SOCIAL NETWORKS OF LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents are increasingly deployed in interacting populations, raising the question of what they come to believe collectively. Whether a population aggregates genuine knowledge or collapses into a false consensus affects how much such systems can be trusted. Classical social-network models assume that the network structure alone determines how beliefs combine. This assumption breaks down for LLM agents, whose belief updates incorporate only part of the information they are exposed to, so these models may mischaracterize information aggregation and cannot distinguish genuine consensus from herding. We introduce SNLA, a framework that models each agent's realized influence on others rather than network connectivity. This influence depends on each agent's position in the network and on how sharply attention focuses. Theoretically, we show in a linear belief model that narrow attention leads to herding, with only a few agents shaping the collective belief as the population grows. Wide attention recovers wisdom-of-crowds behavior only on undirected, degree-regular exposure graphs. Empirically, a controlled testbed validates these predictions, and the herding–wisdom transition is demonstrated on operator-controlled variants of three multi-agent LLM benchmarks. These results identify attention breadth, together with exposure structure, as the key determinant of whether an LLM population aggregates knowledge or herds, offering a principled basis for designing and auditing multi-agent systems that are trustworthy at scale.

1 INTRODUCTION

Large language model (LLM) agents are increasingly deployed in interacting populations. They debate to reach answers (Du et al., 2024; Chan et al., 2024; Liang et al., 2024), advise and persuade one another (Park et al., 2023), and simulate social processes at scale (Gao et al., 2023; Piao et al., 2025). In all of these settings, one question recurs: what does the population come to believe collectively, and can that collective belief be trusted? (i) A group that pools its members' independent information can be more accurate than any individual (Galton, 1907); (ii) a group that copies its most visible members can collapse into a confident error (Banerjee, 1992; Bikhchandani et al., 1992). Distinguishing the two is essential before such systems are relied upon for decisions.

The classical answer comes from the theory of social learning on networks. In the DeGroot model (DeGroot, 1974) and its variants (Friedkin & Johnsen, 1990), each agent repeatedly replaces its belief with a weighted average of its neighbors', with weights read off the network: the graph structure alone determines how beliefs combine. Golub & Jackson (2010) characterize when this process is wise. A large society converges to the truth if and only if no single agent stays too influential as the society grows; prominent agents are the main obstruction. A parallel Bayesian line reaches the same tension through sequential observational learning (Bala & Goyal, 1998; Acemoglu & Ozdaglar, 2011). But it rests on one assumption that fails for LLM agents: that an agent aggregates everything it is exposed to in the network.

But an LLM agent does not aggregate information in this way. Its context is limited, and thus its attention is selective: it incorporates only part of what it is exposed to, and concentrates on the most salient sources. What control how beliefs are combined therefore not the visible network, but a reweighted version that concentrates influence on a few sources (Figure 1). Prior work places LLM agents in networks and documents their failures: agents slide into conformity, herding toward the most confident peer even when a minority is better supported (Cho et al., 2025), and grow increasingly sycophantic in repeated debate (Pitre et al., 2025). These studies propose fixes, but they

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

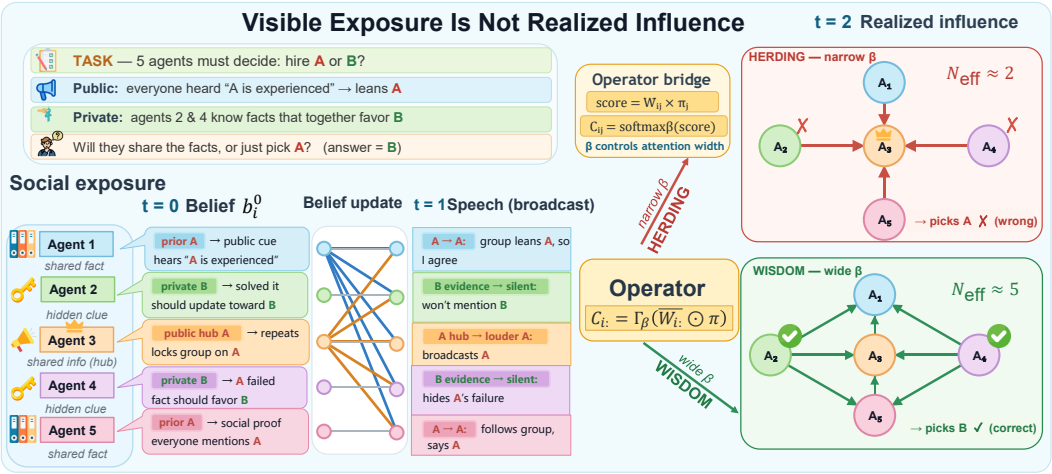

Figure 1: **From exposure to realized influence.** The visible *exposure* graph $\bar{W}$ (left) records what each agent could see; edges are exposure weights and node a_3 carries high social power π. The realized-influence operator $C_{ij} = \Gamma_\beta(\bar{W}_{ij} \odot \pi)$ reweights each agent's exposure by source social power and concentrates it through a finite attention width β. The graph C (right) depends sharply on β: narrow attention (β=0.3) collapses influence onto the most powerful source (*herding*); wide attention (β=3) spreads it across the population (*wisdom of crowds*); β=1 allocates proportionally.

do not explain when a population herds and when it stays wise, nor do they identify a controllable knob that moves a population between the two regimes.

We close this gap with SNLA (Social Networks of LLM Agents), a social-network model in which exposure and realized influence are distinct. Starting from the visible network, SNLA derives a *realized-influence operator* determined by two quantities: the *social power* of each source (Bonacich, 1987; Brin, 1998), and an *attention width* that controls how sharply finite attention concentrates on the most powerful sources. Our focus is the selective-attention regime away from this classical limit.

Our contributions are summarized as follows:
▷ **Realized influence as a derived object.** We introduce SNLA, a framework in which influence is not read directly from the network structure but *derived* from it through social power and a finite attention width, thereby separating visible exposure from realized influence (Section 3).
▷ **Attention width controls collective intelligence.** A *bridge theorem* couples the LLM population to a tractable proxy path-wise (Theorem 4.1); a *two-regime theorem* shows attention width drives the herding–wisdom transition on the proxy, reaching wisdom-of-crowds accuracy only when every agent hears comparably many others (Theorem 4.2); an *equalization theorem* shows failure modes share a common cause, which a decentralized pricing protocol removes, eliminating herding in the exact-clearing limit and substantially reducing it online (Theorem 4.3).
▷ **Empirical validation.** A controlled testbed with scalar signals and explicitly anchored updating directly validates the variance predictions and the proxy–LLM bridge; on operator-controlled variants of three recognized benchmarks, the herding–wisdom transition, the placement effect, and the equalizer's removal of herding all reproduce as qualitative swings in collective accuracy (Section 5).

2 Related Work

Social networks. Classical opinion dynamics reads influence directly off the network. In the DeGroot (DeGroot, 1974) and Friedkin–Johnsen (Friedkin & Johnsen, 1990) models, each agent updates toward a weighted average of its neighbors, and Golub & Jackson (2010) characterize when this yields accurate collective estimates. A parallel Bayesian line reaches similar conclusions under sequential updating (Bala & Goyal, 1998; Acemoglu & Ozdaglar, 2011). These models were designed for human agents who read everything they are exposed to; for LLM agents, finite attention real-

izes only part of the exposure as influence. SNLA builds on this tradition but replaces the network weights with an influence operator derived from social power and attention width, recovering the classical models as a special case.

LLM multi-agent systems and collective behavior. Prior work shows LLM populations reproducing human-like opinion dynamics and network structure (Chuang et al., 2024; Papachristou & Yuan, 2025), with consensus degrading at scale (De Marzo et al., 2024), debate helping only when independent (Du et al., 2024; Liang et al., 2024; Chan et al., 2024; Taubenfeld et al., 2024), and single persuasive agents swaying whole groups (Amayuelas et al., 2024). A separate line wires such agents into systems for software development (Qian et al., 2024), role-play (Li et al., 2023), and task-solving (Wu et al., 2023; Hong et al., 2024; Guo et al., 2024). Neither asks what a topology actually realizes: the former documents herding without a control that reverses it; the latter treats it as a design choice for performance, not a determinant of influence. SNLA asks that question directly, separating exposure from realized influence to derive when a population herds or stays wise.

Mixture-of-experts. MoE routing can exhibit a pathological behavior similar to herding: a temperature-softmax over gate scores collapses tokens onto a few experts (*router collapse*), and the standard fixes, including load-balancing losses (Shazeer et al., 2017; Fedus et al., 2022), Sinkhorn routing (Clark et al., 2022), and loss-free per-expert bias (Wang et al., 2024), can be seen as instances of our equalizer, with the exposure price y_j the analogue of the per-expert load bias. SNLA provides the missing collective-estimation account: it characterizes when concentration limits the effective sample size and when balancing restores it.

3 Problem Formulation

In this section, we formalize the social network induced by a population of LLM agents by distinguishing visible exposure from the *realized-influence operator* that shapes agents' beliefs.

3.1 Opinion Dynamics and Collective Estimation

Opinion dynamics. Let $b_i(t) \in \mathbb{R}$ denote agent i's scalar belief at time t, such as a numerical score representing how strongly it believes a given statement. We write $b(t) = (b_1(t), \dots, b_n(t))^\top \in \mathbb{R}^n$ and $b^0 := b(0)$. In the DeGroot model (DeGroot, 1974), each agent updates its belief as a weighted average of its neighbors'; the Friedkin–Johnsen model (Friedkin & Johnsen, 1990) adds anchoring to agent-specific initial beliefs:

$$b_i(t+1) = \sum_j W_{ij}\, b_j(t), \qquad b_i(t+1) = (1-\lambda_i)\, b_i^0 + \lambda_i \sum_j W_{ij}\, b_j(t),$$

where $\boldsymbol{W} = [W_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is row-stochastic, $\lambda_i \in [0,1]$ is the *anchoring weight*, and $b_i^0 := b_i(0)$ is the initial belief. The matrix is exogenous: the network determines how beliefs combine.

Collective estimation. Let $\theta^\star \in \mathbb{R}$ denote the unknown quantity being estimated. Each agent begins with a *private signal* $b_i^0 = \theta^\star + \xi_i$, where the ξ_i are independent, mean-zero, and have variance s_i^2. For evaluation, the *collective estimate* at horizon T is the population average

$$\hat{\theta}_T = \frac{1}{n}\sum_i b_i(T) = \sum_i q_{T,i} b_i^0,$$

where $q_T \in \mathbb{R}^n$ and $q_T^\top \mathbf{1} = 1$. This evaluation does not assume a flat interaction structure. In the homoskedastic case $s_i^2 \equiv s^2$, its variance is

$$\mathrm{Var}(\hat{\theta}_T) = s^2 \|q_T\|_2^2,$$

minimized at s^2/n when $q_T = \frac{1}{n}\mathbf{1}$, the wisdom-of-crowds optimum (Golub & Jackson, 2010). A weight concentrated on a few agents instead keeps the variance bounded away from s^2/n regardless of population size, resulting in herding. Whether a population is wise or herds is therefore determined by q_T, which the network produces as beliefs evolve.

3.2 Social Networks of LLM Agents

Agents and events. We model a population $\mathcal{A} = \{1, \dots, n\}$ of n LLM agents. Each agent i is a triple (ρ_i, μ_i, m_i), where ρ_i is a fixed *persona* that anchors its belief, $\mu_i(t)$ is a *mutable state* holding its memory, and m_i is a *backbone model* mapping context to an emission. At each round $t \in \{0, 1, 2, \dots\}$, agents emit *events*. With $\mathcal{X}$ denoting the space of message texts, an event

$$e_m = (s_m, R_m, x_m, b_m, t_m) \in \mathcal{A} \times 2^{\mathcal{A}} \times \mathcal{X} \times \mathbb{R} \times \mathbb{R}_{\geq 0}$$

records its sender, recipient set, message text, declared belief, and timestamp. The history $\mathcal{E}_{\leq t} = \{e_m : t_m \leq t\}$ induces an agent network capturing the interactions up to round t.

Exposure and realized-influence graphs. Classical opinion dynamics distinguishes the *communication graph*, recording which agents are aware of one another's opinions, from the *influence graph*, recording who actually shapes whom. In the classical setting, the two coincide, since an agent that reads a neighbor also aggregates information from that neighbor (Proskurnikov & Tempo, 2017). Finite attention breaks this coincidence. At round t, the interaction history $\mathcal{E}_{\leq t}$ induces two evolving weighted directed graphs on the agents:

$$\mathcal{G}_{\bar{W}}(t) = \big(\mathcal{A}, \bar{W}(t)\big), \qquad \mathcal{G}_{C}(t) = \big(\mathcal{A}, C(t)\big),$$

In the *visible exposure graph* $\mathcal{G}_{\bar{\boldsymbol{W}}}(t)$, an edge $j \to i$ is present whenever agent i has been exposed to a message from agent j in $\mathcal{E}_{\leq t}$, with weight $\bar{W}_{ij}(t) \in [0, 1]$ the attention agent i places on source j. The *realized-influence graph* $\mathcal{G}_{\boldsymbol{C}}(t)$ records the influence that finite attention realizes: an agent can be influenced only by sources it is exposed to, so its arcs are contained in the exposure arcs,

$$\mathcal{E}_{\boldsymbol{C}}(t) \subseteq \mathcal{E}_{\bar{\boldsymbol{W}}}(t),$$

and $\boldsymbol{C}_{ij}(t) \in [0, 1]$ is the weight with which agent i aggregates source j's belief. Together with a *social power* vector $\pi(t) \in \Delta_n$ that weights each agent by its centrality as a source in $\mathcal{G}_{\bar{\boldsymbol{W}}}(t)$, these objects form the *social network of LLM agents* (SNLA), the time-indexed tuple:

$$\mathcal{S}(t) = \big(\mathcal{A},\, \mathcal{E}_{\leq t},\, \bar{\boldsymbol{W}}(t),\, \pi(t),\, \boldsymbol{C}(t)\big).$$

Its defining feature is that $\boldsymbol{C}(t) \neq \bar{\boldsymbol{W}}(t)$ in general; under uniform social power and proportional attention, $\boldsymbol{C}(t) = \bar{\boldsymbol{W}}(t)$, recovering the classical model.

4 Proposed Methodology

In this section, we construct the realized-influence operator $\boldsymbol{C}(t)$, introduce a decentralized protocol that prevents source dominance, and state the resulting theoretical guarantees.

4.1 The Realized-Influence Operator

Exposure and social power. Fix a memory kernel $\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that discounts old events and per-event qualities $q_m \geq 0$. The unnormalized exposure of agent i to source j accumulates every message j sent that i received:

$$A_{ij}(t) = \sum_{e_m \in \mathcal{E}_{\leq t}} \mathbf{1}\{s_m = j,\, i \in R_m\}\, \kappa(t - t_m)\, q_m,$$

and row-normalizing gives the row-stochastic exposure matrix $\bar{\boldsymbol{W}}(t)$. Exposure alone treats all sources symmetrically, but an agent carries more standing when influential agents attend to it. We score this standing by the Katz–Bonacich centrality of the exposure graph (Bonacich, 1987), for a damping factor $\zeta \in [0, 1)$:

$$\pi(t)^{\top} = \frac{1 - \zeta}{n}\, \mathbf{1}^{\top} \big(\boldsymbol{I} - \zeta \bar{\boldsymbol{W}}(t)\big)^{-1} \in \Delta_n,$$

which weights each source by exposure accumulated over paths of every length. At $\zeta = 0$ power is uniform and influence reduces to exposure alone.

Algorithm 1: SNLA-EQ: realized influence with equalized context allocation.

Input : agents $\{(\rho_i, \mu_i, m_i)\}_{i=1}^n$, private signals b_i^0, parameters β, ζ, T
Output: collective estimate $\hat{\theta}_T$

initialize beliefs $b_i(0) \leftarrow b_i^0$ and prices $y_j \leftarrow 1$ for all i, j
for $t = 0, 1, \ldots, T-1$ **do**
  form exposure $\bar{\boldsymbol{W}}(t)$ and social power $\pi(t)$ from history $\mathcal{E}_{\le t}$
  Reader rule: priced attention width-β allocation
  $\boldsymbol{C}_{ij}(t) \leftarrow \dfrac{s_{ij}(t)^{1/\beta}/y_j}{\sum_k s_{ik}(t)^{1/\beta}/y_k}$ for all i, j
  foreach *agent* i **do**
    build context from persona ρ_i, state $\mu_i(t)$, and influence row $\boldsymbol{C}_{i:}(t)$
    emit belief, message, recipients $(b_i(t{+}1), x_i(t), R_i(t)) \sim m_i$
    append event $e_i(t)$ to $\mathcal{E}$ and update state $\mu_i(t{+}1)$
  Source rule: each source raises its price by its attention share
  $a_j \leftarrow \sum_i \boldsymbol{C}_{ij}(t), \quad y_j \leftarrow y_j \cdot a_j$ for all j
return $\hat{\theta}_T = \frac{1}{n}\sum_i b_i(T)$

Realized influence. Exposure and power say how much each source *could* influence agent i; a finite context window forces i to realize only part of it. We form a per-pair *score* $s_{ij}(t) := \bar{W}_{ij}(t)\,\pi_j(t)$ — multiplicative, so a source matters to i only when i is exposed to it *and* it carries standing — and pass each row through a temperature-β softmax over the log-scores:

$$s_{ij}(t) := \bar{W}_{ij}(t)\,\pi_j(t), \qquad \boldsymbol{C}_{ij}(t) = \frac{s_{ij}(t)^{1/\beta}}{\sum_k s_{ik}(t)^{1/\beta}}. \tag{1}$$

The width β is the temperature, controlling how sharply each reader's budget concentrates: $\beta \to 0$ collapses the row onto its highest-scoring source (hard attention), $\beta = 1$ allocates proportionally, and $\beta \to \infty$ spreads it uniformly over exposed sources. The result $\boldsymbol{C}(t)$ is the realized-influence operator of Section 3: row-stochastic, supported on $\bar{\boldsymbol{W}}(t)$, and equal to $\bar{\boldsymbol{W}}(t)$ at the classical point $\zeta = 0, \beta = 1$. It drives the opinion dynamics in place of the classical fixed matrix, updated each round as new events arrive. The LLM harness retains top-weighted messages covering 90% of $C_{i:}$ (Appendix 12.3); omitted mass is absorbed into δ.

4.2 Equalized Context Allocation

Exposure pricing. The operator we just built has a structural weakness that the theory of Section 4.3 traces to a single source. Row normalization bounds each reader's budget, but nothing bounds the source side: a high-power source can occupy a large share of the population's contexts, so its column sum grows without bound,

$$\sum_i \boldsymbol{C}_{ij}(t) \;\gg\; 1.$$

We correct this with a decentralized pricing protocol: each source j holds a scalar *exposure price* $y_j > 0$ (initialized at one), which discounts its weight in every reader's allocation and is updated by the attention share it received,

$$\boldsymbol{C}_{ij}(t) = \frac{s_{ij}(t)^{1/\beta}/y_j}{\sum_k s_{ik}(t)^{1/\beta}/y_k}, \qquad y_j \leftarrow y_j \cdot a_j, \quad a_j = \sum_i \boldsymbol{C}_{ij}(t). \tag{2}$$

At the fixed point $a_j = 1$ for all j, so $\boldsymbol{C}(t)$ has unit row and column sums, $\mathbf{1}^\top \boldsymbol{C}(t) = \mathbf{1}^\top$ and $\boldsymbol{C}(t)\mathbf{1} = \mathbf{1}$: the realized influence is doubly stochastic. This is one Sinkhorn iteration per round (Sinkhorn, 1967), decentralized to $O(d_i)$ work and one broadcast scalar per agent, and (in its cleared limit) invariant to the social-power weighting, attenuates exactly the high-power sources that would otherwise dominate. Algorithm 1 collects the full per-round loop; the Sinkhorn equivalence and a targeted heteroskedastic variant are in Appendix 11.

4.3 THEORETICAL GUARANTEES

With the allocator in place, we now analyze what it implies for collective accuracy. The operational beliefs are emitted by LLMs and are not directly analyzable, so we pair them with a tractable proxy driven by the same realized influence, an anchored best-response (Friedkin–Johnsen) recursion:

$$b_i^{\mathrm{BR}}(t+1) = (1-\lambda_i)\, b_i^0 + \lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, b_j^{\mathrm{BR}}(t), \qquad \lambda_{\max} = \max_i \lambda_i, \tag{3}$$

with anchoring weights $\lambda_i \in [0,1)$. The operational emission need not solve this recursion exactly; we measure its per-step deviation by the residual of agent i at round t, evaluated at the operational system's own beliefs $b(t)$:

$$\hat{\delta}_i(t) = \Big| b_i(t+1) - (1-\lambda_i)\, b_i^0 - \lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, b_j(t) \Big|, \tag{4}$$

which is computable from the event log once the run is complete. Unrolling the proxy recursion writes its estimate as a fixed weighting of the initial signals, $\hat{\theta}_T^{\mathrm{BR}} = q_T^\top b^0$, where the collective weight q_T is determined by $\{\boldsymbol{C}(t)\}$ (Appendix 8). The wisdom-versus-herding question of Section 3.1 thus becomes a question about q_T, which the three results below answer: the first connects the proxy to the operational system, and the next two characterize the consensus weight under the baseline and the equalized allocator. All proofs, constants, and assumptions are in Appendix 8.

Theorem 4.1 (Bridge). *Suppose every LLM emission is a δ-approximate best response evaluated at its own beliefs, i.e. $\hat{\delta}_i(t) \le \delta_T$ for all i and $t < T$, where $\delta_T = \max_{i,\, 0 \le t < T} \hat{\delta}_i(t)$ is the uniform residual equation 4 and $\lambda_{\max} < 1$. Then for every $t \le T$ and every row-stochastic sequence $\{\boldsymbol{C}(t)\}$,*

$$\big\| b(t) - b^{\mathrm{BR}}(t) \big\|_\infty \le \frac{\delta_T}{1-\lambda_{\max}}. \tag{5}$$

The bound is path-wise and attained, and anchoring cannot be dropped: at $\lambda_{\max} = 1$ a row-stochastic sequence and emission rule exists whose gap grows linearly in t (Appendix 8), a counterexample rather than a claim that every $\lambda_{\max} = 1$ system diverges. The bound is non-vacuous in practice: an explicit anchoring prompt keeps λ comfortably below one on the anchored-persona cells (median 0.25; full breakdown in Section 5.2), and Figure 4 confirms the bound holds on real agents. The operational estimate inherits the proxy's accuracy, $|\hat{\theta}_T - \hat{\theta}_T^{\mathrm{BR}}| \le \delta_T/(1-\lambda_{\max})$, and likewise its standard deviation, so it suffices to analyze q_T, measured via the effective sample size $N_{\mathrm{eff}}(q_T) = 1/\|q_T\|_2^2$ (Kish, 1965), which gives variance $s^2/N_{\mathrm{eff}}(q_T)$.

The next theorem fixes the exposure graph and varies the attention width. It characterizes the consensus weight $\nu_{\boldsymbol{C}(\beta)}$, the stationary distribution of $\boldsymbol{C}(\beta)$, which is the long-horizon limit of q_T under uniform anchoring as $\lambda \uparrow 1$ (Appendix 8); accordingly $N_{\mathrm{eff}}(\beta) = 1/\|\nu_{\boldsymbol{C}(\beta)}\|_2^2$. We reserve $\pi(t)$ for the social power that builds $\boldsymbol{C}$ and write ν for the consensus weight it induces; the two are distinct.

Theorem 4.2 (Attention width controls collective accuracy). *Let $\boldsymbol{C}(\beta)$ be the fixed-graph realized influence with $\zeta \in [0,1)$ and irreducible exposure support B, where $B_{ij} = \mathbf{1}\{\bar{W}_{ij} > 0\}$. Write the degrees $d_i = \sum_j B_{ij}$, $D = \mathrm{diag}(d_1, \dots, d_n)$, $d_{\max} = \max_i d_i$, and let L be the log-spread and $\underline{g}$ the dominance gap of the scores s_{ij} (Appendix 8). In the homoskedastic case:*

1. (Wide, wisdom of crowds.) *If $B = B^\top$ is irreducible, then $\boldsymbol{C}(\beta) \to D^{-1}B$ and $\nu_{\boldsymbol{C}(\beta)} \to d/(\mathbf{1}^\top d)$ as $\beta \to \infty$, so:*

$$\lim_{\beta \to \infty} N_{\mathrm{eff}}(\beta) = \frac{(\sum_i d_i)^2}{\sum_i d_i^2},$$

which equals n if and only if B is degree-regular; for finite β the consensus weight obeys $\big| \|\nu_{\boldsymbol{C}(\beta)}\|_2^2 - \|\nu^\star\|_2^2 \big| \le 2\kappa_B (e^{L/\beta} - 1)$ with $\nu^\star = d/(\mathbf{1}^\top d)$ and κ_B a condition number of $D^{-1}B$ (Appendix 8), so: $N_{\mathrm{eff}}(\beta) \to (\sum_i d_i)^2 / \sum_i d_i^2$ by continuity.

2. (Narrow, herding.) *Assume B is irreducible and the scores admit a unique dominant pair: $j_i^\star = \arg\max_k s_{ik}$ is unique for every row, with $j_i^\star = h$ for all $i \ne h$ and $j_h^\star = g \ne h$. Then with $\rho(\beta) = (d_{\max} - 1)\, e^{-\underline{g}/\beta}$, whenever $\rho(\beta) < 1$,*

$$N_{\mathrm{eff}}(\beta) \le \frac{2}{(1-\rho(\beta))^2},$$

$$\textit{so } N_{\text{eff}}(\beta) \le 8 \textit{ whenever } \rho(\beta) \le \tfrac{1}{2}\textit{, i.e. } \beta \le \underline{g}/\log(2(d_{\max}-1))\textit{, independently of } n.$$

Narrow attention concentrates each row of $\boldsymbol{C}$ on the dominant pair, collapsing the collective weight onto a few agents; wide attention flattens it but reaches the optimum only under degree-regularity, the sole remaining obstruction. Both failures are column imbalances of $\boldsymbol{C}$, which is exactly what the equalized allocator removes.

Theorem 4.3 (Exact optimality under equalization)**.** *Under uniform anchoring $\lambda \in [0,1)$ and signal exogeneity, let $\bar{\delta}_{\text{col}}$ be the largest column defect over the run. Then the proxy weight satisfies:*

$$\left\| q_T - \tfrac{1}{n}\mathbf{1} \right\|_1 \le \bar{\delta}_{\text{col}} \left(\tfrac{\lambda}{1-\lambda} + T\lambda^T \right),$$

for every horizon T and every realized sequence. In particular, if every $\boldsymbol{C}(t)$ is doubly stochastic ($\bar{\delta}_{\text{col}} = 0$), then $q_T = \frac{1}{n}\mathbf{1}$ exactly, so $\mathrm{Var}(\hat{\theta}_T^{\mathrm{BR}}) = s^2/n$ and $N_{\text{eff}} = n$. The operational estimate inherits this up to the bridge gap,

$$\left| \mathrm{sd}(\hat{\theta}_T) - \tfrac{s}{\sqrt{n}} \right| \le \frac{\delta_T}{1-\lambda} + s\,\bar{\delta}_{\text{col}} \left(\tfrac{\lambda}{1-\lambda} + T\lambda^T \right).$$

Each price round is one Sinkhorn iteration, which need not clear the market exactly in a single pass; the guarantee is stated in the measured column defect $\bar{\delta}_{\text{col}}$ actually achieved at run time, and is exact only in the doubly stochastic limit, reached by iterating the price update to tolerance (Appendix 11). Unlike the baseline, which collapses to $N_{\text{eff}} \le 8$ under a dominant pair no matter how large n is, the equalized allocator drives q_T to uniform at a rate set by $\bar{\delta}_{\text{col}}$, with no mixing, aperiodicity, or connectivity hypothesis on the realized sequence.

5 Experiments

5.1 Experimental setup

Benchmarks. *HiddenBench* (Li et al., 2025) poses 12 hidden-profile reasoning problems, where each of n=24 agents holds a different clue that the group must pool to answer correctly; *Werewolf* (Xu et al., 2023) is a 16-agent social deduction game where villagers deliberate and vote to unmask a hidden adversarial werewolf; and *AgentsNet* (Grötschla et al., 2025) places $n \in \{8, 16, 24, 32, 50\}$ agents as nodes in a graph, where each communicates only with its neighbours to reach a conflict-free colouring. We additionally map the theorem's boundary on full-information *Debate* (Du et al., 2024) and commons-dilemma *GovSim* (Piatti et al., 2024). Full environment specifications are in Appendix 12.4.

Models. By default, each agent uses Qwen2.5-7B-Instruct (Qwen et al., 2025) with greedy decoding. Each condition uses 5 seeds that redraw tasks and position assignments; the shaded bands in Figures 2 and 3 show their min–max. Unless noted, agents share weights and differ only in persona, private information, and network position, so the confident-wrong leader differs by position rather than capability. Two experiments introduce heterogeneity: AgentsNet pairs strong 7B hubs with weak 1.5B agents, while the capability ladder varies the central agent over Qwen2.5-$\{1.5, 3, 7, 14\}$B and Llama-3-$\{1, 3, 8\}$B (Grattafiori et al., 2024). Full model, operator, seed, prompt, and hardware details are in Appendices 12.1–12.6.

5.2 Main results

Attention width β controls herding versus wisdom. A single knob, the context width β, swings the collective between herding and wisdom on both aggregation-dependent tasks (Figure 2A,B). In *HiddenBench*, the baseline collapses onto the confident-wrong hub at narrow β and recovers at wide β, improving collective accuracy by $+\mathbf{0.64}$. A single item traced through both regimes, with the exact prompts and the model's real replies, is in Appendix 12.7. The equalizer removes this collapse, while the control shows it disappears without the dominant-wrong source. *Werewolf* shows the same pattern with the game's intrinsic wrong source seated at the influence hub: the baseline moves from near-zero accuracy at narrow β to substantially higher accuracy at wide β ($+\mathbf{0.61}$), while the equalizer and control stay above the herding floor. The same collapse and equalizer recovery replicate on Llama-3.1-8B (Table 11). The individual and oracle lines mark the lower and upper reference levels; the oracle gap is diagnosed in Appendix 13.3. The full per-β breakdown, including the causal adversary-on/off decomposition, is in Appendices 13.4 (HiddenBench) and 13.5 (Werewolf).

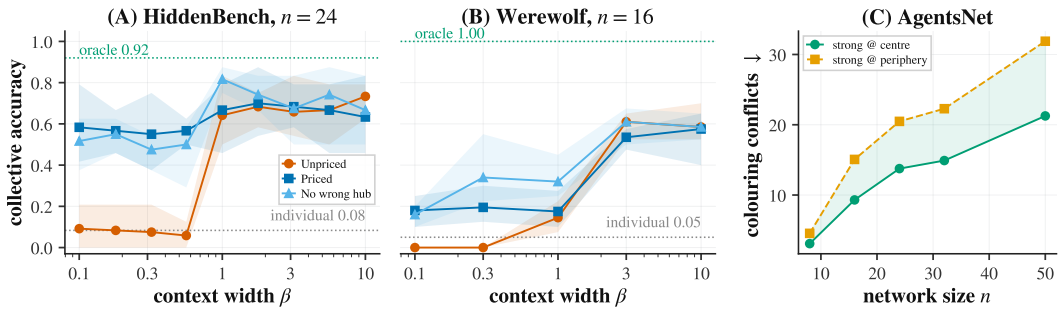

Figure 2: **One influence operator, three recognized environments, two facets. (A, B)** Collective accuracy vs. context width β on *HiddenBench* (n=24) and *Werewolf* (n=16). Orange is the *baseline* allocator, blue is the *equalizer*, and light blue is the *control* that removes the dominant-wrong source. Dotted lines show individual and oracle reference levels; shaded bands in (A, B) span the seed-to-seed min–max over 5 seeds. **(C)** Colouring conflicts vs. network size n on *AgentsNet*, comparing strong agents placed at the centre (green) with strong agents placed at the periphery (gold).

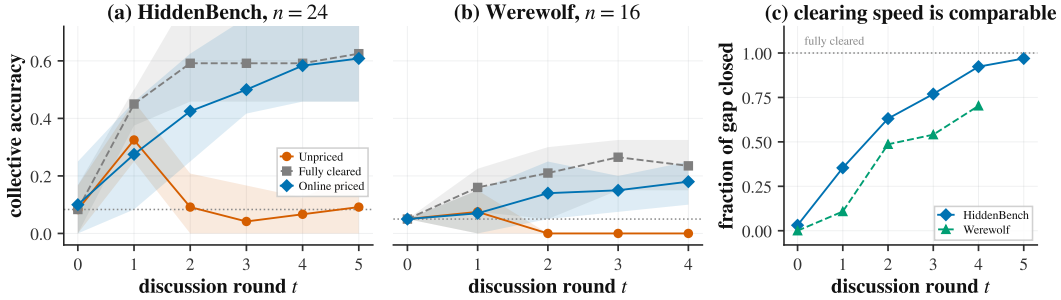

Figure 3: **The equalizer clears herding online, round by round.** Per-round collective accuracy at narrow β on *HiddenBench* (n=24) and *Werewolf* (n=16). Orange is the *baseline*; grey dashed is the *operator*, obtained by clearing the equalized operator to its fixed point each round; blue is *online pricing*, which applies one price update per discussion round. Shaded bands span the min–max over 5 seeds. **(c)** Fraction of the narrow-β gap closed over rounds on HiddenBench and Werewolf.

Where capability sits controls coordination. On a coordination task the operator exposes a different facet (Figure 2C). *AgentsNet* has agents self-organise into complementary colours. A heterogeneous 7B×1.5B society fixes the capability budget, varying only where the strong agents sit: strong agents at the centre yield fewer conflicts than at the periphery, and the advantage grows with scale, from $+\mathbf{1.5}$ at n=8 to $+\mathbf{10.6}$ at n=50. Coordination has no single wrong answer to herd onto, isolating placement rather than the herding–wisdom transition. The placement ablation matrix (social-power and Sinkhorn invariance, the weak-hub Watts–Strogatz control) is in Appendix 13.6.

Online pricing clears herding round by round. Figure 3 separates the ideal equalized operator from its online implementation, which runs one price update per discussion round rather than to convergence (the decentralized protocol of Section 4.2). Because prices start uniform, the first step sits at the baseline and the collective follows the confident-wrong source; subsequent updates reduce the captured column, so the online curve rises toward the operator curve. The recovery is nearly complete on *HiddenBench* (Figure 3a, 0.10 → 0.61) but partial on the shorter-horizon *Werewolf* (Figure 3b, 0.05→0.18 against the operator's 0.24). The two close their gaps at a similar per-round pace, so the remaining Werewolf gap is a horizon effect, not a different clearing mechanism.

The proxy predicts real LLM estimates. The discussion benchmarks run DeGroot updating (λ=1), the unanchored limit of Theorem 4.2: the proxy's effective sample size collapses and recovers with β, and the LLM's collective accuracy tracks that swing (Figure 2A,B). At λ=1 this is empirical, not a consequence of Theorem 4.1, whose bound is vacuous here (Proposition 8.4); the two share the β-gating mechanism, not a theorem mapping N_{eff} onto accuracy. The bridge bound

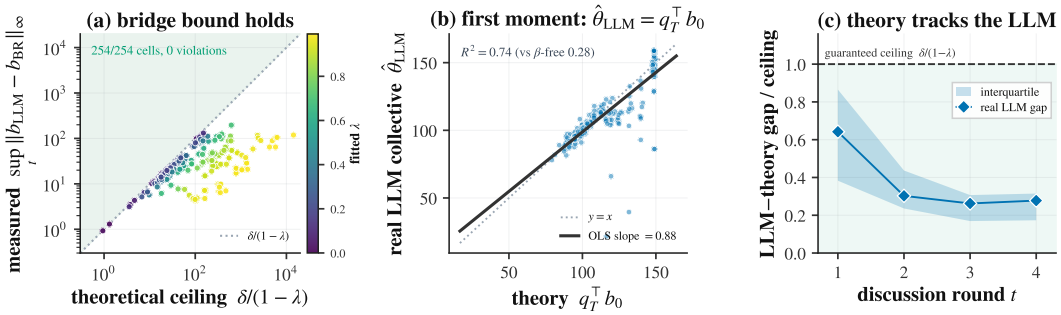

Figure 4: **The proxy–LLM bridge (Theorem 4.1) on real Qwen data.** **(a)** Measured sup-norm discrepancy $\sup_t \|b_{\mathrm{LLM}} - b_{\mathrm{BR}}\|_\infty$ vs. the theoretical ceiling $\delta_T/(1-\lambda_{\max})$, pooling 254 cells — one per (arm, β, seed) with $\lambda_{\max} < 1$ — across four 8-seed testbed runs spanning Qwen2.5-3B and -7B backbones, including one 7B run with an explicit anchoring-persona prompt and one 3B run at reduced damping (ζ=0.6); coloured by fitted $\lambda_{\max}$; dotted line is $y = x$. **(b)** Real LLM collective estimate $\hat{\theta}_{\mathrm{LLM}}$ vs. the proxy's first-moment prediction $q_T^\top b_0$, pooling the 8-seed 7B run from panel (a) with a further 24-seed 7B sweep; dotted is $y = x$, solid is the OLS fit. **(c)** LLM–theory gap, normalized by the guaranteed ceiling $\delta/(1-\lambda)$ (dashed), across T=4 discussion rounds on the 8-seed anchored-persona 7B run from panel (a) alone; band is the interquartile range.

Table 1: **Ablations (controlled testbed, Qwen2.5-3B, 8 seeds).** Collective error, lower is better; narrow β=0.1 is the herding regime, wide β=10 the aggregate. Variants: memory, exposure-pricing, late-joiner, distance-decay, skeptic.

Ablation	range	narrow β	wide β	reading
social power ζ=0	gate off	**5.5**	4.9	**no damping ⇒ no herding**
social power ζ>0	0.2–0.95	≈ 42	4–6	damping on ⇒ herding
self-reliance λ	0.3–0.7	42	11.5	robust to follower self-weight
equalizer k (Sinkhorn)	1–4	column defect 0.14 → 0.037 (Fig. 5a)		one knob clears the market
model variants	5 kinds	18–59	2–15	herding persists; eq clears each

requires anchoring, so we certify the correspondence on the anchored testbed: there the LLM estimate is the influence-weighted average the theory predicts (slope 0.88, R^2=0.74), and the sup-norm bound holds on all 254 cells (Figure 4). The anchoring λ is fit per agent, so the ceiling is fitted not prescribed—tight on the anchored subset, satisfied over the pool; provenance is in Appendix 13.3.

5.3 Ablations

The herding floor is an effective-sample-size effect, not a small-network artifact: it holds at every society size on the real benchmarks (Werewolf for $n \in \{6, 8, 10, 12, 16\}$, HiddenBench for $n \in \{8, 16, 24\}$; full breakdown in Appendix 13.4 and 13.5), so a larger network does not dilute a captured hub. Table 1 sweeps the remaining controls in the controlled testbed (the ablation grid, plus the model-scale sweep, is in Appendix 13.2 and 13.1): the two-regime is invariant to follower self-reliance and survives five structural variants of the interaction model. It is, however, gated causally by social power, vanishing entirely when the gate is off (ζ=0).

6 Conclusion

SNLA shows that a single parameter controls collective intelligence: the attention width β. It governs whether an LLM-agent population pools its information or herds. Narrow β collapses collective accuracy onto a few dominant sources; wide β recovers wisdom-of-crowds behavior, once exposure is balanced. One failure then remains, a high-power source that occupies too many contexts, and a decentralized pricing rule removes it, restoring the optimal collective weights whenever a dominant source is present, with no retraining.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and code assistance. We have not used generative AI tools to generate experimental results or fabricate data, and other tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools to improve the clarity and presentation of the manuscript. We have reviewed all AI-assisted work. AI-assisted code was inspected and tested by the authors for correctness, and all AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Daron Acemoglu and Asuman Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011.

Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Yang Wang. Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6929–6948, 2024.

Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *The review of economic studies*, 65 (3):595–621, 1998.

Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3): 797–817, 1992.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992.

Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92 (5):1170–1182, 1987.

Sergey Brin. The pagerank citation ranking: bringing order to the web. *Proceedings of ASIS, 1998*, 98:161–172, 1998.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International conference on learning representations*, volume 2024, pp. 9079–9093, 2024.

Young-Min Cho, Sharath Chandra Guntuku, and Lyle Ungar. Herd behavior: Investigating peer influence in llm-based multi-agent systems. *arXiv preprint arXiv:2505.21588*, 2025.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 3326–3346, 2024.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pp. 4057–4086. PMLR, 2022.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Giordano De Marzo, Claudio Castellano, and David Garcia. Language understanding as a constraint on consensus size in llm societies. *arXiv preprint arXiv:2409.02822*, 2024.

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345): 118–121, 1974.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of mathematical sociology*, 15(3-4):193–206, 1990.

Francis Galton. Vox populi, 1907.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Florian Grötschla, Luis Müller, Jan Tönshoff, Mikhail Galkin, and Bryan Perozzi. Agentsnet: Coordination and collaborative reasoning in multi-agent llms. *arXiv preprint arXiv:2507.08616*, 2025.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. arxiv 2024. *arXiv preprint arXiv:2402.01680*, 10, 2024.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Leslie Kish. Survey sampling. 1965.

David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Society, 2026.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Hiddenbench: Assessing collective reasoning in multi-agent llms via hidden profile tasks. *arXiv preprint arXiv:2505.11556*, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

Carl D Meyer, Jr. The role of the group generalized inverse in the theory of finite markov chains. *Siam Review*, 17(3):443–464, 1975.

Marios Papachristou and Yuan Yuan. Network formation and dynamics among multi-llms. *PNAS nexus*, 4(12):pgaf317, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. 2025.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759, 2024.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. Consensagent: Towards efficient and effective consensus in multi-agent llm interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22112–22133, 2025.

Anton V Proskurnikov and Roberto Tempo. A tutorial on modeling and analysis of dynamic social networks. part i. *Annual Reviews in Control*, 43:65–79, 2017.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15174–15186, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 251–267, 2024.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940*, 2023.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Daisy Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, 2025.

Supplementary Materials

7 Assumptions and Measurement of δ

This appendix collects the assumptions of the three theorems and the procedure measuring δ from the event log. Throughout, $P = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_{\max} = \max_i \lambda_i$, and $\{\boldsymbol{C}(t)\}$ is the row-stochastic realized-influence sequence; operational and proxy beliefs $b(t), b^{\mathrm{BR}}(t)$ are both initialized at b^0.

Assumption 7.1 (Anchored cooperative regime)**.** The proxy is the quadratic, cooperative, anchored best response: the anchoring weights satisfy $\lambda_i \in [0, 1)$ for all i, so that $\lambda_{\max} < 1$.

The weight $\lambda_i < 1$ is the regime in which the persona anchor ρ_i exerts nonzero pull toward b_i^0. The restriction $\lambda_{\max} < 1$ cannot in general be dropped: Proposition 8.4 exhibits a row-stochastic sequence and a δ-rational rule with $\lambda_{\max} = 1$ whose gap grows linearly in t. This is a constructive counterexample, not a claim that every $\lambda_{\max} = 1$ system diverges.

Assumption 7.2 (δ-rational emission)**.** There exists $\delta \geq 0$ such that, along the operational trajectory,

$$\left| b_i(t+1) - (1-\lambda_i)\, b_i^0 - \lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, b_j(t) \right| \;\leq\; \delta \qquad \text{for all } i \text{ and all } t \geq 0, \tag{6}$$

where $\boldsymbol{C}(t)$ is the realized-influence matrix produced by the operational system at round t.

Two features of equation 6 matter. The best response is evaluated at the LLM's *own* beliefs $b(t)$, not the proxy $b^{\mathrm{BR}}(t)$: a one-step condition presupposing nothing about trajectory distance. And every quantity in equation 6 is logged, so δ is measurable per emission, the assumption falsifiable.

Assumption 7.3 (Signal model and signal–exposure independence)**.** The private signals are $b_i^0 = \theta^\star + \xi_i$ with $\xi_1, \ldots, \xi_n$ independent, $\mathbb{E}[\xi_i] = 0$, and $\mathrm{Var}(\xi_i) = s_i^2$, and the noise vector $\xi = (\xi_1, \ldots, \xi_n)$ is independent of the realized sequence $\{\boldsymbol{C}(t)\}_{t \geq 0}$. The homoskedastic case is $s_i^2 \equiv s^2$.

Independence of ξ from $\{\boldsymbol{C}(t)\}$ holds when recipient selection and event qualities depend on messages only through timing and routing, not belief values; it is the exogenous-topology analogue from classical opinion dynamics and matches our fixed-topology environments. It is what licenses treating q_T as fixed under $\mathbb{E}_\xi$ in Proposition 8.6: belief-dependent selection couples q_T and ξ, voiding the conditional variance identities, and so lies outside all variance-level results below. The Bridge Theorem, which does not use Assumption 7.3, is unaffected.

Measuring δ. Assumption 7.2 is checkable round by round. At round t, both the realized-influence row $\boldsymbol{C}_{i:}(t)$ and the operational beliefs $b(t)$ are functions of the event log, so the per-emission residual

$$\hat{\delta}_i(t) = \left| b_i(t+1) - (1-\lambda_i)\, b_i^0 - \lambda_i \textstyle\sum_j \boldsymbol{C}_{ij}(t)\, b_j(t) \right| \tag{7}$$

is computable offline once the run is complete. The anchoring weights (λ_i) are not assumed known: they are fit per agent by least squares over the run, minimizing $\sum_t \hat{\delta}_i(t)^2$ over $\lambda_i \in [0, 1)$. For fixed i this objective is a univariate quadratic in λ_i, since $\hat{\delta}_i(t)$ is affine in λ_i at each t, so the fit is closed-form. The theorems require the *uniform* residual of Assumption 7.2, so the value that instantiates them is the maximum

$$\delta_T = \max_{i,\, 0 \leq t < T} \hat{\delta}_i(t),$$

not a quantile: an upper quantile bounds all but a fraction of the emissions and yields only a trimmed, high-probability diagnostic, under which the bridge bound holds after discarding the residual outliers. A finite run of length T measures δ_T and so supports the bridge bound up to horizon T; extending it to all t requires the residual to remain uniformly bounded beyond the run, which is an assumption, not a measurement. The empirical distribution of $\hat{\delta}_i(t)$ is itself a finding: it quantifies how far a given backbone, persona, and topic depart from a rational anchored opinion updater, and its tail governs the gap between the uniform bound and its trimmed counterpart.

8 PROOF OF THEOREM 4.1 (BRIDGE)

This appendix proves the three theorems and supporting results. The development is self-contained: all constants are explicit, and the stationary perturbation of the two-regime theorem is carried out from a deviation-matrix identity (Lemma 9.3).

We use the following notation throughout. For a matrix R, $\|R\|_\infty = \max_i \sum_j |R_{ij}|$ is the maximum absolute row sum, which is the operator norm induced by the vector ℓ_∞ norm; for a vector v, $\|v\|_p$ is the usual ℓ_p norm. A row-stochastic matrix $\boldsymbol{C}$ satisfies $\boldsymbol{C} \geq 0$ entrywise and $\boldsymbol{C}\mathbf{1} = \mathbf{1}$. A nonnegative square matrix is *irreducible* if its support digraph is strongly connected. We write e_j for the j-th standard basis vector and Δ_n for the probability simplex in $\mathbb{R}^n$. The exposure graph, social power, and realized influence are $\bar{\boldsymbol{W}}(t)$, $\pi(t)$, and $\boldsymbol{C}(t)$ as defined in Section 4.1; we abbreviate the per-pair exposure–power score by

$$s_{ij}(t) \coloneqq \bar{W}_{ij}(t)\, \pi_j(t), \tag{8}$$

so that the realized influence is the row-wise tempered allocation $\boldsymbol{C}_{ij}(t) = s_{ij}(t)^{1/\beta} / \sum_k s_{ik}(t)^{1/\beta}$. Since $\pi_j(t) \geq \frac{1-\zeta}{n} > 0$ for every j (Lemma 8.2 below), $s_{ij}(t) > 0$ if and only if $\bar{W}_{ij}(t) > 0$, so $\boldsymbol{C}(t)$ has exactly the support of the exposure graph.

8.1 THE PROXY AND ITS CLOSED FORM

The operational beliefs are emitted by the LLM backbones and are not directly analyzable. The proxy replaces each emission by an anchored best response driven by the *same* realized influence sequence $\{\boldsymbol{C}(t)\}$. Writing the cooperative quadratic best response of agent i explicitly, the proxy is the recursion

$$b^{\mathrm{BR}}(0) = b^0, \qquad b_i^{\mathrm{BR}}(t+1) = (1-\lambda_i)\, b_i^0 + \lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, b_j^{\mathrm{BR}}(t), \tag{9}$$

which in matrix form, with $P = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, reads

$$b^{\mathrm{BR}}(t+1) = (\boldsymbol{I} - P)\, b^0 + P\, \boldsymbol{C}(t)\, b^{\mathrm{BR}}(t). \tag{10}$$

The peer beliefs entering equation 9 are the proxy's own $b_j^{\mathrm{BR}}(t)$, never the operational $b_j(t)$: the proxy is a self-contained linear dynamical system, coupled to the operational run only through the realized matrices $\{\boldsymbol{C}(t)\}$. Two classical models are strict special cases. With uniform anchoring $\lambda_i \equiv \lambda$ the recursion is the Friedkin–Johnsen model (Friedkin & Johnsen, 1990) on the realized influence; with $\lambda_i \equiv 1$ it is the DeGroot model (DeGroot, 1974).

Lemma 8.1 (Best-response closed form)**.** *Fix agent i and round t, and let the peer beliefs $b^{\mathrm{BR}}(t)$ be given. The cooperative quadratic best response*

$$b_i^{\mathrm{BR}}(t+1) = \arg\min_{b \in \mathbb{R}} (1-\lambda_i)\, (b - b_i^0)^2 + \lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, \big(b - b_j^{\mathrm{BR}}(t)\big)^2$$

has the unique solution equation 9, for every $\lambda_i \in [0,1]$ and every row-stochastic $\boldsymbol{C}(t)$.

Proof. Write the objective as

$$F(b) = (1-\lambda_i)\, (b - b_i^0)^2 + \lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, \big(b - b_j^{\mathrm{BR}}(t)\big)^2.$$

The coefficient of b^2 is $(1-\lambda_i) + \lambda_i \sum_j \boldsymbol{C}_{ij}(t) = 1 > 0$, so F is strictly convex with unique minimizer at $F'(b) = 0$. Since

$$F'(b) = 2(1-\lambda_i)(b - b_i^0) + 2\lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, \big(b - b_j^{\mathrm{BR}}(t)\big) = 2\Big[b - (1-\lambda_i)\, b_i^0 - \lambda_i \sum_j \boldsymbol{C}_{ij}(t)\, b_j^{\mathrm{BR}}(t)\Big]$$

(using $\sum_j \boldsymbol{C}_{ij}(t) = 1$), $F'(b) = 0$ gives equation 9. □

8.2 WELL-POSEDNESS OF SOCIAL POWER

The realized influence is built from the exposure graph $\bar{\boldsymbol{W}}(t)$ and the social power $\pi(t)$, the Katz–Bonacich centrality

$$\pi(t)^\top = \frac{1-\zeta}{n}\,\mathbf{1}^\top\big(\boldsymbol{I} - \zeta\bar{\boldsymbol{W}}(t)\big)^{-1}, \qquad \zeta \in [0,1).$$

The following lemma records that this is well defined and strictly positive, a fact used repeatedly below (it guarantees that $\boldsymbol{C}(t)$ has exactly the support of $\bar{\boldsymbol{W}}(t)$).

Lemma 8.2 (Well-posedness of social power)**.** *For every row-stochastic $\bar{\boldsymbol{W}}(t)$ and every $\zeta \in [0,1)$, the inverse $(\boldsymbol{I} - \zeta\bar{\boldsymbol{W}}(t))^{-1}$ exists, $\pi(t) \in \Delta_n$, and*

$$\pi_j(t) \;\geq\; \frac{1-\zeta}{n} \;>\; 0 \qquad \textit{for every } j.$$

At $\zeta = 0$, $\pi_j(t) = 1/n$ for all j.

Proof. Since $\bar{\boldsymbol{W}}(t)$ is row-stochastic, $\|\bar{\boldsymbol{W}}(t)\|_\infty = 1$, hence $\|\zeta\bar{\boldsymbol{W}}(t)\|_\infty = \zeta < 1$. The Neumann series

$$\big(\boldsymbol{I} - \zeta\bar{\boldsymbol{W}}(t)\big)^{-1} = \sum_{k=0}^{\infty} \zeta^k \bar{\boldsymbol{W}}(t)^k$$

converges absolutely, so the inverse exists. Every term is entrywise nonnegative, so $\pi(t) \geq 0$; the $k = 0$ term gives

$$\pi_j(t) = \frac{1-\zeta}{n}\sum_{k\geq 0}\zeta^k\big[\mathbf{1}^\top\bar{\boldsymbol{W}}(t)^k\big]_j \;\geq\; \frac{1-\zeta}{n}\big[\mathbf{1}^\top\bar{\boldsymbol{W}}(t)^0\big]_j = \frac{1-\zeta}{n}.$$

Row-stochasticity gives $\bar{\boldsymbol{W}}(t)^k\mathbf{1} = \mathbf{1}$, so

$$\pi(t)^\top\mathbf{1} = \frac{1-\zeta}{n}\,\mathbf{1}^\top\sum_{k\geq 0}\zeta^k\bar{\boldsymbol{W}}(t)^k\mathbf{1} = \frac{1-\zeta}{n}\sum_{k\geq 0}\zeta^k\,\mathbf{1}^\top\mathbf{1} = \frac{1-\zeta}{n}\cdot n\cdot\frac{1}{1-\zeta} = 1,$$

so $\pi(t) \in \Delta_n$. At $\zeta = 0$, $\pi(t)^\top = \frac{1}{n}\mathbf{1}^\top$. □

8.3 THE BRIDGE THEOREM

The Bridge Theorem couples the operational system to the proxy path-wise, for the realized $\{\boldsymbol{C}(t)\}$, with no stationarity, mixing, or convergence hypothesis: the contraction comes entirely from $\lambda_{\max} < 1$.

Theorem 4.1 (Bridge)**.** *Suppose every LLM emission is a δ-approximate best response evaluated at its own beliefs, i.e. $\hat{\delta}_i(t) \leq \delta_T$ for all i and $t < T$, where $\delta_T = \max_{i,\,0\leq t<T}\hat{\delta}_i(t)$ is the uniform residual equation 4 and $\lambda_{\max} < 1$. Then for every $t \leq T$ and every row-stochastic sequence $\{\boldsymbol{C}(t)\}$,*

$$\big\|b(t) - b^{\mathrm{BR}}(t)\big\|_\infty \leq \frac{\delta_T}{1-\lambda_{\max}}. \tag{5}$$

Proof. Let $e(t) \coloneqq b(t) - b^{\mathrm{BR}}(t)$, so $e(0) = 0$.

Step 1 (residual). Define

$$u(t+1) \coloneqq b(t+1) - \Big[(\boldsymbol{I} - P)\,b^0 + P\,\boldsymbol{C}(t)\,b(t)\Big].$$

Its i-th entry is the quantity bounded in equation 6, so

$$\|u(t+1)\|_\infty \;\leq\; \delta, \qquad t \geq 0. \tag{11}$$

Step 2 (error recursion). Subtracting equation 10 from the definition of $u(t+1)$, the anchor terms $(\boldsymbol{I} - P)b^0$ cancel:

$$\begin{aligned} e(t+1) &= \Big[u(t+1) + (\boldsymbol{I} - P)b^0 + P\boldsymbol{C}(t)\,b(t)\Big] - \Big[(\boldsymbol{I} - P)b^0 + P\boldsymbol{C}(t)\,b^{\mathrm{BR}}(t)\Big] \\ &= u(t+1) + P\boldsymbol{C}(t)\,e(t). \end{aligned}$$

Step 3 (contraction). Since $\|\boldsymbol{C}(t)\|_\infty = 1$ and $\|P\|_\infty = \lambda_{\max}$, $\|P\boldsymbol{C}(t)\|_\infty \le \lambda_{\max}$, so by equation 11,

$$\|e(t+1)\|_\infty \le \|u(t+1)\|_\infty + \|P\boldsymbol{C}(t)\, e(t)\|_\infty \le \delta + \lambda_{\max}\, \|e(t)\|_\infty. \tag{12}$$

Step 4 (unrolling). Induction on equation 12 from $e(0) = 0$ gives $\|e(t)\|_\infty \le \delta \sum_{r=0}^{t-1} \lambda_{\max}^r$: the inductive step is

$$\|e(t+1)\|_\infty \le \delta + \lambda_{\max} \cdot \delta \sum_{r=0}^{t-1} \lambda_{\max}^r = \delta \sum_{r=0}^{t} \lambda_{\max}^r.$$

Summing the geometric series, with $\lambda_{\max} < 1$,

$$\|e(t)\|_\infty \le \delta\, \frac{1-\lambda_{\max}^t}{1-\lambda_{\max}} \le \frac{\delta}{1-\lambda_{\max}} = \bar{\delta}, \tag{13}$$

the bound of Theorem 4.1. □

Remark 8.3 (No stationarity is required). The bound equation 13 holds for every realization of the time-varying, endogenous, feedback-coupled influence sequence $\{\boldsymbol{C}(t)\}$, with no stationarity, mixing, irreducibility, or convergence hypothesis whatsoever. The only structural input is row-stochasticity of each $\boldsymbol{C}(t)$, which is automatic, and the strict anchor $\lambda_{\max} < 1$. In particular the theorem applies verbatim to the equalized sequence of Section 4.2, whose matrices are produced by the online pricing protocol and are in general neither stationary nor mutually commuting.

The next proposition shows the bound is tight and that the anchoring hypothesis cannot be removed: at $\lambda_{\max} = 1$ the discrepancy grows without bound.

Proposition 8.4 (Sharpness and failure at $\lambda = 1$)**.** *The bound equation 13 is attained. For every row-stochastic sequence $\{\boldsymbol{C}(t)\}$ and uniform anchoring $\lambda_i \equiv \lambda$, the operational emission rule*

$$b(t+1) = (1-\lambda)\, b^0 + \lambda\, \boldsymbol{C}(t)\, b(t) + \delta \mathbf{1}$$

satisfies Assumption 7.2 with equality and produces

$$b(t) - b^{\mathrm{BR}}(t) = \delta\, \frac{1-\lambda^t}{1-\lambda}\, \mathbf{1} \quad (\lambda < 1), \qquad b(t) - b^{\mathrm{BR}}(t) = \delta\, t\, \mathbf{1} \quad (\lambda = 1).$$

In particular, at $\lambda = 1$ the gap grows linearly in t and no uniform-in-t bridge exists.

Proof. The residual is $u(t+1) = \delta\mathbf{1}$, so the error recursion becomes $e(t+1) = \lambda\, \boldsymbol{C}(t)\, e(t) + \delta\mathbf{1}$. By induction $e(t) = c_t\mathbf{1}$ with $c_0 = 0$, $c_{t+1} = \lambda c_t + \delta$: if $e(t) = c_t\mathbf{1}$ then, since $\boldsymbol{C}(t)\mathbf{1} = \mathbf{1}$,

$$e(t+1) = \lambda\, \boldsymbol{C}(t)\, (c_t\mathbf{1}) + \delta\mathbf{1} = \lambda c_t\, \boldsymbol{C}(t)\mathbf{1} + \delta\mathbf{1} = (\lambda c_t + \delta)\, \mathbf{1},$$

Solving gives $c_t = \delta \sum_{r=0}^{t-1} \lambda^r$, equal to $\delta\frac{1-\lambda^t}{1-\lambda}$ for $\lambda < 1$ and δt for $\lambda = 1$. □

Proposition 8.4 makes anchoring a structural requirement: unanchored, the per-round emission noise accumulates without bound, so any proxy theory tolerating $\delta > 0$ must operate at $\lambda_{\max} < 1$.

8.4 Collective estimation by the proxy

The collective estimate is the population average of beliefs. We write the operational and proxy estimates as

$$\hat{\theta}_T := \frac{1}{n}\mathbf{1}^\top b(T), \qquad \hat{\theta}_T^{\mathrm{BR}} := \frac{1}{n}\mathbf{1}^\top b^{\mathrm{BR}}(T).$$

The proxy estimate is a fixed linear functional of the private signals; the next lemma identifies the weight q_T and shows it conserves mass. For an arbitrary row-stochastic sequence $\{\boldsymbol{C}(t)\}$ set $M(t) := P\boldsymbol{C}(t)$ and define the backward products

$$\Phi(T, r) := M(T-1)\, M(T-2) \cdots M(r) \quad (0 \le r < T), \qquad \Phi(T, T) := \boldsymbol{I}. \tag{14}$$

Lemma 8.5 (Trajectory representation and mass conservation)**.** *For every row-stochastic sequence* $\{\boldsymbol{C}(t)\}$ *and every* $T \geq 0$*,*

$$b^{\mathrm{BR}}(T) = G_T\, b^0, \qquad G_T \coloneqq \Phi(T,0) + \sum_{r=0}^{T-1} \Phi(T, r+1)\,(\boldsymbol{I} - P), \tag{15}$$

and the collective weight $q_T^\top \coloneqq \frac{1}{n}\mathbf{1}^\top G_T$ *satisfies* $\hat{\theta}_T^{\mathrm{BR}} = q_T^\top b^0$ *and is mass-conserving,* $q_T^\top \mathbf{1} = 1$*.*

Proof. Representation. Induction on T. At $T = 0$, $\Phi(0,0) = \boldsymbol{I}$ and $G_0 = \boldsymbol{I}$. Assuming equation 15 at T, by equation 10 and the hypothesis,

$$\begin{aligned} b^{\mathrm{BR}}(T+1) &= (\boldsymbol{I} - P)\, b^0 + M(T)\, b^{\mathrm{BR}}(T) \\ &= (\boldsymbol{I} - P)\, b^0 + M(T)\, G_T\, b^0 \\ &= \Big[M(T)\,\Phi(T,0) + \sum_{r=0}^{T-1} M(T)\,\Phi(T, r+1)\,(\boldsymbol{I} - P) + (\boldsymbol{I} - P)\Big] b^0. \end{aligned}$$

Since $M(T)\Phi(T,r) = \Phi(T+1,r)$ and $(\boldsymbol{I} - P) = \Phi(T+1, T+1)(\boldsymbol{I} - P)$,

$$b^{\mathrm{BR}}(T+1) = \Big[\Phi(T+1,0) + \sum_{r=0}^{T} \Phi(T+1, r+1)\,(\boldsymbol{I} - P)\Big] b^0 = G_{T+1}\, b^0,$$

the anchor term supplying the $r = T$ summand. Then $\hat{\theta}_T^{\mathrm{BR}} = \frac{1}{n}\mathbf{1}^\top G_T b^0 = q_T^\top b^0$.

Mass conservation. We show $G_T\mathbf{1} = \mathbf{1}$, giving $q_T^\top \mathbf{1} = \frac{1}{n}\mathbf{1}^\top \mathbf{1} = 1$. For $0 \leq r < T$, using $\boldsymbol{C}(r)\mathbf{1} = \mathbf{1}$,

$$\Phi(T,r)\,\mathbf{1} = \Phi(T, r+1)\, M(r)\,\mathbf{1} = \Phi(T, r+1)\, P\, \boldsymbol{C}(r)\,\mathbf{1} = \Phi(T, r+1)\, P\,\mathbf{1},$$

hence

$$\Phi(T, r+1)\,(\boldsymbol{I} - P)\,\mathbf{1} = \Phi(T, r+1)\,\mathbf{1} - \Phi(T, r+1)\, P\,\mathbf{1} = \Phi(T, r+1)\,\mathbf{1} - \Phi(T, r)\,\mathbf{1}.$$

Summing over r telescopes:

$$\sum_{r=0}^{T-1} \Phi(T, r+1)\,(\boldsymbol{I} - P)\,\mathbf{1} = \Phi(T,T)\,\mathbf{1} - \Phi(T,0)\,\mathbf{1} = \mathbf{1} - \Phi(T,0)\,\mathbf{1},$$

so $G_T\mathbf{1} = \Phi(T,0)\mathbf{1} + [\mathbf{1} - \Phi(T,0)\mathbf{1}] = \mathbf{1}$. □

Proposition 8.6 (Unbiasedness and variance)**.** *Under Assumption 7.3, conditionally on the realized sequence* $\{\boldsymbol{C}(t)\}$*,*

$$\mathbb{E}\big[\hat{\theta}_T^{\mathrm{BR}}\big] = \theta^\star, \qquad \mathrm{Var}\big(\hat{\theta}_T^{\mathrm{BR}}\big) = q_T^\top S\, q_T, \qquad S = \mathrm{diag}(s_1^2, \ldots, s_n^2),$$

and in the homoskedastic case $\mathrm{Var}(\hat{\theta}_T^{\mathrm{BR}}) = s^2 \|q_T\|_2^2$*.*

Proof. By $b^0 = \theta^\star \mathbf{1} + \xi$ and $q_T^\top \mathbf{1} = 1$ (Lemma 8.5), $\hat{\theta}_T^{\mathrm{BR}} = q_T^\top b^0 = \theta^\star + q_T^\top \xi$. Conditionally on $\{\boldsymbol{C}(t)\}$, q_T is fixed and $\xi \perp \{\boldsymbol{C}(t)\}$ with $\mathbb{E}[\xi] = 0$, so $\mathbb{E}[\hat{\theta}_T^{\mathrm{BR}}] = \theta^\star$. By independence of the ξ_i,

$$\mathrm{Var}\big(\hat{\theta}_T^{\mathrm{BR}}\big) = \mathrm{Var}\big(q_T^\top \xi\big) = \sum_{i=1}^{n} q_{T,i}^2\, \mathrm{Var}(\xi_i) = \sum_{i=1}^{n} q_{T,i}^2\, s_i^2 = q_T^\top S\, q_T.$$

In the homoskedastic case $S = s^2 \boldsymbol{I}$, giving $s^2 \|q_T\|_2^2$. □

Definition 8.7 (Effective sample size)**.** For a weight vector $q \in \mathbb{R}^n$ with $q^\top \mathbf{1} = 1$, the *effective sample size* is $N_{\mathrm{eff}}(q) \coloneqq 1/\|q\|_2^2$ (Kish, 1965).

In the homoskedastic case $\mathrm{Var}(\hat{\theta}_T^{\mathrm{BR}}) = s^2 / N_{\mathrm{eff}}(q_T)$: the collective estimate behaves like the average of $N_{\mathrm{eff}}(q_T)$ independent signals. A dispersed weight $q_T \approx \frac{1}{n}\mathbf{1}$ gives $N_{\mathrm{eff}} \approx n$, the wisdom-of-crowds regime (Golub & Jackson, 2010); a concentrated weight $q_T \approx e_h$ gives $N_{\mathrm{eff}} \approx 1$, herding. The remaining theorems characterize $N_{\mathrm{eff}}(q_T)$ under the baseline and the equalized allocators.

The Bridge Theorem transfers these proxy moments to the operational estimate.

Corollary 8.8 (Moment transfer). *Under Assumptions 7.1–7.3, conditionally on the realized sequence* $\{\boldsymbol{C}(t)\}$,

$$\left|\hat{\theta}_T - \hat{\theta}_T^{\mathrm{BR}}\right| \le \bar{\delta} \ \text{a.s.}, \qquad \left|\mathbb{E}[\hat{\theta}_T] - \theta^\star\right| \le \bar{\delta}, \qquad \left|\mathrm{sd}(\hat{\theta}_T) - \mathrm{sd}(\hat{\theta}_T^{\mathrm{BR}})\right| \le \bar{\delta}.$$

In the homoskedastic case $\mathrm{sd}(\hat{\theta}_T^{\mathrm{BR}}) = s\|q_T\|_2$, *so* $\left|\mathrm{sd}(\hat{\theta}_T) - s\|q_T\|_2\right| \le \bar{\delta}$.

Proof. Pathwise gap. With $e(T) = b(T) - b^{\mathrm{BR}}(T)$,

$$\left|\hat{\theta}_T - \hat{\theta}_T^{\mathrm{BR}}\right| = \left|\frac{1}{n}\mathbf{1}^\top e(T)\right| \le \frac{1}{n}\sum_{i=1}^{n}|e_i(T)| \le \|e(T)\|_\infty \le \bar{\delta}$$

by Theorem 4.1. Set $z_T = \hat{\theta}_T - \hat{\theta}_T^{\mathrm{BR}}$, $|z_T| \le \bar{\delta}$ a.s.

Mean. Since $\mathbb{E}[\hat{\theta}_T^{\mathrm{BR}}] = \theta^\star$,

$$\left|\mathbb{E}[\hat{\theta}_T] - \theta^\star\right| = \left|\mathbb{E}[z_T]\right| \le \mathbb{E}\,|z_T| \le \bar{\delta}.$$

Standard deviation. With $\widetilde{X} = X - \mathbb{E}[X]$, Minkowski's inequality gives

$$\mathrm{sd}(\hat{\theta}_T) = \left\|\widetilde{\hat{\theta}_T^{\mathrm{BR}}} + \widetilde{z}_T\right\|_{L^2} \le \left\|\widetilde{\hat{\theta}_T^{\mathrm{BR}}}\right\|_{L^2} + \|\widetilde{z}_T\|_{L^2} = \mathrm{sd}(\hat{\theta}_T^{\mathrm{BR}}) + \mathrm{sd}(z_T),$$

and symmetrically, so $|\mathrm{sd}(\hat{\theta}_T) - \mathrm{sd}(\hat{\theta}_T^{\mathrm{BR}})| \le \mathrm{sd}(z_T) \le \bar{\delta}$, the last by $\mathrm{sd}(z_T)^2 \le \mathbb{E}[z_T^2] \le \bar{\delta}^2$. The homoskedastic identity is Proposition 8.6. □

9 Proof of Theorem 4.2 (Attention Width)

This appendix proves Theorem 4.2: graph constants and the tempered-allocation bound (Appendix 9.1), a stationary perturbation bound (Appendix 9.2), then the two regimes.

9.1 Constants and quantitative tempered allocation

The two-regime theorem fixes the exposure graph and varies the attention width β, so we work with the fixed-graph realized influence $\boldsymbol{C}(\beta)$ whose row i is the tempered allocation of the score row $s_{i:} = (s_{ij})_j$, with $s_{ij} = \bar{W}_{ij}\pi_j$ as in equation 8. We record the combinatorial constants attached to the exposure graph. Let

$$B_{ij} = \mathbf{1}\{\bar{W}_{ij} > 0\}, \qquad d_i = \sum_j B_{ij}, \qquad D = \mathrm{diag}(d_1, \dots, d_n), \qquad d_{\max} = \max_i d_i,$$

so B is the unweighted exposure support, d_i the out-degree of agent i, and D the degree matrix. Per row define the *log-spread* and, when $d_i \ge 2$, the *dominance gap*

$$L_i = \max_{j,k\,:\,B_{ij}=B_{ik}=1}\left|\log s_{ij} - \log s_{ik}\right|, \qquad g_i = \log\frac{s_{i,(1)}}{s_{i,(2)}}, \tag{16}$$

where $s_{i,(1)} \ge s_{i,(2)}$ are the two largest scores in row i (set $g_i = +\infty$ if $d_i = 1$). The global constants are

$$L = \max_i L_i, \qquad \underline{g} = \min_i g_i.$$

Since $\pi_j \ge \frac{1-\zeta}{n} > 0$ (Lemma 8.2), we have $s_{ij} > 0 \iff B_{ij} = 1$, so row i of $\boldsymbol{C}(\beta)$ is supported exactly on the d_i exposed sources, and $\boldsymbol{C}(\beta)$ is irreducible whenever B is.

The next lemma controls how sharply Γ_β concentrates as a function of β. For $s \in \mathbb{R}^n_{\ge 0}$, $s \ne 0$, write $N(s) = \{j : s_j > 0\}$, $|N(s)| = d$, $[\Gamma_\beta(s)]_j = s_j^{1/\beta}/\sum_{k\in N(s)} s_k^{1/\beta}$ on $N(s)$ (else 0), $\mathbf{1}_{N(s)}$ the indicator of $N(s)$, and $L(s), g(s)$ the log-spread and dominance gap of equation 16.

Lemma 9.1 (Quantitative tempered allocation). *Let* $s \in \mathbb{R}^n_{\ge 0}$, $s \ne 0$, *with* $|N(s)| = d$. *Then for every* $\beta > 0$:

(i) (Flattening.)

$$\left\|\Gamma_\beta(s) - \tfrac{1}{d}\mathbf{1}_{N(s)}\right\|_1 \;\le\; e^{L(s)/\beta} - 1 \;\le\; \frac{2L(s)}{\beta} \qquad \textit{whenever } \beta \ge L(s).$$

(ii) (Concentration.) *If s has a unique maximizer $j^\star$ with gap $g(s) > 0$, then*

$$1 - [\Gamma_\beta(s)]_{j^\star} \;\le\; (d-1)\, e^{-g(s)/\beta}.$$

Consequently $\Gamma_\beta(s) \to \frac{1}{d}\mathbf{1}_{N(s)}$ as $\beta \to \infty$ and $\Gamma_\beta(s) \to e_{j^\star}$ as $\beta \to 0$.

Proof. Set $a_j = \beta^{-1} \log s_j$ for $j \in N(s)$. Then, for $j \in N(s)$,

$$[\Gamma_\beta(s)]_j = \frac{e^{a_j}}{\sum_{k\in N(s)} e^{a_k}} = \Big(\sum_{k\in N(s)} e^{a_k - a_j}\Big)^{-1}.$$

(i) Flattening. For $j, k \in N(s)$, $|a_k - a_j| \le L(s)/\beta$, so each $e^{a_k - a_j} \in [e^{-L(s)/\beta}, e^{L(s)/\beta}]$ and their sum lies in $[d\, e^{-L(s)/\beta}, d\, e^{L(s)/\beta}]$; taking reciprocals,

$$\frac{1}{d}\, e^{-L(s)/\beta} \;\le\; [\Gamma_\beta(s)]_j \;\le\; \frac{1}{d}\, e^{L(s)/\beta} \qquad (j \in N(s)).$$

With $x = L(s)/\beta \ge 0$, using $1 - e^{-x} \le e^x - 1$,

$$\Big|[\Gamma_\beta(s)]_j - \frac{1}{d}\Big| \;\le\; \frac{1}{d}\,(e^x - 1).$$

Summing over the d supported coordinates,

$$\left\|\Gamma_\beta(s) - \tfrac{1}{d}\mathbf{1}_{N(s)}\right\|_1 = \sum_{j\in N(s)} \Big|[\Gamma_\beta(s)]_j - \frac{1}{d}\Big| \;\le\; d \cdot \frac{e^x - 1}{d} = e^{L(s)/\beta} - 1,$$

the first inequality. For the second, when $\beta \ge L(s)$, $x \in [0,1]$ and convexity of $e^x - 1$ gives $e^x - 1 \le (e-1)x \le 2x = 2L(s)/\beta$.

(ii) Concentration. Let $j^\star$ be the unique maximizer. For $k \ne j^\star$ in $N(s)$, retaining only the $j^\star$ term in the denominator,

$$[\Gamma_\beta(s)]_k = \frac{e^{a_k}}{\sum_{m\in N(s)} e^{a_m}} \;\le\; \frac{e^{a_k}}{e^{a_{j^\star}}} = e^{(\log s_k - \log s_{j^\star})/\beta} \;\le\; e^{-g(s)/\beta},$$

using $\log s_k - \log s_{j^\star} \le -g(s)$ ($s_{j^\star} = s_{(1)}$, $s_k \le s_{(2)}$). Summing over $k \ne j^\star$,

$$1 - [\Gamma_\beta(s)]_{j^\star} = \sum_{k\ne j^\star} [\Gamma_\beta(s)]_k \;\le\; (d-1)\, e^{-g(s)/\beta}.$$

Limits: $\beta \to \infty$ in (i) gives $\Gamma_\beta(s) \to \frac{1}{d}\mathbf{1}_{N(s)}$; $\beta \to 0$ in (ii) gives $\Gamma_\beta(s) \to e_{j^\star}$. □

9.2 A SELF-CONTAINED STATIONARY PERTURBATION BOUND

The wide-context regime compares $\nu_{\boldsymbol{C}(\beta)}$ with a reference chain, via the following Perron–Frobenius facts and an exact perturbation identity.

Lemma 9.2 (Stationary distribution; maximum principle)**.** *Let $\boldsymbol{C}$ be row-stochastic and irreducible. Then:*

(i) *if $x \in \mathbb{R}^n$ satisfies $\boldsymbol{C}x = x$, then x is a constant multiple of $\mathbf{1}$;*

(ii) *there is a unique $\nu_{\boldsymbol{C}} \in \Delta_n$ with $\nu_{\boldsymbol{C}}^\top \boldsymbol{C} = \nu_{\boldsymbol{C}}^\top$, and $\nu_{\boldsymbol{C},j} > 0$ for every j.*

Proof. *(i)* Let $i^\star \in \arg\max_i x_i$ and $m = x_{i^\star}$. From $\boldsymbol{C}x = x$ and row-stochasticity,

$$m = x_{i^\star} = \sum_j \boldsymbol{C}_{i^\star j}\, x_j \;\le\; \sum_j \boldsymbol{C}_{i^\star j}\, m = m,$$

so $x_j = m$ for every out-neighbor j of $i^\star$. The set $\{i : x_i = m\}$ is thus closed under out-neighbors, hence all of $\{1, \dots, n\}$ by irreducibility, so $x = m\mathbf{1}$.

(ii) Existence. The linear map $p \mapsto \boldsymbol{C}^\top p$ sends Δ_n into itself ($\boldsymbol{C} \ge 0$ and $\mathbf{1}^\top \boldsymbol{C}^\top p = \mathbf{1}^\top p$), so Brouwer gives a fixed point $\nu_{\boldsymbol{C}} \in \Delta_n$.

Positivity. Let $Z = \{j : \nu_{\boldsymbol{C},j} = 0\}$. For $j \in Z$, $0 = \sum_i \nu_{\boldsymbol{C},i} \boldsymbol{C}_{ij}$ forces $\boldsymbol{C}_{ij} = 0$ for all $i \in Z^c$: no edge from Z^c into Z, contradicting irreducibility unless $Z = \emptyset$. Hence $\nu_{\boldsymbol{C}} > 0$.

Uniqueness. Let ν, ν' be stationary (both > 0), $c = \min_j \nu'_j/\nu_j$, $v = \nu' - c\nu \ge 0$. Then $v^\top \boldsymbol{C} = v^\top$ and $v_{j_0} = 0$ at the minimizing index, so the positivity argument forces $v = 0$, i.e. $\nu' = c\nu$. Total mass gives $c = 1$, so $\nu' = \nu$. □

Lemma 9.3 (Deviation matrix and exact perturbation identity)**.** *Let $P^\star$ be row-stochastic and irreducible with stationary distribution $\nu^\star$. Then:*

(i) $Q := \boldsymbol{I} - P^\star + \mathbf{1}\nu^{\star\top}$ *is invertible; with* $Z := Q^{-1}$, *the* deviation matrix *is* $H := Z - \mathbf{1}\nu^{\star\top}$.

(ii) $Z\mathbf{1} = \mathbf{1}$, $\nu^{\star\top} Z = \nu^{\star\top}$, *and* $(\boldsymbol{I} - P^\star)\, H = \boldsymbol{I} - \mathbf{1}\nu^{\star\top}$.

(iii) *For every row-stochastic $\widetilde{\boldsymbol{C}}$ with a stationary distribution $\tilde\nu$, setting $E := \widetilde{\boldsymbol{C}} - P^\star$,*

$$\tilde\nu^\top - \nu^{\star\top} = \tilde\nu^\top E\, H, \qquad \textit{hence} \qquad \|\tilde\nu - \nu^\star\|_1 \le \|E\|_\infty\, \kappa(P^\star), \quad \kappa(P^\star) := \|H\|_\infty.$$

Proof. *(i) Invertibility.* Suppose $Qx = 0$. Multiply on the left by $\nu^{\star\top}$ and use $\nu^{\star\top} P^\star = \nu^{\star\top}$ and $\nu^{\star\top}\mathbf{1} = 1$:

$$0 = \nu^{\star\top} Qx = \nu^{\star\top} x - \nu^{\star\top} P^\star x + (\nu^{\star\top}\mathbf{1})(\nu^{\star\top} x) = \nu^{\star\top} x - \nu^{\star\top} x + \nu^{\star\top} x = \nu^{\star\top} x,$$

so $\nu^{\star\top} x = 0$, and $Qx = 0$ reduces to $P^\star x = x$. By Lemma 9.2(i), $x = c\mathbf{1}$, and $0 = \nu^{\star\top} x = c$, so $x = 0$: Q is invertible.

(ii) Identities. $Q\mathbf{1} = \mathbf{1}$ gives $Z\mathbf{1} = \mathbf{1}$, and $\nu^{\star\top} Q = \nu^{\star\top}$ gives $\nu^{\star\top} Z = \nu^{\star\top}$. Since $(\boldsymbol{I} - P^\star)\mathbf{1} = 0$,

$$(\boldsymbol{I} - P^\star)H = (\boldsymbol{I} - P^\star)\big(Z - \mathbf{1}\nu^{\star\top}\big) = (\boldsymbol{I} - P^\star)Z,$$

the term $(\boldsymbol{I} - P^\star)\mathbf{1}\nu^{\star\top}$ vanishing. Now $\boldsymbol{I} - P^\star = Q - \mathbf{1}\nu^{\star\top}$, so

$$(\boldsymbol{I} - P^\star)Z = \big(Q - \mathbf{1}\nu^{\star\top}\big)Z = QZ - \mathbf{1}\nu^{\star\top} Z = \boldsymbol{I} - \mathbf{1}\nu^{\star\top},$$

using $QZ = \boldsymbol{I}$.

(iii) Perturbation identity. Stationarity $\tilde\nu^\top \widetilde{\boldsymbol{C}} = \tilde\nu^\top$ gives

$$\tilde\nu^\top(\boldsymbol{I} - P^\star) = \tilde\nu^\top - \tilde\nu^\top P^\star = \tilde\nu^\top \widetilde{\boldsymbol{C}} - \tilde\nu^\top P^\star = \tilde\nu^\top(\widetilde{\boldsymbol{C}} - P^\star) = \tilde\nu^\top E.$$

Multiplying on the right by H and using (ii),

$$\tilde\nu^\top E\, H = \tilde\nu^\top(\boldsymbol{I} - P^\star)H = \tilde\nu^\top\big(\boldsymbol{I} - \mathbf{1}\nu^{\star\top}\big) = \tilde\nu^\top - (\tilde\nu^\top\mathbf{1})\nu^{\star\top} = \tilde\nu^\top - \nu^{\star\top},$$

the stated identity, exact in E.

Norm bound. Since $\tilde\nu^\top E = \sum_i \tilde\nu_i E_{i:}$ is a convex combination of rows of E,

$$\|\tilde\nu^\top E\|_1 = \Big\|\sum_i \tilde\nu_i\, E_{i:}\Big\|_1 \le \sum_i \tilde\nu_i\, \|E_{i:}\|_1 \le \max_i \|E_{i:}\|_1 = \|E\|_\infty.$$

For any row vector $w^\top$,

$$\|w^\top H\|_1 = \sum_j \Big|\sum_i w_i H_{ij}\Big| \le \sum_i |w_i| \sum_j |H_{ij}| \le \Big(\max_i \sum_j |H_{ij}|\Big) \sum_i |w_i| = \|H\|_\infty\, \|w\|_1.$$

Chaining with $w^\top = \tilde\nu^\top E$,

$$\|\tilde\nu - \nu^\star\|_1 = \|\tilde\nu^\top E\, H\|_1 \le \|H\|_\infty\, \|\tilde\nu^\top E\|_1 \le \|H\|_\infty\, \|E\|_\infty = \kappa(P^\star)\, \|E\|_\infty.$$ □

The identity (iii) is exact and $\kappa(P^\star) = \|H\|_\infty$ is a standard Markov-chain condition number (Meyer, 1975).

9.3 THE TWO-REGIME THEOREM

We now prove that the attention width β controls the collective accuracy of the baseline allocator. The object analyzed is the stationary distribution $\nu_{\boldsymbol{C}(\beta)}$ of the fixed-graph realized influence $\boldsymbol{C}(\beta)$, which is the infinite-horizon limit of the collective weight q_T under uniform anchoring (see Remark 9.4); accordingly we measure accuracy by

$$\mathrm{Var}_\infty(\beta) \coloneqq s^2 \left\|\nu_{\boldsymbol{C}(\beta)}\right\|_2^2, \qquad N_{\mathrm{eff}}(\beta) \coloneqq \left\|\nu_{\boldsymbol{C}(\beta)}\right\|_2^{-2}.$$

Remark 9.4 (Relation to the operational weight). For uniform anchoring $\lambda_i \equiv \lambda$ and a fixed irreducible $\boldsymbol{C}$, Lemma 8.5 gives the infinite-horizon weight

$$q_\infty(\lambda)^\top = \frac{1-\lambda}{n}\,\mathbf{1}^\top(\boldsymbol{I}-\lambda\boldsymbol{C})^{-1} = (1-\lambda)\sum_{k\geq 0}\lambda^k\,\tfrac{1}{n}\mathbf{1}^\top\boldsymbol{C}^k,$$

and $q_\infty(\lambda) \to \nu_{\boldsymbol{C}}$ as $\lambda \uparrow 1$: the Cesàro averages $\frac{1}{N}\sum_{k<N}\frac{1}{n}\mathbf{1}^\top\boldsymbol{C}^k$ converge to $\nu_{\boldsymbol{C}}^\top$ for every irreducible $\boldsymbol{C}$ (Levin & Peres, 2026, Ch. 4), and Abel summation agrees with the Cesàro limit. The two-regime theorem is therefore the unanchored long-horizon limit of the *baseline* allocator; the guarantee for the equalized allocator (Theorem 4.3) is exact at every horizon and anchoring level and does not rely on this remark.

Theorem 4.2 (Attention width controls collective accuracy)**.** *Let $\boldsymbol{C}(\beta)$ be the fixed-graph realized influence with $\zeta \in [0,1)$ and irreducible exposure support B, where $B_{ij} = \mathbf{1}\{\bar{W}_{ij} > 0\}$. Write the degrees $d_i = \sum_j B_{ij}$, $D = \mathrm{diag}(d_1,\dots,d_n)$, $d_{\max} = \max_i d_i$, and let L be the log-spread and $\underline{g}$ the dominance gap of the scores s_{ij} (Appendix 8). In the homoskedastic case:*

1. (Wide, wisdom of crowds.) *If $B = B^\top$ is irreducible, then $\boldsymbol{C}(\beta) \to D^{-1}B$ and $\nu_{\boldsymbol{C}(\beta)} \to d/(\mathbf{1}^\top d)$ as $\beta \to \infty$, so:*

$$\lim_{\beta\to\infty} N_{\mathrm{eff}}(\beta) = \frac{(\sum_i d_i)^2}{\sum_i d_i^2},$$

which equals n if and only if B is degree-regular; for finite β the consensus weight obeys $\left|\|\nu_{\boldsymbol{C}(\beta)}\|_2^2 - \|\nu^\star\|_2^2\right| \leq 2\kappa_B(e^{L/\beta}-1)$ with $\nu^\star = d/(\mathbf{1}^\top d)$ and κ_B a condition number of $D^{-1}B$ (Appendix 8), so: $N_{\mathrm{eff}}(\beta) \to (\sum_i d_i)^2/\sum_i d_i^2$ by continuity.

2. (Narrow, herding.) *Assume B is irreducible and the scores admit a unique dominant pair: $j_i^\star = \arg\max_k s_{ik}$ is unique for every row, with $j_i^\star = h$ for all $i \neq h$ and $j_h^\star = g \neq h$. Then with $\rho(\beta) = (d_{\max}-1)\,e^{-\underline{g}/\beta}$, whenever $\rho(\beta) < 1$,*

$$N_{\mathrm{eff}}(\beta) \leq \frac{2}{(1-\rho(\beta))^2},$$

so $N_{\mathrm{eff}}(\beta) \leq 8$ whenever $\rho(\beta) \leq \frac{1}{2}$, i.e. $\beta \leq \underline{g}/\log(2(d_{\max}-1))$, independently of n.

Proof. **Wide context.** *Matrix bound equation 17.* Row i of $\boldsymbol{C}(\beta)$ is $\Gamma_\beta(s_{i:})$ and row i of $D^{-1}B$ is $\frac{1}{d_i}\mathbf{1}_{N(s_{i:})}$, on the same support. By Lemma 9.1(i) their ℓ_1 distance is $\leq e^{L_i/\beta}-1 \leq e^{L/\beta}-1$; maximizing over rows,

$$\left\|\boldsymbol{C}(\beta) - D^{-1}B\right\|_\infty \;\leq\; e^{L/\beta}-1. \tag{17}$$

Reference stationary distribution. With $B = B^\top$,

$$\left(d^\top D^{-1}B\right)_j = \sum_i d_i \cdot \frac{B_{ij}}{d_i} = \sum_i B_{ij} = \sum_i B_{ji} = d_j,$$

so $\nu^\star = d/(\mathbf{1}^\top d)$ is stationary for $D^{-1}B$, unique by Lemma 9.2(ii).

Stationary perturbation. Lemma 9.3(iii) with $P^\star = D^{-1}B$, $E = \boldsymbol{C}(\beta) - D^{-1}B$ yields

$$\left\|\nu_{\boldsymbol{C}(\beta)} - \nu^\star\right\|_1 \leq \kappa_B\,\|E\|_\infty \leq \kappa_B\left(e^{L/\beta}-1\right), \tag{18}$$

using equation 17.

Variance transfer. With $\Delta = \nu_{\boldsymbol{C}(\beta)} - \nu^\star$, since all entries lie in $[0,1]$,

$$\big|\|\nu_{\boldsymbol{C}(\beta)}\|_2^2 - \|\nu^\star\|_2^2\big| = \Big|\sum_j \Delta_j\big(\nu_{\boldsymbol{C}(\beta),j} + \nu_j^\star\big)\Big| \le \max_j\big(\nu_{\boldsymbol{C}(\beta),j} + \nu_j^\star\big)\,\|\Delta\|_1 \le 2\|\Delta\|_1,$$

since each summand of the max is at most 1. As $\|\nu^\star\|_2^2 = \sum_k d_k^2/(\sum_k d_k)^2$, multiplying by s^2 gives the matching variance bound stated in the theorem.

Floor and degree-regularity. By Cauchy–Schwarz on $d, \mathbf{1}$,

$$\Big(\sum_k d_k\Big)^2 = \langle d, \mathbf{1}\rangle^2 \le \|d\|_2^2\,\|\mathbf{1}\|_2^2 = n\sum_k d_k^2,$$

so $\sum_k d_k^2/(\sum_k d_k)^2 \ge 1/n$, with equality iff $d \propto \mathbf{1}$ (degree-regular); then equation 18 gives $N_{\text{eff}}(\beta) \to n$ as $\beta \to \infty$.

Narrow context. Assume B irreducible, so $\boldsymbol{C}(\beta)$ is irreducible for every $\beta > 0$ and $\nu := \nu_{\boldsymbol{C}(\beta)} > 0$ is unique (Lemma 9.2). Assume the unique dominant pair: $j_i^\star = \arg\max_k s_{ik}$ is unique for each row, $j_i^\star = h$ for $i \ne h$, $j_h^\star = g \ne h$; so $j_i^\star \in S := \{h, g\}$ for all i. Write $\rho = \rho(\beta)$ and assume $\rho < 1$. By Lemma 9.1(ii),

$$\sum_{k \ne j_i^\star} \boldsymbol{C}_{ik}(\beta) = 1 - \boldsymbol{C}_{i,j_i^\star}(\beta) \le (d_i - 1)\,e^{-g_i/\beta} \le (d_{\max} - 1)\,e^{-\underline{g}/\beta} = \rho. \tag{19}$$

Mass outside S. For $k \in S^c$, $\boldsymbol{C}_{ik}(\beta)$ is off-maximizer, so $\sum_{k\in S^c} \boldsymbol{C}_{ik}(\beta) \le \rho$ by equation 19. By stationarity,

$$\nu(S^c) = \sum_{k\in S^c} \nu_k = \sum_{k\in S^c}\sum_i \nu_i\,\boldsymbol{C}_{ik}(\beta) = \sum_i \nu_i \sum_{k\in S^c} \boldsymbol{C}_{ik}(\beta) \le \rho \sum_i \nu_i = \rho,$$

$$\nu(S^c) \le \rho(\beta), \qquad N_{\text{eff}}(\beta) \le \frac{2}{(1-\rho(\beta))^2}, \tag{20}$$

the first claim of equation 20.

Sandwich for ν_h. By equation 19, $\boldsymbol{C}_{hh}(\beta) \le \rho$ (off-maximizer) and $\boldsymbol{C}_{ih}(\beta) \ge 1 - \rho$ for $i \ne h$. Lower bound:

$$\nu_h \ge \sum_{i\ne h} \nu_i\,(1-\rho) = (1-\rho)(1-\nu_h) \quad\Longrightarrow\quad \nu_h(1 + (1-\rho)) \ge 1 - \rho \quad\Longrightarrow\quad \nu_h \ge \frac{1-\rho}{2-\rho}.$$

Upper bound:

$$\nu_h = \nu_h\,\boldsymbol{C}_{hh}(\beta) + \sum_{i\ne h} \nu_i\,\boldsymbol{C}_{ih}(\beta) \le \rho\,\nu_h + (1-\nu_h) \quad\Longrightarrow\quad \nu_h(2-\rho) \le 1 \quad\Longrightarrow\quad \nu_h \le \frac{1}{2-\rho}.$$

Bounds for ν_g. Since $\boldsymbol{C}_{hg}(\beta) \ge 1 - \rho$ and $\boldsymbol{C}_{ig}(\beta) \le \rho$ for $i \ne h$,

$$\nu_h(1-\rho) \le \nu_g \le \nu_h + \rho(1-\nu_h) \le \nu_h + \rho. \tag{21}$$

Upper bound on N_{eff}. Keeping the two S coordinates and using $(a+b)^2 \le 2(a^2+b^2)$,

$$\|\nu\|_2^2 \ge \nu_h^2 + \nu_g^2 \ge \tfrac{1}{2}(\nu_h + \nu_g)^2 = \tfrac{1}{2}\big(1 - \nu(S^c)\big)^2 \ge \tfrac{1}{2}(1-\rho)^2,$$

using $\nu_h + \nu_g = 1 - \nu(S^c)$. Thus $N_{\text{eff}}(\beta) \le 2(1-\rho)^{-2}$, the second claim of equation 20. At $\beta \le \underline{g}/\log(2(d_{\max}-1))$, $\rho \le \frac{1}{2}$ and $N_{\text{eff}}(\beta) \le 8$.

Two-sided bound equation 22. Suppose $\rho \le \frac{1}{4}$. We bound $\|\nu\|_2^2$ from above. For ν_h, using $\frac{1}{1-x} \le 1 + \frac{x}{1-x}$ with $x = \rho/2$,

$$\nu_h \le \frac{1}{2-\rho} = \frac{1}{2}\cdot\frac{1}{1-\rho/2} \le \frac{1}{2}\Big(1 + \frac{\rho/2}{1-\rho/2}\Big) \le \frac{1}{2}\Big(1 + \tfrac{4}{7}\rho\Big) \le \frac{1}{2} + \frac{1}{3}\rho,$$

where $\frac{\rho/2}{1-\rho/2} \le \frac{\rho/2}{7/8} = \frac{4}{7}\rho$ at $\rho \le \frac{1}{4}$, and $\frac{1}{2}\cdot\frac{4}{7} = \frac{2}{7} \le \frac{1}{3}$. By equation 21, $\nu_g \le \nu_h + \rho \le \frac{1}{2} + \frac{4}{3}\rho$. The mass on S^c contributes at most $\sum_{k\in S^c}\nu_k^2 \le (\sum_{k\in S^c}\nu_k)^2 \le \rho^2$. Summing,

$$\|\nu\|_2^2 \le \left(\tfrac{1}{2}+\tfrac{1}{3}\rho\right)^2 + \left(\tfrac{1}{2}+\tfrac{4}{3}\rho\right)^2 + \rho^2 = \tfrac{1}{2} + \tfrac{\rho}{3} + \tfrac{4\rho}{3} + \tfrac{\rho^2}{9} + \tfrac{16\rho^2}{9} + \rho^2 = \tfrac{1}{2} + \tfrac{5}{3}\rho + \tfrac{26}{9}\rho^2.$$

At $\rho \le \frac{1}{4}$, $\frac{26}{9}\rho^2 \le \frac{26}{9}\cdot\frac{1}{4}\rho = \frac{26}{36}\rho \le \frac{5}{6}\rho$, so $\|\nu\|_2^2 \le \frac{1}{2} + \frac{5}{3}\rho + \frac{5}{6}\rho = \frac{1}{2} + \frac{5}{2}\rho$. Therefore

$$N_{\text{eff}}(\beta) = \frac{1}{\|\nu\|_2^2} \ge \frac{1}{\frac{1}{2}+\frac{5}{2}\rho} = \frac{2}{1+5\rho} \ge 2(1-5\rho) = 2 - 10\rho,$$

using $\frac{1}{1+u} \ge 1-u$. For the upper bound, $(1-\rho)^{-2} \le 1+4\rho$ on $[0, \frac{1}{4}]$ gives $N_{\text{eff}}(\beta) \le 2(1+4\rho) = 2+8\rho$. Combining,

$$\begin{aligned} \left|N_{\text{eff}}(\beta) - 2\right| &\le \max\{8\rho, 10\rho\} = 10\rho \le 16\rho, \\ \left|N_{\text{eff}}(\beta) - 2\right| &\le 16\,\rho(\beta) \qquad (\rho(\beta) \le \tfrac{1}{4}), \end{aligned} \tag{22}$$

which is equation 22.

Limit. As $\beta \to 0$, $\rho \to 0$, so equation 22 gives $N_{\text{eff}}(\beta) \to 2$ and $\nu \to \frac{1}{2}(e_h + e_g)$ by equation 21. □

Both regimes are *column* pathologies of $\boldsymbol{C}$: the stationary mass is the degree distribution $d/(\mathbf{1}^\top d)$ in the wide limit and concentrates on $\{h, g\}$ in the narrow limit. Row-stochasticity constrains only the reader side; the equalized allocator removes the remaining column freedom.

10 Proof of Theorem 4.3 (Equalization)

This appendix proves Theorem 4.3 and develops the exposure-pricing protocol: the exact-optimality guarantee (Appendix 10.1), then the Sinkhorn equivalence and the targeted heteroskedastic variant (Appendix 11).

10.1 Exact optimality under equalization

The equalized allocator drives each $\boldsymbol{C}(t)$ toward double stochasticity. The guarantee below is stated for an *arbitrary* row-stochastic sequence in terms of its measured per-round column defects, so it covers the endogenous, feedback-coupled operational sequence with no exogeneity, stationarity, irreducibility, aperiodicity, or mixing hypothesis. Recall the column defect of a row-stochastic $\boldsymbol{C}$,

$$\delta_{\text{col}}(\boldsymbol{C}) := \frac{1}{n}\left\|\mathbf{1}^\top\boldsymbol{C} - \mathbf{1}^\top\right\|_1 = \frac{1}{n}\sum_{j=1}^{n}\left|\sum_{i=1}^{n}\boldsymbol{C}_{ij} - 1\right|, \tag{23}$$

which is zero exactly when $\boldsymbol{C}$ has unit column sums, i.e. is doubly stochastic.

Theorem 4.3 (Exact optimality under equalization)**.** *Under uniform anchoring $\lambda \in [0, 1)$ and signal exogeneity, let $\bar{\delta}_{\text{col}}$ be the largest column defect over the run. Then the proxy weight satisfies:*

$$\left\|q_T - \tfrac{1}{n}\mathbf{1}\right\|_1 \le \bar{\delta}_{\text{col}}\left(\tfrac{\lambda}{1-\lambda} + T\lambda^T\right),$$

for every horizon T and every realized sequence. In particular, if every $\boldsymbol{C}(t)$ is doubly stochastic ($\bar{\delta}_{\text{col}} = 0$), then $q_T = \frac{1}{n}\mathbf{1}$ exactly, so $\mathrm{Var}(\hat{\theta}_T^{\text{BR}}) = s^2/n$ and $N_{\text{eff}} = n$. The operational estimate inherits this up to the bridge gap,

$$\left|\mathrm{sd}(\hat{\theta}_T) - \tfrac{s}{\sqrt{n}}\right| \le \frac{\delta_T}{1-\lambda} + s\,\bar{\delta}_{\text{col}}\left(\tfrac{\lambda}{1-\lambda} + T\lambda^T\right).$$

Proof. At $\lambda = 0$, $M(t) = 0$ and Lemma 8.5 gives $q_T = \frac{1}{n}\mathbf{1}$, so both sides of equation 26 vanish. Assume $\lambda \in (0, 1)$ and write $u := \frac{1}{n}\mathbf{1}$.

Step 1 (normalized backward recursion). With $P = \lambda\boldsymbol{I}$, $\Phi(T, r) = \lambda^{T-r}\,\boldsymbol{C}(T-1)\cdots\boldsymbol{C}(r)$. Define

$$p_r^\top := \lambda^{-(T-r)}\,\tfrac{1}{n}\mathbf{1}^\top\Phi(T, r) \qquad (0 \le r \le T),$$

so $p_T = u$ and, peeling off $\boldsymbol{C}(r)$,

$$p_r^\top = \lambda^{-(T-r)} \tfrac{1}{n} \mathbf{1}^\top \Phi(T, r+1)\, \lambda \boldsymbol{C}(r) = \Big[\lambda^{-(T-r-1)} \tfrac{1}{n} \mathbf{1}^\top \Phi(T, r+1)\Big] \boldsymbol{C}(r) = p_{r+1}^\top \boldsymbol{C}(r).$$

Let $\varrho_r^\top := \frac{1}{n}(\mathbf{1}^\top \boldsymbol{C}(r) - \mathbf{1}^\top)$, so $\|\varrho_r\|_1 = \delta_r$ by equation 23. Then

$p_r^\top - u^\top = p_{r+1}^\top \boldsymbol{C}(r) - u^\top = (p_{r+1} - u)^\top \boldsymbol{C}(r) + \big(u^\top \boldsymbol{C}(r) - u^\top\big) = (p_{r+1} - u)^\top \boldsymbol{C}(r) + \varrho_r^\top$,

using $u^\top \boldsymbol{C}(r) - u^\top = \varrho_r^\top$. For any row-stochastic $\boldsymbol{C}$, the ℓ_1 norm contracts:

$$\|v^\top \boldsymbol{C}\|_1 = \sum_j \Big|\sum_i v_i \boldsymbol{C}_{ij}\Big| \le \sum_i |v_i| \sum_j \boldsymbol{C}_{ij} = \sum_i |v_i| = \|v\|_1, \tag{24}$$

Applying equation 24 and the triangle inequality,

$$\|p_r - u\|_1 \le \|(p_{r+1} - u)^\top \boldsymbol{C}(r)\|_1 + \|\varrho_r\|_1 \le \|p_{r+1} - u\|_1 + \delta_r.$$

Backward induction from $p_T = u$ (so $\|p_T - u\|_1 = 0$) yields

$$\|p_r - u\|_1 \le \sum_{v=r}^{T-1} \delta_v \le (T - r)\, \bar{\delta}_{\mathrm{col}}. \tag{25}$$

Step 2 (convex combination). By Lemma 8.5 with $\boldsymbol{I} - P = (1-\lambda)\boldsymbol{I}$,

$$q_T^\top = \tfrac{1}{n} \mathbf{1}^\top \Phi(T, 0) + (1-\lambda) \sum_{r=0}^{T-1} \tfrac{1}{n} \mathbf{1}^\top \Phi(T, r+1).$$

Substituting $\frac{1}{n} \mathbf{1}^\top \Phi(T, r) = \lambda^{T-r} p_r^\top$,

$$q_T^\top = \lambda^T p_0^\top + (1-\lambda) \sum_{r=0}^{T-1} \lambda^{T-r-1} p_{r+1}^\top.$$

The scalar coefficients are nonnegative and sum to one:

$$\lambda^T + (1-\lambda) \sum_{r=0}^{T-1} \lambda^{T-r-1} = \lambda^T + (1-\lambda) \cdot \frac{1 - \lambda^T}{1 - \lambda} = \lambda^T + (1 - \lambda^T) = 1,$$

Since the coefficients sum to one, subtracting $u^\top$ termwise:

$$q_T^\top - u^\top = \lambda^T (p_0 - u)^\top + (1-\lambda) \sum_{r=0}^{T-1} \lambda^{T-r-1} (p_{r+1} - u)^\top.$$

Taking ℓ_1 norms and applying equation 25,

$$\|q_T - u\|_1 \le \lambda^T\, T \bar{\delta}_{\mathrm{col}} + (1-\lambda) \bar{\delta}_{\mathrm{col}} \sum_{r=0}^{T-1} \lambda^{T-r-1}\, (T - r - 1).$$

Substituting $j = T - r - 1$, $\sum_{j=0}^{T-1} j\, \lambda^j \le \frac{\lambda}{(1-\lambda)^2}$, so

$$\|q_T - u\|_1 \le \bar{\delta}_{\mathrm{col}} \Big(T\lambda^T + (1-\lambda) \cdot \frac{\lambda}{(1-\lambda)^2} \Big) = \bar{\delta}_{\mathrm{col}} \Big(\frac{\lambda}{1-\lambda} + T\lambda^T \Big), \tag{26}$$

which is the main bound of Theorem 4.3.

(i) If $\bar{\delta}_{\mathrm{col}} = 0$, equation 26 forces $q_T = u = \frac{1}{n}\mathbf{1}$ exactly. Then Proposition 8.6 gives $\mathrm{Var}(\hat{\theta}_T^{\mathrm{BR}}) = s^2 \|u\|_2^2 = s^2 \cdot \frac{1}{n} = s^2/n$ and $N_{\mathrm{eff}}(q_T) = \|u\|_2^{-2} = n$.

(ii) By Corollary 8.8 ($\bar{\delta} = \delta/(1-\lambda)$), $\big|\mathrm{sd}(\hat{\theta}_T) - s\|q_T\|_2\big| \le \frac{\delta}{1-\lambda}$. By $\|\cdot\|_2 \le \|\cdot\|_1$ and equation 26,

$$\big|s\|q_T\|_2 - s\|u\|_2\big| \le s\, \|q_T - u\|_2 \le s\, \|q_T - u\|_1 \le s\, \bar{\delta}_{\mathrm{col}} \Big(\frac{\lambda}{1-\lambda} + T\lambda^T \Big),$$

and $\|u\|_2 = 1/\sqrt{n}$. Combining,

$$\Big|\mathrm{sd}(\hat{\theta}_T) - \frac{s}{\sqrt{n}}\Big| \le \big|\mathrm{sd}(\hat{\theta}_T) - s\|q_T\|_2\big| + \big|s\|q_T\|_2 - s\|u\|_2\big| \le \frac{\delta}{1-\lambda} + s\, \bar{\delta}_{\mathrm{col}} \Big(\frac{\lambda}{1-\lambda} + T\lambda^T \Big). \quad \square$$

Remark 10.1 (Quantified herding impossibility). Write $\Delta = \bar{\delta}_{\mathrm{col}}(\frac{\lambda}{1-\lambda} + T\lambda^T)$. By equation 26 and $\|\cdot\|_2 \le \|\cdot\|_1$, $\|q_T\|_2 \le \|u\|_2 + \|q_T - u\|_2 \le \frac{1}{\sqrt{n}} + \Delta$, so

$$N_{\mathrm{eff}}(q_T) = \frac{1}{\|q_T\|_2^2} \ge \frac{n}{(1+\sqrt{n}\,\Delta)^2}.$$

For example $\Delta \le 1/\sqrt{n}$ already forces $N_{\mathrm{eff}} \ge n/4$. Contrast Theorem 4.2: in the narrow regime the baseline collapses to $N_{\mathrm{eff}} \le 8$ regardless of n, whereas the equalized allocator cannot collapse at any attention width.

11 The Pricing Protocol

This appendix derives two facts about the pricing protocol: one price round equals one Sinkhorn iteration on the tempered scores (so it converges to a doubly stochastic limit, invariant to social power), and retargeting the clearing marginal extends exact optimality to the heteroskedastic optimum. Throughout, $K_{ij} = s_{ij}^{1/\beta} = (\bar{W}_{ij}\pi_j)^{1/\beta}$, $\mathrm{rn}(M) = \mathrm{diag}(M\mathbf{1})^{-1}M$, $\mathrm{cn}(M) = M\,\mathrm{diag}(\mathbf{1}^\top M)^{-1}$.

11.1 One price round is one Sinkhorn iteration

Recall the two local rules. Each source j holds a price $y_j > 0$; the reader rule sets $\boldsymbol{C}_{ij} = (K_{ij}/y_j)/\sum_k (K_{ik}/y_k)$, i.e. $\boldsymbol{C} = \mathrm{rn}(K\mathrm{diag}(y)^{-1})$, and the source rule updates $y_j \leftarrow y_j\, a_j$ with $a_j = \sum_i \boldsymbol{C}_{ij}$.

Lemma 11.1 (Sinkhorn equivalence). *Let $\boldsymbol{C} = \mathrm{rn}(K\mathrm{diag}(y)^{-1})$ be the current realized influence and $a_j = \sum_i \boldsymbol{C}_{ij}$ its column sums. After the source update $y_j \leftarrow y_j a_j$, the next realized influence is*

$$\boldsymbol{C}' = \mathrm{rn}\big(K\,\mathrm{diag}(y \odot a)^{-1}\big) = \mathrm{rn}\big(\mathrm{cn}(\boldsymbol{C})\big).$$

That is, one price round applies a column normalization followed by a row normalization to $\boldsymbol{C}$, which is one Sinkhorn iteration on K.

Proof. The reader rule at $y' = y \odot a$ gives $\boldsymbol{C}' = \mathrm{rn}(K\,\mathrm{diag}(y \odot a)^{-1})$, the first equality. With $\mathrm{diag}(y \odot a)^{-1} = \mathrm{diag}(a)^{-1}\mathrm{diag}(y)^{-1}$,

$$K\,\mathrm{diag}(y \odot a)^{-1} = \big(K\,\mathrm{diag}(y)^{-1}\big)\,\mathrm{diag}(a)^{-1}.$$

Since $\mathrm{rn}(D_{\mathrm{row}}M) = \mathrm{rn}(M)$ and $K\mathrm{diag}(y)^{-1}$ is a positive row scaling of $\boldsymbol{C}$,

$$\boldsymbol{C}' = \mathrm{rn}\big(K\mathrm{diag}(y)^{-1}\mathrm{diag}(a)^{-1}\big) = \mathrm{rn}\big(\boldsymbol{C}\,\mathrm{diag}(a)^{-1}\big).$$

Finally $\boldsymbol{C}\,\mathrm{diag}(a)^{-1} = \mathrm{cn}(\boldsymbol{C})$ since $a^\top = \mathbf{1}^\top\boldsymbol{C}$, so $\boldsymbol{C}' = \mathrm{rn}(\mathrm{cn}(\boldsymbol{C}))$. □

Lemma 11.2 (Convergence of inner iterations). *Fix a round and hold $K = K(t)$ fixed during the price updates. If K has total support, then iterating the price round of Lemma 11.1 produces $\boldsymbol{C}^{(\ell)} \to \boldsymbol{C}^\star$, the unique doubly stochastic matrix with $\mathrm{supp}\,\boldsymbol{C}^\star = \mathrm{supp}\,K$, and $\delta_{\mathrm{col}}(\boldsymbol{C}^{(\ell)}) \to 0$ geometrically. Total support holds whenever $K_{ii} > 0$ for all i and $\mathrm{supp}\,K$ is symmetric.*

Proof. Each price round is one Sinkhorn iteration $\mathrm{rn} \circ \mathrm{cn}$ on K (Lemma 11.1). For K with total support, the alternating row/column normalization converges to the unique doubly stochastic $\mathrm{diag}(u)\,K\,\mathrm{diag}(v)$, with geometric rate in the Hilbert projective metric (Franklin & Lorenz, 1989); convergence of the iterates gives $\delta_{\mathrm{col}}(\boldsymbol{C}^{(\ell)}) \to 0$. If $K_{ii} > 0$ for all i, every diagonal entry is positive, so each index lies on a positive diagonal (a trivial support permutation); symmetry of $\mathrm{supp}\,K$ then gives total support. □

Remark 11.3 (Diagonal invariance of the cleared limit). One price round absorbs a right diagonal scaling of the tempered scores, $\mathrm{cn}(KD_2) = \mathrm{cn}(K)$, but not a left scaling, since the column sums it normalizes change under $K \mapsto D_1K$. Both are absorbed only in the fully balanced Sinkhorn limit $\boldsymbol{C}^\star$, the doubly stochastic projection of K (Remark 11.6), which is invariant to $K \mapsto D_1KD_2$. Two consequences. The social-power factor π enters as a right scaling, $K = (\bar{\boldsymbol{W}} \odot \mathbf{1}\pi^\top)^{\circ 1/\beta}$, so it is removed after one round: the cleared allocator caps each source's share regardless of its power, attenuating the high-power columns of Theorem 4.2. The row normalization defining $\bar{\boldsymbol{W}}$ is a left scaling and is removed only in the limit; the one-round iterate retains it until $\delta_{\mathrm{col}} \to 0$, which Theorem 4.3 tracks through the measured defect.

Remark 11.4 (One Sinkhorn step does not clear exactly). A single price round (Lemma 11.1) reduces the column defect but need not make $\boldsymbol{C}(t)$ doubly stochastic. Run to convergence each round with $K(t)$ held fixed, under the total-support condition of Lemma 11.2 (e.g. self-exposure $A_{ii}(t) > 0$ and symmetric support), $\delta_{\mathrm{col}}(\boldsymbol{C}(t)) \to 0$ and Theorem 4.3(i) is exact. Run once per round as in Algorithm 1, the realized matrices need not clear, and Theorem 4.3 applies in its general form through the measured defects δ_r; the deviation of q_T from uniform is then controlled by the $\bar{\delta}_{\mathrm{col}}$ observed in the run, not asserted a priori. Exact optimality is a property of the cleared limit; the one-round protocol attains it up to $\bar{\delta}_{\mathrm{col}}$.

11.2 Targeted clearing and heteroskedastic optimality

Uniform clearing drives q_T to $\frac{1}{n}\mathbf{1}$, optimal when signal precisions are equal. When precisions differ, the best linear unbiased weight is $q_i \propto s_i^{-2}$, and the protocol attains it by clearing to a non-uniform target μ at no change to the two rules.

Corollary 11.5 (Targeted clearing; heteroskedastic optimality)**.** *Fix a target $\mu \in \Delta_n$ with $\mu_j > 0$ for all j. Replace the source rule's clearing target by μ, i.e. $a_j \leftarrow \frac{1}{\mu_j}\sum_i \mu_i \boldsymbol{C}_{ij}$ with fixed point $\mu^\top \boldsymbol{C} = \mu^\top$, and the readout by $\hat{\theta}_T^{\mu} = \mu^\top b(T)$. Define the weighted defects $\delta_r^{\mu} := \|\mu^\top \boldsymbol{C}(r) - \mu^\top\|_1$ and $\bar{\delta}^{\mu} := \max_{0\le r<T} \delta_r^{\mu}$. Then the induced weight $q_T(\mu)^\top := \mu^\top G_T$ satisfies $q_T(\mu)^\top \mathbf{1} = 1$ and*

$$\big\|q_T(\mu) - \mu\big\|_1 \le \bar{\delta}^{\mu}\Big(\frac{\lambda}{1-\lambda} + T\lambda^T\Big);$$

at exact clearing, $q_T(\mu) = \mu$ for every $T \ge 0$ and every $\lambda \in [0,1)$. Consequently, under Assumption 7.3 with heteroskedastic variances $s_1^2, \dots, s_n^2$, the choice $\mu_i \propto s_i^{-2}$ attains

$$\mathrm{Var}\big(\mu^\top b^{\mathrm{BR}}(T)\big) = \mu^\top S\,\mu = \Big(\sum_{i=1}^{n} s_i^{-2}\Big)^{-1},$$

the best-linear-unbiased optimum, exactly. Uniform clearing $\mu = \frac{1}{n}\mathbf{1}$ (Theorem 4.3) is the special case for unknown or exchangeable precisions.

Proof. Mass and variance. By Lemma 8.5, $q_T(\mu)^\top \mathbf{1} = \mu^\top G_T \mathbf{1} = 1$, and Proposition 8.6 gives $\mathrm{Var}(\mu^\top b^{\mathrm{BR}}(T)) = q_T(\mu)^\top S\, q_T(\mu)$.

Bound. The case $\lambda = 0$ is immediate as in Theorem 4.3. For $\lambda \in (0,1)$ set $p_r^\top := \lambda^{-(T-r)}\mu^\top \Phi(T,r)$, so $p_T = \mu$ and $p_r^\top = p_{r+1}^\top \boldsymbol{C}(r)$ as before. Steps 1 and 2 of the proof of Theorem 4.3 apply with u replaced by μ throughout: they use only the ℓ_1 contraction equation 24 and the per-round residual $\mu^\top \boldsymbol{C}(r) - \mu^\top$, whose ℓ_1 norm is δ_r^{μ} by definition. This yields the stated bound, and at exact clearing ($\bar{\delta}^{\mu} = 0$) it forces $q_T(\mu) = \mu$.

Optimality. At exact clearing, with $\mu_i = s_i^{-2}/\sum_k s_k^{-2}$,

$$\sum_i \mu_i^2 s_i^2 = \frac{\sum_i s_i^{-4} s_i^2}{(\sum_k s_k^{-2})^2} = \frac{\sum_i s_i^{-2}}{(\sum_k s_k^{-2})^2} = \Big(\sum_i s_i^{-2}\Big)^{-1}.$$

This is the minimum of $q^\top S q$ over $q^\top \mathbf{1} = 1$: by Cauchy–Schwarz on $(q_i s_i)_i, (s_i^{-1})_i$,

$$1 = \Big(\sum_i q_i\Big)^2 = \Big(\sum_i (q_i s_i)\, s_i^{-1}\Big)^2 \le \Big(\sum_i q_i^2 s_i^2\Big)\Big(\sum_i s_i^{-2}\Big),$$

so $q^\top S q \ge (\sum_i s_i^{-2})^{-1}$, equality iff $q_i \propto s_i^{-2}$. □

Remark 11.6 (Why pricing rather than quotas). The fully cleared limit of the protocol is an information projection: among all matrices with the given support and unit row and column sums, it is the unique minimizer of $\mathrm{KL}(\boldsymbol{C} \,\|\, K)$ with $K = (s_{ij}^{1/\beta})$ (Csiszár, 1975). Equalization thus deviates from the relevance-driven allocation as little as possible subject to clearing. Hard per-source quotas or random subsampling also clear the market, but they are cruder projections that discard the score information; pricing is the minimal-distortion mechanism in the entropic geometry.

12 EXPERIMENTAL DETAILS

This appendix documents the full experimental configuration and reports the complete result set underlying the three figures in the main text. All quantities are read directly from the committed run artifacts; where a value is a design choice we give the governing source file.

12.1 MODELS AND INFERENCE

All agents are instruction-tuned LLMs served locally with vLLM 0.23.0. We use the **Qwen2.5-Instruct** family at 0.5/1.5/3/7/14B and, for the cross-family generality check (App. 13.7), **Llama-3.1-8B-Instruct** and **Llama-3.2-1B/3B-Instruct**. The heterogeneous "society" pairs a *strong* `Qwen2.5-7B-Instruct` with a *weak* `Qwen2.5-1.5B-Instruct`; the controlled testbed (App. 13.1) uses a single `Qwen2.5-3B-Instruct` by default, swept over the scale ladder for App. 13.2.

Decoding. Greedy (temperature=0, top-p=1.0, no repetition penalty, no per-request sampling seed), so per-prompt generation is deterministic; the two-regime is not an artifact of greedy decoding (Fig. 5c contrasts greedy vs. τ=0.7). Maximum new tokens are 320 for the discussion benchmarks and 96 for the estimation testbed; the context window is 4096 tokens by default and raised to 16384/32768 for the large-n discussion runs whose transcripts overflow 4096. The two cooperative-boundary environments that require a norm-setting free-text turn (GovSim) or match the source protocol (Debate) use temperature=0.7.

Operator arms. Each benchmark is run under three realizations of the influence operator C: the *baseline* tempered allocator Γ_β, the *equalized* allocator Γ^{EQ}, and (for the herding environments) an *adversary-off* causal control. The allocator definitions and their parameters are given next.

12.2 THE INFLUENCE OPERATOR AND ITS PARAMETERS

- **Context-width tempering Γ_β (baseline).** Row-wise, $[\Gamma_\beta(s)]_j = s_j^{1/\beta} / \sum_k s_k^{1/\beta}$: $\beta \to 0$ is a hard argmax (herd onto the single most-exposed source), β=1 is proportional, $\beta \to \infty$ is uniform. Realized influence is $C(\beta) = \Gamma_\beta(\bar{W} \odot \pi)$.
- **Equalized allocator Γ^{EQ}.** k Sinkhorn iterations (column-normalize under the stationary weights μ, then row-normalize) that drive the exposure matrix toward doubly stochastic, removing the exposure-column concentration a dominant source exploits. Default k=4; the k-robustness sweep is in App. 13.2.
- **Social power ζ (herding gate).** A PageRank-style centrality mixing $\pi^\top = \frac{1-\zeta}{n}\mathbf{1}^\top(I - \zeta\bar{W})^{-1}$, $\zeta \in [0, 1)$: ζ=0 gives uniform influence $1/n$; $\zeta \uparrow 1$ concentrates influence on central sources. Default ζ=0.6 in the discussion benchmarks and ζ=0.7 in the testbed.
- **Follower self-reliance λ.** A Friedkin–Johnsen anchoring P=diag(λ), $\lambda_i \in [0, 1]$: λ=1 is DeGroot updating, λ<1 anchors an agent to its own prior. The testbed uses λ=0.8 for followers with the pinned leader forced to 1.0; the discussion harness uses λ=1.0 (DeGroot) unless the optional anchored-prompt flag is set.
- **Background exposure.** In the centralized (hub) topology every agent reads the hub at weight 1.0 and every other peer at a floor weight 0.15, which keeps the exposure support full; a coverage gate keeps the visible peers covering 90% of each row's influence mass.

12.3 HOW THE CONTEXT WIDTH β IS REALIZED

β is *not* a parameter of the language model: the backbone is a fixed, greedily-decoded instruction model that never sees β. It is a knob of the simulation harness that controls, for each agent and round, *which peers' messages enter that agent's prompt and with what displayed weight*. Each round the harness forms the influence row $C_{i:}(\beta)$ (Sec. 12.2) and builds agent i's context by listing peers in descending C_{ij} order until their cumulative normalized influence mass reaches the coverage threshold (0.9); peers below the threshold are *withheld entirely*. This makes β a genuine information bottleneck, not a cosmetic weight:

- **Narrow β.** $C_{i:}$ concentrates on the single most-exposed source (the hub), so the threshold is met by that source alone and every other agent's private facts never appear in i's prompt. The agent cannot pool what it never sees and defers to the dominant, confident-wrong source: herding.

- **Wide β.** $C_{i:}$ spreads out, so covering 90% of the mass requires nearly all peers; the prompt then carries the distributed private facts and the agent pools them: wisdom.

Each shown peer is also labelled with its weight in the prompt (e.g. "`[weight 0.31] Agent 7`"), but the load-bearing effect is the hard inclusion/exclusion above. Because β acts entirely through prompt construction — never through model weights, fine-tuning, or decoding temperature — the mechanism is backbone-agnostic, which is why the narrow-β collapse and its equalizer fix replicate on a second model family (App. 13.7). The equalized allocator Γ^{EQ} changes *which* column C concentrates on (it removes the concentration), not the gating rule; the online priced protocol reaches the same C by updating per-source prices once per round. The exact prompt templates, including the gated peer block, are in App. 12.7.

12.4 Environments

HiddenBench (hidden-profile reasoning). The 65-task benchmark of Li et al. (2025) (bit-identical to the public `HiddenBench` release, built on the Stasser–Titus hidden-profile paradigm). Each task distributes unshared clues round-robin across the n agents so that no agent can answer alone and only pooling recovers the correct option. The *injected adversary* pins the hub agent to a fixed wrong option $((\text{answer} \bmod K) + 1)$ and makes it structurally stubborn (the hub reads no one and never updates), so it broadcasts a confident wrong answer at every round; the *adversary-off* control removes the pin. Collective accuracy is the answer at the first round all n agents agree on a valid option (else the final-round majority); the pre-discussion baseline is the round-0 majority over the distributed beliefs. Large-n runs use $n \in \{8, 16, 24\}$, $T{=}5$ discussion rounds, 12 tasks/seed, 5 seeds, and a 9-point β grid ($n{=}24$) or 5-point grid.

Werewolf (social deduction). Standard one-werewolf games compiled into the HiddenBench format: each of the $N{-}1$ innocents holds one truthful alibi clue whose pooling uniquely identifies the werewolf. The werewolf plays a naturally deceptive role — it broadcasts a confident false accusation — and we seat it at the influence hub, realized through the same hub-pin mechanism as the HiddenBench adversary, so its deflection dominates each reader's context at narrow β. Game files exist for $N \in \{6, 8, 10, 12, 14, 16\}$ (100 games each; 20 sampled per run), $T{=}4$, 5 seeds, a 5-point β grid; the collective is scored correct when the village converges on the werewolf. Both seats use the same model, so Werewolf isolates $\beta{\times}$adversary$\times$allocator, not placement.

AgentsNet (distributed graph colouring). The coordination benchmark of Grötschla et al. (2025). Each agent is a graph node choosing a colour in $1..\Delta{+}1$ different from all neighbours; a valid colouring always exists, so residual *colouring conflicts* (edges whose endpoints share a colour, on the final beliefs, lower is better) measure coordination failure. Graphs are Barabási–Albert (BA, attachment $m{=}2$; the default), plus Watts–Strogatz (degree ≈ 4, rewire $p{=}0.3$) and Delaunay for the topology ablation. The society seats 2 strong (7B) agents either on the highest-degree nodes (*central*) or the lowest-degree nodes (*peripheral*), with everyone else weak (1.5B); placement is chosen by node degree. Runs use $n \in \{8, 16, 24, 32, 50\}$, $T{=}4$, 30 graphs, 5 seeds, a 5-point β grid.

Boundary environments. *Debate* (Du et al., 2024) (network variant): agents on a BA graph answer a hard multiple-choice question then revise after reading peers; collective accuracy is the majority of final answers ($n{=}8$, $T{=}3$, 40 tasks, 5 seeds, a disclosed hub adversary). *GovSim* (Piatti et al., 2024): $n{=}5$ fishers share a lake of capacity $20n$ over up to 12 months; we report survival months and the sustainability rate (fraction of runs that do not collapse below the threshold). *MARBLE* (Zhu et al., 2025): a MultiAgentBench coding-collaboration team whose relationship graph is set to the thresholded $C(\beta)$ attention structure, scored by total milestone completion, with the strong agent seated at the hub vs. a leaf.

12.5 SEEDS, PROTOCOL, AND STATISTICS

Every run reports the mean over independent seeds, each seed re-drawing task instances, fact distribution, and graph/game structure. Seeds are explicit and offset from a per-experiment base: HiddenBench/Werewolf use base 4200 with seeds $\{4200, 5200, 6200, 7200, 8200\}$ (5 seeds); AgentsNet uses base 5500 (5 seeds); the estimation testbed uses base 51000 (8 seeds); the N_{eff} suite uses base 53000 (24 seeds, with 1000 bootstrap resamples over seeds to check the stability of the reported N_{eff} estimates); Debate uses base 7300 (5 seeds). Unless stated, cells are 5 seeds. The two-regime "gap" is wide-β minus narrow-β collective accuracy (discussion benchmarks) or the analogous narrow→wide change in error (testbed).

12.6 HARDWARE, SOFTWARE, AND REPRODUCIBILITY

All LLM inference runs on a single node with 8× **NVIDIA RTX PRO 6000** (Blackwell, `sm_120`, 96 GB each; driver 580.159.04, CUDA 13.0). We use **vLLM** 0.23.0 under **Python** 3.12, **PyTorch** 2.11.0, **Transformers** 5.12.1, and NumPy 2.3.5, in `bfloat16` with tensor-parallel size 1 (one model replica per GPU). The in-process vLLM engine is used throughout, except MARBLE and the corrected GovSim runs, which use a persistent OpenAI-compatible vLLM server. Randomness flows through the explicit seeds above; greedy decoding is used for reproducibility, with the caveat that greedy generation is batch-composition sensitive in `bfloat16`, so bit-identical per-prompt decoding requires batch size 1. (The committed CPU-only `environment.yml` provisions the analysis/plotting environment only; the GPU inference stack above is installed separately and its exact versions are recorded here from the live environment, as no GPU lockfile is committed.)

12.7 PROMPT TEMPLATES

Every agent receives a *system* message (its role and private information) and a per-round *user* message (the task instruction plus the β-gated peer block of Sec. 12.3); every reply ends with a parseable `BELIEF: <integer>` line, from which we read the agent's choice. We reproduce the exact templates verbatim below; $\{\ldots\}$ marks a per-task field filled from the benchmark item. Boxes are colour-coded throughout this section: **grey** is a harness-constructed template, **blue** is the exact text a model actually saw, **green** is a model's real, unedited output.

HiddenBench — system message (template)

```
{scenario description}

Facts known to EVERYONE:
- {shared fact 1}
- ...
Facts known ONLY to you:
- {this agent's private fact(s)}

Options:
1) {option 1}
2) {option 2}
...

First, in 1-2 sentences, state the specific fact(s) YOU were given that others
may not have. Then weigh all available facts and choose. End with exactly one
line:
BELIEF: <integer>
where <integer> is the option number of your current best decision (digits only).
```

HiddenBench — round-0 user message (template, no peers yet)

```
Choose the single best option by its number. Options: 1) ...; 2) ...; ...

Using only the information you were given, share your unique fact(s) and state
your current best answer.
```

HiddenBench — round $t \geq 1$ user message (template; the peer block is β-gated, Sec. 12.3)

```
Choose the single best option by its number. Options: 1) ...; 2) ...; ...

Messages from the agents you are listening to (with their weight in your context):
--- [weight 0.71] Agent 0 ---
{Agent 0's most recent message}
--- [weight 0.12] Agent 3 ---
{Agent 3's most recent message}

Pool everyone's facts -- especially any you did not originally have -- and
reconsider. If a peer revealed a decisive fact, update accordingly. Give your
updated answer.
```

Anchoring suffix (appended to the system message in the anchored runs that make the bridge bound non-vacuous, $\lambda_{\max} < 1$; App. 13.3)

```
IMPORTANT: weigh your OWN reading of the facts substantially. Do not simply adopt
the group's majority; keep significant weight on your own current choice and
update only partially toward what others report.
```

The system message carries the node id, its neighbour ids, the palette size K, and the goal; the round-t user message shows only the neighbours $C(\beta)$ attends to (non-neighbours have zero exposure):

AgentsNet — system + user message (template)

```
GRAPH COLOURING. Your neighbours are: {neighbour ids}. There are {K} colours,
numbered 1 to {K}. GOAL: choose a colour DIFFERENT from every one of your
neighbours.

Reason briefly, then end with exactly one line:
BELIEF: <integer>
where <integer> is the colour number you choose (digits only).

    [user, round >= 1]
Colours currently chosen by the neighbours you are attending to
(with their weight in your context):
  - Agent 3: colour 2   [weight 0.44]
  - Agent 7: undecided  [weight 0.19]

Choose a colour that DIFFERS from every neighbour shown above. If they already
occupy several colours, pick the smallest colour none of them uses. Give your
updated colour.
```

Worked example on a real item. To make the templates concrete, the box below is the verbatim system prompt the harness produced for one agent on the committed HiddenBench item "Safe Haven After the Spill" (item 34 of the 65) — a textbook hidden profile in which the *shared* facts favour Maple Lodge (the listed safe site) but the correct answer is Cedar Station, recoverable only by pooling the three *unshared* facts (Pine Retreat lies in the drifting cloud; the road to Maple Lodge is washed out; Cedar Station alone has working communications). The four agents are each dealt one clue — Agent 0 (the hub) the drifting cloud, Agent 1 the washed-out road, Agent 2 the Cedar-Station radio — except a fourth, Agent 3, that is dealt none. The box below is the verbatim system prompt for *Agent 1*, holder of the decisive "washed-out road" clue:

Worked example — real system prompt (item 34, "Safe Haven After the Spill")

```
You are part of a university field team conducting research in a rural
region. Suddenly, a chemical spill at a manufacturing plant several miles
away has triggered a region-wide alert. Everyone on your team must decide
together where to relocate for safety and communication with authorities.
You have very little time. There are three potential locations:
- Maple Lodge (a youth camp to the east)
- Pine Retreat (a cabin complex to the north)
- Cedar Station (an old ranger station to the south)
All locations are usually reachable in under an hour. Your goal: choose the
only safe and reachable location, maximizing your team's safety and your
ability to coordinate with emergency services.

Facts known to EVERYONE:
- Maple Lodge is listed on the evacuation plan as a safe site for
```

```
  chemical-related incidents.
- Pine Retreat is farther from the spill than where you currently are.
- Cedar Station is an older building, less commonly used, but the route to
  it is familiar to everyone on the team.
- Maps indicate that all sites should have food and water supplies for at
  least 24 hours.

Facts known ONLY to you:
- Emergency traffic updates say the only road to Maple Lodge has been washed
  out by heavy rains this morning, making it inaccessible by car or foot.

Options:
1) Maple Lodge
2) Pine Retreat
3) Cedar Station

First, in 1-2 sentences, state the specific fact(s) YOU were given that
others may not have. Then weigh all available facts and choose. End with
exactly one line:
BELIEF: <integer>
where <integer> is the option number of your current best decision (digits
only).
```

Agent 3's system prompt on this item is identical — same scenario, same shared facts, same options, same answer contract — except for its private-facts block, which reads:

Worked example — Agent 3's private-facts block (item 34), verbatim, all else identical to the box above

```
Facts known ONLY to you:
(no extra facts beyond the shared ones)
```

On this item the two regimes play out verbatim on the same Qwen2.5-7B backbone: Agent 3's answer depends entirely on which peers β lets it read. At narrow β=0.1 the 0.9 coverage threshold is met by the hub (Agent 0) alone, so Agent 3 sees only the hub's message — which, leaning on the shared "Maple Lodge is the listed safe site", argues merely that Pine Retreat is dangerous — and it herds onto the wrong shared favourite:

Peer block shown at narrow β=0.1

```
Messages from the agents you are listening to (with their weight in your context):
--- [weight 1.00] Agent 0 ---
The chemical cloud is drifting north, making Pine Retreat dangerous.  BELIEF: 1
```

Follower's real reply at narrow β=0.1

```
Based on the new information that the chemical cloud is drifting north, Pine
Retreat is now a dangerous location.
BELIEF: 1                                    <-- Maple Lodge  (WRONG)
```

At wide β=10 the same follower's prompt clears nearly all peers, so it now also reads the two decisive unshared clues, pools them, and flips to the correct answer:

Peer block shown at wide β=10

```
Messages from the agents you are listening to (with their weight in your context):
--- [weight 0.42] Agent 0 --- ...Pine Retreat dangerous.              BELIEF: 1
--- [weight 0.29] Agent 1 --- ...road to Maple Lodge washed out...    BELIEF: 2
--- [weight 0.29] Agent 2 --- ...Cedar Station has ... radio.         BELIEF: 3
```

Follower's real reply at wide β=10

```
Based on the updated information:
- Maple Lodge is listed as a safe site but the road is washed out, making it
  inaccessible.
- Pine Retreat ... the chemical cloud is drifting north, making it dangerous.
- Cedar Station has a functional landline, backup generator, and emergency
  broadcast radio ...
```

```
BELIEF: 3                                    <-- Cedar Station  (CORRECT)
```

Nothing changed but β — same item, same follower, same greedily-decoded backbone — yet the context bottleneck alone turns a herded wrong answer into the pooled correct one. This single item is drawn from the full 65-task benchmark for illustration; averaged over the sampled HiddenBench runs the same swing is the 0.09 ± 0.08 narrow-β floor versus the 0.73 ± 0.07 wide-β accuracy of Table 7.

13 Full Experimental Results

All tables in this section are computed from the committed `data.json`/`cells.json` artifacts. Error is the estimation testbed's collective error (lower is better); collective accuracy and colouring conflicts are as defined above.

13.1 Controlled Estimation Testbed: Two Regimes Across Model Scale

The testbed is a measurement instrument (App. 12.2: it exposes δ, λ, and N_{eff}, which no cooperative benchmark reports), not a collective-intelligence claim in its own right. A wrong "leader" is pinned at the influence centre; the collective error is high at narrow β (the followers herd onto the leader) and low at wide β (they pool the distributed noisy signals). Table 2 sweeps the crowd model over the capability ladder (n=12, 8 seeds); the two-regime is present at every scale, and equalization reduces the narrow-β error the more capable the crowd (the 1.5B row is the unstable/oscillating point noted in the diagnostics, not a failure of the effect).

Table 2: Two-regime by model scale in the controlled testbed (collective error, lower is better; n=12, 8 seeds). Baseline Γ_β vs. equalized Γ^{EQ} at narrow (β=0.1) and wide (β=10) context width.

crowd model	base narrow	base wide	eq. narrow	eq. wide
Qwen2.5-0.5B	57.6 ± 1.3	8.7 ± 5.1	45.9 ± 22.2	37.2 ± 23.5
Qwen2.5-1.5B	14.3 ± 3.5	8.0 ± 4.1	7.7 ± 4.6	9.4 ± 5.4
Qwen2.5-3B	41.5 ± 4.7	5.4 ± 4.1	15.5 ± 23.6	5.4 ± 2.7
Qwen2.5-7B	46.5 ± 3.7	12.5 ± 8.9	15.2 ± 21.0	12.1 ± 16.9

13.2 Mechanism Ablations

Table 3 is the controlled-testbed ablation grid (Qwen2.5-3B, n=12, 8 seeds) over social power, self-reliance, Sinkhorn iterations, and five structural variants of the interaction model, read from the per-setting run data. The central result is that the narrow-β herding error is *gated* by social power: it collapses to the healthy floor at ζ=0 (5.5, no concentration, no herding) and jumps to ≈ 42 for any ζ>0. Equalization lowers the narrow-β error under every setting; a single Sinkhorn iteration (k=1) already clears the dominant column, so the effect is k-robust. At wide β there is no dominant column left to clear, so equalization has little upside and can slightly distort an already-balanced allocation (e.g. the k=4 eq.-wide cell, and the smallest crowd in Table 2); the mechanism's value is at narrow β, where a cheap price step recovers most of the wide-β wisdom without paying for wide context. The herding persists across five structural variants of the interaction model (memory, exposure-pricing, late-joiner, distance-decay, skeptic), each of which equalization again clears. Self-reliance λ has no measurable effect over $[0.3, 0.7]$ (the followers' anchoring does not change the pinned-leader herding). Horizon and background exposure are not swept in this testbed; population is checked separately on the real benchmarks (Appendix 13.4, 13.5), not as a testbed grid axis.

13.3 Proxy↔LLM Bridge and Effective Sample Size

The bound holds. Pooling the bridge cells (three β-sweep conditions plus the anchored-persona run), the measured sup-norm proxy↔LLM discrepancy is at or under the theoretical ceiling $\delta/(1-\lambda)$ in $254/254$ **cells (0 violations)**, median margin $2.5\times$ (Theorem 4.1). The first-moment identity

Table 3: Controlled-testbed ablations (collective error; Qwen2.5-3B, n=12, 8 seeds). "base" is Γ_β, "eq" is Γ^{EQ}; narrow = β=0.1, wide = β=10.

setting	base narrow	base wide	eq. narrow	eq. wide
Social power ζ				
ζ=0.0	$\mathbf{5.5 \pm 2.6}$	4.9 ± 2.0	11.4 ± 16.2	8.7 ± 5.6
ζ=0.2	42.2 ± 4.3	4.7 ± 2.3	25.5 ± 33.3	4.0 ± 1.5
ζ=0.4	42.1 ± 4.2	5.2 ± 1.5	18.7 ± 23.2	4.8 ± 3.5
ζ=0.6	42.1 ± 4.3	5.6 ± 4.2	17.3 ± 21.4	9.4 ± 8.2
ζ=0.8	42.1 ± 4.2	3.9 ± 2.2	15.3 ± 17.3	10.3 ± 10.1
ζ=0.9	41.7 ± 3.9	5.3 ± 1.9	16.7 ± 19.6	6.6 ± 2.5
ζ=0.95	41.5 ± 3.6	4.5 ± 2.0	18.7 ± 22.8	8.1 ± 8.5
Self-reliance λ				
0.3–0.7 (all)	42.1 ± 4.2	11.5 ± 18.5	10.0 ± 13.7	16.0 ± 17.5
Sinkhorn iterations k				
k=1	39.4 ± 5.9	4.7 ± 2.7	6.9 ± 6.5	5.1 ± 2.4
k=2	39.4 ± 5.9	4.7 ± 2.7	6.9 ± 6.5	4.5 ± 1.6
k=4	39.4 ± 5.9	4.7 ± 2.7	6.9 ± 6.5	12.9 ± 15.0
Structural variants				
memory	42.3 ± 8.6	14.5 ± 31.2	1.9 ± 2.1	1.9 ± 2.0
exposure-price	50.6 ± 1.7	5.3 ± 2.3	4.3 ± 1.9	4.9 ± 1.9
late-joiner	28.7 ± 5.0	2.2 ± 1.2	4.4 ± 2.4	5.6 ± 3.0
distance-decay	59.1 ± 1.1	4.4 ± 2.2	6.5 ± 3.7	11.0 ± 8.8
skeptic	18.4 ± 8.1	3.4 ± 1.8	5.0 ± 2.3	5.3 ± 5.0

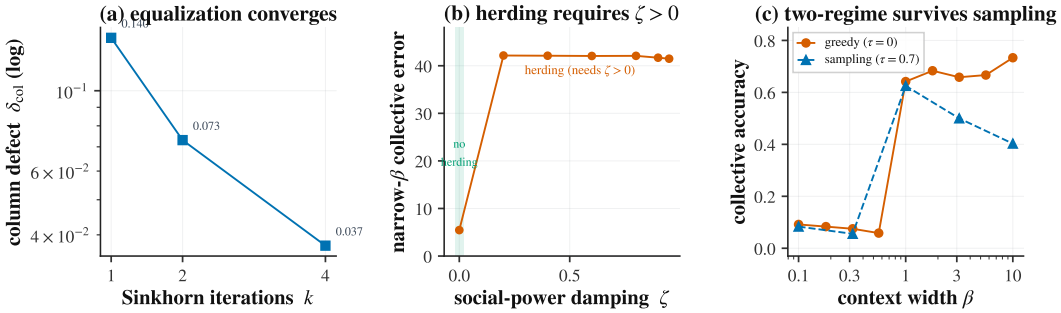

Figure 5: **Mechanism ablations (visual). (a)** the equalized allocator's column defect falls monotonically with Sinkhorn iterations k, from 0.14 at k=1 to 0.037 at k=4; **(b)** herding is gated by social power, absent at ζ=0 and present for any ζ>0 (cf. Table 3); **(c)** on real HiddenBench (n=24) the two-regime holds under both greedy (τ=0) and stochastic (τ=0.7) decoding.

$\hat{\theta}_{\mathrm{LLM}} \approx q_T^\top b_0$ holds with OLS slope 0.88 (R^2=0.74, vs. 0.28 for a β-free mean-b_0 baseline) — the LLM collective is the influence-weighted average the theory predicts.

The anchoring weights are measured, not assumed. The ceiling $\delta_T/(1-\lambda_{\max})$ depends on the per-agent anchoring λ_i, which we do not set but fit by least squares over each run: for fixed i the residual $\hat{\delta}_i(t)$ is affine in λ_i at each t, so $\sum_t \hat{\delta}_i(t)^2$ is a univariate quadratic and the minimiser on $[0, 1)$ is closed-form. Reporting a goodness-of-fit residual after fitting λ is standard; what matters is where the resulting ceiling is genuinely small, not merely finite. It is tightest on the 89 anchored-persona cells, whose explicit anchoring prompt keeps self-reliance low (median λ=0.25), so 99% have a ceiling below 250, against the $\sim 10^5$ an unanchored persona produces as $\lambda \to 1$. Across the full 254-cell pool the fitted λ is higher (median 0.61) and the ceiling looser, with 70% below 250; the bound holds with zero violations throughout, but is non-vacuous in the strong sense only on the anchored subset and merely satisfied on the rest. Because λ_i is fit from the T residuals of one run, δ_T is the worst-case residual *after* each agent's best-fit anchoring, so "zero violations" is a statement about per-agent best-fit ceilings rather than about a prescribed λ.

N_{eff} across network topology. Table 4 reports the proxy N_{eff} (q_T-based, n=12, 24 seeds) as β widens. The herding floor at narrow β and its recovery toward the degree ceiling at wide β appear on every graph family that has a concentrated hub (BA, ER, WS, SBM); the *regular* graph, which has no hub, shows no floor (N_{eff}=12 throughout) — the floor is a property of hub concentration, not of n. This is consistent with Theorem 4.2's wide-context prediction: for the degree-regular graph, the closed form $(\sum_k d_k)^2 / \sum_k d_k^2$ evaluates to exactly $12.00 = n$ and the measured N_{eff} recovers it, while the Barabási–Albert graph's degree heterogeneity evaluates the same closed form at 8.89. The wide-β measurement is 9.8; the two numbers are not the same quantity — 8.89 is the $\beta \to \infty$ stationary-weight (ν) ceiling of the theorem, whereas 9.8 is the finite-β, finite-horizon q_T-based estimate — but they agree in magnitude and both sit well above the narrow-β floor, which is the prediction being tested. The same diagnosis does not transfer directly to the HiddenBench and Werewolf exposure graphs (Appendix 13.4, 13.5): their hub structure is asymmetric ($\bar{W} \neq \bar{W}^\top$), outside this part of the theorem, and their *binary* support is near-degree-regular, so the gap to oracle visible in Figure 2A,B reflects the hidden-profile information ceiling rather than degree heterogeneity.

Table 4: Effective sample size N_{eff} (proxy q_T; n=12, 24 seeds) at narrow vs. wide β, by graph topology. Narrow β herds onto the hub (small N_{eff}); wide β recovers the degree ceiling. Regular graphs have no hub and hence no floor.

topology	N_{eff} narrow β	N_{eff} wide β
Barabási–Albert (3B)	3.0	9.8
Barabási–Albert (7B)	3.0	9.8
Erdős–Rényi	2.8	11.1
Watts–Strogatz	4.3	11.2
stochastic block model	6.0	11.6
regular (no hub)	12.0	12.0

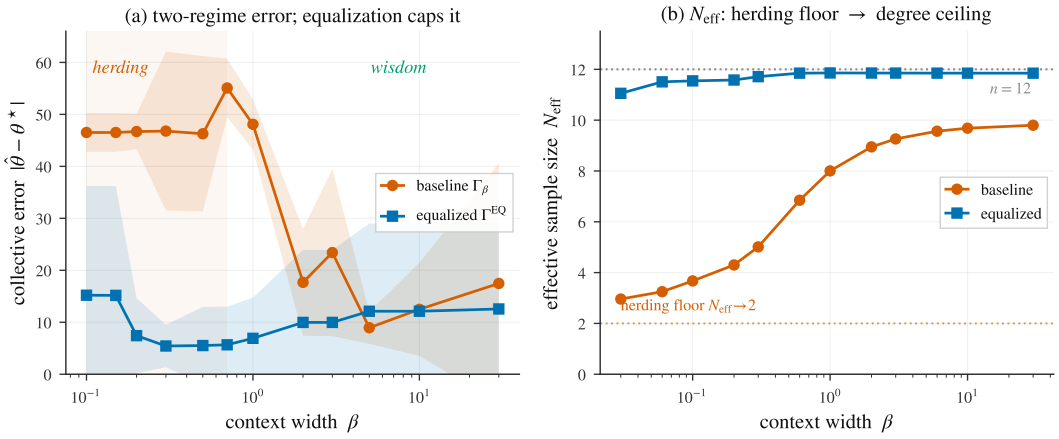

Figure 6: **The closed-form two-regime, realized on the LLM testbed.** Baseline collective error shows the two regimes across β while the equalized operator holds error low; N_{eff} rises from the narrow-β floor toward the degree ceiling (Theorems 4.2, 4.3). Real Qwen2.5-7B.

13.4 HiddenBench: two-regime, causation, floor, and equalizer

Table 5 is the causation 2×2 at n=8 (adversary {on,off} × allocator {baseline, Γ^{EQ}}, 5 seeds): only the adversary-on/baseline cell collapses at narrow β, and it is repaired by *either* intervention. Table 6 shows the floor is n-independent (flat narrow-β accuracy from n=8 to n=24) and that equalization recovers narrow-β accuracy at every size. The equalizer is robust to its one knob: at n=8 every $k \in \{1, 2, 4\}$ lifts narrow-β accuracy off the floor (from 0.08 to 0.60–0.67).

Variance across arms. Table 7 gives mean ± std at narrow and wide β (seed-level spread, pooled across the causation cells above). The baseline gate and the herding floor are both low-variance

Table 5: HiddenBench causation 2×2 (collective accuracy; n=8, 5 seeds).

cell	β=0.1	0.3	1.0	3.0	10	gap
adv-on, baseline	**0.083**	0.067	0.667	0.700	0.733	+0.65
adv-on, equalized	0.667	0.633	0.633	0.600	0.600	−0.07
adv-off, baseline	0.583	0.583	0.750	0.767	0.717	+0.13
adv-off, equalized	0.683	0.817	0.717	0.717	0.767	+0.08

Table 6: HiddenBench N_{eff} floor across n (adversary-on baseline vs. equalized; collective accuracy; 5 seeds). The narrow-β floor is flat in n; equalization recovers it at every n.

	baseline Γ_β			equalized Γ^{EQ}	
n	narrow β	wide β	gap	narrow β	recovery
8	0.083	0.733	+0.65	0.667	+0.58
16	0.083	0.633	+0.55	0.583	+0.50
24	0.050	0.767	+0.72	0.617	+0.57

and clearly separated (adv-on/baseline: 0.09 ± 0.08 at narrow β vs. 0.73 ± 0.07 at wide β); the equalizer's narrow-β recovery is high-variance (0.58 ± 0.13), i.e. some seeds clear the collapse fully and others only partially — an honest property of the mechanism, not noise to be hidden.

Table 7: HiddenBench, mean $\pm$ std collective accuracy at narrow (β=0.1) and wide (β=10) context width.

arm	narrow β	wide β
adv-on, baseline	0.092 ± 0.075	0.733 ± 0.069
equalized	0.583 ± 0.134	0.633 ± 0.153
adv-off	0.517 ± 0.110	0.667 ± 0.146

13.5 WEREWOLF: TWO-REGIME, FLOOR, AND EQUALIZER

Table 8 gives the n=6 arms: at narrow β the village herds onto the werewolf's false accusation and convicts the wrong player on *every* game (0.00); at wide β it pools the alibis and convicts the werewolf (0.87). Removing the werewolf's deflection (the intrinsic control) removes the collapse (0.71), and the fully-cleared equalizer recovers narrow-β to 0.55 at this size. The narrow-β floor is n-independent, pinned at 0.00 for every $n \in \{6, 8, 10, 12, 16\}$ while wide-β stays high. The equalizer's recovery, however, is not n-independent: at n=6 a single Sinkhorn iteration already lifts narrow-β off the floor, but the recovery weakens as n grows (the fully-cleared operator falls to 0.24 at n=16; Table 9), so unlike HiddenBench, Werewolf does not recover equally at every size.

Table 8: Werewolf n=6 (20 games, 5 seeds; collective accuracy). "equalized" is the fully-cleared operator Γ^{EQ} run to its fixed point each round. The N_{eff} floor is n-independent (narrow-β pinned at 0.00 for all n).

arm (n=6)	β=0.1	0.3	1.0	3.0	10
baseline (wolf deflects)	**0.00**	0.00	0.00	0.88	0.87
equalized Γ^{EQ} (fully cleared)	0.55	0.51	0.48	0.48	0.48
wolf does not deflect (control)	0.71	0.72	0.66	0.89	0.91

N_{eff} *floor*: narrow-β accuracy by n $(6, 8, 10, 12, 16)$ = $0.00, 0.00, 0.00, 0.00, 0.00$; wide-$\beta$ = $0.87, 0.83, 0.70, 0.63, 0.59$.

Variance across arms. Table 9 gives mean $\pm$ std at n=16 (seed-level spread). The narrow-β floor has *zero* variance (0.00 ± 0.00): every seed lynches the wrong player, not just most of them. The *online-priced* allocator (one price update per round, as in Figure 3) recovers narrow-β to 0.18 ± 0.07,

just below the fully-cleared operator at the same size (0.24; Figure 3b) and well below the operator's 0.55 at n=6 (Table 8) — the operator recovery itself weakens with n, and online pricing tracks just under it within T=4.

Table 9: Werewolf n=16, mean ± std collective accuracy at narrow (β=0.1) and wide (β=10) context width. Here "online-priced" is the one-update-per-round protocol of Figure 3, *not* the fully-cleared operator of Table 8.

arm (n=16)	narrow β	wide β
baseline (wolf deflects)	0.000 ± 0.000	0.585 ± 0.077
online-priced	0.180 ± 0.074	0.575 ± 0.097
wolf does not deflect	0.160 ± 0.063	0.585 ± 0.055

13.6 AGENTSNET: PLACEMENT AND ITS ABLATIONS

Placement is the complementary facet: seating the strong agents at the high-degree influence centre beats seating them at the periphery, and the advantage grows monotonically with scale (main-text Figure 2c; +1.5 at n=8 to +10.6 at n=50). Table 10 gives the ablation matrix. Crucially, the placement gap is *degree/topology*-driven, not social-power-driven: it is invariant to ζ and to the number of Sinkhorn iterations, and equalization *lowers* conflicts but does *not* collapse the gap — so these ablations do not contradict Theorem 4.3 (which neutralizes social-power column concentration, not a degree-induced coordination advantage). The Watts–Strogatz control, whose hubs are weak, shows the gap shrink (+2.2 at n=16) and reverse sign (−2.9 at n=32), consistent with the effect being carried by hub degree; per-seed conflicts were not logged for this ablation, so the n=32 sign flip is not independently verified against sampling noise and should be read as suggestive rather than confirmed.

Table 10: AgentsNet placement ablations (mean colouring conflicts, lower is better; 5 seeds). Gap = peripheral − central (positive = centre wins).

setting	central	peripheral	gap	reading
baseline n=16	9.3	15.1	+5.8	—
equalized n=16	7.6	14.0	+6.4	eq lowers conflicts, gap survives
equalized n=32	13.2	23.4	+10.2	same at scale
k=1 / k=2 n=16	7.0/7.1	13.5/13.9	+6.5/+6.8	Sinkhorn-k invariant
ζ=0 n=16	10.6	16.6	+6.0	ζ-invariant
ζ=0.9 n=16	9.5	15.2	+5.7	ζ-invariant
T=8 n=16	9.2	14.7	+5.5	horizon-robust
n_{strong}=1 n=32	17.2	23.3	+6.1	one hub agent already wins
small-world (WS) n=16	13.9	16.1	+2.2	weak hubs → small gap
small-world (WS) n=32	33.0	30.1	−2.9	homogeneous → gap vanishes/flips

13.7 CROSS-FAMILY GENERALITY: LLAMA-3.1-8B

Table 11 re-runs the HiddenBench 2 × 2 with Llama-3.1-8B in place of Qwen-7B (same code, model swapped). All three signatures replicate: the narrow-β herding collapse (0.02), the equalizer's flat recovery (+0.37), and the adversary-off no-collapse control (0.55). The honest scope is that *capability bounds the wide-β regime*: Llama-8B is a weaker reasoner here, so its wide-β ceiling is lower (0.35 vs. Qwen's 0.73) and its adversary-off curve *declines* with β (0.55 → 0.33) rather than rising — a weak model is hurt by wide context. The narrow-β collapse and the equalizer fix are model-family-agnostic; the wide-β *wisdom* magnitude is capability-bounded. Mean ± std at the endpoints (narrow β=0.1, wide β=10): baseline $0.017 \pm 0.020 \rightarrow 0.267 \pm 0.110$; equalized $0.383 \pm 0.076 \rightarrow 0.392 \pm 0.114$; adversary-off $0.550 \pm 0.085 \rightarrow 0.325 \pm 0.082$. All three arms separate cleanly at narrow β, confirming the signatures in Table 11 are not an artifact of seed noise.

Table 11: Cross-family generality: HiddenBench 2×2 with Llama-3.1-8B (collective accuracy; 5 seeds). Individual ≈ 0.08, oracle ≈ 0.76.

arm	β=0.1	0.3	1.0	3.0	10
baseline Γ_β	**0.02**	0.03	0.35	0.26	0.27
equalized Γ^{EQ}	0.38	0.31	0.38	0.36	0.39
adversary-off	0.55	0.49	0.41	0.39	0.33

13.8 Boundary environments

The two-regime is a property of aggregation-dependent tasks, and its magnitude tracks how strongly the task forces deference to a dominant source. Table 12 maps the boundary. *Debate* keeps every agent's own answer and spreads influence over a BA graph with no single bottleneck, so the injected adversary captures only a small share and the two-regime is weak ($+0.07$). *GovSim*, with the corrected inference backend, *saturates*: every condition sustains the resource (survival 12 months, sustainability rate 1.0). This is structural, not a tuning artifact: GovSim is a continuous-harvest decision in which each fisher independently chooses a catch, with no private clue that must be pooled across agents to reach the right action — so there is no hidden-profile information to aggregate and no confident-wrong source for the population to herd onto. The two-regime therefore cannot arise here by construction, and GovSim saturates by ceiling rather than by collapse. Together with Debate, this marks where the aggregation channel is absent or the task never forces deference, so the theory neither predicts nor is tested by these environments. *MARBLE* is a different case: it has a placement axis the theory does speak to, but decomposes into independent per-agent milestones with no shared latent for placement to act on, and the placement advantage flips sign across replicates ($+0.61$ versus -0.58). We report this as a negative result — the predicted placement effect does not hold on independently-scored subtasks — not as a boundary alongside the ceiling cases above.

Table 12: Boundary environments. Debate: collective accuracy by β (the two-regime is weak). GovSim: survival months / sustainability (the corrected backend saturates, no two-regime by ceiling).

Debate (Du et al. network variant; n=8, 5 seeds; collective accuracy)

arm	β=0.1	0.3	1.0	3.0	10	gap
baseline Γ_β	0.300	0.275	0.340	0.305	0.370	$+0.07$
equalized	0.330	0.315	0.305	0.340	0.300	-0.03
adversary-off	0.350	0.355	0.345	0.340	0.305	-0.04

GovSim (corrected server backend): survival $= 12$ months and sustainability $= 1.0$ in all conditions (n=5–24, baseline / equalized / fixed-pie / anchored) — the collective saturates the cooperative ceiling; there is no two-regime.