

## Student Submission Form:

*Please complete the following sections and include this as the cover page for your submission.*

***By including this form as part of your submission you are agreeing to the Plagiarism and Anonymisation Statements below.***

<b>Module:</b>	PhD 11 Economic Theory
<b>Candidate Number / BGN:</b>	2302B
<b>Submission Deadline:</b>	Friday, 30 <sup>th</sup> January 2026, 12:00 PM
<b>Actual Word Count:</b>	4,937 words (no Appendix)

### **Statement 1: Plagiarism**

I confirm that this is entirely my own work and has not previously been submitted for assessment, and I have read and understood the University's and Faculty's definition of Plagiarism (please see links below):

[Plagiarism and Academic Misconduct](#)

[Plagiarism and correct referencing in dissertations and essays](#)

### **Statement 2: Anonymisation of Work**

I confirm that I have taken all reasonable steps to ensure that all submitted files for assessment have been anonymised and do not contain any identifiable information to me. For more information, please refer to the UIS help page link below which contains useful information within the 'Anonymisation' section:

<https://help.uis.cam.ac.uk/service/security/glossary#a>

## 1. INTRODUCTION

"I think it's much more interesting to live not knowing than to have answers which might be wrong. I have approximate answers and possible beliefs and different degrees of uncertainty about different things, but I am not absolutely sure of anything and there are many things I don't know anything about, such as whether it means anything to ask why we're here. I don't have to know an answer. I don't feel frightened not knowing things, by being lost in a mysterious universe without any purpose, which is the way it really is as far as I can tell."

— Richard P. Feynman

"...in the modern world the stupid are cocksure while the intelligent are full of doubt."

— Bertrand Russell

Knowledge is fundamental to economic prosperity and human progress. From Arrow (1962)'s seminal work on learning-by-doing to the modern theory of endogenous growth (Howitt & Aghion, 1998; Romer, 1990), economists have long recognized that the accumulation and diffusion of knowledge drive technological innovation, productivity growth, and social welfare. Yet despite its centrality, knowledge diffusion remains imperfectly understood, particularly when individuals systematically misjudge what they know.

Recent decades have witnessed an expansion of platforms enabling rapid information exchange. Social media, online forums, and digital networks have dramatically reduced the costs of communication and persuasion. While these technologies promise to democratize knowledge and accelerate learning, they have also coincided with the proliferation of misinformation, polarization, and epistemic fragmentation (Acemoglu et al., 2024; Levy & Razin, 2019). Understanding when social learning enhances collective knowledge and when it leads societies astray has become a question of profound economic and social importance.

**Motivating examples:** The phenomenon of false convergence, where groups coordinate on incorrect beliefs despite initial exposure to truth, manifests in everyday social learning. Consider a football training session. A coach explains a complex tactical drill, but players grasp only fragments of the instructions. As they begin executing the drill, their partial understandings interact through observation and communication. Players who appear most confident in their interpretation—even when mistaken—exert disproportionate influence on teammates who are unsure. Gradually, the group converges to a coordinated pattern of play that differs systematically from what the coach prescribed. The players are not learning from each other's actions or outcomes, but rather from the information each claims to possess about the drill. Only when the coach intervenes, does the team's collective understanding shift back toward truth.

Similarly, when learning a traditional Scottish dance, a group of non-Scottish participants may initially misinterpret the instructor's demonstration. Confident dancers propagate incorrect

steps to uncertain peers, and the group quickly synchronizes on an inaccurate choreography. Again, learning occurs through the direct transmission of claims about what the dance steps should be, not through observing which dancers appear more successful. Periodic corrections from the instructor eventually steer the group toward the authentic dance.

Such examples involve small groups, simple tasks, and rapid feedback. But what happens when the stakes are higher, the groups larger, the information more complex, and feedback sparse or unreliable? When does false consensus persist indefinitely? How long does convergence to truth take when it occurs at all? These questions take on particular urgency in domains such as the diffusion of scientific claims through social media and "influencers" (particularly serious nowadays in the area of nutrition), conversations about economic policy, judicial proceedings, hiring decisions, and political discourse. In these settings, agents learn not by observing outcomes or inferring from actions, but by directly exchanging claims about what they sincerely believe to be true. And in each case, misjudgments of one's own knowledge—overconfidence among the uninformed and underconfidence among experts—can systematically distort who influences whom.

This work departs fundamentally from the canonical social learning literature, in which agents observe others' actions and update beliefs by inferring the private information those actions reveal (Acemoglu & Ozdaglar, 2011; Banerjee, 1992; Bikhchandani et al., 1992). In those models, learning is indirect: actions serve as noisy signals of private information. Here, by contrast, learning is direct. Agents explicitly communicate what they claim to know, and listeners adopt these claims based on the speaker's perceived credibility. Confidence, rather than the informativeness of observed actions, determines influence. A highly confident but misinformed agent can persuade many listeners, while a knowledgeable but diffident expert may fail to convince anyone, even when both would take identical actions in a learning-from-actions framework.

A large body of empirical evidence shows that such confidence distortions are systematic rather than idiosyncratic. Research across psychology, education, and behavioral economics documents a robust, non-monotonic relationship between actual knowledge and perceived competence over the learning process. Individuals with limited knowledge tend to substantially overestimate their abilities, a phenomenon known as the Dunning–Kruger effect (Kruger & Dunning, 1999). This overconfidence reflects genuine miscalibration rather than strategic exaggeration. Lacking skill, individuals also lack the metacognitive capacity to recognize their deficiencies (Dunning, 2011). As individuals acquire more knowledge and confront the complexity of a domain, many transition into a phase of underconfidence, doubting their abilities despite substantial expertise (Clance & Imes, 1978; Sakulku, 2011). Only at high levels of mastery does confidence become well calibrated, with experts accurately assessing both their competence and its limits (Ericsson et al., 2007; Teo et al., 2023).

These metacognitive patterns have direct implications for social learning. When confidence determines persuasiveness (Anderson & Kilduff, 2009; Kennedy et al., 2011), overconfident novices exert disproportionate influence despite limited knowledge, while knowledgeable but

self-doubting individuals fail to shape beliefs. Importantly, confidence evolves endogenously with learning rather than being a fixed trait. Longitudinal evidence shows that miscalibration declines only slowly with experience and often persists despite feedback (Kausel et al., 2021). Individuals are typically unaware of their bias and cannot directly correct for it.

We incorporate these empirical regularities into a formal model of knowledge diffusion in networks and lattices. Agents possess knowledge over a continuous domain partitioned into discrete facts. They meet randomly according to a Poisson process, form bilateral links, and exchange information. Communication is sincere: agents truthfully transmit what they believe to be correct. There is no strategic deception. A receiver adopts a sender's claim when she finds the sender sufficiently credible, where credibility is inferred from the sender's subjective confidence rather than from objective accuracy.

We model confidence as an endogenous, non-monotonic function of cumulative knowledge, capturing overconfidence among novices, underconfidence at intermediate skill levels, and calibration among experts. Confidence evolves mechanically with learning, and agents do not learn about their own bias. This structure allows us to study how metacognitive distortions shape aggregate knowledge dynamics in two environments: a full-truth benchmark, where all claims correspond to genuine facts and ignorance is the only obstacle to learning, and a sparse-truth setting, where some claims are false or unverifiable and errors may persist as absorbing states. **OUR SUBJECTIVE PERCEPTION OF KNOWLEDGE FUNCTION CAN SERVE AS A (ENDOGENOUS) BELIEF UPDATING RULE-ALTERNATIVE TO BAYESIAN BELIEF UPDATING- IN ANY ENVIRONMENT THAT ALSO ALLOWS FOR ZERO-PROBABILITY SPACES.**

We establish three main results. First, in full-truth environments, social learning converges almost surely to complete knowledge under minimal connectivity assumptions, despite heterogeneous confidence and non-monotonic dynamics. Second, in sparse-truth environments, false beliefs act as epistemic contaminants, and the presence of overconfident agents accelerates misinformation diffusion. Third, introducing a truth-revelation mechanism (modeling fact-checking or institutional correction) restores convergence to correct knowledge only when agents retain memory of revealed truth. Without memory, the system converges instead to a stationary distribution with persistent error, highlighting the critical role of institutional memory in stabilizing knowledge.

Our framework sheds light on why misinformation persists despite widespread access to factual information, when social networks enhance versus degrade collective learning, and how epistemic biases generate inefficiencies in labor markets, political discourse, and legal decision-making that cannot be resolved through incentives alone. By integrating empirically grounded miscalibration into a tractable model of networked learning, this work offers a unified account of confidence, persuasion, and the diffusion of knowledge.

The remainder of the paper proceeds as follows. Section 2 reviews related literature. Section 3 presents the model. Section 4 analyzes the full-truth benchmark. Section 5 studies sparse-truth environments. Section 6 introduces truth revelation and contrasts dynamics with

and without memory. Section 7 discusses extensions and policy implications. All proofs are relegated to the Appendix.

## 2. LITERATURE REVIEW

A robust empirical finding across psychology, education, and economics is that individuals systematically misjudge their own competence. Low-performing individuals tend to substantially overestimate their abilities, a phenomenon known as the Dunning–Kruger effect. First documented by Kruger and Dunning (1999), this pattern has been replicated across domains ranging from education and medicine to reasoning and decision-making. Importantly, overconfidence among low performers reflects genuine metacognitive miscalibration, not strategic exaggeration. Individuals who lack skill also lack the ability to recognize their own deficiencies (Dunning, 2011).

Subsequent empirical work confirms both the prevalence and persistence of this bias. Poor performers consistently overestimate their performance, while high performers are comparatively well calibrated (Knof et al., 2024). Even with feedback, miscalibration declines slowly and unevenly, suggesting that confidence errors stem from epistemic limitations rather than noise or learning frictions (Pennycook et al., 2017). A growing body of evidence attributes these patterns to metacognitive insensitivity: novices struggle to distinguish good from bad performance in themselves and others (McIntosh et al., 2019). Formalizing this intuition, Jansen et al. (2021) show that even rational agents with limited diagnostic information can exhibit Dunning–Kruger–type patterns.

Crucially, miscalibration is non-monotonic. As individuals acquire more knowledge, many transition from overconfidence to underconfidence, entering a “humility valley” in which they underestimate their abilities despite substantial competence (Crollen et al., 2011). This phenomenon aligns with evidence on impostor syndrome among high-achieving individuals (Clance & Imes, 1978; Sakulku, 2011). At high levels of mastery, however, self-assessment becomes accurate again: experts tend to recognize both their competence and its limits (Ericsson et al., 2007; Teo et al., 2023). Together, these findings support a non-monotonic relationship between knowledge and confidence, motivating the functional form of confidence in our model.

These metacognitive biases have direct implications for social learning and persuasion. Confidence strongly predicts influence, often independently of accuracy (Anderson & Kilduff, 2009; Kennedy et al., 2011). As a result, overconfident but poorly informed individuals may disproportionately shape beliefs, while knowledgeable but self-doubting agents fail to correct errors. Empirical evidence from educational and organizational settings confirms that overconfidence predicts poorer outcomes and that miscalibrated experts can propagate low-quality information (Feld et al., 2017; Sulphey & Senan, 2025).

Our work contributes to the growing literature on misspecified learning, which studies how agents learn and interact when their internal models are incorrect or incomplete. Bohren and Hauser (2024) provide a unified framework showing that misspecified beliefs can persist even with abundant data. Related work demonstrates that heterogeneity in misspecification can gen-

erate persistent disagreement and inefficiencies (Bohren & Hauser, 2021), while over- and underreaction to information can produce systematic distortions (Ba et al., 2024). Cho and Libgober (2025) and He and Libgober (2025) further show that misspecified learning rules may be evolutionarily stable when they confer strategic advantages. OUR WORK DIFFERS SINCE IN THIS LITERATURE, ASYMPTOTIC LEARNING IS ABOUT THE LONG-RUN BELIEFS ABOUT THE STATE RATHER THAN LONG-RUN TRUE KNOWLEDGE AND ALIGNED SUBJECTIVE PERCEPTION OF KNOWLEDGE.

This work instantiates a novel and empirically grounded form of misspecification: agents misperceive their own knowledge rather than the environment itself. Confidence evolves endogenously as a deterministic function of accumulated knowledge, generating overconfidence, underconfidence, and eventual calibration without agents being aware of their bias. This metacognitive misspecification operates at the individual level but scales into collective outcomes through social interaction.

TALK ABOUT PAPERS THAT DEAL WITH OVERCONFIDENCE (SEE LIST FROM "APPARENT OVERCONFIDENCE" JUAN DUBRA)

Our analysis is closely related to Arieli et al. (2025), who study how excessive confidence (“condescension”) can both accelerate and obstruct social learning. We complement their work by modeling the origin of confidence distortions—metacognitive bias rather than strategic choice—and by studying how these distortions interact with network structure, information quality, and institutional correction mechanisms. In particular, we characterize how the cardinality of overconfident agents, the granularity of the knowledge domain, and the sparseness of truth jointly determine convergence outcomes.

Methodologically, our framework differs from canonical social learning models (Acemoglu et al., 2011; Banerjee, 1992; Bikhchandani et al., 1992) in which agents infer information from actions. Here, agents directly transmit claims, and persuasion depends on perceived credibility rather than revealed behavior. This distinction is crucial since miscalibration is directly observable through influence and thus directly shapes learning dynamics. Instead, our analysis builds on tools from statistical mechanics (Blume, 1993; Liggett, 1985), modeling knowledge diffusion as a continuous-time stochastic process and characterizing convergence and absorbing states in networks and lattices.

Finally, our work contributes to the literature on misinformation and belief persistence (Acemoglu et al., 2021; Levy & Razin, 2019). In contrast to models emphasizing strategic manipulation, polarization, or echo chambers, misinformation in our model arises purely from epistemic miscalibration among sincere agents. Overconfidence accelerates diffusion but increases error; connectivity improves learning in full-truth environments but amplifies misinformation when truth is sparse. Institutional correction restores accuracy only when agents retain memory, highlighting a fundamental tradeoff between speed and accuracy in decentralized learning.

By integrating empirically documented metacognitive biases into a tractable model of networked learning with misspecification, this paper provides a unified account of how individual

miscalibration scales into collective epistemic failure, even in the absence of strategic behavior or conflicting incentives.

### 3. LEARNING MODEL AND DYNAMICS

#### 3.1. The Model

Consider a finite population of agents indexed by  $i \in V = \{1, \dots, N\}$ . Time is continuous and indexed by  $t \in [0, \infty)$ . Agents meet over time and form a network.

We model interactions using an unweighted, undirected social network  $G = (V, E, t)$ , where  $V$  is a finite set of nodes (agents) with  $|V| = N$ , and  $E$  is a set of unordered pairs  $(u, v) \in V \times V$ , so that  $(u, v) = (v, u)$ . This symmetry implies that agents are simultaneously potential senders and receivers of information. Consequently, communication takes the form of bilateral discussions with the idea to convince or persuade one another that they are right in what they claim.

In each period  $t$ , the network is represented by a symmetric adjacency matrix  $A_t = [g_{t,ij}]_{i,j \in V}$ . A link between agents  $i$  and  $j$  exists if  $g_{t,ij} = 1$ , while  $g_{t,ij} = 0$  indicates the absence of a connection. We assume that agents do not communicate with themselves, which amounts to setting  $g_{t,ii} = 0$  for all time periods.

Let  $d_i(A_t) = \sum_j g_{t,ij}$  denote  $i$ 's degree at time  $t$ , which represents the number of connections of agent  $i$  in the network at time  $t$ .

**Definition 3.1** (Path). A path from  $i$  to  $j$  in  $G = (V, E, t)$  is a sequence of agents  $i = i_0, i_1, \dots, i_m = j$  such that  $(i_k, i_{k+1}) \in E_t$  for all  $k$ .

**Definition 3.2.**  $G = (V, E, t)$  is connected if there exists a path between every pair of agents  $i, j \in V$ .

**Definition 3.3** (Eventual connectedness). The network process is eventually connected if

$$\mathbb{P}[\exists T \text{ s.t. } G = (V, E, t) \text{ is connected } \forall_{t \geq T}] = 1$$

#### 3.2. Knowledge

The domain of all potentially knowable facts is the nonatomic unit interval  $X = [0, 1]$ . We partition  $X$  into  $P \in \mathbb{N}$  equal-length intervals with strictly positive measure such that

$$X = \bigcup_{l=1}^P I_l, \quad I_l = [\frac{l-1}{P}, \frac{l}{P})$$

where  $I_l$  is the  $l^{\text{th}}$  partition, and with the convention that  $I_P = [\frac{P-1}{P}, 1]$ . Note that points in the unit interval represent primitive informational units, so that knowledge can be viewed as a continuum of fine-grained elements. Partitioning the unit interval allows us to aggregate these elements into coherent knowledge domains. For example, if one partition corresponds to

classical mechanics, an agent does not need to acquire each underlying element individually. Instead, by learning the partition as a whole, she is treated as having broadly mastered the subject. This abstraction captures the idea that knowledge is fundamentally granular, yet acquired, transmitted, and evaluated in practice at a higher, thematic level.

Each agent  $i$  is endowed with a binary knowledge field

$$k_{i,t} : X \rightarrow \{0, 1\}$$

which is constant on each partition element  $I_l$ . Hence,  $k_{i,t}$  is a measurable step function and can be interpreted as a (non-probabilistic) knowledge density. For any fact indexed by  $x \in X$ ,

$$k_{i,t}(x) = \begin{cases} 1 & \text{if agent } i \text{ knows fact } x \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

The objective knowledge level of agent  $i$  at period  $t$  is defined as the average of her knowledge field:

$$K_{i,t} \equiv \int_0^1 k_{i,t}(x) dx = \frac{1}{P} \sum_{k=1}^P k_{i,t}(I_k) \quad (3.1)$$

Thus  $K_{i,t} \in [0, 1]$  represents the fraction of facts known by agent  $i$ . Moreover, agents may misperceive their own knowledge. Let

$$\beta_{i,t} = g(K_{i,t})$$

denote agent  $i$ 's subjective belief about her own knowledge at  $t$ . Importantly, agents are unaware of their own biases; thus, over- or underestimation of knowledge is not a deliberate choice but an unintended (endogenous) outcome of the learning process. Specifically,

$$\beta_{i,t} = r \sin 2\pi K_{i,t} + K_{i,t} \quad (3.2)$$

where  $r = \frac{1}{\alpha}$  indexes the degree of miscalibration, with  $\alpha > 1.36$  governing the degree of metacognitive discipline<sup>1</sup>. This functional form conveniently captures the three metacognitive states within a single smooth mapping on the unit interval:

1.  $g(0) = 0$  such that ignorance is recognized.
2.  $0 < K_{i,t} < \frac{1}{2}$  such that learning initially increases confidence and lead to overestimation.
3.  $\frac{1}{2} < K_{i,t} < 1$  such that learning reveals "ignorance"; humility valley.

---

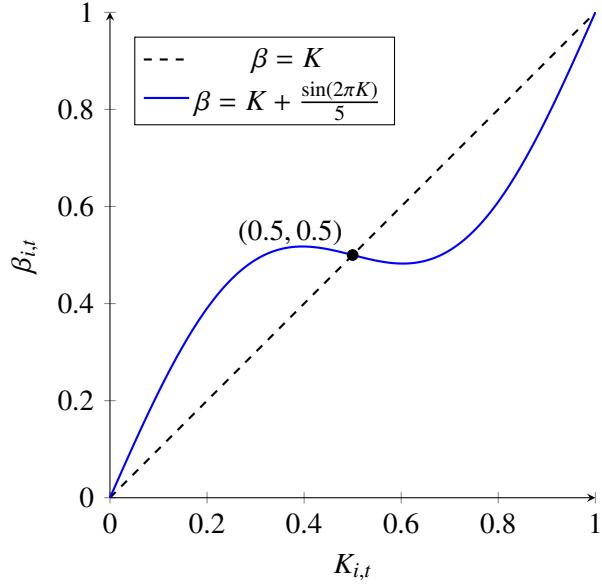
<sup>1</sup>The lower bound on  $\alpha$  ensures that  $\beta_{i,t} \in [0, 1]$  for all  $K_{i,t} \in [0, 1]$ .

4.  $\beta_{i,t} \rightarrow K_{i,t}$  as  $K_{i,t} \rightarrow 1$  such that "experts" regain confidence and are accurate in their beliefs.

Lower values of  $\alpha$  correspond to stronger miscalibration, generating pronounced overconfidence at low knowledge levels and underestimation at intermediate levels. As  $\alpha$  increases, subjective confidence increasingly aligns with objective knowledge, and in the limit  $\alpha \rightarrow \infty$ , agents are perfectly calibrated:  $\beta_{i,t} \rightarrow K_{i,t}$  (see Figure 3.1).

**Figure 3.1**

*Subjective confidence as a function of objective knowledge ( $\alpha = 5$ ).*



Define agent  $i$ 's bias as  $b_{i,t} \equiv \beta_{i,t} - K_{i,t}$ . Let

$$O_t \equiv \{i \in V : b_{i,t} > 0\} = \{i \in V : \beta_{i,t} > K_{i,t}\}$$

$$U_t \equiv \{i \in V : b_{i,t} < 0\} = \{i \in V : \beta_{i,t} < K_{i,t}\}$$

$$C_t \equiv \{i \in V : b_{i,t} = 0\} = \{i \in V : \beta_{i,t} = K_{i,t}\}$$

be the sets of agents who overestimate, underestimate, and accurately assess their knowledge, respectively, at each time period  $t$ , such that  $V = O_t \cup U_t \cup C_t$ . Note that agents' subjective perceptions of their own knowledge are independent of the truth. An individual may be highly knowledgeable yet hold a false belief (e.g., believing that  $2 + 2 = 3$ ), while at the same time being acutely aware of the breadth and complexity of knowledge and of the distinction between knowing and understanding. This awareness may lead her to underestimate or accurately assess her own knowledge and to remain open to others' information.

There exists an objective, agent-independent truth function chosen by Nature at  $t = 0$

$$k^* : X \rightarrow \{0, 1\}$$

unknown to all agents, such that

$$k^*(x) = \begin{cases} 1 & \text{if fact } x \text{ is true and knowable,} \\ 0 & \text{if fact } x \text{ is false, meaningless or nonexistent.} \end{cases}$$

for all  $t$ . This asymmetry between truth and ignorance is essential: without it, persuasion would be epistemically neutral<sup>2</sup>. Note that with epistemically neutral persuasion we mean

$$\mathbb{E}[k_{i,t}|\text{persuasion}] = \mathbb{E}[k_{i,t}|\text{no persuasion}]$$

such that learning and unlearning would be symmetric and that truth and falsehood would be identical. Basically the idea of the asymmetry is that if persuasion could just as easily destroy knowledge as create it, then talking would not make society any wiser on average.

When two agents interact, persuasion depends on relative perceived competence. Let

$$f : [-1, 1] \rightarrow [0, 1]$$

be an increasing, bounded function satisfying  $f(0) = \frac{1}{2}$ . We interpret  $f(\beta_{i,t} - \beta_{j,t})$  as the probability that agent  $j$  adopts agent  $i$ 's claim. Given these conditions, a canonical choice is the logistic sigmoid function:

$$f(\beta_{i,t} - \beta_{j,t}) = \frac{1}{1 + e^{-(\beta_{i,t} - \beta_{j,t})}} \quad (3.3)$$

Thus, more confident agents are more persuasive, even if their confidence is unwarranted. Authority is therefore endogenous, rather than assigned exogenously to experts<sup>3</sup>. When both agents coincide in their own subjective knowledge, they are indifferent whether to be "persuaded" or not. For simplicity, we will assume they are not persuaded and stick with their own claim.

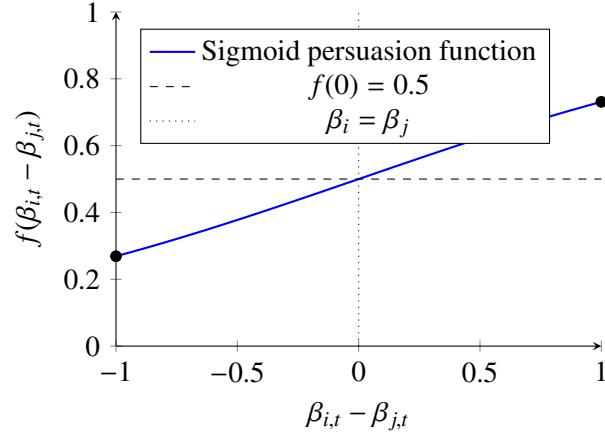
---

<sup>2</sup>Facts are either true or not true, but  $k_{i,t}(x) = 0$  means the agent lacks knowledge regardless of whether the fact is true or not true.

<sup>3</sup>Since  $\beta_{i,t}, \beta_{j,t} \in [0, 1]$ , the restricted domain is given by  $\beta_{i,t} - \beta_{j,t} \in [-1, +1]$  and thus,  $f(-1) \approx 0.27$  and  $f(+1) \approx 0.73$ . That is, the sigmoid is almost linear in this region (see Figure 3.2). That implies that persuasion is never impossible and never certain. Agents never have overwhelming authority.

**Figure 3.2**

Persuasion probability as a function of relative perceived competence.



Note that we assume agent  $i$  takes as true what agent  $j$  claims to know, that is,  $j$ 's subjective perception of her own knowledge  $\beta_{j,t}$ .<sup>4</sup> Fix an interval  $I_l$ . If agents  $i$  and  $j$  interact and

$$k_{i,t}(I_l) = 1, \quad k_{j,t}(I_l) = 0,$$

then agent  $j$  updates according to

$$k_{j,t+1}(I_l) \leftarrow \begin{cases} 1, & \text{w.p. } f(\beta_{i,t} - \beta_{j,t}) \\ k_{j,t}(I_l), & \text{otherwise} \end{cases} \quad (3.4)$$

If persuasion succeeds<sup>5</sup>:

- If  $k^*(I_l) = 1$ :  $j$  gains true knowledge.
- If  $k^*(I_l) = 0$ :  $j$  gains false knowledge.

Whenever  $k_{j,t}(I_l)$  switches from 0 to 1,  $K_{j,t}$  increases by  $\frac{1}{p}$  and  $\beta_{j,t}$  is updated according to Equation 3.2. Note that social interaction can only move information/knowledge from  $0 \rightarrow 1$ ; persuasion always adds knowledge to the information set, never removes it.

**Key tension.** Overconfident agents with high  $\beta_{i,t}$  but low  $K_{i,t}$  may be highly persuasive despite possessing little true knowledge.

<sup>4</sup>To clarify, agent  $i$  observes her own knowledge state  $(k_{i,t}, K_{i,t}, \beta_{i,t})$ . She does not observe agent  $j$ 's knowledge variables  $(k_{j,t}, K_{j,t})$ , but upon interaction (once a link  $(i, j)$  is formed), she observes  $\beta_{j,t}$ . The truth function  $k^*$  is unobserved by all agents. Because the subjectivity rule 3.2 is not known to agents, they cannot infer agent  $j$ 's underlying knowledge level  $K_{j,t}$  from the observed value of  $\beta_{j,t}$ .

<sup>5</sup>Technically, it would be with probability  $\lambda_{i,j} f(\beta_{i,t} - \beta_{j,t})$ , but we leave it as it is for notational convenience.

**Definition 3.4** (Interval-wide consensus). Agents reach consensus on interval  $I_l$  at time  $T$  if

$$k_{i,T}(I_l) = k_{j,T}(I_l), \quad \forall_{i,j \in V}$$

possibly with false consensus if  $k_{i,T}(I_l) \neq k^*(I_l)$ .

**Definition 3.5** (Pointwise agreement). Agents  $i$  and  $j$  agree if

$$k_{i,t}(x) = k_{j,t}(x), \quad \text{for almost every (a.e.) } x \in X$$

For each  $i \in V$  and interval  $l \in \{1, \dots, P\}$ , we consider a random initialization such that

$$k_{i,0}(I_l) \sim \text{Bernoulli}(p_l), \quad \text{i.i.d. across agents}$$

For simplicity, we assume homogeneous difficulty such that  $p_l$  is uniform:  $p_l = p$ , for all  $l$ <sup>6</sup>.

**Definition 3.6** (Persuasion event). At time  $t$ , on edge  $(i, j) \in E_t$ , for interval  $I_l$ , persuasion from  $i$  to  $j$  occurs if and only if:

- $k_{i,t}(I_l) = 1$  ( $i$  "knows" the fact).
- $k_{j,t}(I_l) = 0$  ( $j$  is ignorant).
- Persuasion probability  $f(\beta_{i,t} - \beta_{j,t})$  succeeds.

Upon persuasion,  $k_{j,t}(I_l) \leftarrow 1$ .

These properties of persuasion are:

- Irreversibility: once  $k_{j,t}(I_l) = 1$ , it remains 1 forever (knowledge is absorbing; there is no forgetting or unlearning).
- Asymmetry: if  $k_{i,t}(I_l) = 0$  and  $k_{j,t}(I_l) = 1$ , no persuasion from  $i$  to  $j$  can occur (persuasion can only occur from knowledgeable to ignorant agents).
- Unidirectionality: on an edge  $(i, j)$ , persuasion may occur from  $i \rightarrow j$  or  $j \rightarrow i$ , depending on their knowledge states, but not simultaneously for the same interval  $I_l$ . To see this, consider the case in which agents  $i$  and  $j$  both claim to "know" interval  $I_l$ . If  $k^*(I_l) = 1$ , since truth is unique and universal, then both agents possess the same correct knowledge (e.g., both know that  $2 + 2 = 4$  in the reals). By contrast, if  $k^*(I_l) = 0$ , the situation is fundamentally different: regardless of whether the agents' beliefs coincide or differ—agent  $i$  may "know" that  $2 + 2 = 3$  while agent  $j$  "knows" that  $2 + 2 = 5$  in the reals—both hold false knowledge. In this case, bidirectional persuasion is irrelevant, since no exchange between agents can produce true knowledge; what matters is not which false belief is held, but that all beliefs are incorrect. When agent is ignorant, he simply cannot transmit information.

---

<sup>6</sup>One could allow  $p_l$  to vary with  $l$ , interpreting larger indices as more advanced facts.

### 3.3. Dynamics

Similar to Jackson et al. (2023), we model communication opportunities in continuous time using a homogeneous Poisson random arrival (point) process. This approach implies that meetings between agents occur asynchronously. Almost surely, only one dyadic meeting occurs at any instant, and that the intensity of interactions between any given pair of agents is time independent.

For each unordered pair of distinct agents  $(i, j)$ , meetings occur according to a Poisson process with constant intensity  $\lambda_{i,j} > 0$ . Let  $N_{i,j}(t)$  denote the number of meetings between agents  $i$  and  $j$  up to time  $t$ . Homogeneity implies stationarity of increments, so the expected number of meetings between  $i$  and  $j$  over any time interval of length  $\Delta t$  is  $\lambda_{i,j}\Delta t$ . Thus,  $\lambda_{i,j}$  represents the expected meeting frequency per unit of time for that pair<sup>7</sup>.

Because interactions are undirected, meetings are modeled at the dyadic level. For each unordered pair  $(i, j)$  there is a single Poisson arrival process, rather than two directed processes<sup>8</sup>. Since  $\lambda_{i,j}$  is constant over time for all pairs, the arrival process is stationary. A special benchmark case is a symmetric network in which all pairs interact at the same rate, that is,  $\lambda_{i,j} = \lambda$  for all  $t$  and  $i \neq j$ .

Whenever a meeting occurs at the dyadic level, the corresponding link becomes active. If no link previously existed, the meeting results in link formation and activation; if the link already exists, the meeting simply reactivates the existing connection. Agents always consent to meetings, as they are uninformed about each other's knowledge states. As meetings accumulate over time, the set of active edges grows monotonically, and under standard connectivity conditions, the network almost surely becomes complete at some finite time  $T$ . That is, the network is monotonically growing such that:

$$E_{t_1} \subseteq E_{t_2}, \quad \forall_{t_1 \leq t_2}$$

### 3.4. Communication

Persuasion in the model is local in content, not global. Agents do not attempt to persuade each other about their aggregate knowledge levels. Instead, persuasion occurs interval by interval. Communication unfolds according to the following protocol:

---

<sup>7</sup>For any time interval  $[s, t]$  with  $t > s$ ,

$$\mathbb{E}[N_{i,j}(t) - N_{i,j}(s)] = \int_s^t \lambda_{ij} du = \lambda_{ij}(t-s), \quad \text{for } (t-s) = \Delta t$$

where  $(t-s)$  is the length of the interval.

<sup>8</sup>For example, if  $N = 5$ , there are  $\binom{5}{2} = \frac{5!}{2!3!} = 10$  distinct unordered pairs of agents. Since the network is undirected,  $(i, j) = (j, i)$ . We therefore assume that independent Poisson processes govern meetings at the level of these unordered pairs.

1. **Network formation:** links between agents are formed in continuous time according to a Poisson process. At  $t = 0$ , the network starts from an empty configuration with zero edges and the dynamics start once the first meeting happens. Each Poisson arrival corresponds to the activation of a single edge, and once formed, edges persist over time. Subsequent Poisson processes operate on the topology generated in the previous step, so that each realization builds on the existing network structure.
2. **Topic selection:** upon the formation/activation of an edge  $(i, j)$ , an interval  $I_l$  is drawn uniformly at random from  $\{I_1, \dots, I_P\}$ . This interval represents the topic discussed during that interaction.
3. **Local persuasion attempt:** agents compare their knowledge on the selected interval. If  $k_{i,t}(I_l) = 1$  and  $k_{j,t}(I_l) = 0$ , then agent  $i$  attempts to persuade agent  $j$  according to the persuasion rule 3.6.

This protocol implies that information transmission is: (1) sequential, as interactions occur one edge at a time; (2) local, as persuasion concerns a single fact at each interaction; and (3) path-dependent, since the evolving network structure shapes future communication opportunities.

**Definition 3.7** (Almost-sure convergence to full knowledge). The system converges almost surely (a.s.) to full knowledge if

$$\mathbb{P}[\exists T < \infty \text{ s.t. } k_{i,T}(I_l) = k^*(I_l), \forall_{i,l}] = 1$$

## 4. RESULTS

### 4.1. Benchmark case: Full truth

We begin with a benchmark environment in which all potentially knowable facts are true. Formally,

$$k^*(x) = 1, \quad \forall_{x \in X}$$

Thus every interval corresponds to meaningful facts, and ignorance is the only obstacle to knowledge. Epistemic consistency requires that agents cannot know false facts:

$$k_{i,t}(x) \leq k^*(x), \quad \forall_{x \in X}$$

Under full truth, this restriction implies that knowledge is veridical by construction, and agents know that. In particular,

$$k_{i,t}(x) = 1 \Rightarrow \text{agent } i \text{ truly knows } x \text{ at } t, \quad k_{i,t}(x) = 0 \Rightarrow \text{agent } i \text{ is ignorant of } x \text{ at } t$$

**Definition 4.1** (Full knowledge). The network achieves full knowledge at time  $T$  if

$$k_{i,T}(x) = 1, \quad \forall_{i \in V}, \quad \text{for a.e. } x \in X$$

which is equivalent to

$$K_{i,T} = 1, \quad \forall_{i \in V}$$

Let

$$K_{i,t} = \int_0^1 k_{i,t}(x) dx, \quad K^* = \int_0^1 k^*(x) dx.$$

Full knowledge implies  $K_{i,t} = K^*$  for all  $i$ . In the benchmark case,  $K^* = 1$ .

**Proposition 4.1** (Almost-sure convergence to full knowledge). *Suppose that*

1.  $k^*(x) = 1, \forall_{x \in X}$ ,
2. *Communication intensities satisfy:  $\lambda_{i,j} \geq \lambda > 0, \forall_{i,j,t}$ ,*
3. *Initial knowledge is i.i.d., such that  $k_{i,0}(I_l) \sim \text{Bernoulli}(p)$  with  $p > 0$ .*
4. *For all  $l \in \{1, \dots, P\}$ :  $\exists i$  such that  $k_{i,0}(I_l) = 1$ .*

Then with probability one, there exists  $T < \infty$  such that

1. (Convergence)

$$K_{i,t} = 1 \quad \forall_{i \in V}, \forall_{t \geq T}.$$

2. (Stability) the state  $\mathbf{k}^* = (k_{1,T}(x) = 1, \dots, k_{N,T}(x) = 1, \quad \forall_x)$  is absorbing.

The second condition means that no pair is permanently disconnected or arbitrarily unlikely to meet, where  $\lambda$  is a uniform positive lower bound. That is, every pair of agents communicates infinitely often, almost surely.

Note that this result is closely related to a steady state in statistical-mechanical models of ferromagnetism: the ferromagnetic phase, in which all spins align and the system exhibits complete consensus (Ising, 1925; Sznajd-Weron, 2005).

**Remark 4.2** (Speed of convergence). *Fix the number of agents, communication intensities  $\lambda_{i,j} \geq \lambda > 0, \forall_{i,j,t}$ , and the initial knowledge distribution. Let  $T_P$  denote the (random) time to reach full knowledge when the knowledge domain is partitioned into  $P$  intervals. Then:*

$$\mathbb{E}[T_P] \quad \text{is weakly increasing in } P$$

The intuition is that more intervals implies that there are more facts to learn and thus, longer convergence time. Each interval  $I_l$  must be independently transmitted across the network.

## 4.2. Sparse truth

We now consider an environment in which only a subset of the knowledge domain corresponds to true facts. Let

$$K^*(x) = \mathbf{1}_{x \in T}$$

where  $T \subseteq [0, 1]$  is a measurable set. Thus, facts indexed by  $x \in T$  are true and potentially knowable, while facts indexed by  $x \notin T$  are false, meaningless, or nonexistent. The total mass of truth is

$$K^* \equiv \int_0^1 K^*(x) dx = |T| < 1$$

Unlike the full-truth benchmark, agents may now agree on many claims and still be wrong by omission or commission. In this environment, full (correct) knowledge requires both coverage and precision. An agent must know all true facts and reject all false ones:

$$k_{i,t}(x) = \begin{cases} 1, & \text{if } x \in T \\ 0, & \text{if } x \notin T \end{cases}$$

Thus, full knowledge is no longer characterized by  $k_{i,t}(x) = 1$  almost everywhere, but by exact alignment with the truth function  $k^*$ . At the level of partitions, truth is specified as follows. For each interval  $l \in \{1, \dots, P\}$ :

$$k^*(I_l) \sim \text{Bernoulli}(\rho)$$

where  $\rho \in (0, 1)$  is the "truth density". This specification implies that the model permits false persuasion. Since agents are not allowed to "unlearn", once  $k_{j,t}(I_l)$  switches from 0 to 1, the false fact persists indefinitely. The system contains no intrinsic mechanism for truth verification.

**Definition 4.2** (Full (correct) knowledge). The network achieves full (correct) knowledge at time  $T$  if

$$k_{i,T}(x) = k^*(x), \quad \forall_{i \in V}, \quad \text{for a.e. } x \in X$$

Equivalently,

$$k_{i,T}(I_l) = k^*(I_l), \quad \forall_{i \in V}, \forall_{l \in \{1, \dots, P\}}.$$

**Proposition 4.3** (Full (correct) knowledge under sparse truth). *Suppose that  $K^* < 1$  (sparse truth), and the network becomes connected almost surely. If the initial knowledge distribution satisfies*

$$k_{i,0}(I_l) = 0 \quad \text{for all } i \in V \text{ and all } l \text{ such that } k^*(I_l) = 0,$$

*then with probability one there exists  $T < \infty$  such that*

1. (Convergence)

$$\mathbb{P}[k_{i,T}(x) = k^*(x), \quad \text{a.e. } \forall_i] = 1$$

2. (Stability) the state  $\mathbf{k}^* = \{k_{i,T}(I_l) = k^*(I_l), \forall_{i,l}\}$  is absorbing.

\*The intuition is that since all agents are ignorant about what is not true, and we do not allow for people to make up "stories", this result is analogous to the full-truth case.

**Remark 4.4.** *False knowledge is an absorbing contamination. Under irreversible persuasion, any false belief present initially prevents convergence to full correct knowledge with probability one.*

Intuitively, in the absence of any correction mechanism, once the system is initially "contaminated" with false knowledge, the fraction of agents who overestimate their knowledge from below becomes irrelevant: full correct knowledge is never attained.

**Proposition 4.5** (Misinformation Spread Under Overconfidence). *Suppose  $K^* < 1$ , the network becomes connected, and:*

1. Initial false beliefs:  $\exists l \in \{1, \dots, P\}$  with  $k^*(I_l) = 0$  and  $\exists i$  with  $k_{i,0}(I_l) = 1$ .

2. Overconfidence concentration:  $|O(0)| \geq \theta N$  for some  $\theta \in (0, 1)$
3. False belief holders are overconfident: If  $k_{i,0}(I_l) = 1$  and  $k^*(I_l) = 0$ , then  $i \in O(0)$ .

Then:

1. (Non-convergence)  $\mathbb{P}[\text{system reaches full correct knowledge}] = 0$ .
2. (Scaling with partitions) The expected number of intervals with false consensus is at least:

$$\mathbb{E} \left[ \sum_{l=1}^P \mathbf{1}\{k_i(I_l) \neq k^*(I_l) \text{ for some } i\} \right] \geq (1 - \rho) P h(\theta)$$

where  $\rho$  is truth density and  $h(\theta)$  is increasing in  $\theta$ .

3. (Speed of contamination) If  $\theta > 1/2$ , false beliefs spread faster than true beliefs: the expected time for false consensus on an interval with  $k^*(I_l) = 0$  is shorter than for true consensus on an interval with  $k^*(I_l) = 1$ .

**Proposition 4.6** (Overconfidence Amplifies Misinformation). Fix  $P, N$ , network dynamics, and initial knowledge. Let  $M(t) = \frac{1}{NP} \sum_{i,l} k_i(I_l)(1 - k^*(I_l))$  be the average false knowledge at time  $t$ . If  $g(K)$  satisfies the non-monotonic confidence condition in 3.2, then:

$$\frac{dE[M(t)]}{dt} \text{ is increasing in } |O(t)|$$

for small  $t$ .

That is, more overconfident agents leads to faster accumulation of false beliefs initially.

### 4.3. Truth revelation

Now suppose there exists a correction mechanism. The baseline interaction remains almost unchanged: a Poisson meeting occurs between agents  $i$  and  $j$ , an interval  $I_l$  is selected and (external) persuasion proceeds but now from  $0 \leftrightarrow 1$ . However, after the interaction, one of the two agents chosen uniformly at random receives a truth signal regarding the interval they discussed,  $I_l$ . This truth signal reveals  $k^*(I_l)$  to her. Specifically, (1) with probability  $\varphi \in (0, 1)$ , Nature reveals  $k^*(I_l)$  to one agent chosen uniformly at random from the Poisson-activated edge  $(i, j)$ ; (2) the revelation is private: the other agent does not observe it; (3) the truth signal reveals only the binary state  $k^*(I_l) \in \{0, 1\}$ , not the semantic content of  $I_l$ <sup>9</sup>.

---

<sup>9</sup>That is, if  $k^*(I_l) = 1$ , the signal reveals that possessing knowledge of  $I_l$  corresponds to having true knowledge, while if  $I_l$  is currently unknown, acquiring it would result in true knowledge. Conversely if  $k^*(I_l) = 0$ .

The only way an agent can transmit and persuade  $k_{i,t}(I_l) = 0$  is when  $k_{i,t-1}(I_l) = 1$ , which implies she got verified information  $v_{i,t-1} = 1$  such that  $k^*(I_l) = 0^{10}$ . Whenever two agents can transmit information to each other, the persuasion rules are computed for both.

**Table 4.1**

*Truth revelation and voluntary correction*

Agent belief $k_{i,t}(I_l)$	Truth $k^*(I_l)$	Agent response
0	1	No immediate change
0	0	No change
1	1	Lock-in
1	0	Voluntary reversal to 0

Table 4.1 summarizes agents' responses under the correction mechanism. Suppose agent  $i$  receives a truth signal for interval  $I_l$  after the interaction with agent  $j$ :

1. **Ignorance of a true fact:** If  $k_{i,t}(I_l) = 0$  and  $k^*(I_l) = 1$ , agent  $i$  learns that her belief/claim is incorrect but cannot update immediately. Knowledge acquisition requires social transmission, so she must wait to meet a more confident agent than her and acquire  $I_l$  through the standard persuasion dynamics. Self-learning is not allowed<sup>11</sup>. Once agent  $i$  eventually acquires  $I_l$ , she becomes immune to persuasion on this interval: having observed the truth, she will neither voluntarily revert to  $k_{i,t}(I_l) = 0$  nor be externally persuaded to do so.
2. **Correct ignorance:** If  $k_{i,t}(I_l) = 0$  and  $k^*(I_l) = 0$ , agent  $i$ 's belief is correct. Upon receiving the truth signal, she confirms this correctness and records  $I_l$  as verified knowledge. Although she is correct, she remains susceptible to future persuasion from  $0 \rightarrow 1$  if she was memoryless. In this section we assume agents have memory, so agent  $i$  becomes immune to future persuasion on  $I_l$  and will never switch to  $k_{i,t}(I_l) = 1$ .
3. **Correct belief:** If  $k_{i,t}(I_l) = 1$  and  $k^*(I_l) = 1$ , agent  $i$  learns that her belief is correct. Observing the truth renders agent  $i$  immune to future persuasion on  $I_l$ : she will never revert to  $k_{i,t}(I_l) = 0$ , neither voluntarily nor externally.
4. **False belief and voluntary correction:** If  $k_{i,t}(I_l) = 1$  and  $k^*(I_l) = 0$ , the truth signal reveals that agent  $i$  holds a false belief. She voluntarily revises her belief and switches to  $k_{i,t}(I_l) = 0$ . "Unlearning" is only allowed internally; it is a personal and voluntary action. Due to memory, agent  $i$  permanently records the falsity of  $I_l$  and will never adopt  $k_{i,t}(I_l) = 1$  in future interactions, and can spread  $k_{i,t}(I_l) = 0$ .

---

<sup>10</sup>Note that when  $k_{i,t}(I_l) = 0$  and agent  $i$  has not received a truth revelation, the value 0 represents ignorance; she has no information about  $I_l$ . Once the falsehood of  $I_l$  is revealed, the same value 0 reflects informed knowledge: the agent knows that  $I_l$  is false but still does not have content. Transition  $1 \rightarrow 0$  is only possible when agent  $i$  had information about  $I_l$  but she realized is false; she has information, content to share.

<sup>11</sup>The reason for this assumption is that observing is different from understanding. There is truth revelation, but agents need content of the fact, and this takes time.

Throughout, we assume agents are truth-oriented: once the true state of an interval is revealed, they update their knowledge accordingly and do not knowingly deviate from the truth. Knowledge reversals can occur only through truth revelation, not through persuasion. Moreover, in this section we assume that agents perfectly recall truth signals received in previous periods.

Let the state space for each agent  $i$  and interval  $l$  be:  $(k_{i,t}(I_l), v_{i,t}(I_l)) \in \{0, 1\}^2$  where  $k_{i,t}(I_l)$  is  $i$ 's knowledge and  $v_{i,t}(I_l)$  is the verification status. Let also for each agent  $i$

$$K_{i,t}^T \equiv \frac{1}{P} \sum_l^P |\{l : k_{i,t}(I_l) = k^*(I_l)\}|, \quad K_{i,t}^E \equiv \frac{1}{P} \sum_{l=1}^P \max\{k_{i,t}(I_l), v_{i,t}(I_l)\}$$

be her true knowledge and her total epistemic exposure, respectively. For each interval there are four possible states:

**Table 4.2**

*Knowledge and Verification States for Interval  $I_l$*

Knowledge status	Verification status	
	0 (unverified)	1 (verified)
0 (ignorant)	Ignorant, unverified	Ignorant but verified
1 (claims knowledge)	Unverified knowledge	Verified knowledge

In this case,

$$\beta_{i,t} = g(K_{i,t}^E) = r \sin 2\pi K_{i,t}^E + K_{i,t}^E \quad (4.1)$$

Given this, the transition dynamics are summarized in Figure 1. Note that with memory, there are two absorbing states:  $(1, 1)$  verified truth  $k^*(I_l) = 1$  and  $(0, 1)$  with  $k^*(I_l) = 0$  verified rejection of falsehood. Then,

**Definition 4.3** (Full correct knowledge). The network achieves full (correct) knowledge at time  $T$  if

$$\forall_i, \forall_l : (k_{i,T}(I_l), v_{i,T}(I_l)) = \begin{cases} (1, 1), & \text{if } k^*(I_l) = 1 \vee k^*(I_l) = 0 \\ (0, 1), & \text{if } k^*(I_l) = 0 \end{cases}$$

Note that truth revelation changes  $K_{i,t}^E$ , which changes  $\beta_{i,t}$ , which feeds back into persuasion. So agents who learn truth gain authority and confidence becomes earned over time. This creates endogenous epistemic hierarchy, not assumed experts.

**Proposition 4.7** (Almost-Sure Convergence with Truth Revelation and Memory). *Suppose truth revelation occurs with probability  $\varphi > 0$ , agents have perfect memory, and the network becomes connected a.s. Then:*

$$P \left[ \exists T < \infty : (k_i(I_l), v_i(I_l)) = \begin{cases} (1, 1) & \text{if } k^*(I_l) = 1 \vee k^*(I_l) = 0 \\ (0, 1) & \text{if } k^*(I_l) = 0 \end{cases} \quad \forall_{i,l} \text{ and } \forall_{l \geq T} \right] = 1$$

*That is, with probability one, the system converges to full correct knowledge.*

#### 4.3.1. Memoryless agents

In this case, agents do not retain truth revelations across periods. That is, observing the truth leads to belief revision but not immunity to future persuasion; subsequent belief changes depend only on the current belief and realized interactions. As a result, in the first case described above, agent  $i$  learns the true state at time  $t$ , but to remain correct she must continue to learn or be persuaded in subsequent periods. From period  $t + 1$  onward, she no longer remembers her revelation at time  $t$ . Consequently, to “lock in” this correct knowledge, she must again experience the third case, in which the truth is revealed and confirms that  $k_{i,t}(I_l) = 1$ .

More generally, without memory, correction is not absorbing and agents do not acquire permanent epistemic immunity. By contrast, in the memory case, the system admits absorbing states: full and correct knowledge is achievable almost surely, and long-run learning is guaranteed. In the memoryless setting, full correct knowledge is generally not stable, and the process converges, if at all, to a stationary distribution rather than a fixed point. See Figure 2 to visualize the dynamics of the memoryless case.

**Proposition 4.8** (Failure of Convergence without Memory). *Suppose truth revelation occurs with probability  $\varphi > 0$ , but agents are memoryless. Then full correct knowledge is not reached almost surely. Instead, the system converges to a stationary distribution  $\pi$  where:*

$$\pi(\text{all agents correct on all intervals}) < 1$$

See Appendix C for an extension to statistical mechanics and a connection to interacting particle systems.

## BIBLIOGRAPHY

- Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, 78(4), 1201–1236.
- Acemoglu, D., Egorov, G., & Sonin, K. (2021). Institutional change and institutional persistence. In *The handbook of historical economics* (pp. 365–389). Elsevier.
- Acemoglu, D., & Ozdaglar, A. (2011). Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1), 3–49.
- Acemoglu, D., Ozdaglar, A., & Siderius, J. (2024). A model of online misinformation. *Review of Economic Studies*, 91(6), 3117–3150.
- Anderson, C., & Kilduff, G. J. (2009). Why do dominant personalities attain influence in face-to-face groups? the competence-signaling effects of trait dominance. *Journal of personality and social psychology*, 96(2), 491.
- Arieli, I., Babichenko, Y., Müller, S., Pourbabae, F., & Tamuz, O. (2025). The hazards and benefits of condescension in social learning. *Theoretical Economics*, 20(1), 27–56.
- Arrow, K. J. (1962). The economic implications of learning by doing. *The review of economic studies*, 29(3), 155–173.
- Ba, C., Bohren, J. A., & Imas, A. (2024). Over-and underreaction to information. Available at SSRN 4274617.
- Banerjee, A. V. (1992). A simple model of herd behavior. *The quarterly journal of economics*, 107(3), 797–817.
- Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5), 992–1026.
- Blume, L. E. (1993). The statistical mechanics of strategic interaction. *Games and economic behavior*, 5(3), 387–424.
- Bohren, J. A., & Hauser, D. N. (2021). Learning with heterogeneous misspecified models: Characterization and robustness. *Econometrica*, 89(6), 3025–3077.
- Bohren, J. A., & Hauser, D. N. (2024). Misspecified models in learning and games. *Annual Review of Economics*, 17.
- Cho, I.-K., & Libgober, J. (2025). Learning underspecified models. *Journal of Economic Theory*, 106015.
- Clance, P. R., & Imes, S. A. (1978). The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *Psychotherapy: Theory, research & practice*, 15(3), 241.
- Crollen, V., Castronovo, J., & Seron, X. (2011). Under-and over-estimation. *Experimental psychology*.
- Dunning, D. (2011). The dunning–kruger effect: On being ignorant of one’s own ignorance. In *Advances in experimental social psychology* (pp. 247–296, Vol. 44). Elsevier.

- Ericsson, K., Roring, R. W., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High ability studies*, 18(1), 3–56.
- Feld, J., Sauermann, J., & De Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of behavioral and experimental economics*, 68, 18–24.
- Granovsky, B. L., & Madras, N. (1995). The noisy voter model. *Stochastic Processes and their applications*, 55(1), 23–43.
- Harris, T. E. (1974). Contact interactions on a lattice. *The Annals of Probability*, 2(6), 969–988.
- He, K., & Libgober, J. (2025). Misspecified learning and evolutionary stability. *Journal of Economic Theory*, 106082.
- Howitt, P., & Aghion, P. (1998). Capital accumulation and innovation as complementary factors in long-run growth. *Journal of Economic Growth*, 3(2), 111–130.
- Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1), 253–258.
- Jackson, M. O., Yariv, L., Snowberg, E., & Nei, S. (2023). *A hedonic model of the dynamic formation of networks and homophily* [Work in progress].
- Jansen, R., Rafferty, A., & Griffiths, T. (2021). A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *nature human behaviour*, 5 (6), 756–763.
- Kausel, E. E., Carrasco, F., Reyes, T., Hirmas, A., & Rodríguez, A. (2021). Dynamic overconfidence: A growth curve and cross lagged analysis of accuracy, confidence, overestimation and their relations. *Thinking & Reasoning*, 27(3), 417–444.
- Kennedy, A., LaVail, K., Nowak, G., Basket, M., & Landry, S. (2011). Confidence about vaccines in the united states: Understanding parents’ perceptions. *Health affairs*, 30(6), 1151–1159.
- Knof, H., Berndt, M., & Shiozawa, T. (2024). Prevalence of dunning-kruger effect in first semester medical students: A correlational study of self-assessment and actual academic performance. *BMC Medical Education*, 24(1), 1210.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6), 1121.
- Levy, G., & Razin, R. (2019). Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics*, 11(1), 303–328.
- Liggett, T. M. (1985). *Interacting particle systems* (Vol. 2). Springer.
- McIntosh, R. D., Fowler, E. A., Lyu, T., & Della Sala, S. (2019). Wise up: Clarifying the role of metacognition in the dunning-kruger effect. *Journal of Experimental Psychology: General*, 148(11), 1882.
- Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic bulletin & review*, 24(6), 1774–1784.
- Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, 98(5, Part 2), S71–S102.

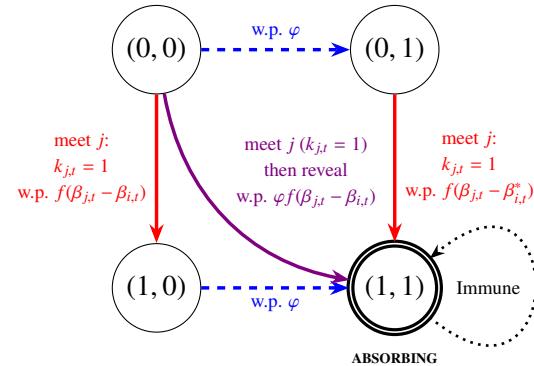
- Sakulku, J. (2011). The impostor phenomenon. *The Journal of Behavioral Science*, 6(1), 75–97.
- Sulphey, M., & Senan, N. A. M. (2025). Miscalibrations in self-evaluation: The influence of dunning–kruger effect among educators. *Higher Education for the Future*, 23476311251348773.
- Sznajd-Weron, K. (2005). Sznajd model and its applications. *arXiv preprint physics/0503239*.
- Teo, N. Q., Lim, T., & Tong, E. M. (2023). The humble estimate: Humility predicts higher self-assessment accuracy. *British Journal of Social Psychology*, 62(1), 561–582.

A.

**Figure 1**

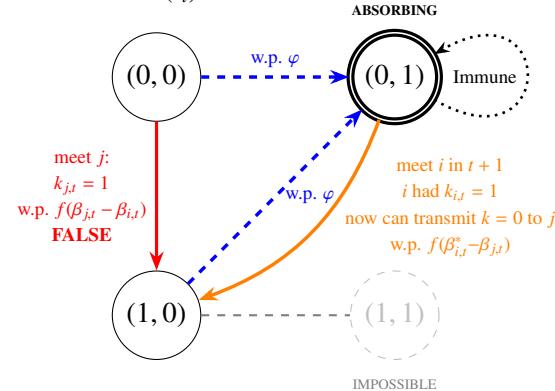
Complete state transition dynamics with memory. **Left:** Dynamics when  $I_l$  is true ( $k^*=1$ ). All paths lead to absorbing state  $(1, 1)$  where all agents have verified true knowledge. **Right:** Dynamics when  $I_l$  is false ( $k^*=0$ ). False knowledge can spread (red arrow), but revelation allows unlearning. Agents in  $(0, 1)$  who previously had the false belief can persuade others to unlearn (orange arrow). Absorbing state is  $(0, 1)$  where all agents know  $I_l$  is false.

**Case:  $k^*(I_l) = 1$**



**(a) Absorbing state is  $(1, 1)$**

**Case:  $k^*(I_l) = 0$**

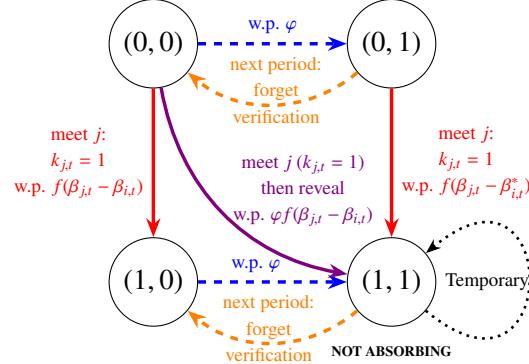


**(b) Absorbing state is  $(0, 1)$**

**Figure 2**

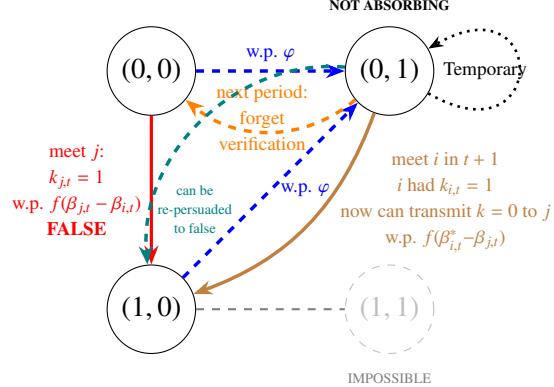
Complete state transition dynamics with memoryless agents. **Left:** Dynamics when  $I_l$  is true ( $k^* = 1$ ). Verification is temporary—agents in  $(0, 1)$  or  $(1, 1)$  forget verification in next period (orange dashed arrows) and return to  $(0, 0)$  or  $(1, 0)$ . No absorbing states. **Right:** Dynamics when  $I_l$  is false ( $k^* = 0$ ). Agents can learn false beliefs (red), be temporarily corrected via revelation (blue dashed), but forget verification (orange dashed) and can be re-persuaded to false beliefs (teal dashed). Agents in  $(0, 1)$  who previously had false belief can still persuade others to unlearn (brown arrow), but this correction is also temporary.

**Case:  $k^*(I_l) = 1$**



**(a) No absorbing state (verification forgotten)**

**Case:  $k^*(I_l) = 0$**



**(b) No absorbing state (can be re-persuaded)**

**Figure 3**

Comparison of memory vs. memoryless dynamics

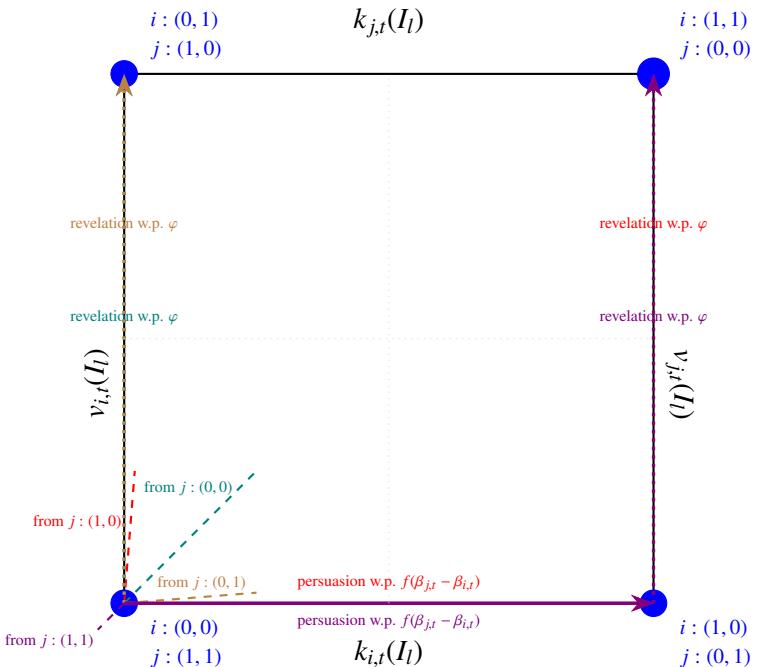
Feature	Memory	Memoryless
Absorbing states exist?	Yes	No
$(1, 1)$ with $k^*(I_l) = 1$	Absorbing (immune)	Temporary (forget)
$(0, 1)$ with $k^*(I_l) = 0$	Absorbing (immune)	Temporary (forget)
Can be re-persuaded to false belief?	No (if verified)	Yes (after forgetting)
Long-run convergence	To truth	Stationary distribution
Verification persists?	Forever	Only current period
<b>Key implication</b>		
	Truth eventually wins	Error persists indefinitely
	Almost-sure convergence	Cycles between states
	Institutional memory critical	No permanent learning

## B. UNFINISHED

In this section, we illustrate the interaction between agents  $i$  and  $j$  using an Edgeworth box. Note, however, that these graphs are primarily illustrative and are most useful in settings with continuous individual knowledge levels. This representation is "convenient" and feasible because communication occurs bilaterally, with only one interaction taking place at a time.

**Figure 4**

Edgeworth box representation for  $k^*(I_l) = 1$ . Agent  $i$ 's state  $(k_{i,t}, v_{i,t})$  is shown in standard orientation; agent  $j$ 's state would be read from inverted axes. Dashed lines show partner's initial state. Solid arrows show each agent's transitions from persuasion. Dotted arrows show transitions from revelation.



### C. Connection to interacting particle systems

This section establishes the formal connection between our knowledge diffusion model in a lattice and the theory of interacting particle systems (IPS)<sup>12</sup>. We show that our dynamics can be represented as a continuous-time Markov chain with an explicit generator, allowing us to apply powerful techniques from IPS theory to characterize convergence, stationary distributions, and scaling limits.

Our model operates on a growing complete graph where agents meet pairwise according to independent Poisson processes. This differs from many classical IPS models studied on spatial lattices  $\mathbb{Z}^d$  where agents interact only with spatial neighbors. However, the mathematical framework is identical: what matters for IPS theory is that we have a continuous-time Markov chain on a finite or countable state space with well-defined transition rates. The complete graph structure simplifies analysis (ensuring connectivity) while preserving the essential features of decentralized social learning. Extensions to spatial networks are discussed in Subsection 3.6.

For a single interval  $I_l$  at time  $t$ , we represent the population state as a configuration

$$\eta_t = (\eta_t(1), \eta_t(2), \dots, \eta_t(N)) \in \mathcal{S}^N \quad (2)$$

where  $\mathcal{S} = \{0, 1, 2, 3\}$  encodes the joint knowledge-verification state of each agent:

$$\eta_t(i) = 0 \Leftrightarrow (k_{i,t}(I_l), v_{i,t}(I_l)) = (0, 0) \quad (\text{ignorant, unverified}) \quad (3)$$

$$\eta_t(i) = 1 \Leftrightarrow (k_{i,t}(I_l), v_{i,t}(I_l)) = (0, 1) \quad (\text{ignorant, verified truth state}) \quad (4)$$

$$\eta_t(i) = 2 \Leftrightarrow (k_{i,t}(I_l), v_{i,t}(I_l)) = (1, 0) \quad (\text{knowledge, unverified truth state}) \quad (5)$$

$$\eta_t(i) = 3 \Leftrightarrow (k_{i,t}(I_l), v_{i,t}(I_l)) = (1, 1) \quad (\text{knowledge, verified truth state}) \quad (6)$$

The full state space is  $X = \mathcal{S}^N = \{0, 1, 2, 3\}^N$ , with  $|X| = 4^N$  possible configurations<sup>13</sup>. For notational convenience, we denote by  $\eta^{i,s}$  the configuration obtained from configuration  $\eta$  by changing agent  $i$ 's state to  $s \in \mathcal{S}$ <sup>14</sup>:

$$\eta^{i,s}(j) = \begin{cases} s & \text{if } j = i \\ \eta(j) & \text{if } j \neq i \end{cases} \quad (7)$$

---

<sup>12</sup>See Liggett (1985); and Blume (1993) for an extension to strategic environments.

<sup>13</sup>Note that for all pieces of knowledge, the full state space is given by  $X = \mathcal{S}^{N \times P} = \{0, 1, 2, 3\}^{N \times P}$ , with  $|X| = 4^{N \times P}$  possible configurations.

<sup>14</sup>Note that in a continuous-time Markov model, transitions look like: “Agent  $i$  updates her state from what it was to something else.” but only one agent changes at a time, and everyone else stays exactly the same. So we take the current configuration  $\eta$ , and change only agent  $i$  to state  $s$ .

### C.1. The Infinitesimal Generator

The knowledge diffusion process  $\{\eta_t\}_{t \geq 0}$  is a continuous-time Markov chain on  $X$ . Its law is uniquely determined by the infinitesimal generator<sup>15</sup>  $\Omega$ , which acts on functions  $\phi : X \rightarrow \mathbb{R}$  according to<sup>16</sup>:

$$\Omega\phi(\eta) = \sum_{\eta' \in X} c(\eta, \eta')[\phi(\eta') - \phi(\eta)] \quad (8)$$

where  $c(\eta, \eta')$  is the transition rate from configuration  $\eta$  to  $\eta'$ . The function  $\phi$  is a test function that specifies the value of some quantity when the system is in configuration  $\eta$ <sup>17</sup>. It is simply a "measurement" as the system evolves.

The product  $\Omega\phi(\eta)$  tells us the instantaneous rate of change of  $\phi$  when the system is in state  $\eta$ . There are two equivalent ways to describe how a Markov process evolves in time:

1. Globally in time via the semigroup  $(T_t)_{t \geq 0}$

$$(T_t\phi)(\eta) = \mathbb{E}_\eta[\phi(X_t)]$$

2. Infinitesimally in time via a differential operator  $\Omega$  that tells us what is the first-order effect of letting the process run for a very small time.

**Definition 4.** Let  $(T_t)_{t \geq 0}$  be a strongly continuous contraction semigroup on  $C_0(E)$  (consisting on continuous functions that vanish at infinity). The infinitesimal generator  $\Omega$  is defined by

$$\Omega\phi := \lim_{\Delta t \rightarrow 0} \frac{T_{\Delta t}\phi - \phi}{\Delta t}$$

for all  $\phi$  such that the limit exists in the sup norm. The set of such functions is the domain  $D(\Omega)$ .

<sup>15</sup>This is the continuous-time analogue of the transition matrix of a discrete-time Markov chain.

<sup>16</sup>Note that the infinitesimal operator, instead of directly specifying the probability of being in state  $\eta'$  at time  $t$ , the generator tells us how fast any observable quantity changes as the system evolves.

<sup>17</sup>For instance, how many agents know fact  $I_l$ ?

$$\phi(\eta) = \sum_{i=1}^N \mathbb{I}_{\{\eta(i) \in \{2, 3\}\}}$$

which counts agents in states 2 or 3. That is, those with  $k_i(I_l) = 1$ ; or is agent 5 in state  $(1, 1)$ ?

$$\phi(\eta) = \mathbb{I}_{\{\eta(5)=3\}}$$

which returns 1 if yes, 0 if no; or what fraction of agents are overconfident?

$$\phi(\eta) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{\beta(\eta(i)) > K(\eta(i))\}}$$

or total knowledge in the system:

$$\phi(\eta) = \sum_{i=1}^N \sum_{l=1}^P k_i(I_l)$$

where knowing  $\Omega$  determines  $(T_t)$  uniquely by Hille–Yosida’s theorem. Then, for  $\phi \in D(\Omega)$ ,

$$\Omega\phi(\eta) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{E}_\eta[\phi(\eta_{\Delta t})\eta_0 = \eta] - \phi(\eta)}{\Delta t}$$

and so  $\Omega\phi(\eta)$  is the instantaneous rate of change of the expected value of  $\phi(\eta_{\Delta t})$  when starting from  $\eta$ . Note that since  $C_0(E)$  is a complete normed space, it is a Banach space.

Note also that a Feller process is a Markov process whose semigroup behaves well with respect to topology. That is, small changes in the starting point produce small changes in expectations of nice functions.

**Definition 5.** Let  $E$  be locally compact, separable metric. A Markov process is a Feller process if:

1. State-space regularity

$$T_t(C_0(E)) \subset C_0(E)$$

2. Strong continuity

$$\lim_{t \rightarrow 0} \|T_t\phi - \phi\|_\infty = 0$$

Generators of Feller processes satisfy:

1. Positive maximum principle: if  $\phi \in D(\Omega)$  and  $\phi$  attains a global maximum at  $\eta_0$ , then

$$\Omega\phi(\eta_0) \leq 0$$

2. Hille–Yosida conditions: an operator  $\Omega$  generates a Feller semigroup if and only if (1)  $\Omega$  is densely defined, (2)  $\Omega$  is closed, (3) resolvents are positive contractions and (4) positive maximum principle holds.

Let us recall the reader that a semigroup of operators is a family  $(T_t)_{t \geq 0}$  such that:

1. Identity at zero:  $T_0 = \mathbb{I}$
2. Composition rule:  $T_{t+s} = T_t \circ T_s \quad \forall_{t,s \geq 0}$

That is, running the evolution for time  $s$ , then for time  $t$ , is the same as running it once for time  $t + s$ . And each  $T_t$  is a contraction on  $C_0(E)$  since

$$\|T_t\phi\|_\infty \leq \|\phi\|_\infty \quad \forall \phi \in C_0(E)$$

such that  $T_t$  does not amplify the size of functions: expectations don’t blow up.

### Transition Rates: Memory Case with $k^*(I_l) = 1$

We decompose the generator into two components:  $\Omega = \Omega^{\text{persuasion}} + \Omega^{\text{revelation}}$ . For the persuasion transitions, when edge  $(i, j)$  activates (at rate  $\lambda_{i,j}$  according to the Poisson process), persuasion may occur depending on the states of both agents. We know from Equation 3.3 that the persuasion probability is given by:

$$f_{j,i}(\eta) := f(\beta(\eta(j)) - \beta(\eta(i))) \quad (9)$$

where  $\beta(s)$  is the confidence level associated with state  $s$ , as defined in Equation 4.1. However, a crucial complication arises. In the full model, confidence depends on an agent's total epistemic exposure  $K_{i,t}^E$  across all intervals, not just the state on interval  $I_l$ . For exposition, we make a simplifying assumption for the single-interval analysis: treat confidence as depending only on the state  $s \in \{0, 1, 2, 3\}$  for interval  $I_l$ , with:

$$\beta(s) \approx \begin{cases} \beta_0 & \text{if } s = 0 \quad (\text{baseline confidence for ignorant agents}) \\ \beta_0 + \Delta\beta_v & \text{if } s = 1 \quad (\text{enhanced confidence from verification}) \\ \beta_0 + \Delta\beta_k & \text{if } s = 2 \quad (\text{confidence from unverified knowledge}) \\ \beta_0 + \Delta\beta_k + \Delta\beta_v & \text{if } s = 3 \quad (\text{confidence from verified knowledge}) \end{cases} \quad (10)$$

where  $\beta_0$  represents confidence from knowledge on other intervals, and  $\Delta\beta_k, \Delta\beta_v > 0$  are increments from knowledge and verification on  $I_l$ . This approximation allows us to analyze single-interval dynamics while acknowledging that the full model exhibits richer coupling across intervals through the endogenous confidence function  $\beta_{i,t} = g(K_{i,t}^E)$ .

For simplicity in exposition, assume symmetric meetings:  $\lambda_{i,j} = \lambda$  for all  $i \neq j$ . The persuasion component is:

$$\Omega^{\text{persuasion}} \phi(\eta) = \lambda \sum_{i \neq j} \sum_{s,r,s'} \mathbb{I}_{\{\eta(i)=s, \eta(j)=s'\}} \cdot p_{ss'}^r \cdot [\phi(\eta^{i,r}) - \phi(\eta)] \quad (11)$$

where  $p_{ss'}^r$  is the probability of transitioning from state  $s$  to state  $r$  given partner in state  $s'$ . The explicit rates for key transitions (agent  $i$  changes state) are:

$$c(\eta, \eta^{i,2}) = \lambda \sum_{j: \eta(j) \in \{2,3\}} \mathbb{I}_{\{\eta(i)=0\}} \cdot f_{j,i}(\eta) \quad (\text{ignorant learns from knowledgeable}) \quad (12)$$

$$c(\eta, \eta^{i,3}) = \lambda \sum_{j: \eta(j) \in \{2,3\}} \mathbb{I}_{\{\eta(i)=1\}} \cdot f_{j,i}(\eta) \quad (\text{verified ignorant acquires content}) \quad (13)$$

After each meeting, one agent receives a truth signal with probability  $\varphi$ . This adds:

$$\Omega^{\text{revelation}} \phi(\eta) = \frac{\lambda\varphi}{2} \sum_{i \neq j} \left[ \sum_{r \in R(\eta(i))} [\phi(\eta^{i,r}) - \phi(\eta)] \right] \quad (14)$$

where  $R(s)$  is the set of states reachable from  $s$  via revelation:

$$R(s) = \begin{cases} \{1\} & \text{if } s = 0 \quad (\text{ignorant learns truth exists or doesn't}) \\ \emptyset & \text{if } s = 1 \quad (\text{already verified}) \\ \{3\} & \text{if } s = 2 \quad (\text{unverified knowledge gets verified}) \\ \emptyset & \text{if } s = 3 \quad (\text{already verified}) \end{cases} \quad (15)$$

The explicit revelation rates are:

$$c(\eta, \eta^{i,1}) = \frac{\lambda\varphi}{2} \sum_{j \neq i} \mathbb{I}_{\{\eta(j)=0\}} \quad (\text{ignorant gets verification}) \quad (16)$$

$$c(\eta, \eta^{i,3}) = \frac{\lambda\varphi}{2} \sum_{j \neq i} \mathbb{I}_{\{\eta(j)=2\}} \quad (\text{knowledge gets verified}) \quad (17)$$

Hence, the complete generator for  $k^*(I_l) = 1$  (memory case):

$$\begin{aligned} \Omega\phi(\eta) = \lambda \sum_{i \neq j} \left[ & \mathbb{I}_{\{\eta(i)=0, \eta(j)\geq 2\}} \cdot f_{j,i}(\eta) \cdot [\phi(\eta^{i,2}) - \phi(\eta)] \right. \\ & + \mathbb{I}_{\{\eta(i)=1, \eta(j)\geq 2\}} \cdot f_{j,i}(\eta) \cdot [\phi(\eta^{i,3}) - \phi(\eta)] \\ & + \frac{\varphi}{2} \mathbb{I}_{\{\eta(i)=0\}} \cdot [\phi(\eta^{i,1}) - \phi(\eta)] \\ & \left. + \frac{\varphi}{2} \mathbb{I}_{\{\eta(i)=2\}} \cdot [\phi(\eta^{i,3}) - \phi(\eta)] \right] \end{aligned} \quad (18)$$

### Transition Rates: Memory Case with $k^*(I_l) = 0$

The persuasion component is given by Equation 11, and the explicit rates for key transitions (agent  $i$  changes state) are:

$$c(\eta, \eta^{i,2}) = \lambda \sum_{j: \eta(j) \in \{2,3\}} \mathbb{I}_{\{\eta(i)=0\}} \cdot f_{j,i}(\eta) \quad (\text{ignorant learns from knowledgeable}) \quad (19)$$

$$c(\eta, \eta^{i,3}) = \lambda \sum_{j: \eta(j) \in \{2,3\}} \mathbb{I}_{\{\eta(i)=1\}} \cdot f_{j,i}(\eta) \quad (\text{verified ignorant acquires content}) \quad (20)$$

$$c(\eta, \eta^{i,1}) = \lambda \sum_{j: \eta(j) \in \{0,1,2,3\}} \mathbb{I}_{\{\eta(i)=3\}} \cdot f_{j,i}(\eta) \quad (\text{voluntary correction}) \quad (21)$$

Again, after each meeting, one agent receives a truth signal with probability  $\varphi$ . The explicit revelation rates are:

$$c(\eta, \eta^{i,1}) = \frac{\lambda\varphi}{2} \sum_{j \neq i} \mathbb{I}_{\{\eta(j)=0\}} \quad (\text{ignorant gets verification}) \quad (22)$$

$$c(\eta, \eta^{i,3}) = \frac{\lambda\varphi}{2} \sum_{j \neq i} \mathbb{I}_{\{\eta(j)=2\}} \quad (\text{knowledge gets verified}) \quad (23)$$

Hence, the complete generator for  $k^*(I_l) = 0$  (memory case):

$$\begin{aligned} \Omega f(\eta) = \lambda \sum_{i \neq j} & \left[ \mathbb{I}_{\{\eta(i)=0, \eta(j) \geq 2\}} \cdot f_{j,i}(\eta) \cdot [f(\eta^{i,2}) - f(\eta)] \right. \\ & + \mathbb{I}_{\{\eta(i)=1, \eta(j) \geq 2\}} \cdot f_{j,i}(\eta) \cdot [f(\eta^{i,3}) - f(\eta)] \\ & + \mathbb{I}_{\{\eta(i)=3, \eta(j) \geq 0\}} \cdot f_{j,i}(\eta) \cdot [f(\eta^{i,1}) - f(\eta)] \\ & + \frac{\varphi}{2} \mathbb{I}_{\{\eta(i)=0\}} \cdot [f(\eta^{i,1}) - f(\eta)] \\ & \left. + \frac{\varphi}{2} \mathbb{I}_{\{\eta(i)=2\}} \cdot [f(\eta^{i,3}) - f(\eta)] \right] \end{aligned} \quad (24)$$

### Transition Rates: Memoryless Case with $k^*(I_l) = 0$

In the memoryless setting, verification does not confer permanent immunity. We work with a simplified state space  $\tilde{\mathcal{S}} = \{0, 1\}$  where:

$$\eta_t(i) = 0 \Leftrightarrow k_{i,t}(I_l) = 0 \quad (\text{ignorant}) \quad (25)$$

$$\eta_t(i) = 1 \Leftrightarrow k_{i,t}(I_l) = 1 \quad (\text{holds belief, possibly false}) \quad (26)$$

The generator becomes:

$$\begin{aligned} \Omega f(\eta) = \lambda \sum_{i \neq j} & \left[ \mathbb{I}_{\{\eta(i)=0, \eta(j)=1\}} \cdot f_{ji}(\eta) \cdot [f(\eta^{i,1}) - f(\eta)] \right. \\ & \left. + \varphi \cdot \mathbb{I}_{\{\eta(i)=1\}} \cdot [f(\eta^{i,0}) - f(\eta)] \right] \end{aligned} \quad (27)$$

The first term represents false belief transmission (infection), and the second term represents correction via revelation (recovery). This is the structure of a contact process (Liggett, 1985) with state-dependent infection rate.

### Voter Model

Consider the simplified setting: no verification ( $v_i \equiv 0$ ), symmetric persuasion ( $f(\beta_i - \beta_j) = 1/2$ ), and  $k^*(I_l) = 1$ . The state space reduces to  $\{0, 1\}^N$  (knowledge or ignorance), and the generator becomes:

$$\Omega f(\eta) = \frac{\lambda}{2} \sum_{i \neq j} [\mathbb{I}_{\{\eta(i)=0, \eta(j)=1\}} + \mathbb{I}_{\{\eta(i)=1, \eta(j)=0\}}] \cdot [f(\eta^{i,\eta(j)}) - f(\eta)] \quad (28)$$

This is exactly the voter model on the complete graph with rate  $\lambda/2$  per edge. By Theorem 2.18 of Liggett (1985), the voter model on a finite connected graph reaches consensus almost surely in finite time: all agents eventually adopt the same state (0 or 1), with probabilities determined by initial conditions.

Even with heterogeneous confidence  $\beta_i \neq \beta_j$ , Proposition 4.1 (convergence to full knowledge) can be proved via comparison with the voter model using coupling techniques (see Sub-section 3.4).

## Contact Process

In the memoryless case with  $k^*(I_l) = 0$  and no persuasion asymmetry, our model resembles the contact process (Harris, 1974):

- **State space:**  $\{0, 1\}^N$  (ignorant or misinformed)
- **Infection:**  $0 \rightarrow 1$  at rate  $\lambda \sum_{j:\eta(j)=1} f_{ji}(\eta)$  (learns false belief)
- **Recovery:**  $1 \rightarrow 0$  at rate  $\varphi$  (revelation corrects belief)

The key parameter is the effective infection rate:

$$\lambda_{\text{eff}} = \frac{\lambda \bar{f}}{\varphi} \quad (29)$$

where  $\bar{f}$  is the average persuasion probability.

In the standard contact process on  $\mathbb{Z}^d$ , there exists a critical value  $\lambda_c$  such that:

- If  $\lambda_{\text{eff}} < \lambda_c$ : the infection dies out (convergence to truth)
- If  $\lambda_{\text{eff}} > \lambda_c$ : the infection persists (stationary distribution with endemic misinformation)

Proposition 4.8 (memoryless agents reach stationary distribution with error) is the finite-network analog of survival in the contact process. The persistence of misinformation depends on the ratio  $\lambda \bar{f}/\varphi$ .

## Biased Voter Model with Mutation

The memory case with  $\varphi > 0$  can be viewed as a voter model with spontaneous opinion changes. In the voter model literature, this is called a "voter model with mutation" or "noisy voter model" (Granovsky & Madras, 1995).

The key insight is that revelation acts as a bias toward truth. States 3 (verified knowledge when  $k^* = 1$ ) and 1 (verified rejection when  $k^* = 0$ ) are absorbing, creating a drift toward the correct configuration.

## C.2. Coupling and Monotonicity

A powerful technique from IPS theory is coupling (Liggett, 1985, Chapter I). Constructing two processes  $\{\eta_t\}$  and  $\{\xi_t\}$  on the same probability space such that their relationship is preserved over time.

**Definition .6** (Stochastic Domination). For configurations  $\eta, \xi \in \{0, 1\}^N$ , write  $\eta \leq \xi$  if  $\eta(i) \leq \xi(i)$  for all  $i$ . A process  $\{\eta_t\}$  is stochastically dominated by  $\{\xi_t\}$  if  $\eta_0 \leq \xi_0$  implies  $\eta_t \leq \xi_t$  for all  $t \geq 0$  almost surely.

**Lemma .9** (Monotonicity of Knowledge Diffusion). *Suppose  $f(\beta_i - \beta_j)$  is increasing in both  $\beta_i$  and  $\beta_j$ , and  $\varphi = 0$  (no revelation). Then the knowledge diffusion process  $\{\eta_t\}$  restricted to  $\{0, 2\}^N$  is attractive: it can be coupled with itself such that  $\eta_0 \leq \xi_0$  implies  $\eta_t \leq \xi_t$  for all  $t$ .*

To prove convergence to full knowledge, couple  $\{\eta_t\}$  with a voter model  $\{\xi_t\}$  where  $\xi_t(i) \in \{0, 1\}$  and all persuasion succeeds with probability 1/2. Since the voter model reaches consensus a.s., and  $\eta_t \geq \xi_t$  (knowledge accumulation is faster in  $\eta_t$ ), we conclude  $\eta_t \rightarrow$  full knowledge a.s.

## C.3. Mean Field Limit

For large populations ( $N \rightarrow \infty$ ), the evolution of the empirical distribution can be approximated by deterministic ordinary differential equations (ODEs). Define the occupation densities:

$$\rho_s(t) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{\eta_t(i)=s\}}, \quad s \in \{0, 1, 2, 3\} \quad (30)$$

**Theorem .10** (Mean Field Approximation, informal). *As  $N \rightarrow \infty$  with  $\lambda \rightarrow \infty$  such that  $\lambda/N \rightarrow \gamma$ , the occupation densities converge to the solution of:*

$$\frac{d\rho_0}{dt} = -\gamma\rho_0(\rho_2 + \rho_3)\bar{f}(\rho) + \text{revelation terms} \quad (31)$$

$$\frac{d\rho_2}{dt} = +\gamma\rho_0(\rho_2 + \rho_3)\bar{f}(\rho) - \varphi\rho_2 \quad (32)$$

$$\frac{d\rho_3}{dt} = +\varphi\rho_2 \quad (33)$$

where  $\bar{f}(\rho)$  is the average persuasion probability given distribution  $\rho$ .

This allows **explicit computation** of convergence speed  $E[T_P]$  in Proposition 4.2 by solving the mean field ODEs?????

## C.4. Ergodic Theory and Stationary Distributions

For the memoryless case, the Markov chain  $\{\eta_t\}$  on finite state space  $\{0, 1\}^N$  has a unique stationary distribution  $\pi$  if the chain is irreducible and aperiodic.

**Proposition .11** (Existence of Stationary Distribution). *Suppose  $\varphi > 0$  and  $\lambda > 0$ . Then the memoryless process with  $k^*(I_l) = 0$  is irreducible on  $\{0, 1\}^N$  and has a unique stationary distribution  $\pi$  satisfying:*

$$\sum_{\eta} \pi(\eta) c(\eta, \eta') = \pi(\eta') \sum_{\eta} c(\eta', \eta) \quad (34)$$

In special cases (e.g., symmetric persuasion, homogeneous network), the stationary distribution may satisfy detailed balance:

$$\pi(\eta) c(\eta, \eta') = \pi(\eta') c(\eta', \eta), \quad \forall \eta, \eta' \quad (35)$$

This would give an explicit form:

$$\pi(\eta) \propto \exp(-\beta \mathcal{H}(\eta)) \quad (36)$$

where  $\mathcal{H}(\eta)$  is a Hamiltonian (energy function) and  $\beta$  is an inverse temperature parameter. This is the Gibbs distribution studied in statistical mechanics (Blume, 1993).

In the memoryless setting, verification does not confer permanent immunity. We work with a simplified state space  $\tilde{\mathcal{S}} = \{0, 1\}$  where:

$$\eta_t(i) = 0 \Leftrightarrow k_{i,t}(I_l) = 0 \quad (\text{ignorant}) \quad (37)$$

$$\eta_t(i) = 1 \Leftrightarrow k_{i,t}(I_l) = 1 \quad (\text{holds belief, possibly false}) \quad (38)$$

The generator becomes:

$$\Omega f(\eta) = \lambda \sum_{i \neq j} \left[ \mathbb{1}_{\{\eta(i)=0, \eta(j)=1\}} \cdot f_{ji}(\eta) \cdot [f(\eta^{i,1}) - f(\eta)] + \varphi \cdot \mathbb{1}_{\{\eta(i)=1\}} \cdot [f(\eta^{i,0}) - f(\eta)] \right]$$

(39)

The first term represents **false belief transmission** (infection), and the second term represents **correction via revelation** (recovery). This is precisely the structure of a **contact process** (Liggett, 1985) with state-dependent infection rate.

## C.5. Connection to Classical IPS Models

### Voter Model

Consider the simplified setting: no verification ( $v_i \equiv 0$ ), symmetric persuasion ( $f(\beta_i - \beta_j) = 1/2$ ), and  $k^*(I_l) = 1$ . The state space reduces to  $\{0, 1\}^N$  (knowledge or ignorance), and the generator becomes:

$$\Omega f(\eta) = \frac{\lambda}{2} \sum_{i \neq j} [\mathbb{1}_{\{\eta(i)=0, \eta(j)=1\}} + \mathbb{1}_{\{\eta(i)=1, \eta(j)=0\}}] \cdot [f(\eta^{i,\eta(j)}) - f(\eta)] \quad (40)$$

This is exactly the **voter model** on the complete graph with rate  $\lambda/2$  per edge. By Theorem 2.18 of Liggett (1985), the voter model on a finite connected graph reaches **consensus almost surely** in finite time: all agents eventually adopt the same state (0 or 1), with probabilities determined by initial conditions.

**Implication for our model:** Even with heterogeneous confidence  $\beta_i \neq \beta_j$ , Proposition 4.1 (convergence to full knowledge) can be proved via **comparison with the voter model** using coupling techniques (see Subsection 3.4).

**Note on network structure:** Your model operates on a \*\*complete graph\*\* (or growing complete graph) where any pair  $(i, j)$  can meet via the Poisson process. This differs from the classical voter model on spatial lattices  $\mathbb{Z}^d$ , but the mathematical structure is identical—what matters is that the graph is connected, not its spatial geometry.

## Contact Process

In the memoryless case with  $k^*(I_l) = 0$  and no persuasion asymmetry, our model resembles the **contact process** (Harris, 1974):

- **State space:**  $\{0, 1\}^N$  (ignorant or misinformed)
- **Infection:**  $0 \rightarrow 1$  at rate  $\lambda \sum_{j:\eta(j)=1} f_{ji}(\eta)$  (learns false belief)
- **Recovery:**  $1 \rightarrow 0$  at rate  $\varphi$  (revelation corrects belief)

The key parameter is the **effective infection rate**:

$$\lambda_{\text{eff}} = \frac{\lambda \bar{f}}{\varphi} \quad (41)$$

where  $\bar{f}$  is the average persuasion probability.

In the standard contact process on  $\mathbb{Z}^d$ , there exists a **critical value**  $\lambda_c$  such that:

- If  $\lambda_{\text{eff}} < \lambda_c$ : the infection dies out (convergence to truth)
- If  $\lambda_{\text{eff}} > \lambda_c$ : the infection persists (stationary distribution with endemic misinformation)

**Implication for our model:** Proposition 4.8 (memoryless agents reach stationary distribution with error) is the finite-network analog of survival in the contact process. The persistence of misinformation depends on the ratio  $\lambda \bar{f}/\varphi$ .

## Biased Voter Model with Mutation

The memory case with  $\varphi > 0$  can be viewed as a **voter model with spontaneous opinion changes**. In the voter model literature, this is called a "voter model with mutation" or "noisy voter model" (Granovsky & Madras, 1995).

The key insight: revelation acts as a **bias toward truth**. States 3 (verified knowledge when  $k^* = 1$ ) and 1 (verified rejection when  $k^* = 0$ ) are absorbing, creating a drift toward the correct configuration.

## C.6. Coupling and Monotonicity

A powerful technique from IPS theory is **coupling** (Liggett, 1985, Chapter I): constructing two processes  $\{\eta_t\}$  and  $\{\xi_t\}$  on the same probability space such that their relationship is preserved over time.

**Definition .7** (Stochastic Domination). For configurations  $\eta, \xi \in \{0, 1\}^N$ , write  $\eta \leq \xi$  if  $\eta(i) \leq \xi(i)$  for all  $i$ . A process  $\{\eta_t\}$  is **stochastically dominated** by  $\{\xi_t\}$  if  $\eta_0 \leq \xi_0$  implies  $\eta_t \leq \xi_t$  for all  $t \geq 0$  almost surely.

**Lemma .12** (Monotonicity of Knowledge Diffusion). *Suppose  $f(\beta_i - \beta_j)$  is increasing in both  $\beta_i$  and  $\beta_j$ , and  $\varphi = 0$  (no revelation). Then the knowledge diffusion process  $\{\eta_t\}$  restricted to  $\{0, 2\}^N$  is **attractive**: it can be coupled with itself such that  $\eta_0 \leq \xi_0$  implies  $\eta_t \leq \xi_t$  for all  $t$ .*

*Proof sketch.* Use the **basic coupling** (Liggett, 1985, Theorem I.2.9): drive both processes with the same Poisson clocks and same uniform random variables for persuasion. When edge  $(i, j)$  activates:

- If  $\eta(i) \leq \xi(i)$  and  $\eta(j) \leq \xi(j)$ , then  $\beta_i(\eta) \leq \beta_i(\xi)$  (monotonicity of confidence)
- Therefore  $f(\beta_j(\eta) - \beta_i(\eta)) \leq f(\beta_j(\xi) - \beta_i(\xi))$  (monotonicity of  $f$ )
- This ensures  $\eta_t \leq \xi_t$  is preserved after the interaction

□

**Application to Proposition 4.1:** To prove convergence to full knowledge, couple  $\{\eta_t\}$  with a **voter model**  $\{\xi_t\}$  where  $\xi_t(i) \in \{0, 1\}$  and all persuasion succeeds with probability 1/2. Since the voter model reaches consensus a.s., and  $\eta_t \geq \xi_t$  (knowledge accumulation is faster in  $\eta_t$ ), we conclude  $\eta_t \rightarrow$  full knowledge a.s.

## C.7. Mean Field Limit

For large populations ( $N \rightarrow \infty$ ), the evolution of the empirical distribution can be approximated by deterministic ordinary differential equations (ODEs). Define the **occupation densities**:

$$\rho_s(t) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{\eta_i(i)=s\}}, \quad s \in \{0, 1, 2, 3\} \quad (42)$$

**Theorem .13** (Mean Field Approximation, informal). *As  $N \rightarrow \infty$  with  $\lambda \rightarrow \infty$  such that  $\lambda/N \rightarrow \gamma$ , the occupation densities converge to the solution of:*

$$\frac{d\rho_0}{dt} = -\gamma\rho_0(\rho_2 + \rho_3)\bar{f}(\rho) + \text{revelation terms} \quad (43)$$

$$\frac{d\rho_2}{dt} = +\gamma\rho_0(\rho_2 + \rho_3)\bar{f}(\rho) - \varphi\rho_2 \quad (44)$$

$$\frac{d\rho_3}{dt} = +\varphi\rho_2 \quad (45)$$

where  $\bar{f}(\rho)$  is the average persuasion probability given distribution  $\rho$ .

This allows **explicit computation** of convergence speed  $E[T_P]$  in Proposition 4.2 by solving the mean field ODEs.

## C.8. Ergodic Theory and Stationary Distributions

For the memoryless case, the Markov chain  $\{\eta_t\}$  on finite state space  $\{0, 1\}^N$  has a unique stationary distribution  $\pi$  if the chain is irreducible and aperiodic.

**Proposition .14** (Existence of Stationary Distribution). *Suppose  $\varphi > 0$  and  $\lambda > 0$ . Then the memoryless process with  $k^*(I_l) = 0$  is irreducible on  $\{0, 1\}^N$  and has a unique stationary distribution  $\pi$  satisfying:*

$$\sum_{\eta} \pi(\eta)c(\eta, \eta') = \pi(\eta') \sum_{\eta} c(\eta', \eta) \quad (46)$$

*Proof.* **Irreducibility:** From any configuration  $\eta$ , the all-ignorant state  $\mathbf{0} = (0, \dots, 0)$  can be reached via consecutive revelations (probability  $\varphi^N > 0$ ). From  $\mathbf{0}$ , any other configuration  $\eta'$  can be reached via persuasion events. Thus all states communicate.

**Aperiodicity:** The process has self-loops (e.g., revelation when already in state 0), so the period is 1.

By standard Markov chain theory (**norris1998markov**), a unique stationary distribution exists.  $\square$

**Characterization via detailed balance:** In special cases (e.g., symmetric persuasion, homogeneous network), the stationary distribution may satisfy **detailed balance**:

$$\pi(\eta)c(\eta, \eta') = \pi(\eta')c(\eta', \eta), \quad \forall \eta, \eta' \quad (47)$$

This would give an explicit form:

$$\pi(\eta) \propto \exp(-\beta\mathcal{H}(\eta)) \quad (48)$$

where  $\mathcal{H}(\eta)$  is a Hamiltonian (energy function) and  $\beta$  is an inverse temperature parameter. This is the Gibbs distribution studied in statistical mechanics (Blume, 1993).

## C.9. Extensions and Open Questions

1. **Multiple intervals:** The full model has state space  $\{0, 1, 2, 3\}^{N \times P}$ . Intervals do NOT evolve independently because confidence  $\beta_{i,t} = g(K_{i,t}^E)$  depends on epistemic exposure across all intervals. This creates interesting correlation structure: when an agent gains knowledge on one interval, her confidence increases, making her more persuasive on ALL other intervals. This coupling is a distinctive feature of your model absent from standard IPS.
2. **Spatial structure (potential extension):** If agents were placed on a lattice  $\mathbb{Z}^d$  and only interacted with spatial neighbors (rather than the complete graph in your current model), techniques from **percolation theory** ([grimmett1999percolation](#)) could characterize when misinformation clusters grow to infinity. This would require modifying your Poisson meeting process to respect spatial constraints:  $\lambda_{ij} = \lambda \cdot \mathbb{1}_{\{\|i-j\|=1\}}$  for nearest neighbors only.
3. **Phase transitions:** Does there exist a critical revelation rate  $\varphi_c$  such that:
  - $\varphi < \varphi_c$ : misinformation dominates ( $\pi(\text{many errors}) > 1/2$ )
  - $\varphi > \varphi_c$ : truth dominates ( $\pi(\text{many errors}) < 1/2$ )

This parallels phase transitions in the Ising model and contact process.

4. **Convergence rates:** Can we obtain **exponential convergence** bounds of the form:

$$\|\mu_t - \pi\|_{TV} \leq C e^{-t/\tau} \quad (49)$$

where  $\tau$  is the mixing time? Techniques include spectral gap estimates and coupling arguments.

5. **Scaling limits:** In the limit  $N \rightarrow \infty$ ,  $P \rightarrow \infty$  simultaneously, the process may converge to a **measure-valued diffusion** on the space of probability measures over  $[0, 1]$ . This connects to **Fleming-Viot processes** ([ethier1993fleming](#)).

## C.10. Summary

The knowledge diffusion model developed in Section 3 can be rigorously formulated as an interacting particle system with explicit generator  $\Omega$ . This formulation:

- Connects our model to classical IPS (voter model, contact process)
- Enables application of powerful techniques (coupling, mean field limits, ergodic theory)
- Provides rigorous foundations for Propositions 4.1, 4.7, and 4.8
- Suggests natural extensions (spatial structure, phase transitions, scaling limits)

In the following sections, we leverage these connections to prove our main results and characterize the long-run behavior of knowledge diffusion with overconfident agents.