta - \rho_{ij}\frac{\sqrt{\sigma_{jj}}}{\sqrt{\sigma_{ii}}}(\theta_i - \theta)\right)^2 = (\theta_j - \theta)^2 - 2\rho_{ij}\frac{\sqrt{\sigma_{jj}}}{\sqrt{\sigma_{ii}}}(\theta_j - \theta)(\theta_i - \theta) + \rho_{ij}^2\frac{\sigma_{jj}}{\sigma_{ii}}(\theta_i - \theta)^2.$$

Substituting back, Equation (B.1) becomes

$$-\frac{1}{2(1-\rho_{ij}^2)}\left[\frac{(\theta_j - \theta)^2}{\sigma_{jj}} - \frac{2\rho_{ij}}{\sqrt{\sigma_{jj}\sigma_{ii}}}(\theta_j - \theta)(\theta_i - \theta) + \frac{(\theta_i - \theta)^2}{\sigma_{ii}}\right].$$

Define

$$A = \frac{1}{1-\rho_{ij}^2}\left(\frac{1}{\sigma_{jj}} + \frac{1}{\sigma_{ii}} - \frac{2\rho_{ij}}{\sqrt{\sigma_{jj}\sigma_{ii}}}\right); \quad B = \frac{1}{1-\rho_{ij}^2}\left(\frac{\theta_j}{\sigma_{jj}} + \frac{\theta_i}{\sigma_{ii}} - \frac{\rho_{ij}}{\sqrt{\sigma_{jj}\sigma_{ii}}}(\theta_j + \theta_i)\right);$$
$$\text{and} \quad C = \frac{1}{1-\rho_{ij}^2}\left(\frac{\theta_j^2}{\sigma_{jj}} - \frac{2\rho_{ij}\theta_j\theta_i}{\sqrt{\sigma_{jj}\sigma_{ii}}} + \frac{\theta_i^2}{\sigma_{ii}}\right),$$

so that Equation (B.1) becomes $-\frac{1}{2}\left[A\theta^2 - 2B\theta + C\right]$. Integrating over θ eliminates the latent parameter and yields the marginal density of θ_j conditional on θ_i.

$$\int_{\mathbb{R}} \exp\left[-\tfrac{1}{2}\left(A\theta^2 - 2B\theta + C\right)\right] d\theta = \sqrt{\tfrac{2\pi}{A}} \exp\frac{1}{2}\left[\tfrac{B^2}{A} - C\right].$$

Thus the marginal density of θ_j given θ_i is proportional to $\exp[-\frac{1}{2}(C - \frac{B^2}{A})]$, and direct simplification yields

$$C - \frac{B^2}{A} = \frac{(\theta_j - \theta_i)^2}{\sigma_{ii} + \sigma_{jj} - 2\rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Hence,

$$\theta_j|\theta_i \sim \mathcal{N}\left(\theta_i,\ \sigma_{ii} + \sigma_{jj} - 2\rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}\right),$$

so the conditional variance is $\sigma_{ii} + \sigma_{jj} - 2\rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}$, or equivalently $\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}$. □

Proof of Proposition 2.1

Proof. The proof proceeds in two parts. The first establishes that truthful reporting is the unique equilibrium for rational users. The second shows that, for any specification of the function g, truthful reporting also constitutes an equilibrium for naive users.

Naive users. The argument for naive users proceeds in three steps. First, the message that maximizes engagement also maximizes learning. Second, the message maximizing within–the-platform utility coincides with the one that maximizes expected engagement. Third, truthful reporting maximizes within–the-platform utility. The second step provides the link between the two optimization problems: it connects the learning-maximization result in the first step with the utility-maximization argument in the third, ensuring that truthful reporting jointly maximizes engagement, learning, and utility.

First, note that a naive user who cares only about learning will choose the message that maximizes engagement. For such users, engagement and learning are monotonically related: given any algorithm and message profile, if the user reads weakly more posts ($e_i' \geq e_i$), her posterior variance weakly decreases, and learning weakly improves. Hence, under any algorithm, the message that maximizes engagement also maximizes learning. This follows directly from Bayesian updating: conditional on the algorithm, observing more signals implies a weakly lower posterior variance.

Next, the message that maximizes the expected within–the-platform utility when $m_j = \theta_j$ also maximizes the expected engagement. By definition of a random variable with finite support, the expected engagement is

$$\mathbb{E}\left[e_i|m_i, \theta_i, \mathscr{F}_i\right] = \sum_{k=1}^{n} k \Pr[e_i = k|m_i, \theta_i, \mathscr{F}_i]. \tag{B.2}$$

With the continuation function $g(\cdot)$, [1]

$$\Pr[e_i = k|m_i, \theta_i, \mathscr{F}_i] = \Big(1 - g(u_i(m_i, k))\Big) \prod_{r=1}^{k-1} g(u_i(m_i, r)) \quad \text{for } 1 \le k \le n-1;$$

$$\Pr[e_i = n|m_i, \theta_i, \mathscr{F}_i] = \prod_{r=1}^{n-1} g(u_i(m_i, r)).$$

Therefore,

$$\mathbb{E}[e_i|m_i, \theta_i, \mathscr{F}_i] = \mathbb{E}_{\theta_{-i}}\left[\sum_{k=1}^{n-1} k\Big(1-g(u_i(m_i,k))\Big) \prod_{r=1}^{k-1} g(u_i(m_i,r)) + n\prod_{r=1}^{n-1} g(u_i(m_i,r))\right].$$

Here, $\mathbb{E}_{\theta_{-i}}[\cdot]$ denotes the expectation over the other users' signals, conditional on θ_i, while keeping the realized engagement path e_i fixed. Differentiating expected engagement with respect to m_i, and applying the chain and product rules, yields

$$\begin{aligned}\frac{\partial}{\partial m_i}\mathbb{E}[e_i|m_i, \theta_i, \mathscr{F}_i] = \mathbb{E}_{\theta_{-i}}\Bigg[\sum_{k=1}^{n-1} k\Bigg(& (1 - g(u_i(m_i,k))) \prod_{r=1}^{k-1} g(u_i(m_i,r)) \sum_{s=1}^{k-1} \frac{g'(u_i(m_i,s))}{g(u_i(m_i,s))} \frac{\partial u_i(m_i,s)}{\partial m_i} \\ & - g'(u_i(m_i,k)) \frac{\partial u_i(m_i,k)}{\partial m_i} \prod_{r=1}^{k-1} g(u_i(m_i,r))\Bigg) \\ & + n \prod_{r=1}^{n-1} g(u_i(m_i,r)) \sum_{s=1}^{n-1} \frac{g'(u_i(m_i,s))}{g(u_i(m_i,s))} \frac{\partial u_i(m_i(s))}{\partial m_i}\Bigg]. \end{aligned} \tag{B.3}$$

The next step is to show that truthful reporting constitutes a best reply to truthful reporting for user i. When all other users report truthfully, $m_{-i} = \theta_{-i}$, the within-the-platform utility of user i at engagement level k is given by

$$u_i(m_i, k; \theta_{-i}, \mathscr{F}) = -\beta(m_i - \theta_i)^2 - (1-\beta) \sum_{j \in \mathscr{F}^k} (m_i - \theta_j)^2.$$

The first derivative of $u_i(m_i, k; \theta_{-i}, \mathscr{F})$ with respect to m_i is

$$\frac{\partial u_i(m_i, k; \theta_{-i}, \mathscr{F})}{\partial m_i} = -2\beta(m_i - \theta_i) - 2(1-\beta) \sum_{j \in \mathscr{F}^k} (m_i - \theta_j).$$

Define $z_j := m_i - \theta_j$. By Lemma B.1, $\mathbb{E}[\theta_j|\theta_i] = \theta_i$, so that $\mathbb{E}[z_j|\theta_i] = m_i - \theta_i$, which equals zero under truthful reporting. All expressions are evaluated at $m_i = \theta_i$, so the sincerity term $-\beta(m_i - \theta_i)^2$ vanishes, and u_i, g, and their

[1] We write $u_i(m_i, k)$ for the within-the-platform utility when user i sends message m_i and reads up to k posts of the algorithm.

derivatives depend only on the squared deviations z_j^2. Hence the function $f(z_j) := \frac{\partial u_i(m_i,k;\theta_j,\mathscr{F})}{\partial m_i}$ is odd in z_j, while the remaining terms are even. Defining $h(z_j) := u_i(m_i = \theta_i, k; \theta_j, \mathscr{F})$, implies $h(-z_j) = h(z_j)$, so h is an even function of z_j. The same property holds for $g(h(z_j))$ and $g'(h(z_j))$. Since the product of even functions is even, all terms in Equation (B.3) that do not involve $f(z_j)$ are even functions. Moreover, the product of an even and an odd function is odd. Therefore, when Equation (B.3) is expressed in terms of z_j, each summand is an odd function of z_j.

As the expectation of any symmetric odd function is zero, Equation (B.3) is zero when $m_{-i} = \theta_{-i}$ and $m_i = \theta_i$. Therefore, $m_i = \theta_i$ is a best reply to $m_{-i} = \theta_{-i}$; truthful reporting is a best reply to truthful reporting. Hence, evaluated at $m_i = \theta_i$, the *total* derivative of expected engagement with respect to the message m_i is zero. Truthful reporting satisfies the first–order condition for maximizing expected engagement.

However, this condition alone does not guarantee that expected utility is maximized, since the latter expectation is taken over both e_i and θ_j. To complete the argument, it remains to show that, for any given feed, if all other users report truthfully, user i maximizes expected within-the-platform utility by reporting truthfully as well, that is, $m_i = \theta_i$. Fix θ_i and assume $m_j = \theta_j$ for all $j \neq i$. The within-the-platform expected utility is

$$\mathbb{E}[u_i(m_i, \theta_i; m_{-i})|\theta_i, \mathscr{F}_i] = \alpha\mathbb{E}[e_i|\theta_i, \mathscr{F}_i] - \beta\mathbb{E}[(\theta_i - m_i)^2|\theta_i] - (1-\beta)\mathbb{E}\left[\sum_{j\in\mathscr{F}_i^{e_i}} (m_i - \theta_j)^2 \,\middle|\, \theta_i, \mathscr{F}_i\right].$$

The second term simplifies to $-\beta(m_i - \theta_i)^2$. For the third term, the quadratic decomposition $(m_i - \theta_j)^2 = (m_i - \theta_i)^2 + (\theta_j - \theta_i)^2 + 2(m_i - \theta_i)(\theta_i - \theta_j)$ is applied to separate the sincerity deviation, the intrinsic difference between users, and the cross term. Taking conditional expectations given θ_i and using $\mathbb{E}[\theta_j|\theta_i] = \theta_i$, the cross term vanishes and, for any engagement length k,

$$\mathbb{E}\left[\sum_{j\in\mathscr{F}_i^k} (m_i - \theta_j)^2 \middle| \theta_i, \mathscr{F}_i\right] = k(m_i - \theta_i)^2 + \sum_{j\in\mathscr{F}_i^k} \mathbb{E}\Big[(\theta_j - \theta_i)^2|\theta_i, \mathscr{F}_i\Big].$$

Hence, irrespective of how e_i is distributed,

$$\mathbb{E}\left[\sum_{j\in\mathscr{F}_i^{e_i}}(m_i-\theta_j)^2\middle|\theta_i,\mathscr{F}_i\right]=\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big](m_i-\theta_i)^2+\mathbb{E}\left[\sum_{j\in\mathscr{F}_i^{e_i}}(\theta_j-\theta_i)^2\middle|\theta_i,\mathscr{F}_i\right].$$

Substituting back gives

$$\begin{aligned}\mathbb{E}\Big[u_i(m_i,\theta_i;m_{-i})|\theta_i,\mathscr{F}_i\Big]=\alpha\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big]-\Big[\beta+(1-\beta)\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big]\Big](m_i-\theta_i)^2\\-(1-\beta)\mathbb{E}\left[\sum_{j\in\mathscr{F}_i^{e_i}}(\theta_j-\theta_i)^2\middle|\theta_i,\mathscr{F}_i\right].\end{aligned}$$

Fixing other players' strategies, let the expected utility be denoted by $q(m_i)$. Its first derivative is

$$\begin{aligned}q'(m_i)=\alpha\frac{\partial}{\partial m_i}\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big]-2\Big[\beta+(1-\beta)\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big]\Big](m_i-\theta_i)\\-(1-\beta)\frac{\partial}{\partial m_i}\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big](m_i-\theta_i)^2.\end{aligned}$$

The first term is zero at $m_i=\theta_i$ because truthful reporting maximizes the expected engagement. The remaining two terms are also zero at this point. It therefore suffices to verify that the function is concave at $m_i=\theta_i$ to establish that truthful reporting constitutes a maximum, and hence a best reply. Differentiating q once more yields

$$\begin{aligned}q''(m_i)=\alpha\frac{\partial^2}{\partial m_i^2}\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big]-2\Big[\beta+(1-\beta)\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big]\Big]\\-4(1-\beta)\frac{\partial}{\partial m_i}\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big](m_i-\theta_i)\\-(1-\beta)\frac{\partial^2}{\partial m_i^2}\mathbb{E}\Big[e_i|\theta_i,\mathscr{F}_i\Big](m_i-\theta_i)^2.\end{aligned}$$

Since truthful reporting maximizes expected engagement, the first term is non-positive. The last two terms vanish when $m_i=\theta_i$, and the second term is strictly negative. Consequently, $q''(m_i)<0$ at $m_i=\theta_i$, which establishes that truthful reporting attains a local maximum and therefore constitutes a best reply.

Rational users. Rational user i chooses a message $m_i\in\mathbb{R}$ to maximize her within-the-platform expected utility given her private signal θ_i and the algorithm

$\mathscr{F}$:

$$\mathbb{E}[u_i|\theta_i,\mathscr{F}] = \lambda\Big(\alpha\mathbb{E}[e_i|\theta_i,\mathscr{F}]-\beta(\theta_i-m_i)^2-(1-\beta)\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}}(m_i-m_j(\theta_j))^2\,\Big|\,\theta_i,\mathscr{F}\Big]\Big).$$

Because engagement e_i is chosen at a later stage of the game, the user's message choice m_i does not influence expected engagement. Formally, by the envelope theorem, when e_i is chosen optimally given m_i, the derivative of the maximized value with respect to m_i ignores the indirect effect of m_i on e_i, so that $\frac{\partial}{\partial m_i}\max_{e_i}\mathbb{E}[U_i] = \partial_{m_i}\mathbb{E}[U_i]$ at interior points.[2] Conditional on the algorithm and engagement level, m_i affects only the conformity component of utility and has no impact on learning.

Thus, the optimal m_i maximizes

$$-\beta(\theta_i-m_i)^2-(1-\beta)\Big(e_i m_i^2+\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}}m_j(\theta_j)^2\,\Big|\,\theta_i,\mathscr{F}\Big]-2m_i\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}}m_j(\theta_j)\,\Big|\,\theta_i,\mathscr{F}\Big]\Big).$$

The first-order condition with respect to m_i is

$$\beta\theta_i + (1-\beta)\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}}m_j(\theta_j)\;\Big|\;\theta_i,\mathscr{F}\Big] = m_i. \tag{B.4}$$

Under truthful reporting by all other users, $m_j(\theta_j)=\theta_j$ for every $j\neq i$, and the right-hand side of Equation (B.4) is maximized at $m_i=\theta_i$. Hence, truthful reporting constitutes a best reply and therefore an equilibrium.

Uniqueness of the equilibrium for rational users follows by successive substitution. For any user ℓ, define $D_\ell := \beta + (1-\beta)e_\ell$. Since the first-order condition holds for every user, each user's best reply can be expressed in terms of the best replies of her neighbors. Iterating this substitution for neighbors, and for neighbors of neighbors up to depth m, yields:

[2] Boundary cases $e_i \in \{1, n\}$ can be treated separately and do not affect the main argument.

$$
\begin{aligned}
m_i &= \frac{\beta}{D_i}\theta_i + \frac{1-\beta}{D_i}\mathbb{E}_i\Big[\sum_{j\in\mathscr{F}_i^{e_i}}\Big(\frac{\beta}{D_j}\theta_j + \frac{1-\beta}{D_j}\mathbb{E}_j\Big[\sum_{l\in\mathscr{F}_j^{e_j}} m_l(\theta_l)\Big]\Big)\Big] \\
&= \frac{\beta}{D_i}\theta_i + \frac{1-\beta}{D_i}\mathbb{E}_i\Big[\sum_{j\in\mathscr{F}_i^{e_i}}\frac{\beta}{D_j}\theta_j\Big] + \frac{1-\beta}{D_i}\mathbb{E}_i\Big[\sum_{j\in\mathscr{F}_i^{e_i}}\frac{1-\beta}{D_j}\mathbb{E}_j\Big[\sum_{l\in\mathscr{F}_j^{e_j}}\frac{\beta}{D_l}\theta_l\Big]\Big] \\
&\quad + \frac{1-\beta}{D_i}\mathbb{E}_i\Big[\sum_{j\in\mathscr{F}_i^{e_i}}\frac{1-\beta}{D_j}\mathbb{E}_j\Big[\sum_{l\in\mathscr{F}_j^{e_j}}\frac{1-\beta}{D_l}\mathbb{E}_l\Big[\sum_{q\in\mathscr{F}_l^{e_l}} m_q(\theta_q)\Big]\Big]\Big] + \cdots \\
&= \frac{\beta}{D_i}\theta_i + \sum_{r=1}^{m-1}\mathbb{E}_i\Big[\sum_{j_1\in\mathscr{F}_i^{e_i}}\frac{1-\beta}{D_i}\sum_{j_2\in\mathscr{F}_{j_1}^{e_{j_1}}}\frac{1-\beta}{D_{j_1}}\cdots\sum_{j_r\in\mathscr{F}_{j_{r-1}}^{e_{j_{r-1}}}}\frac{\beta}{D_{j_r}}\theta_{j_r}\Big] \\
&\quad + \mathbb{E}_i\Big[\sum_{j_1\in\mathscr{F}_i^{e_i}}\frac{1-\beta}{D_i}\sum_{j_2\in\mathscr{F}_{j_1}^{e_{j_1}}}\frac{1-\beta}{D_{j_1}}\cdots\sum_{j_m\in\mathscr{F}_{j_{m-1}}^{e_{j_{m-1}}}}\frac{1-\beta}{D_{j_{m-1}}}\mathbb{E}_{j_m}\Big[\sum_{p\in\mathscr{F}_{j_m}^{e_{j_m}}} m_p(\theta_p)\Big]\Big].
\end{aligned}
\tag{B.5}
$$

Because of improper priors and the law of iterated expectations, $\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}}\theta_j|\theta_i,\mathscr{F}\Big] = e_i\theta_i$. Substituting this identity into the inner expectation in Equation (B.5) allows each θ_j to be replaced by θ_i. Applying the same substitution recursively at every subsequent depth implies that each term of order r is a nonnegative scalar multiple of θ_i.

The remainder of the expression (the third term in Equation (B.5)) converges to zero. In such term, each additional nesting at depth ℓ contributes a factor $\frac{1-\beta}{D_\ell} = \frac{1-\beta}{\beta+(1-\beta)e_\ell}$, and the adjacent sum introduces at most e_ℓ addends. Hence the depth-ℓ layer scales the upstream magnitude by $\frac{(1-\beta)e_\ell}{\beta+(1-\beta)e_\ell} < 1$. Define

$$
\gamma_{\max} := \max_{\ell}\frac{(1-\beta)e_\ell}{\beta+(1-\beta)e_\ell} < 1 \quad \text{(finite since } e_\ell \le n\text{)}.
$$

Under the quadratic loss, admissible strategies satisfy $\mathbb{E}_i[m_j(\theta_j)^2] < \infty$; if reading a message yielded an expected disutility larger than the utility obtainable in isolation ($e_i = 1$), the user would not read it. By the Cauchy–Schwarz inequality, there exists a finite constant $K := \sup_{j\in\mathcal{U}}\mathbb{E}_i[|m_j(\theta_j)|] < \infty$. Since conditional expectations do not increase $\mathcal{L}^1$ norms, the absolute value of the remainder after m nestings is bounded by $\gamma_{\max}^m K \to 0$ as $m \to \infty$. Hence, the third term in Equation (B.5) vanishes, so it reduces to the fixed-point form

$$
m_i = \frac{\beta}{D_i}\theta_i + \frac{1-\beta}{D_i}\mathbb{E}\left[\sum_{j\in\mathscr{F}_i^{e_i}} m_j|\theta_i,\mathscr{F}\right].
$$

Define the iteration

$$m_i^{(r+1)} := \frac{\beta}{D_i}\theta_i + \frac{1-\beta}{D_i}\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}} m_j^{(r)} \,\Big|\, \theta_i, \mathscr{F}\Big], \qquad m^{(0)} := 0.$$

The corresponding operator $T_i(m) := \frac{\beta}{D_i}\theta_i + \frac{1-\beta}{D_i}\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}} m_j|\theta_i, \mathscr{F}\Big]$ is a contraction under the supremum norm with modulus

$$\frac{(1-\beta)e_i}{\beta+(1-\beta)e_i} = 1 - \frac{\beta}{D_i} < 1.$$

Hence, $m^{(r)} \to m^*$ as $r \to \infty$, and m^* is the unique fixed point of T_i. Substituting any fixed point into Equation (B.4) and subtracting the identity $(\beta+(1-\beta)e_i)\theta_i = \beta\theta_i + (1-\beta)\mathbb{E}_i\Big[\sum_{j\in\mathscr{F}_i^{e_i}} \theta_j|\theta_i, \mathscr{F}\Big] = \beta\theta_i + (1-\beta)e_i\theta_i$ yields

$$(\beta+(1-\beta)e_i)\Big(m_i^* - \theta_i\Big) = (1-\beta)\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}} \Big(m_j^* - \theta_j\Big) \,\Big|\, \theta_i, \mathscr{F}\Big].$$

Taking absolute values, the supremum over i gives

$$\|m^* - \theta\|_\infty \le \max_i \frac{(1-\beta)e_i}{\beta+(1-\beta)e_i}\|m^* - \theta\|_\infty.$$

Because the coefficient on the right-hand side is strictly less than one, it follows that $\|m^* - \theta\|_\infty = 0$, and therefore $m_i^* = \theta_i$ for all i. Truthful reporting is thus the unique equilibrium among rational users. □

Lemma B.2 (Truthful reporting is the only g-robust equilibrium)**.** *Suppose users are naive and the continuation rule $g : \mathbb{R} \to (0,1)$ is $\mathcal{C}^1$ and non-decreasing in the within-the–platform utility. If a message profile $m = (m_i)_i$ is a BNE for every such g, then $m_i = \theta_i$ for all i.*

Proof. The argument establishes the existence of at least one continuation rule g under which truthful reporting is the unique equilibrium among naive users. Consider a constant continuation rule $g(u) = \frac{1}{2}$ for all $u \in \mathbb{R}$. Because engagement no longer depends on the utility level, it follows that $\frac{\partial \mathbb{E}[e_i|\theta_i,\mathscr{F}]}{\partial m_i} = 0$. Under this rule, user i's optimization problem simplifies to maximizing

$$-\beta(\theta_i - m_i)^2 - (1-\beta)\Big(e_i m_i^2 + \mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}} m_j(\theta_j)^2 \,\Big|\, \theta_i, \mathscr{F}\Big] - 2m_i\mathbb{E}\Big[\sum_{j\in\mathscr{F}_i^{e_i}} m_j(\theta_j) \,\Big|\, \theta_i, \mathscr{F}\Big]\Big).$$

This objective is strictly concave in m_i. By the same reasoning used in Proposition 2.1, the unique best response to truthful reports by all other users is $m_i = \theta_i$. □

Proof of Proposition 2.3

Proof. The result follows in two steps. First, maximizing the probability that user i remains active for one additional period requires showing, in her feed, the message of a user (who has not yet appeared) that minimizes the expected conformity loss relative to i. Second, an algorithm that ranks messages in reverse order of the expected conformity loss to user i is precisely the one that maximizes expected engagement.

Under algorithm $\mathscr{F}$, the probability that user i remains active for one additional period after reading k posts is given by $g(u_i(k, m_i, m_{-i}, \mathscr{F}, \theta_i))$. For notational convenience, denote this probability by $g(u_i(k, \mathscr{F}))$. To maximize retention, the platform selects the k-th user to appear in i's feed as

$$\mathscr{F}_i(k) = \underset{j \in \mathcal{U} \setminus \mathscr{F}_i^{k-1}}{\operatorname{argmax}} \Big\{ \mathbb{E}_p[g(u_i(k, \mathscr{F}))] \Big\}.$$

Since g is increasing in u_i and expectations preserve order, maximizing $\mathbb{E}_p[g(u_i(k, \mathscr{F}))]$ is equivalent to maximizing $\mathbb{E}_p[u_i(k, \mathscr{F})]$. Under truthful reporting, the only component of the within–the–platform utility u_i that the platform can influence is the conformity term. Therefore, the platform effectively selects the k-th user in i's feed according to

$$\begin{aligned} \mathscr{F}_i(k) &= \underset{j \in \mathcal{U} \setminus \mathscr{F}_i^{k-1}}{\operatorname{argmax}} \left\{ -\mathbb{E}_p \left[\sum_{l \in \mathscr{F}_i^{k-1}} (\theta_i - \theta_l)^2 + (\theta_i - \theta_j)^2 \right] \right\} \\ &= \underset{j \in \mathcal{U} \setminus \mathscr{F}_i^{k-1}}{\operatorname{argmax}} \left\{ -\mathbb{E}_p \left[(\theta_i - \theta_j)^2 \right] \right\}. \end{aligned} \tag{B.6}$$

Maximizing the probability of user i staying for one more period is equivalent to minimizing the conformity cost of such subsequent period.

The next step is to show that an algorithm constructed by selecting the next user according to Equation (B.6) maximizes expected engagement. Given $\mathscr{F}$, the probability that user i remains active for at least k periods is $\prod_{j=1}^{k-1} g(u_i(j, \mathscr{F}))$,

and the probability that the user exits precisely after period k is

$$\Big(1 - g(u_i(k, \mathscr{F}))\Big) \prod_{j=1}^{k-1} g(u_i(j, \mathscr{F})).$$

Consider two algorithms, $\mathscr{F}$ and $\mathscr{F}'$, that generate identical feeds for user i except for the position of two users t and t', such that $\mathscr{F}_i(t) = \mathscr{F}'_i(t')$, $\mathscr{F}_i(t') = \mathscr{F}'_i(t)$. Without loss of generality, assume that $-\mathbb{E}_p[(\theta_i - \theta_t)^2] > -\mathbb{E}_p[(\theta_i - \theta_{t'})^2]$, so that $\mathscr{F}$ places earlier in the feed the user who entails the lower expected conformity loss. Also without loss of generality, let $t = 1$ and $t' = 2$; then $g(u_i(1, \mathscr{F})) > g(u_i(1, \mathscr{F}'))$. The objective is to demonstrate that $\mathscr{F}$ yields higher expected engagement, where the formal expression for expected engagement is given by

$$\mathbb{E}_p[e_i|\mathscr{F}] = \sum_{r=1}^{n} \left[r\mathbb{E}_p \left[(1 - g(u_i(r, \mathscr{F}))) \prod_{k=1}^{r-1} g(u_i(k, \mathscr{F})) \right] \right].$$

By construction, $\mathbb{E}_p[g(u_i(r, \mathscr{F}))] = \mathbb{E}_p[g(u_i(r, \mathscr{F}'))]$ for all $r \geq 2$. Finally, as $g(u_i(1, \mathscr{F})) > g(u_i(1, \mathscr{F}'))$,

$$\begin{aligned}\mathbb{E}_p[e_i|\mathscr{F}] &= \sum_{r=1}^{n} \left[r\mathbb{E}_p \left[(1 - g(u_i(r, \mathscr{F}))) \prod_{k=1}^{r-1} g(u_i(k, \mathscr{F})) \right] \right] \\ &\geq \sum_{r=1}^{n} \left[r\mathbb{E}_p \left[(1 - g(u_i(r, \mathscr{F}'))) \prod_{k=1}^{r-1} g(u_i(k, \mathscr{F}')) \right] \right] = \mathbb{E}_p[e_i|\mathscr{F}'],\end{aligned}$$

where the inequality can be shown to be true by induction. Finally, consider any algorithm $\mathscr{F}$. It has been shown that, for any pair of users in the feed it induces, reversing their order according to their expected conformity loss strictly increases expected engagement. Iterating this operation until no further improvement is possible yields the platform-optimal algorithm $\mathscr{P}$. □

Lemma B.3. *The posterior distribution of θ conditional on $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}$ is*

$$\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}} \sim \mathcal{N}\left(\frac{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}} \boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}} \mathbb{1}^t}, \frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}} \mathbb{1}^t} \right),$$

where $\mathbb{1}$ is a vector of ones of dimension n, $\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}$ is the restriction of $\boldsymbol{\Sigma}$ to the users in $\mathscr{F}_i^{e_i}$, and $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}$ is the vector of their private signals.

Proof. Assume for simplicity that the signals observed by user i in her person-

alized feed $\mathscr{F}_i^{e_i}$ are $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}} = \{\theta_1, \ldots, \theta_{e_i}\}$. By the properties of the multivariate normal distribution, $(\theta_1, \ldots, \theta_{e_i}) \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}})$. The posterior distribution of θ conditional on $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}$ is proportional to the likelihood function,

$$g(\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}) \propto \left(2\pi \det(\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}})\right)^{-1/2} \exp\left[-\tfrac{1}{2}\left(\theta^2 \mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\mathbb{1}^t - 2\theta\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}} + \boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}^t \boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right)\right].$$

Multiplying by the constant $\sqrt{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}\sqrt{\det(\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}})}$ and completing the square in θ gives

$$g(\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}) = \sqrt{\frac{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\theta - \frac{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\mathbb{1}^t}}{\sqrt{\frac{1}{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\mathbb{1}^t}}}\right)^2\right].$$

This is the density of a normal random variable with the stated mean and variance. □

Proof of Proposition 2.5

Proof. Suppose that the platform implements $\mathscr{F}$ with $r(\mathscr{F}, e^*(\mathscr{F})) > d(\mathscr{F}, e^*(\mathscr{F}))$. Since the focal user is a democrat, there exists $k < e^*(\mathscr{F})$ with $d(\mathscr{F}, k) = r(\mathscr{F}, k)$; let k_M be the largest such k. Build $\mathscr{F}'$ by keeping the first k_M positions identical to $\mathscr{F}$ and flipping each label thereafter.

For any $t \le k_M$, the prefixes are identical, so $U(\mathscr{F}', t) = U(\mathscr{F}, t)$. For $t > k_M$, denote by (d_t, r_t) the cumulative counts under $\mathscr{F}$ in the first t positions. By construction, the corresponding counts under $\mathscr{F}'$ are (r_t, d_t), since the increments beyond k_M are exchanged. Because $\mathrm{Var}[\theta|d, r] = \mathrm{Var}[\theta|r, d]$, it follows that

$$U(\mathscr{F}', t) - U(\mathscr{F}, t) = 2\lambda(1-\beta)x(r_t - d_t) \ge 0,$$

where the inequality holds because $r_t \ge d_t$ for all $t \in \{k_M + 1, \ldots, e^*(\mathscr{F})\}$ by the definition of k_M and the assumption that $r(\mathscr{F}, e^*(\mathscr{F})) > d(\mathscr{F}, e^*(\mathscr{F}))$. Hence, $U(\mathscr{F}', t) \ge U(\mathscr{F}, t)$ for every t, with equality for all $t \le k_M$. Therefore, $\max_t U(\mathscr{F}', t) \ge \max_t U(\mathscr{F}, t)$ and consequently $e^*(\mathscr{F}') \ge e^*(\mathscr{F})$. □

Proof of Proposition 2.6

Proof. Before establishing this proposition, two auxiliary lemmas are stated and proved.

Lemma B.4 (Local swap and the best prefix length)**.** *Fix a feed $\mathscr{F}$ and a position k such that $\mathscr{F}^{k-1}$ has composition (d, r) and the next two items are (R, D). Let $\mathscr{F}'$ be obtained from $\mathscr{F}$ by swapping them to (D, R). If $U(d+1, r) \leq U(d, r+1)$, then $e^*(\mathscr{F}') \geq e^*(\mathscr{F})$.*

Proof. For all $m \leq k-1$, the two algorithms yield identical prefixes, so $U(\mathscr{F}^m) = U((\mathscr{F}')^m)$. At $m = k$, the prefixes differ for the first time, and by the construction of $\mathscr{F}'$, $U((\mathscr{F}')^k) \leq U(\mathscr{F}^k)$. At $m = k+1$, both orders reach the same composition $(d+1, r+1)$, implying $U((\mathscr{F}')^{k+1}) = U(\mathscr{F}^{k+1})$; for all $m \geq k+2$, the compositions again coincide. Hence, only the positions k and $k+1$ affect engagement.

Let $M := U(\mathscr{F}, e^*(\mathscr{F}))$ denote the user's maximal within–the-platform utility under $\mathscr{F}$. By assumption,

$$U(d+1, r) \leq U(d, r+1),$$

and therefore $U(d+1, r) \leq M$. It follows that the engagement level under $\mathscr{F}'$ is weakly higher, that is, $e^*(\mathscr{F}') \geq e^*(\mathscr{F})$. □

Lemma B.5 (Explicit form of the swap condition)**.** *Under Equation (2.5), the local comparison $U(d+1, r) \leq U(d, r+1)$ is equivalent to*

$$\frac{\lambda(1-\beta)}{1-\lambda} \leq \frac{-\sigma^2(1-x)(d-r)(1+x(d+r))}{(4xdr - dx + d + 3rx + r - x + 1)(4xdr + 3dx + d - rx + r - x + 1)}. \tag{B.7}$$

For $x \in (0, 1)$, the denominator on the right-hand side is strictly positive. Consequently, the right-hand side has sign $-\operatorname{sgn}(d-r)$ and equals zero when $d = r$.

The previous lemma shows that the condition defined by Lemma B.4 will occur for small values of λ (when the user wants to learn) or for large values of β (when conformity matters little). Starting from any feed $\mathscr{F}$, consider its prefix $\mathscr{F}^{e^*(\mathscr{F})}$. Whenever an adjacent pair (R, D) appears at composition (d, r) satisfying Equation (B.7), the pair is replaced by (D, R). After finitely many such replacements, the procedure yields a feed $\widetilde{\mathscr{F}}$ satisfying $e^*(\widetilde{\mathscr{F}}) \geq e^*(\mathscr{F})$. Each admissible local swap weakly increases e^* by Lemma B.4. By Lemma B.5, if

the condition in Equation (B.7) holds for an (R, D) pair at composition (d, r), it continues to hold when that pair is shifted one position to the right. The process must terminate, since every swap strictly reduces the number of (R, D) inversions preceding the first index where Equation (B.7) ceases to hold. For any subset of the feed not affected by a swap, the value of d remains unchanged; at a swapped position, replacing R with D weakly increases d. Maximizing e^* over all feeds then delivers the desired property. **Proof of Proposition 2.1**

Proof. For any algorithm $\mathscr{F}$, the platform's expectation over user i's engagement is a finite real number even if platform size grows asymptotically large. If platform size is n, expected engagement is given by

$$\mathbb{E}_p[e_i] = \mathbb{E}_p\left[\sum_{r=1}^{n}\left[r(1-g(u_i(r,\mathscr{F})))\prod_{k=1}^{r-1} g(u_i(k,\mathscr{F}))\right]\right].$$

Note that there exist constants $0 < \underline{g} < \overline{g} < 1$ such that $g(x) \in (\underline{g}, \overline{g})$ for all $x \in \mathbb{R}$.[3]

$$\begin{aligned}\mathbb{E}_p[e_i] =& \mathbb{E}_p\left[\sum_{r=1}^{n}\left[r\prod_{k=1}^{r-1} g(u_i(k,\mathscr{F}))\right] - \sum_{r=1}^{n}\left[rg(u_i(r,\mathscr{F}))\prod_{k=1}^{r-1} g(u_i(k,\mathscr{F}))\right]\right]\\ \leq& \mathbb{E}_p\left[\sum_{r=1}^{n} r\overline{g}^{r-1} - \sum_{r=1}^{n} r\underline{g}^{r}\right] = \sum_{r=1}^{n} r\overline{g}^{r-1} - \sum_{r=1}^{n} r\underline{g}^{r}.\end{aligned}$$

Since this inequality holds for all $n \in \mathbb{N}$, taking limits implies that, as platform size grows asymptotically large, user i's expected engagement remains finite:

$$\lim_{n\to\infty} \mathbb{E}_p[e_i] \leq \lim_{n\to\infty}\left[\sum_{r=1}^{n} r\overline{g}^{r-1} - \sum_{r=1}^{n} r\underline{g}^{r}\right] = \frac{1}{(1-\overline{g})^2} - \frac{\underline{g}}{(1-\underline{g})^2}. \tag{B.8}$$

Define $k_{pi} := \min\left\{\tilde{k} \in \mathbb{N} : \tilde{k} \geq \frac{1}{(1-\overline{g})^2} - \frac{\underline{g}}{(1-\underline{g})^2}\right\}$. The inequality $k_{pi} > 1$ follows from

$$\frac{1}{(1-\overline{g})^2} - \frac{\underline{g}}{(1-\underline{g})^2} > 1 \quad \Longleftrightarrow \quad (2\overline{g} - \underline{g} - \overline{g}^2) + (2\overline{g}\underline{g}^2 - 2\overline{g}\underline{g}) + (\underline{g}\overline{g}^2 - \overline{g}^2\underline{g}^2) > 0,$$

which holds because $0 < \underline{g} < \overline{g} < 1$. An analogous argument, replacing the platform's expectation with the user's expectation, establishes the existence of

[3] For notational convenience, the continuation probability after the final post is set to zero, $g(u_i(n,\mathscr{F})) := 0$, so that the general expression $\Pr(e_i = r) = (1-g(u_i(r,\mathscr{F})))\prod_{k<r} g(u_i(k,\mathscr{F}))$ also captures the terminal mass at $r = n$.

k_u.

Finally, define $k_p := \max_i k_{pi}$. This quantity is finite because it is bounded from above: taking the largest values of $\overline{g}$ and $\underline{g}$ across users and applying Equation (B.8) yields an upper bound for all k_{pi}. Hence k_p is a natural number. □

□

Proof of Proposition 2.7

Proof. To establish the result, note first that as the platform expands, it becomes arbitrarily likely to find k users whose opinions are almost perfectly correlated with that of user i. Let ρ_{ij} denote the correlation between users i and j. Since each new user's correlation with i is drawn from a distribution with full support on $[-1, 1]$, the probability of observing $\rho_{ij} > 1 - \varepsilon$ is strictly positive for any $\varepsilon > 0$. By the law of large numbers, the number of users satisfying $\rho_{ij} > 1 - \varepsilon$ therefore increases linearly in n. Consequently, for every pair $\varepsilon, \gamma > 0$, there exists $\bar{n} \in \mathbb{N}$ such that, for all $n > \bar{n}$, the probability that user i has at least k neighbors $j_1, \ldots, j_k$ with $\rho_{ij_r} > 1 - \varepsilon$ for all $r \in \{1, \ldots, k\}$ exceeds $1 - \gamma$.

Applying the Cauchy-Schwarz inequality to the correlations between user i and any two of her neighbors, say j_r and j_l, yields $\rho_{j_r,j_l} \geq \rho_{j_r,i}\rho_{j_l,i} - \sqrt{(1-\rho_{j_r,i}^2)(1-\rho_{j_l,i}^2)}$. Conditional on the event that both users satisfy $\rho_{j_r,i} > 1 - \varepsilon$ and $\rho_{j_l,i} > 1 - \varepsilon$, the right-hand side is bounded below by $1 - 4\varepsilon + 2\varepsilon^2$. Since each of these events occurs with probability at least $1 - \gamma$,

$$\Pr\left[\rho_{j_r,j_l} \geq 1 - 4\varepsilon + 2\varepsilon^2\right] \geq (1-\gamma)^2 \quad \text{for all } j_r, j_l.$$

Let $X_n := \sum_{j=1}^{n} \mathbf{1}\{\rho_{j,i} > 1 - \varepsilon\}$. Under independence across j and with $\Pr[\rho_{j,i} > 1 - \varepsilon] = p > 0$, the law of large numbers implies $\Pr[X_n \geq e_i(n)] \to 1$ as $n \to \infty$. On the event $\{X_n \geq e_i(n)\}$, the feed can be composed entirely of users satisfying $\rho_{j,i} > 1 - \varepsilon$; by the Cauchy–Schwarz bound, every pair in the feed then satisfies $\rho_{r,l} \geq 1 - \delta$ with $\delta = 4\varepsilon - 2\varepsilon^2$. Hence, for all sufficiently large n,

$$\Pr\left[\mathbf{A}(n) \leq \mathbf{\Sigma}_{\mathscr{P}_i^{e_i(n)}}\right] \geq 1 - \gamma,$$

and this probability converges to one as $n \to \infty$.

Now, for each platform size n, engagement levels may vary, but it is always bounded by some k as already shown. The corresponding platform optimal feed is denoted by $\mathscr{P}_i^{e_i(n)}$. Let $\delta = 4\varepsilon - 2\varepsilon^2$. For every $\delta > 0$ and $\gamma > 0$, there exists some $\tilde{n} \in \mathbb{N}$ such that, for all $n > \tilde{n}$, the set of k users defined above satisfies $Pr[\rho_{j_r,j_l} > 1-\delta] > 1-\gamma$ for all $j_r, j_l \in \{j_1, \ldots, j_k\}$, which follows by choosing ε sufficiently small. For any given engagement level $e_i(n) \leq k$, the $e_i(n)$ users displayed in user i's feed are selected from this set. Define the associated $e_i(n) \times e_i(n)$ matrix $\mathbf{A} = \sigma^2\Big[(1-\delta)\mathbb{1}\mathbb{1}^\top + \delta I_e\Big]$, which represents the covariance matrix of size $e_i(n)$ signals with equal pairwise correlation $1-\delta$. Since, with probability at least $1-\gamma$, every pair of users in the feed has correlation above $1-\delta$, it follows that $\Pr\left[\mathbf{A}(n) \leq \mathbf{\Sigma}_{\mathscr{P}_i^{e_i(n)}}\right] \geq 1-\gamma$, where $\mathbf{A}(n) \leq \mathbf{\Sigma}_{\mathscr{P}_i^{e_i(n)}}$ denotes elementwise inequality. Although correlations across different pairs are not independent, this bound suffices to establish that, with probability arbitrarily close to one for large n, the actual covariance matrix $\mathbf{\Sigma}_{\mathscr{P}_i^{e_i(n)}}$ dominates $\mathbf{A}(n)$ elementwise. Now, we need an auxiliary result:

Lemma B.6 (Matrix comparison)**.** *Let $e \in \mathbb{N}$ and $\sigma^2 > 0$. For any $\delta \in (0,1)$, define the $e \times e$ equicorrelated covariance matrix $\mathbf{A} = \sigma^2\Big[(1-\delta)\mathbb{1}\mathbb{1}^\top + \delta\mathbf{I}_e\Big]$, whose off–diagonal correlations equal $1-\delta$. Let $\mathbf{\Sigma}$ be any symmetric positive definite matrix with the same diagonal σ^2 and off–diagonal elements satisfying $\mathbf{\Sigma}_{rs} \geq \sigma^2(1-\delta)$ for all $r \neq s$. Write $\mathbf{\Sigma} = \mathbf{A} + \mathbf{E}$, where $\mathbf{E}$ is symmetric with zero diagonal and nonnegative off–diagonal entries. If the norm of the matrix $\mathbf{E}$ satisfies $\|\mathbf{E}\|_{\text{op}} < \sigma^2\delta$,[4] then both matrices are invertible and*

$$\mathbb{1}^\top\mathbf{\Sigma}^{-1}\mathbb{1} - \mathbb{1}^\top\mathbf{A}^{-1}\mathbb{1} \leq \frac{e}{\sigma^4\delta^2}\frac{\|\mathbf{E}\|_{\text{op}}}{1 - \|\mathbf{E}\|_{\text{op}}/(\sigma^2\delta)}. \tag{B.9}$$

In particular, if $\|\mathbf{E}\|_{\text{op}} = o(1)$ (for instance, when all pairwise correlations in $\mathbf{\Sigma}$ converge to $1-\delta$), then

$$\mathbb{1}^\top\mathbf{\Sigma}^{-1}\mathbb{1} \leq \mathbb{1}^\top\mathbf{A}^{-1}\mathbb{1} + o(1), \quad \textit{where } o(1) \to 0 \textit{ as } n \to \infty.$$

[4]The notation $\|\cdot\|_{\text{op}}$ denotes the *operator* (or *spectral*) norm of a matrix. For a symmetric matrix $\boldsymbol{A}$, it equals the largest absolute value of its eigenvalues:

$$\|\boldsymbol{A}\|_{\text{op}} = \max_{\|x\|_2=1} \|\boldsymbol{A}x\|_2 = \lambda_{\max}(|\boldsymbol{A}|).$$

Intuitively, it measures the maximum factor by which the matrix $\boldsymbol{A}$ can stretch a unit vector. In the proof, this norm is used to compare covariance matrices and to express convergence or dominance in terms of their maximal directional deviation.

Proof. Since $\|\mathbf{A}^{-1}\|_{\text{op}}\|\mathbf{E}\|_{\text{op}} < 1$, we can apply the Neumann series expansion and obtain: $\boldsymbol{\Sigma}^{-1} = (\mathbf{A}+\mathbf{E})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1} - \cdots$. Subtracting and bounding term by term gives

$$|\mathbb{1}^\top(\boldsymbol{\Sigma}^{-1} - \mathbf{A}^{-1})\mathbb{1}| \le \sum_{k\ge 1} \|\mathbf{A}^{-1}\|_{\text{op}}^{k+1}\|\mathbf{E}\|_{\text{op}}^{k}\|\mathbb{1}\|_2^2 = \frac{\|\mathbf{A}^{-1}\|_{\text{op}}^2\|\mathbf{E}\|_{\text{op}}}{1 - \|\mathbf{A}^{-1}\|_{\text{op}}\|\mathbf{E}\|_{\text{op}}}\|\mathbb{1}\|_2^2.$$

For $\mathbf{A} = \sigma^2[(1-\delta)\mathbb{1}\mathbb{1}^\top + \delta I_e]$, the eigenvalues are $\lambda_1 = \sigma^2(\delta + (1-\delta)e)$ along $\mathbb{1}$ and $\lambda_2 = \cdots = \lambda_e = \sigma^2\delta$ on its orthogonal complement, so $\|\mathbf{A}^{-1}\|_{\text{op}} = (\sigma^2\delta)^{-1}$. Substituting this and $\|\mathbb{1}\|_2^2 = e$ yields Equation (B.9). Finally, because $\mathbf{E}$ has zero diagonal and $0 \le \mathbf{E}_{rs} \le \sigma^2\delta$, its norm operator satisfies $\|\mathbf{E}\|_{\text{op}} \le e\sigma^2\delta$. Thus, the smallness condition can always be enforced by taking δ sufficiently small or restricting to the high–correlation event. □

Therefore, on the high–correlation event (which occurs with probability at least $1-\gamma$ for all sufficiently large n), $\mathbf{A}(n) \le \boldsymbol{\Sigma}_{\mathscr{P}_i^{e_i(n)}}$ and

$$\mathbb{1}^\top\boldsymbol{\Sigma}^{-1}_{\mathscr{P}_i^{e_i(n)}}\mathbb{1} \le \mathbb{1}^\top\mathbf{A}(n)^{-1}\mathbb{1} + o(1),$$

where $o(1) \to 0$ as $n \to \infty$ (for any fixed $\delta > 0$). Taking reciprocals (hence reversing the inequality since all terms are positive) implies

$$\frac{1}{\mathbb{1}^\top\mathbf{A}(n)^{-1}\mathbb{1} + o(1)} \le \frac{1}{\mathbb{1}^\top\boldsymbol{\Sigma}^{-1}_{\mathscr{P}_i^{e_i(n)}}\mathbb{1}} = \text{Var}\left[\theta\middle|\boldsymbol{\theta}_{\mathscr{P}_i^{e_i(n)}}\right] \le \sigma^2.$$

Moreover, since $\mathbb{1}^\top\mathbf{A}(n)^{-1}\mathbb{1} = \frac{e_i(n)}{\sigma^2\{1+(e_i(n)-1)(1-\delta)\}}$, it follows that, with probability at least $1-\gamma$,

$$\frac{\sigma^2[1+(e_i(n)-1)(1-\delta)]}{e_i(n)} - o(1) \le \text{Var}\left[\theta\middle|\boldsymbol{\theta}_{\mathscr{P}_i^{e_i(n)}}\right] \le \sigma^2.$$

Letting $n \to \infty$ (for any fixed $\delta > 0$), the lower bound converges to $\sigma^2(1-\delta)$; subsequently letting $\delta \to 0$ yields

$$\operatorname*{plim}_{n\to\infty} \text{Var}\left[\theta\middle|\boldsymbol{\theta}_{\mathscr{P}_i^{e_i(n)}}\right] = \sigma^2.$$

□

Proof of Proposition 2.8

Proof. The comparison focuses on the large–n expected utilities of $\mathscr{P}$ and $\mathscr{R}$ for a naive user i. Consider first the platform–optimal algorithm $\mathscr{P}$. By Proposition 2.7, as the number of users n grows, the feed constructed under $\mathscr{P}$ selects accounts satisfying $\rho_{ij} \to 1$ while the engagement level e_i remains finite. Consequently, the conformity component of utility becomes negligible: $\lim_{n\to\infty} \mathbb{E}\Big[\sum_{j\in\mathscr{P}_i^{e_i}} (\theta_i - \theta_j)^2 \Big| \mathscr{P}\Big] = 0$.

Moreover, the posterior variance under $\mathscr{P}$ converges to the prior, so the learning term equals $(1-\lambda)\sigma^2$ in the limit. Hence

$$\lim_{n\to\infty} \mathbb{E}\Big[U_i(e_i, m_i, m_{-i}, a_i, \mathscr{P}, \theta_i, \theta) \Big| \mathscr{P}\Big] = \lambda\alpha\mathbb{E}[e_i|\mathscr{P}] - (1-\lambda)\sigma^2.$$

Next, consider the reverse–chronological algorithm $\mathscr{R}$. Under $\mathscr{R}$, users are not ordered by similarity, so the conformity term need not vanish in the limit. Using the identity $\mathbb{E}\Big[(\theta_i - \theta_j)^2\Big] = 2\sigma^2(1-\rho_{ij})$, and noting that the self–term $j = i$ contributes zero,

$$\mathbb{E}\left[\sum_{j\in\mathscr{R}_i^{e_i}} (\theta_i - \theta_j)^2 \Big| \mathscr{R}\right] = 2\sigma^2\mathbb{E}\left[\sum_{j\in\mathscr{R}_i^{e_i}\setminus\{i\}} (1-\rho_{ij}) \Big| \mathscr{R}\right].$$

If, under $\mathscr{R}$, $\mathbb{E}[\rho_{ij}] = 0$, then

$$\mathbb{E}\left[\sum_{j\in\mathscr{R}_i^{e_i}} (\theta_i - \theta_j)^2 \Big| \mathscr{R}\right] = 2\sigma^2\mathbb{E}[e_i - 1|\mathscr{R}].$$

Regarding learning, the posterior–variance component equals $(1-\lambda)\sigma^2\mathbb{E}\Big[\frac{1}{e_i}|\mathscr{R}\Big]$ (see Equation (2.9)). Therefore,

$$\lim_{n\to\infty} \mathbb{E}\Big[U_i \Big| \mathscr{R}\Big] = \lambda\Big(\alpha\mathbb{E}[e_i|\mathscr{R}] - 2(1-\beta)\sigma^2\mathbb{E}[e_i - 1|\mathscr{R}]\Big) - (1-\lambda)\sigma^2\mathbb{E}\Big[\tfrac{1}{e_i}|\mathscr{R}\Big].$$

To complete the argument, the two expected utilities can now be compared directly. User i weakly prefers the platform–optimal algorithm $\mathscr{P}$ to the reverse–chronological algorithm $\mathscr{R}$ in the large–n limit if and only if

$$\lambda\alpha\mathbb{E}[e_i|\mathscr{P}] - (1-\lambda)\sigma^2 \;\geq\; \lambda\Big(\alpha\mathbb{E}[e_i|\mathscr{R}] - 2(1-\beta)\sigma^2\mathbb{E}[e_i - 1|\mathscr{R}]\Big) - (1-\lambda)\sigma^2\mathbb{E}\Big[\tfrac{1}{e_i}|\mathscr{R}\Big].$$

Rearranging terms gives

$$\lambda \;\geq\; \frac{\sigma^2\Big(1-\mathbb{E}\Big[\frac{1}{e_i}|\mathscr{R}\Big]\Big)}{\alpha(\mathbb{E}[e_i|\mathscr{P}]-\mathbb{E}[e_i|\mathscr{R}])+2(1-\beta)\sigma^2\mathbb{E}[e_i-1|\mathscr{R}]+\sigma^2\Big(1-\mathbb{E}\Big[\frac{1}{e_i}|\mathscr{R}\Big]\Big)}.$$

Since the platform–optimal algorithm $\mathscr{P}$ maximizes engagement, it follows that $\mathbb{E}[e_i|\mathscr{P}] \geq \mathbb{E}[e_i|\mathscr{R}]$. Moreover, because $e_i \geq 1$, we have $\mathbb{E}\Big[\frac{1}{e_i}|\mathscr{R}\Big] \leq 1$. Taking the maximum over all users $i \in \mathcal{U}$ yields the stated threshold condition. □

Proof of Proposition 2.9

Proof. Let $U_i(k)$ denote user i's expected utility after reading the first k items of her feed $\mathscr{R}_i^k$, where the order of the remaining items is uniformly random. At depth k, the continuation decision compares the expected marginal gain from reading one more item with the expected marginal cost. Reading one additional item increases the within–the-platform engagement component by $\lambda\alpha$.

Disagreement term. Let $\mathscr{R}_i^k$ denote the set of $n-k-1$ accounts user i has not yet read. Because the next post comes from a random user $j \in \mathscr{R}_i^k$, each such user appears with probability $1/(n-k-1)$. Hence, the expected disagreement cost of the next post is the uniform average of the conditional variances of the remaining signals:

$$\mathbb{E}_i\Big[(\theta_j-\theta_i)^2|\mathscr{R}_i^k\Big] = \frac{1}{n-k-1}\sum_{j\in\mathscr{R}_i^k}\mathbb{E}_i\Big[(\theta_j-\theta_i)^2|\mathscr{R}_i^k\Big].$$

Learning term. The learning component improves by the expected reduction in posterior variance when moving from k to $k+1$ signals, i.e.,

$$(1-\lambda)\Big(\mathbb{E}_i[\mathrm{Var}(\theta|\mathscr{R}_i^k)]-\mathbb{E}_i[\mathrm{Var}(\theta|\mathscr{R}_i^{k+1})]\Big).$$

Marginal comparison. Combining these components, the expected change in utility from reading one more post is

$$\begin{aligned}U_i(k+1)-U_i(k) =& \lambda\alpha-\frac{\lambda(1-\beta)}{n-k-1}\sum_{j\in\mathscr{R}_i^k}\mathbb{E}_i\Big[(\theta_j-\theta_i)^2|\mathscr{R}_i^k\Big]\\ &+(1-\lambda)\Big(\mathbb{E}_i[\mathrm{Var}(\theta|\mathscr{R}_i^k)]-\mathbb{E}_i[\mathrm{Var}(\theta|\mathscr{R}_i^{k+1})]\Big).\end{aligned}$$

It is therefore optimal to continue scrolling if and only if $U_i(k{+}1) - U_i(k) > 0$, which yields exactly the inequality in the statement. If the inequality fails, the user stops. □

Proof of Proposition 2.10

Proof. The argument proceeds in two steps: conformity is analyzed first, followed by learning. Under joint normality of signals with common variance σ^2, it holds that $\mathbb{E}\big[(\theta_i - \theta_j)^2\big] = 2\sigma^2(1-\rho_{ij})$.

Consider first the conformity component under $\mathscr{B}$. Let $k := \min\big\{\tilde{k} \in \mathbb{N} : \tilde{k} \geq (1-\bar{g})^{-2} - \underline{g}(1-\underline{g})^{-2}\big\}$, as defined in the proof of Proposition 2.7. For any $\varepsilon, \nu > 0$, there exists $\bar{n}$ such that for all $n > \bar{n}$, one can identify a subset of k users, indexed $\{1,\dots,k\}$, satisfying $\Pr\big[\rho_{ij} \geq 1-\varepsilon\big] \geq 1-\nu$ for each $j = 1,\dots,k$. Likewise, for every $\delta, \varphi > 0$ there exists $\tilde{n}$ such that for all $n > \tilde{n}$ there is a user l with $Pr\big[\rho_{il} \leq -1+\delta\big] \geq 1-\varphi$.

Fix any engagement level e_i and take $n \geq \max\{\bar{n}, \tilde{n}\}$. Define the $\mathscr{B}$–prefix as $\mathscr{B}_i^{e_i} = \{l, 1, \dots, e_i - 1\}$, where the indices in $\{1,\dots,e_i-1\}$ are drawn from the k–pool of highly correlated users (so that $e_i - 1 \leq k$) and the index l corresponds to a user outside that pool, with opposite orientation to i. Then, as $n \to \infty$, $\operatorname{plim}_{n\to\infty} \rho_{ij} = 1$ for all $j \in \{1,\dots,e_i-1\}$, and $\operatorname{plim}_{n\to\infty} \rho_{il} = -1$.

Using the law of iterated expectations and decomposing $\mathscr{B}$ into the "similar block" plus the out-group user l,

$$\begin{aligned}\mathbb{E}\left[\sum_{j\in\mathscr{B}_i^{e_i}} (\theta_i-\theta_j)^2 \middle| \mathscr{B}\right] &= \mathbb{E}\left[\sum_{j\in\mathscr{P}_i^{e_i-1}} (\theta_i-\theta_j)^2 \middle| \mathscr{P}\right] + \mathbb{E}\big[(\theta_i-\theta_l)^2\big] \\ &= 2\sigma^2\mathbb{E}\left[\sum_{j\in\mathscr{P}_i^{e_i-1}} \big(1-\rho_{ij}\big) \middle| \mathscr{P}\right] + 2\sigma^2\mathbb{E}[1-\rho_{il}].\end{aligned}$$

By bounded convergence, the first expectation tends to 0 (since $\rho_{ij} \to 1$ for all j in the similar block), while the second tends to $4\sigma^2$ (since $\rho_{il} \to -1$). Hence, as $n \to \infty$,

$$\mathbb{E}\left[\sum_{j\in\mathscr{B}_i^{e_i}} (\theta_i-\theta_j)^2 \middle| \mathscr{B}\right] = \underbrace{o(1)}_{\text{similar block}} + \underbrace{4\sigma^2}_{\text{out-group } l}.$$

Because Proposition 2.7 implies that the conformity loss under $\mathscr{P}$ vanishes,

the asymptotic difference in conformity losses between $\mathscr{B}$ and $\mathscr{P}$ equals a constant $4\sigma^2$. Attention now turns to the learning component under $\mathscr{B}$. Under $\mathscr{P}$, Proposition 2.7 further establishes that learning vanishes asymptotically, that is, $\mathrm{Var}[\theta|\mathscr{P}_i^{e_i}] \xrightarrow[n\to\infty]{} \sigma^2$. Intuitively, as the feed becomes perfectly homogeneous, additional signals are almost identical to the user's own belief and therefore convey no new information.

Under $\mathscr{B}$, by contrast, user i also observes one maximally dissimilar signal θ_l with $\rho_{il} \to -1$. In a two-signal Gaussian system (θ_i, θ_l) with correlation ρ_{il} and variance σ^2, the posterior variance of θ given both signals equals

$$\mathrm{Var}[\theta|\theta_i, \theta_l] = \frac{\sigma^2(1+\rho_{il})}{2}.$$

As $\rho_{il} \to -1$, the denominator $1+\rho_{il} \to 0$; precision diverges and $\mathrm{Var}[\theta|\theta_i, \theta_l] \xrightarrow[n\to\infty]{} 0$. Since adding further (highly correlated) signals cannot increase posterior variance,

$$\mathrm{Var}[\theta|\{\theta_j : j \in \mathscr{B}_i^{e_i}\}] \leq \mathrm{Var}[\theta|\theta_i, \theta_l] \xrightarrow[n\to\infty]{} 0.$$

Thus, learning under $\mathscr{B}$ becomes asymptotically perfect.

To conclude, utilities under the two algorithms are compared to identify the threshold value of λ that renders the user indifferent between $\mathscr{P}$ and $\mathscr{B}$. The only asymptotically nonvanishing differences between $\mathscr{P}$ and $\mathscr{B}$ are: (i) a constant conformity penalty of $4\sigma^2$ under $\mathscr{B}$, (ii) the learning gain under $\mathscr{B}$, and (iii) the difference in expected engagement. Hence, $\mathscr{B}$ yields higher expected utility if and only if $\lambda\Big(\alpha\big(\mathbb{E}[e_i|\mathscr{P}] - \mathbb{E}[e_i|\mathscr{B}]\big) + 4\sigma^2\Big) \leq (1-\lambda)\sigma^2$, that is,

$$\lambda \leq \frac{\sigma^2}{\alpha\big(\mathbb{E}[e_i|\mathscr{P}] - \mathbb{E}[e_i|\mathscr{B}]\big) + 5\sigma^2}.$$

□

Proof of Proposition 2.11

Proof. Using the definition of the user's expected total utility,

$$\Delta U_i = \lambda(\alpha \Delta e_i - (1-\beta)\Delta C_i) - (1-\lambda)\Delta V_i.$$

Whenever $\alpha\Delta e_i - (1-\beta)\Delta C_i + \Delta V_i \neq 0$, the condition $\Delta U_i \geq 0$ is equivalent to

$$\lambda \geq \frac{\Delta V_i}{\alpha\Delta e_i - (1-\beta)\Delta C_i + \Delta V_i}.$$

If the denominator equals zero, no threshold value is required. The next step is to characterize the signs of the relevant terms under the platform–optimal algorithm.

Engagement. Expected engagement weakly increases as the platform grows, implying $\Delta e_i \geq 0$.

Conformity. When moving from n to $n{+}1$ under $\mathscr{P}$, the feed either remains unchanged or replaces the least–correlated user k (with correlation ρ_{ik}) by a new entrant $(n{+}1)$ satisfying $\rho_{i(n+1)} \geq \rho_{ik}$. Using $\mathbb{E}[(\theta_i - \theta_j)^2] = 2\sigma^2(1-\rho_{ij})$ and noting that all other users are common to both feeds, the conformity term $\mathbb{E}_i\Big[\sum_{j\in\mathscr{P}_i^{e_i}}(\theta_i-\theta_j)^2\Big]$ weakly decreases. Hence $\Delta C_i \leq 0$, with strict inequality if the replacement is strict. It follows that $-\lambda(1-\beta)\Delta C_i \geq 0$.

Learning. The change in the learning term can be decomposed into: (i) the effect of replacing a user with a weakly more correlated one while keeping e_i fixed, and (ii) the effect of any increase in e_i. For case (i), replacing a signal by a weakly more informative one (i.e., with larger $|\rho|$) weakly reduces the posterior variance under Gaussian updating. For case (ii), an increase in e_i adds new signals that may improve or deteriorate inference depending on their informativeness and correlation structure. Hence ΔV_i may be of either sign in general.

Putting these pieces together: $\alpha\Delta e_i - (1-\beta)\Delta C_i \geq 0$, and ΔV_i is potentially sign–ambiguous. Hence, inside the platform utility is always increasing while the effect on learning is ambiguous. The inequality above gives the exact λ-threshold that guarantees $\Delta U_i \geq 0$. □

Proof of Proposition 2.12

Proof. Fix user i and an engagement choice $e_i < n$ (so the interior FOC applies). Under truthful reporting, $m_i = \theta_i$, the sincerity term $-\beta(m_i - \theta_i)^2$ is identically zero and does not depend on the pool size. Thus, the within-the-platform

component of expected utility under algorithm $\mathscr{F}$ is

$$\mathbb{E}[u_i|\theta_i, \mathscr{F}] = \lambda\left(\alpha e_i - (1-\beta)\mathbb{E}\left[\sum_{j\in\mathscr{F}_i^{e_i}} (\theta_i - \theta_j)^2 \middle| \theta_i, \mathscr{F}\right]\right),$$

where the expectation is over the (algorithm-induced) distribution of the e_i accounts shown to i. The expected conformity cost is $\mathbb{E}\Big((\theta_i - \theta_j)^2|\mathscr{F}\Big) = 2\sigma^2\Big(1 - \rho_{ij}\Big)$, so

$$\mathbb{E}[u_i|\mathscr{F}] = \lambda\Big[\alpha e_i - 2(1-\beta)\sigma^2 \sum_{j\in\mathscr{F}_i^{e_i}} \Big(1-\rho_{ij}\Big)\Big] = \lambda\Big[\Big(\alpha - 2(1-\beta)\sigma^2\Big)e_i + 2(1-\beta)\sigma^2 \sum_{j\in\mathscr{F}_i^{e_i}} \rho_{ij}\Big].$$

Consider adding one more user j to the feed. The marginal change in expected within-the-platform utility is $\Delta u_j = \lambda\Big[\alpha - 2(1-\beta)\sigma^2\Big(1-\rho_{ij}\Big)\Big] = \lambda\Big[\alpha - 2(1-\beta)\sigma^2 + 2(1-\beta)\sigma^2\rho_{ij}\Big]$. Hence $\Delta u_j > 0$ if and only if $\rho_{ij} > \rho^\star$, where $\rho^\star := 1 - \frac{\alpha}{2(1-\beta)\sigma^2}$. If $\alpha \geq 4(1-\beta)\sigma^2$, then $\Delta u_j > 0$ for all $j \in [-1, 1]$; otherwise, only sufficiently correlated users ($\rho_{ij} > \rho^\star$) are marginally valuable.

Formally, under the full-support assumption on pairwise covariances, for any fixed $\rho^\star < 1$ the set $\{j : \rho_{ij} > \rho^\star\}$ has positive measure and its expected cardinality increases with n. Let $\mathcal{P}(n)$ denote the platform-optimal algorithm at size n. As n grows, new entrants drawn from the same distribution allow the platform to construct an augmented feed $\mathcal{P}(n+1)$ that weakly dominates $\mathcal{P}(n)$: there exists an augmenting subset of positions such that (i) expected within-platform utility weakly increases and (ii) the posterior variance along the implemented path weakly decreases. Hence, $\mathbb{E}[U_i(n+1)] \geq \mathbb{E}[U_i(n)]$, establishing that rational users always enjoy weak network effects under the platform-optimal design. □

Appendix C

Appendix to Chapter 3

C.1 Proofs

Proof of Proposition 3.1.

Proof. Fix a covered story s. Since

$$\frac{\partial \mathfrak{s}_j}{\partial \tau_{js}} = c_{js},$$

we have

$$\frac{\partial P_j(x)}{\partial \tau_{js}} = P_j(x)(1 - P_j(x))2\lambda c_{js}(x - \mathfrak{s}_j).$$

Therefore,

$$\frac{\partial D_j}{\partial \tau_{js}} = 2\lambda c_{js} A_j.$$

The first-order condition for τ_{js} is

$$\rho \frac{\partial D_j}{\partial \tau_{js}} - c_{js} \left[\kappa(\tau_{js} - \theta_s) + \gamma(\tau_{js} - \psi_j)\right] = 0.$$

Dividing by $c_{js} > 0$ gives

$$2\rho\lambda A_j = \kappa(\tau_{js} - \theta_s) + \gamma(\tau_{js} - \psi_j).$$

Rearranging,

$$\Gamma \tau_{js} = \kappa\theta_s + \gamma\psi_j + 2\rho\lambda A_j,$$

which implies

$$\tau_{js}^* = \frac{\kappa\theta_s + \gamma\psi_j}{\Gamma} + \frac{2\rho\lambda}{\Gamma}A_j = \alpha\theta_s + (1-\alpha)\psi_j + \eta A_j.$$

□

Proof of Proposition 3.2.

Proof. For any story s,

$$\frac{\partial \mathfrak{s}_j}{\partial c_{js}} = \tau_{js}.$$

Hence

$$\frac{\partial P_j(x)}{\partial c_{js}} = P_j(x)(1-P_j(x))\left[n_s + 2\lambda\tau_{js}(x-\mathfrak{s}_j)\right].$$

Integrating over viewers gives

$$\frac{\partial D_j}{\partial c_{js}} = n_s B_j + 2\lambda\tau_{js}A_j.$$

The first-order condition for a covered story is

$$\rho\left[n_s B_j + 2\lambda\tau_{js}A_j\right] - \left[\frac{\kappa}{2}(\tau_{js}-\theta_s)^2 + \frac{\gamma}{2}(\tau_{js}-\psi_j)^2\right] - \zeta c_{js} = \mu_j,$$

where μ_j is the multiplier on $\sum_s c_{js} = 1$.

Substituting the optimal tone

$$\tau_{js}^* = \underline{\tau}_{js} + \eta A_j$$

into the first-order condition yields

$$c_{js}^* = \frac{N_{js} - \mu_j}{\zeta},$$

where

$$N_{js} = \rho B_j n_s + 2\rho\lambda A_j\underline{\tau}_{js} + \frac{2\rho^2\lambda^2}{\Gamma}A_j^2 - \frac{\kappa\Gamma}{2\gamma}\delta_{js}^2.$$

Summing over covered stories and imposing the time constraint,

$$\sum_{s\in\mathcal{S}_j}\frac{N_{js}-\mu_j}{\zeta} = 1,$$

we obtain

$$\mu_j = \bar{N}_j - \frac{\zeta}{|\mathcal{S}_j|}, \qquad \bar{N}_j \equiv \frac{1}{|\mathcal{S}_j|} \sum_{s \in \mathcal{S}_j} N_{js}.$$

Substituting back gives

$$c^*_{js} = \frac{1}{|\mathcal{S}_j|} + \frac{N_{js} - \bar{N}_j}{\zeta}, \qquad s \in \mathcal{S}_j.$$

□

Proof of Corollary 3.1.

Proof. From the first-order and complementary slackness conditions of the coverage problem,

$$c^*_{js} = \max\left\{0, \frac{N_{js} - \mu_j}{\zeta}\right\}.$$

Since $\zeta > 0$, it follows immediately that $c^*_{js} > 0$ if and only if $N_{js} > \mu_j$, while $c^*_{js} = 0$ whenever $N_{js} \leq \mu_j$. □

Proof of Corollary 3.2.

Proof. From Corollary 3.1, story s is covered iff $N_{js} > \mu_j$. Substituting $\bar{\tau}_{js} = \alpha\theta_s + (1-\alpha)\psi_j$ and $\delta_{js} = (1-\alpha)(\psi_j - \theta_s)$ into N_{js} and completing the square in θ_s, we obtain

$$N_{js} = \rho B_j n_s + \Gamma\eta A_j \bar{\tau}_{js} + \frac{\Gamma\eta^2}{2} A_j^2 - \frac{\Gamma\alpha}{2(1-\alpha)} d_{js}^2.$$

The condition $N_{js} > \mu_j$ is therefore equivalent to

$$\rho B_j n_s - \frac{\Gamma\alpha}{2(1-\alpha)} d_{js}^2 > \mu_j - \Gamma\eta A_j \bar{\tau}_{js} - \frac{\Gamma\eta^2}{2} A_j^2,$$

which rearranges to the condition in the statement. Since $\rho B_j > 0$ and $\frac{\Gamma\alpha}{2(1-\alpha)} > 0$, the coverage condition defines an interval in θ_s centered at $\theta_j^{\max}$. Solving for the boundary gives the half-width $\overline{d}_{js}$, which is increasing in n_s and decreasing in α because a higher α raises the credibility cost of framing stories far from their true valence. □

Proof of Proposition 3.3.

Proof. By definition,

$$\mathfrak{s}_j = \sum_{s=1}^{\mathcal{S}} c_{js}\tau_{js}.$$

Substituting the optimal tone from Proposition 3.1,

$$\tau_{js}^* = \alpha\theta_s + (1-\alpha)\psi_j + \eta A_j,$$

we obtain

$$\mathfrak{s}_j^* = \sum_{s=1}^{\mathcal{S}} c_{js}^*\Big(\alpha\theta_s + (1-\alpha)\psi_j + \eta A_j\Big).$$

Since $(1-\alpha)\psi_j + \eta A_j$ does not depend on s and $\sum_s c_{js}^* = 1$, this simplifies to

$$\mathfrak{s}_j^* = \alpha \sum_{s=1}^{\mathcal{S}} c_{js}^*\theta_s + (1-\alpha)\psi_j + \eta A_j.$$

Defining

$$\bar{\theta}_j^* \equiv \sum_{s=1}^{\mathcal{S}} c_{js}^*\theta_s,$$

the first expression follows. The second follows by substituting the coverage formula from Proposition 3.2 into $\bar{\theta}_j^*$. □

Proof of Proposition 3.4.

Proof. Part (i) follows directly from the equilibrium tone equation

$$\tau_{js}^* = \alpha\theta_s + (1-\alpha)\psi_j + \eta A_j,$$

which implies

$$\frac{\partial \tau_{js}^*}{\partial A_j} = \eta > 0.$$

For part (ii), recall that equilibrium coverage satisfies

$$c_{js}^* = \max\left\{\frac{N_{js} - \mu_j}{\zeta}, 0\right\}, \qquad \sum_s c_{js}^* = 1.$$

Fix a marginal change in A_j such that the active set $\mathcal{S}_j$ is locally unchanged.

Then, for every $s \in \mathcal{S}_j$,

$$c^*_{js} = \frac{N_{js} - \mu_j}{\zeta}.$$

Differentiating with respect to A_j gives

$$\frac{\partial c^*_{js}}{\partial A_j} = \frac{1}{\zeta}\left(\frac{\partial N_{js}}{\partial A_j} - \frac{\partial \mu_j}{\partial A_j}\right).$$

Using

$$N_{js} = \rho B_j n_s + \Gamma\eta A_j \underline{\tau}_{js} + \frac{\Gamma\eta^2}{2}A_j^2 - \Gamma\frac{\alpha}{2(1-\alpha)}\delta_{js}^2,$$

I obtain

$$\frac{\partial N_{js}}{\partial A_j} = \Gamma\eta\underline{\tau}_{js} + \Gamma\eta^2 A_j = \Gamma\eta\tau^*_{js}.$$

Hence,

$$\frac{\partial c^*_{js}}{\partial A_j} = \frac{1}{\zeta}\left(\Gamma\eta\tau^*_{js} - \frac{\partial \mu_j}{\partial A_j}\right).$$

Differentiating the resource constraint

$$\sum_{s\in\mathcal{S}_j} c^*_{js} = 1$$

yields

$$\sum_{s\in\mathcal{S}_j} \frac{\partial c^*_{js}}{\partial A_j} = 0,$$

so

$$\frac{\partial \mu_j}{\partial A_j} = \frac{\Gamma\eta}{|\mathcal{S}_j|}\sum_{r\in\mathcal{S}_j}\tau^*_{jr}.$$

Substituting back gives

$$\frac{\partial c^*_{js}}{\partial A_j} = \frac{\Gamma\eta}{\zeta}\left[\tau^*_{js} - \frac{1}{|\mathcal{S}_j|}\sum_{r\in\mathcal{S}_j}\tau^*_{jr}\right].$$

This proves part (ii). For part (iii), recall that

$$N_{js} = \rho B_j n_s + \Gamma\eta A_j \underline{\tau}_{js} + \frac{\Gamma\eta^2}{2}A_j^2 - \Gamma\frac{\alpha}{2(1-\alpha)}\delta_{js}^2.$$

Differentiating with respect to A_j yields

$$\frac{\partial N_{js}}{\partial A_j} = \Gamma\eta\underline{\tau}_{js} + \Gamma\eta^2 A_j = \Gamma\eta(\underline{\tau}_{js} + \eta A_j).$$

Using

$$\tau^*_{js} = \underline{\tau}_{js} + \eta A_j,$$

I obtain

$$\frac{\partial N_{js}}{\partial A_j} = \Gamma\eta\tau^*_{js}.$$

This proves part (iii). Since omission is governed by the threshold condition $N_{js} \leq \mu_j$, a change in A_j can alter the omission set only at points where $N_{js} = \mu_j$. □

Proof of Proposition 3.5.

Proof. Throughout, fix outlet j and consider a marginal change in α holding A_j fixed and keeping the active set $\mathcal{S}_j \equiv \{s : c^*_{js} > 0\}$ locally unchanged.

Part (i). Tone. From Proposition 3.1,

$$\tau^*_{js} = \alpha\theta_s + (1-\alpha)\psi_j + \eta A_j.$$

Differentiating with respect to α gives

$$\frac{\partial \tau^*_{js}}{\partial \alpha} = \theta_s - \psi_j.$$

Hence, if $\theta_s > \psi_j$, tone shifts rightward as α rises, while if $\theta_s < \psi_j$, tone shifts leftward. This establishes part (i).

Part (iii). Net benefit. Recall that

$$N_{js} = \rho B_j n_s + \Gamma\eta A_j\underline{\tau}_{js} + \frac{\Gamma\eta^2}{2}A_j^2 - \Gamma\frac{\alpha}{2(1-\alpha)}\delta^2_{js},$$

where

$$\underline{\tau}_{js} = \alpha\theta_s + (1-\alpha)\psi_j, \qquad \delta_{js} = (1-\alpha)(\psi_j - \theta_s).$$

Since A_j is held fixed, the first and third terms are constant with respect to α, so

$$\frac{\partial N_{js}}{\partial \alpha} = \Gamma \eta A_j \frac{\partial \underline{\tau}_{js}}{\partial \alpha} - \Gamma \frac{\partial}{\partial \alpha}\left[\frac{\alpha}{2(1-\alpha)}\delta_{js}^2\right].$$

Now,

$$\frac{\partial \underline{\tau}_{js}}{\partial \alpha} = \theta_s - \psi_j.$$

Also, using $\delta_{js} = (1-\alpha)(\psi_j - \theta_s)$, we can write

$$\delta_{js}^2 = (1-\alpha)^2(\psi_j - \theta_s)^2.$$

Hence

$$\frac{\alpha}{2(1-\alpha)}\delta_{js}^2 = \frac{\alpha(1-\alpha)}{2}(\psi_j - \theta_s)^2.$$

Differentiating yields

$$\frac{\partial}{\partial \alpha}\left[\frac{\alpha}{2(1-\alpha)}\delta_{js}^2\right] = \frac{1}{2}(\psi_j - \theta_s)^2.$$

Since

$$\frac{\delta_{js}^2}{(1-\alpha)^2} = (\psi_j - \theta_s)^2,$$

this can be rewritten as

$$\frac{\partial}{\partial \alpha}\left[\frac{\alpha}{2(1-\alpha)}\delta_{js}^2\right] = \frac{\delta_{js}^2}{2(1-\alpha)^2}.$$

Therefore,

$$\frac{\partial N_{js}}{\partial \alpha} = \Gamma \eta A_j(\theta_s - \psi_j) - \Gamma \frac{\delta_{js}^2}{2(1-\alpha)^2},$$

which proves part (iii).

Part (ii). Coverage. For each covered story $s \in \mathcal{S}_j$, Proposition 3.2 implies

$$c_{js}^* = \frac{N_{js} - \mu_j}{\zeta},$$

where μ_j is the multiplier on the airtime constraint

$$\sum_{s \in \mathcal{S}_j} c_{js}^* = 1.$$

Differentiating with respect to α gives, for each $s \in \mathcal{S}_j$,

$$\frac{\partial c_{js}^*}{\partial \alpha} = \frac{1}{\zeta}\left[\frac{\partial N_{js}}{\partial \alpha} - \frac{\partial \mu_j}{\partial \alpha}\right].$$

To determine $\partial \mu_j / \partial \alpha$, differentiate the resource constraint:

$$\sum_{s \in \mathcal{S}_j} \frac{\partial c_{js}^*}{\partial \alpha} = 0.$$

Substituting the expression above,

$$\sum_{s \in \mathcal{S}_j} \frac{1}{\zeta}\left[\frac{\partial N_{js}}{\partial \alpha} - \frac{\partial \mu_j}{\partial \alpha}\right] = 0.$$

Multiplying by ζ and rearranging,

$$\frac{\partial \mu_j}{\partial \alpha} = \frac{1}{|\mathcal{S}_j|} \sum_{r \in \mathcal{S}_j} \frac{\partial N_{jr}}{\partial \alpha}.$$

Substituting back,

$$\frac{\partial c_{js}^*}{\partial \alpha} = \frac{1}{\zeta}\left[\frac{\partial N_{js}}{\partial \alpha} - \frac{1}{|\mathcal{S}_j|} \sum_{r \in \mathcal{S}_j} \frac{\partial N_{jr}}{\partial \alpha}\right].$$

This proves part (ii).

Parts (i)–(iii) together imply that a rise in credibility pressure moves tone toward truth story by story, and reallocates coverage toward stories whose net benefit improves relative to the active-set average, $\mathcal{S}_j$. □