A Spectral Model of Collective Active Inference

Jonas Hallgren

February 27, 2026

Abstract

We develop a mathematical framework connecting individual belief updating to collective dynamics in multi-agent systems. Starting from active inference principles for individual agents, we show that when agents share a representational basis for beliefs, their collective dynamics can be analyzed using spectral graph theory.

The key technical contribution: representing beliefs in coefficient space allows us to apply the graph Laplacian to multi-agent belief dynamics. This reveals that coordination patterns correspond to eigenmodes of the communication graph, with the spectral structure determining which patterns of agreement and disagreement can persist.

We extend the basic framework to incorporate costs of belief change, following recent work on variational belief updating. This extended model captures phenomena that pure diffusion cannot, including echo chambers and persistent disagreement despite shared evidence. We demonstrate the approach through three detailed examples showing how network structure and individual parameters jointly determine collective outcomes.

The framework provides a tractable approach to analyzing collective intelligence systems, with potential applications to understanding coordination in hybrid human-AI systems.

1 Introduction

1.1 Motivation

Artificial agents are increasingly participating in human coordination systems—algorithmic trading in markets, content recommendation in social networks, assistance in organizational decision-making. This raises a question: how do we analyze collective intelligence in systems containing both human and artificial agents?

Existing approaches face complementary limitations. Game theory provides equilibrium analysis but scales poorly and misses dynamic processes. Network science describes graph structure but lacks principled models of belief propagation. Agent-based simulation captures rich behaviors but sacrifices analytical tractability.

We need frameworks that connect individual belief updating to collective dynamics while maintaining analytical tractability. This paper develops such a framework by extending active inference to multi-agent systems through spectral graph methods.

1.2 Approach

We build on three foundations:

Active inference [**?**, **?**] provides a principled account of how individual agents form and update beliefs through variational free energy minimization. However, it doesn't directly address multi-agent coordination.

Graph Laplacian methods from spectral graph theory provide tools for analyzing dynamics on networks [?]. These have been applied to consensus problems but typically without connection to principled belief updating.

Variational costs of belief change [?] formalize how cognitive costs, social factors, and pragmatic considerations affect belief updating beyond pure information processing.

Our contribution is connecting these three elements into a coherent framework.

1.3 Development

The framework develops in stages:

Coefficient space representation (Section 2). We show that individual beliefs represented as probability distributions can be efficiently encoded as vectors in a shared coefficient space when agents reason about common aspects of the world. This representation preserves information-theoretic structure while enabling linear algebraic analysis.

Graph Laplacian emergence (Sections 3-4). When agents update beliefs by incorporating information from neighbors, the collective dynamics evolve according to $\frac{dC}{dt} = -LC$ where L is the graph Laplacian. The spectral decomposition $L = \sum_k \lambda_k v_k v_k^\top$ reveals which collective patterns (eigenmodes) can persist. We analyze this in detail and provide examples.

Cost incorporation (Sections 5-6). Pure diffusion assumes costless belief revision. We extend the framework to include costs through parameters that capture openness to evidence (α_i), prior attachment (β_i), and preferred beliefs (c_i^{prior}). This produces dynamics:

$$\frac{dc_i}{dt} = -\alpha_i \sum_j w_{ij}(c_i - c_j) - \beta_i(c_i - c_i^{\text{prior}}) \tag{1}$$

We demonstrate through examples how this captures phenomena like echo chambers and persistent disagreement.

1.4 Limitations and Scope

This framework makes several simplifying assumptions:

1. Beliefs are representable in shared coefficient space
2. Communication structure is captured by fixed graphs
3. Dynamics are continuous and deterministic
4. Agents update through local averaging (possibly modified by costs)

These assumptions enable tractable analysis but limit applicability. The framework is best viewed as one approach to collective intelligence analysis, useful for certain problems but not universal.

We do not claim this is "the" model of collective intelligence. Rather, it's a principled starting point that connects established theories (active inference, spectral graph theory, variational reasoning) and generates testable predictions about multi-agent systems.

1.5 Contributions

Theoretical:

- Extension of active inference to multi-agent systems via shared coefficient spaces
- Derivation of graph Laplacian dynamics from belief updating principles
- Spectral characterization of persistent collective patterns
- Integration of belief change costs through variational framework

Analytical:

- Method for predicting coordination timescales from network structure
- Identification of stable disagreement patterns via eigenmode analysis
- Explanation of echo chambers through combined structural and behavioral analysis

Section 2 develops the representational foundations. Sections 3-4 derive and analyze the basic Laplacian model. Sections 5-6 extend it to include costs and provide detailed examples. Section 7 discusses applications and future directions.

2 Foundations: Individual Active Inference

2.1 The Generative Model and Beliefs

We begin with a single agent reasoning about the world. Following the active inference framework [**?**, **?**], the agent possesses a *generative model*—a probabilistic model of how observations arise from hidden world states.

Definition 2.1 (Generative Model)**.** *An agent's generative model specifies:*

- *Hidden states: $s \in \mathcal{S}$ representing unobserved aspects of the world*
- *Observations: $o \in \mathcal{O}$ representing sensory data*
- *Likelihood: $p(o|s)$ describing how states generate observations*
- *Prior: $p(s)$ encoding expectations about state structure*

The agent does not have direct access to true world states s. Instead, it maintains a *belief distribution* $q(s)$ over possible states.

Definition 2.2 (Belief Distribution)**.** *Agent i's belief at time t is a probability distribution $q_i(s;t)$ over hidden states, satisfying:*

$$q_i(s;t) \geq 0, \quad \sum_{s \in \mathcal{S}} q_i(s;t) = 1 \quad \textit{(discrete)} \tag{2}$$

or for continuous states:

$$q_i(s;t) \geq 0, \quad \int_{\mathcal{S}} q_i(s;t)\, ds = 1 \tag{3}$$

The fundamental quantity in active inference is the *variational free energy* [**?**]:

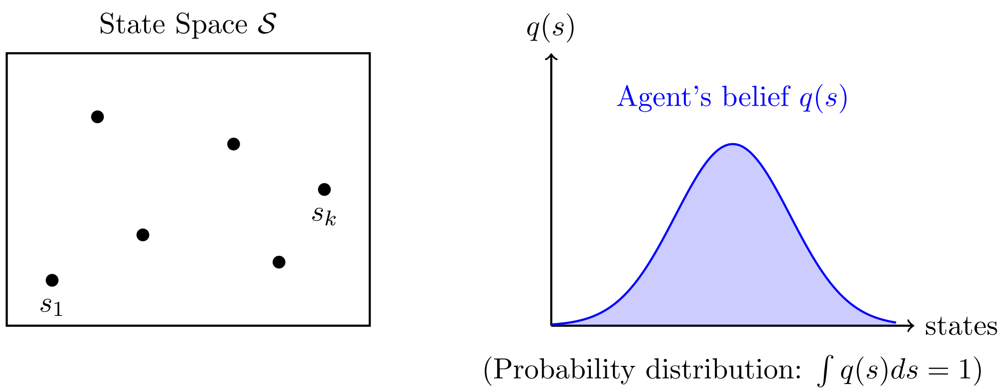

State Space $\mathcal{S}$
s_1
s_k
$q(s)$
Agent's belief $q(s)$
states
(Probability distribution: $\int q(s)ds = 1$)

Figure 1: An agent's belief is a probability distribution over possible world states. The agent cannot observe states directly, so maintains uncertainty expressed through $q(s)$.

Definition 2.3 (Variational Free Energy)**.** *Given observations o, the variational free energy of belief $q(s)$ is:*

$$F[q] = \mathbb{E}_{q(s)}[-\log p(o,s)] + \mathbb{E}_{q(s)}[\log q(s)] \tag{4}$$

Equivalently:

$$F[q] = -\log p(o) + KL(q(s)\|p(s|o)) \tag{5}$$

where KL is the Kullback-Leibler divergence.

Since $-\log p(o)$ (the surprise) does not depend on q, minimizing free energy is equivalent to minimizing the KL divergence between the agent's belief $q(s)$ and the true posterior $p(s|o)$. This provides the normative principle: *agents update beliefs to minimize variational free energy* [**?**].

2.2 Information Geometry: Measuring Belief Distance

How do we measure the difference between two belief distributions? The natural measure from information theory is the Kullback-Leibler divergence [**?**, **?**].

Definition 2.4 (KL Divergence)**.** *The KL divergence from distribution q to distribution p is:*

$$KL(q\|p) = \sum_s q(s)\log\frac{q(s)}{p(s)} \quad \textit{or} \quad \int q(s)\log\frac{q(s)}{p(s)}\,ds \tag{6}$$

measured in nats (natural logarithm) or bits (base-2 logarithm).

Remark 2.5 (Interpretation of KL Divergence)**.** *$KL(q\|p)$ measures the information lost when approximating q with p. It represents:*

- *The excess code length when using distribution p to encode samples from q*
- *The expected log-likelihood ratio between q and p*
- *The information gained when updating from prior p to posterior q*

The KL divergence is asymmetric: $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$ in general. For measuring distances symmetrically, we often use:

$$D(q,p) = \frac{1}{2}[\mathrm{KL}(q\|p) + \mathrm{KL}(p\|q)] \tag{7}$$

The space of probability distributions possesses a natural Riemannian geometry where the metric is given by the Fisher information matrix [**?**, **?**].

Definition 2.6 (Fisher Information Metric)**.** *For a parametric family of distributions $p(s;\theta)$ with parameters $\theta \in \mathbb{R}^d$, the Fisher information matrix is:*

$$g_{ij}(\theta) = \mathbb{E}_{p(s;\theta)}\left[\frac{\partial \log p(s;\theta)}{\partial \theta_i}\frac{\partial \log p(s;\theta)}{\partial \theta_j}\right] \tag{8}$$

This defines a Riemannian metric on parameter space.

The Fisher information metric has a crucial property: for nearby distributions, the Fisher-Rao distance approximates the square root of KL divergence:

Proposition 2.7 (Fisher-Rao and KL Divergence [**?**])**.** *For nearby distributions $p_1 = p(\cdot;\theta)$ and $p_2 = p(\cdot;\theta + d\theta)$:*

$$KL(p_1\|p_2) \approx \frac{1}{2}d\theta^\top g(\theta)d\theta \tag{9}$$

where $g(\theta)$ is the Fisher information matrix.

This establishes information geometry as the natural framework for understanding belief spaces [**?**].

2.3 Parametric Representation: From Distributions to Coefficients

While beliefs are fundamentally probability distributions (constrained to sum to 1), working directly in the infinite-dimensional space of distributions is intractable. The solution: represent beliefs through finite-dimensional parameters.

Definition 2.8 (Parametric Belief Family)**.** *A parametric belief family is a set of distributions $\{q(s;\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$ where:*

- *$\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ are parameters*
- *Each $q(s;\theta)$ is a valid probability distribution*
- *The map $\theta \mapsto q(s;\theta)$ is smooth*

[Common Parametric Families]

- **Gaussian beliefs**: $\theta = (\mu, \sigma^2)$ parameterizes $q(s) = \mathcal{N}(s;\mu,\sigma^2)$
- **Categorical beliefs**: $\theta = (\theta_1, \ldots, \theta_K)$ with $q(s_k) = \frac{\exp(\theta_k)}{\sum_j \exp(\theta_j)}$ (softmax)
- **Exponential family**: $q(s;\theta) = \exp(\theta^\top T(s) - A(\theta))$ where $T(s)$ are sufficient statistics

The key insight: while distributions satisfy $\sum_s q(s) = 1$, the parameters $\theta \in \mathbb{R}^d$ are *unconstrained.* This unconstrained parameter space is what we will work with.

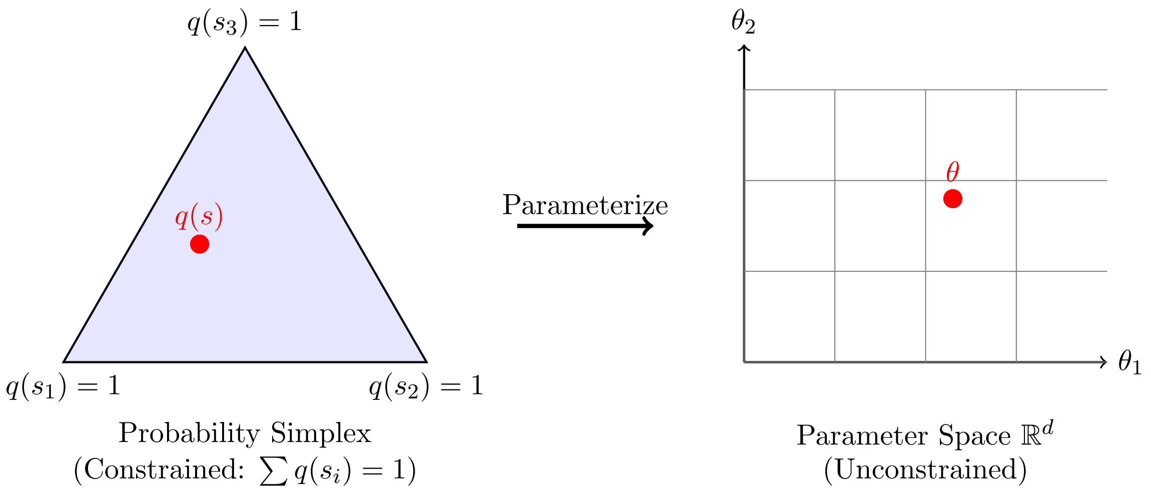


Figure 2: Beliefs are probability distributions (left), constrained to the probability simplex. Through parameterization, we map them to unconstrained parameter space $\mathbb{R}^d$ (right), which is computationally tractable.

2.4 Basis Function Representation

For beliefs about continuous or high-dimensional states, we employ a particularly useful parameterization: representation in a function basis.

Definition 2.9 (Basis Function Representation)**.** *Let $\{\phi_1(s), \phi_2(s), \ldots, \phi_d(s)\}$ be a set of basis functions. An agent's belief can be approximated as:*

$$q(s) \approx \sum_{m=1}^{d} c_m \phi_m(s) \tag{10}$$

where $c = (c_1, \ldots, c_d) \in \mathbb{R}^d$ are coefficients.

For the result to be a valid probability distribution, we require:

$$\sum_{m=1}^{d} c_m \phi_m(s) \geq 0 \quad \forall s, \qquad \int \sum_{m=1}^{d} c_m \phi_m(s)\, ds = 1 \tag{11}$$

Remark 2.10 (Unconstrained Coefficients in Practice)**.** *While mathematically $q(s)$ must be non-negative and normalized, in practice we often:*

1. *Use basis functions that ensure non-negativity (e.g., Gaussian kernels, sigmoids)*
2. *Renormalize after computing: $q(s) = \frac{1}{Z} \sum_m c_m \phi_m(s)$ where $Z = \int \sum_m c_m \phi_m(s) ds$*
3. *Work in log-space: $\log q(s) = \sum_m c_m \phi_m(s)$ then exponentiate and normalize*

In each case, the coefficients $c \in \mathbb{R}^d$ remain unconstrained during inference.

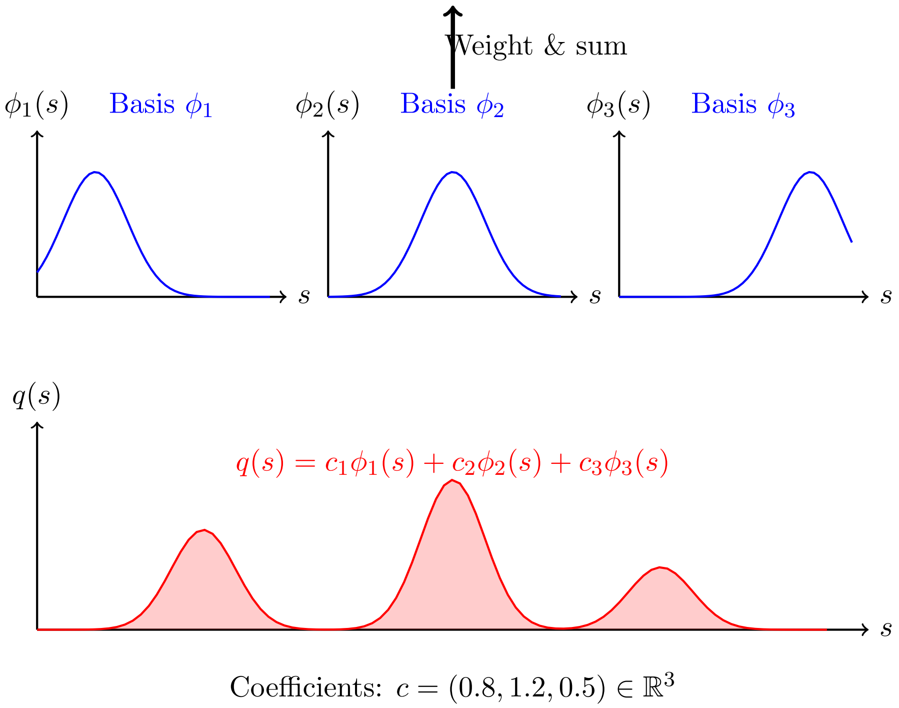


Figure 3: Basis function representation: An agent's belief $q(s)$ is represented as a weighted sum of basis functions $\phi_m(s)$. The weights (coefficients) c_m are unconstrained real numbers.

2.5 The Coefficient Space as Belief Space

This brings us to the central representational choice: we will work in *coefficient space* $\mathbb{R}^d$ rather than probability space.

Definition 2.11 (Coefficient Space Representation)**.** *For an agent with belief $q(s) \approx \sum_m c_m \phi_m(s)$, we represent the agent's state as the coefficient vector:*

$$c = (c_1, c_2, \dots, c_d) \in \mathbb{R}^d \tag{12}$$

The coefficient space is simply $\mathbb{R}^d$ with the standard Euclidean metric.

Why is this useful? Because distances in coefficient space approximate information-theoretic distances:

Proposition 2.12 (Coefficient Distance and KL Divergence)**.** *For beliefs $q_1(s) = \sum_m c_1^m \phi_m(s)$ and $q_2(s) = \sum_m c_2^m \phi_m(s)$ that are sufficiently close, and with appropriately chosen basis functions:*

$$KL(q_1 \| q_2) \approx \frac{1}{2}\|c_1 - c_2\|^2 \tag{13}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$.

Proof sketch. For small perturbations $\delta c = c_1 - c_2$, expand the KL divergence to second order. The Hessian at the expansion point is the Fisher information matrix, which under appropriate basis choice (orthonormal in the Fisher metric) becomes the identity. Full details in Appendix A. □

This coefficient space representation is our fundamental object for analysis. Crucially:

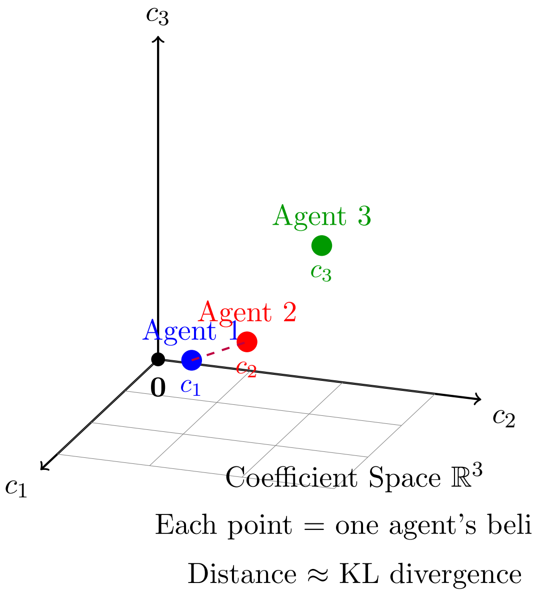


Figure 4: Beliefs represented as points in coefficient space $\mathbb{R}^d$. Each agent i has belief coefficients $c_i \in \mathbb{R}^d$. The Euclidean distance $\|c_i - c_j\|$ approximates the information-theoretic distance (KL divergence) between their beliefs.

- The space is unconstrained: $c \in \mathbb{R}^d$ (no simplex constraints)
- Distances are Euclidean (tractable geometry)
- Yet distances approximate KL divergence (preserves information-theoretic meaning)
- Linear algebra applies (enabling spectral methods)

3 The Shared Generative Space: Collective World Models

3.1 Individual and Collective Coefficient Spaces

Each agent maintains their beliefs in a coefficient space $\mathbb{R}^d$. Simultaneously, there exists a collective coefficient space representing shared knowledge.

3.2 Two Modes of Collective Knowledge: Precision vs Consensus

The mapping from individual to collective beliefs depends on the domain. We illustrate two contrasting cases:

Remark 3.1 (Volume as Uncertainty)**.** *We can interpret the "spread" or "volume" of a belief in coefficient space as representing uncertainty or variance. Formally:*

- *Small volume: High confidence, low variance (e.g., physical constants)*
- *Large volume: Low confidence, high variance (e.g., aesthetic preferences)*

*This connects to the precision (inverse variance) in active inference [**?**].*

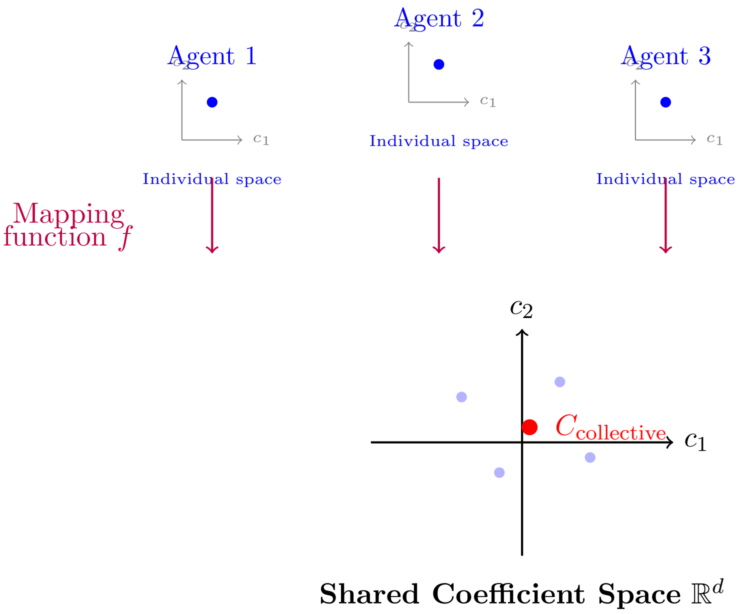


Figure 5: Individual agents maintain beliefs in their own coefficient spaces (top). Through mapping function f, these contribute to a shared collective coefficient space (bottom) representing the group's world model.

3.3 Epistemic Confidence: The Weight of Evidence

Rather than "entrenchment," we speak of *epistemic confidence*—how much evidence supports a particular belief.

Definition 3.2 (Epistemic Confidence)**.** *For a belief represented by coefficient $c \in \mathbb{R}^d$, evolved from prior c_{prior}, the epistemic confidence is:*

$$\Gamma(c) = \|c - c_{prior}\|^2 \tag{14}$$

measuring the squared displacement from prior in coefficient space.

Remark 3.3 (Interpretation)**.** $\Gamma(c)$ *represents:*

- ***Evidence accumulated**: Larger Γ means more evidence has shifted the belief*
- ***Prediction error minimized**: In active inference, moving from prior to posterior minimizes expected free energy [**?**]*
- ***Resistance to change**: Beliefs with large Γ require substantial counter-evidence to reverse*

Crucially, epistemic confidence applies to *both individual and collective beliefs*. An individual scientist can have high confidence in a theory based on personal experiments. A collective can have high confidence in established laws like thermodynamics.

[Individual vs Collective Confidence]

- **Individual**: A researcher who has run 1000 trials has high personal confidence ($\Gamma_{\text{individual}}$ large)
- **Collective**: A theory verified by millions of experiments across decades has high collective confidence ($\Gamma_{\text{collective}}$ large)

Both measure distance from prior, but at different scales.

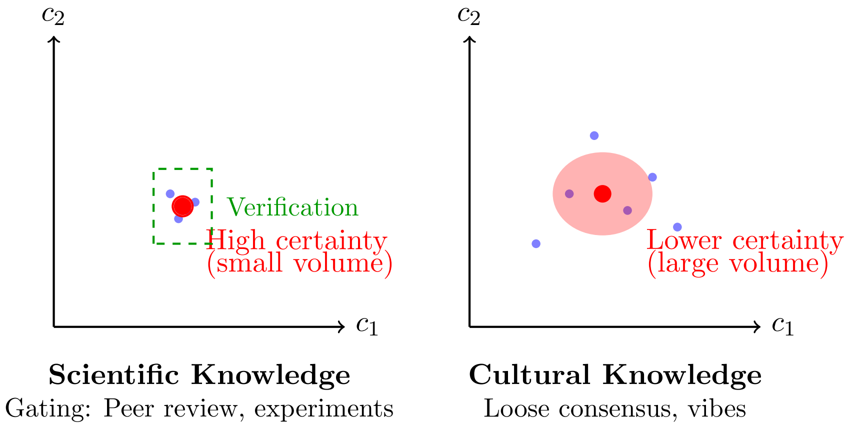


Figure 6: Different domains have different collective dynamics. **Left**: Scientific knowledge with verification gates produces tight consensus (small volume = high certainty). **Right**: Cultural knowledge produces looser consensus (large volume = lower certainty). Volume represents epistemic confidence.

4 Co-Evolution: Individual and Collective Models Over Time

4.1 The Dynamical Picture

The system state evolves according to coupled dynamics:

$$\frac{dc_i}{dt} = F_i(c_i, \{c_j\}_{j\in N_i}, C_{\text{collective}}, \text{evidence}_i) \tag{15}$$

$$\frac{dC_{\text{collective}}}{dt} = G(\{c_1, \dots, c_n\}, C_{\text{collective}}) \tag{16}$$

These dynamics are constrained by the communication graph $G = (V, E, w)$.

4.2 Graph-Constrained Communication

4.3 The Central Question

We have now established the complete setup:

- Individual and collective beliefs in coefficient space $\mathbb{R}^d$
- Epistemic confidence Γ measuring evidence accumulation
- Time evolution with co-dependent dynamics
- Graph structure G constraining communication

This leads to our central question:

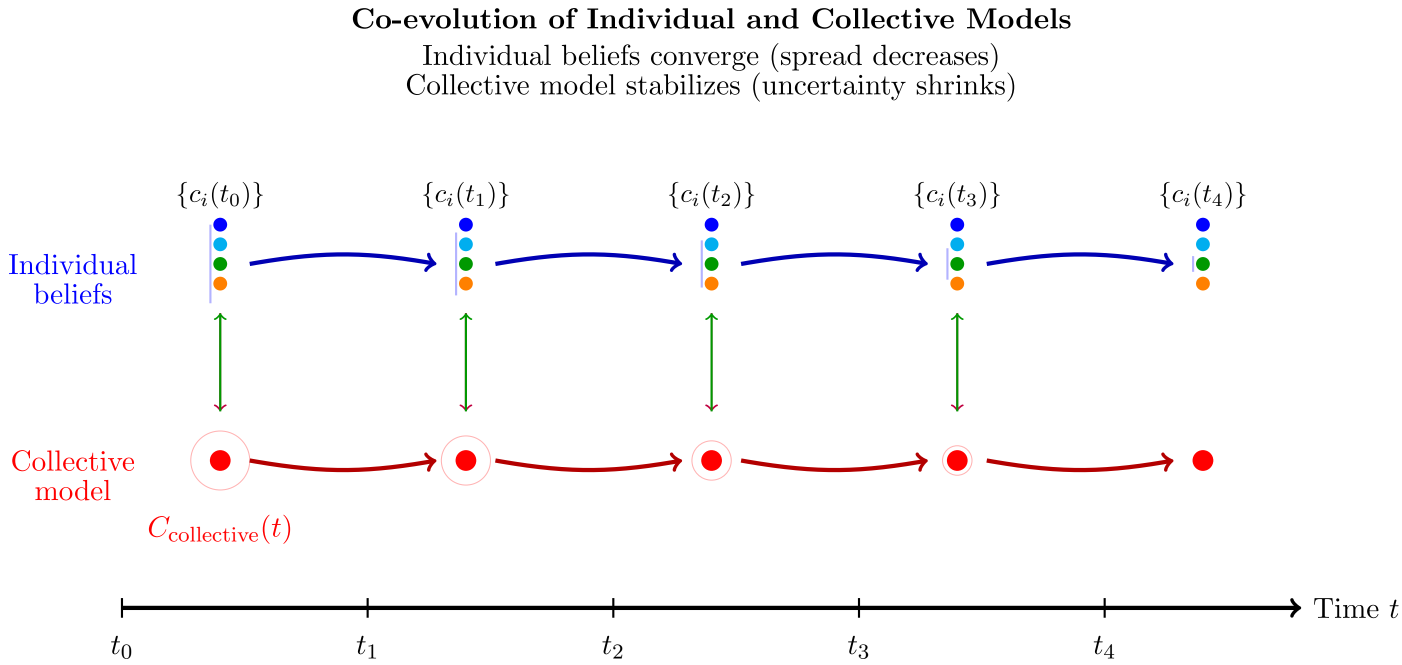

Co-evolution of Individual and Collective Models
Individual beliefs converge (spread decreases)
Collective model stabilizes (uncertainty shrinks)
$\{c_i(t_0)\}$
$\{c_i(t_1)\}$
$\{c_i(t_2)\}$
$\{c_i(t_3)\}$
$\{c_i(t_4)\}$
Individual beliefs
Collective model
$C_{\text{collective}}(t)$
Time t
t_0
t_1
t_2
t_3
t_4

Figure 7: Over time, individual beliefs $\{c_i(t)\}$ (blue points) and collective model $C_{\text{collective}}(t)$ (red point with uncertainty region) co-evolve. The spread of individual beliefs decreases, and the collective model's uncertainty (red circle) shrinks as evidence accumulates.

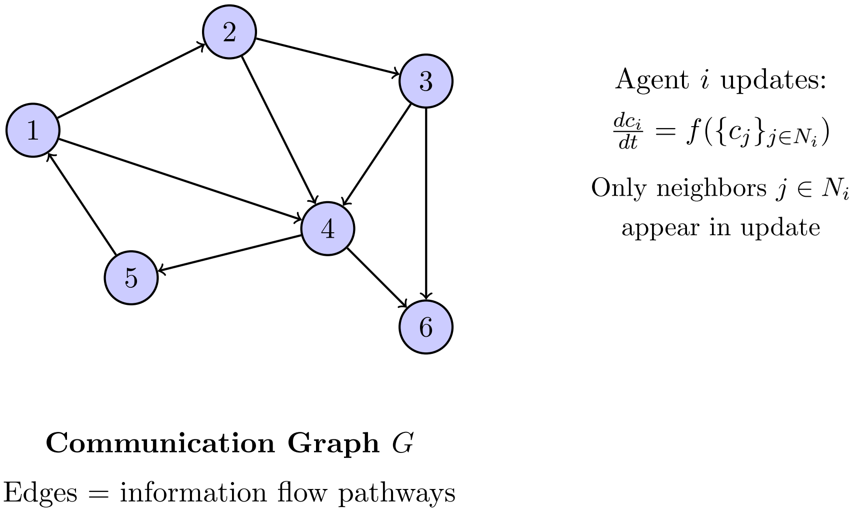

1
2
3
4
5
6
Agent i updates:
$\frac{dc_i}{dt} = f(\{c_j\}_{j \in N_i})$
Only neighbors $j \in N_i$
appear in update
Communication Graph G
Edges = information flow pathways

Figure 8: The graph G determines information flow. Agent i can only incorporate information from neighbors N_i, constraining how beliefs can evolve.

Given: Graph G, initial beliefs $\{c_i(0)\}$, collective model $C_{\text{collective}}(0)$
Question: How do graph structure and dynamics determine:

1. Whether the system reaches consensus?
2. The final collective model $C_{\text{collective}}(\infty)$?
3. The rate of convergence?
4. Which beliefs become collectively confident?

To answer this, we need a mathematical framework for analyzing graph-constrained belief dynamics. This framework is *spectral graph theory*, and the key object that emerges is the *graph Laplacian*.

5 Graph Laplacian Dynamics: Constraining the Collective Space

5.1 Why Spectral Graph Theory?

We have established a dynamical system where individual beliefs $c_i(t) \in \mathbb{R}^d$ and collective model $C_{\text{collective}}(t) \in \mathbb{R}^d$ co-evolve. But we need to understand: *what constraints does the graph structure impose on these dynamics?*

The answer lies in spectral graph theory. The graph G doesn't just determine "who talks to whom"—it fundamentally constrains which collective patterns are possible. Some belief configurations can spread and stabilize across the network; others cannot.

To see this, we start with the simplest physical model: information diffusion.

5.2 The Physical Analogy: Heat Diffusion

The most basic model of information spread is diffusion—the same mathematics that governs heat flow.

[Heat Diffusion on a Network] Imagine a metal network where nodes represent junction points with temperatures $c_i(t)$. Heat flows along edges from hot to cold junctions:

$$\text{Heat flow from } i \to j : \quad \phi_{ij} = w_{ij}(c_i - c_j) \tag{17}$$

The temperature at node i changes according to net flow:

$$\frac{dc_i}{dt} = \sum_{j \in N_i} w_{ij}(c_j - c_i) \tag{18}$$

Eventually, the entire network reaches uniform temperature (consensus).

For beliefs, this models the simplest coordination mechanism: agents average with neighbors, gradually converging toward shared understanding.

5.3 The Graph Laplacian Emerges

This diffusion process can be written compactly using the graph Laplacian operator.

Definition 5.1 (Graph Laplacian)**.** *Given graph $G = (V, E, w)$ with n nodes:*

- *Adjacency matrix:* $A_{ij} = \begin{cases} w_{ij} & \textit{if } (i,j) \in E \\ 0 & \textit{otherwise} \end{cases}$
- *Degree matrix:* $D = \textit{diag}(d_1, \ldots, d_n)$ *where* $d_i = \sum_j A_{ij}$
- *Graph Laplacian:* $L = D - A$

The diffusion dynamics become:

$$\frac{dc}{dt} = -Lc \tag{19}$$

where $c = (c_1, \ldots, c_n)^\top \in \mathbb{R}^n$.

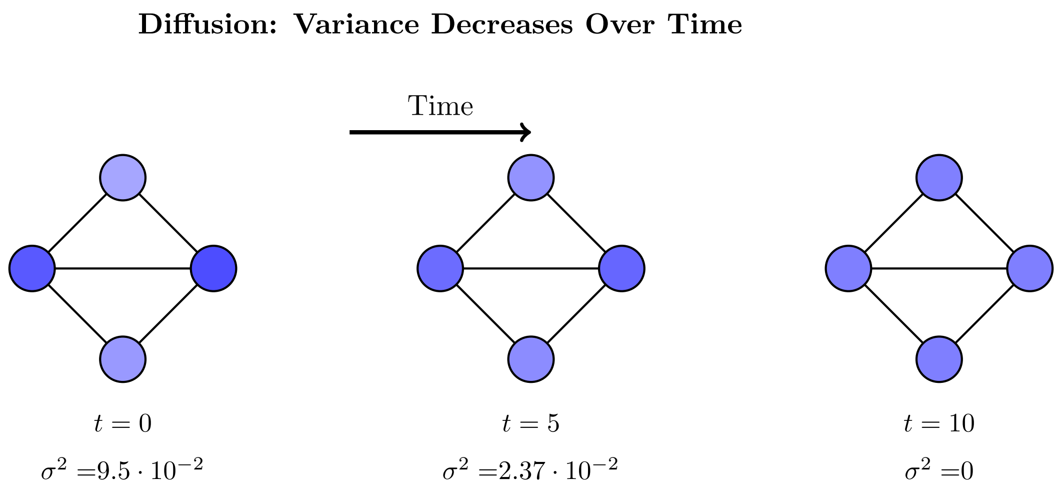

Diffusion: Variance Decreases Over Time
Time
$t = 0$
$\sigma^2 = 9.5 \cdot 10^{-2}$
$t = 5$
$\sigma^2 = 2.37 \cdot 10^{-2}$
$t = 10$
$\sigma^2 = 0$

Figure 9: Information diffusion: Initially diverse beliefs (high variance, $t = 0$) converge toward consensus (low variance, $t = 10$). Color intensity represents belief value; variance decreases monotonically.

Remark 5.2 (Why This Form?). *The Laplacian has the beautiful property:*

$$(Lc)_i = \sum_{j \in N_i} w_{ij}(c_i - c_j) \tag{20}$$

This measures how much node i's value differs from its neighborhood average. The dynamics $\frac{dc}{dt} = -Lc$ *reduce this local variance.*

But the real power comes from the Laplacian's spectral structure.

5.4 Spectral Decomposition: The Eigenbasis

The Laplacian's eigenvectors provide a natural coordinate system for understanding collective patterns.

Theorem 5.3 (Spectral Decomposition [**?**]). *The symmetric Laplacian L has eigendecomposition:*

$$L = \sum_{k=1}^{n} \lambda_k v_k v_k^\top \tag{21}$$

where:

- $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ *(eigenvalues)*
- $v_1, \ldots, v_n$ *(orthonormal eigenvectors)*
- $v_1 = \frac{1}{\sqrt{n}}(1, 1, \ldots, 1)^\top$ *(consensus mode)*

Each eigenvector v_k represents a spatial pattern on the graph—a way beliefs could be distributed across agents.

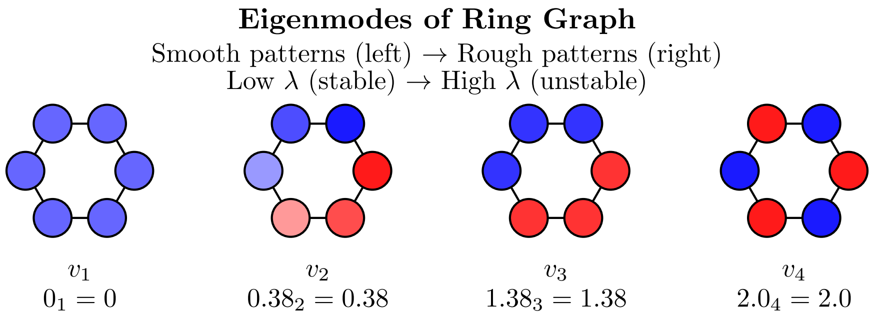

Eigenmodes of Ring Graph
Smooth patterns (left) → Rough patterns (right)
Low λ (stable) → High λ (unstable)
v_1
$0_1 = 0$
v_2
$0.38_2 = 0.38$
v_3
$1.38_3 = 1.38$
v_4
$2.0_4 = 2.0$

Figure 10: Four eigenmodes of a 6-node ring. Mode v_1: uniform (consensus). Mode v_2: smooth gradient. Mode v_3: two clusters. Mode v_4: alternating (roughest). Rougher patterns have larger eigenvalues.

5.5 Eigenvalues as Variance: Not Decay!

Here's the crucial reframing: eigenvalues don't just measure ”decay rate”—they measure *variance across the population* in that mode.

Proposition 5.4 (Eigenvalue as Population Variance)**.** *For a belief pattern c, decompose into eigenmodes: $c = \sum_k \alpha_k v_k$. The variance of c across the population is:*

$$Var(c) = \frac{1}{n}\sum_i (c_i - \bar{c})^2 = \sum_{k=2}^{n} \alpha_k^2 \tag{22}$$

where $\bar{c} = \frac{1}{n}\sum_i c_i = \alpha_1$ (the consensus component).

The eigenvalue λ_k determines how stable variance in mode v_k is:

- *Small λ_k: Mode v_k can sustain variance (stable disagreement pattern)*
- *Large λ_k: Mode v_k cannot sustain variance (pattern dissolves quickly)*

The graph spectrum tells us *which patterns of disagreement are stable.*

5.6 Evolution in Coefficient Space

Now extend to multi-dimensional beliefs where each agent has $c_i \in \mathbb{R}^d$. The system evolves:

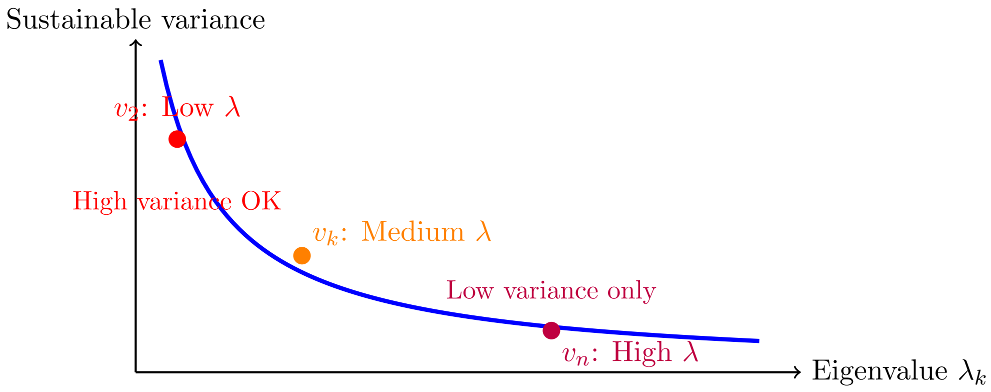
$$\frac{dC}{dt} = -LC \tag{23}$$

where $C \in \mathbb{R}^{n \times d}$ has rows c_i.

Each column (belief component) evolves independently:

$$\frac{dC_{:,m}}{dt} = -LC_{:,m} \tag{24}$$

The spectral decomposition shows which components can maintain variance.



Eigenvalue λ determines sustainable variance in that mode

Figure 11: Inverse relationship between eigenvalue and sustainable variance. Low-λ modes (like v_2) can maintain high population variance (stable disagreement). High-λ modes quickly collapse to consensus.

5.7 Example: Newtonian Physics → Relativity

Consider scientific consensus as a point in coefficient space.

[Paradigm Shift in Physics] **Era 1 (1700-1900)**: Newtonian mechanics

- Strong consensus: $C_{\text{Newton}} = (1, 0, 0, \ldots)$ (low variance across scientists)
- Aligned with $\lambda_1 = 0$ (consensus mode)
- Alternative theories quickly suppressed (high λ modes)

Era 2 (1905-1920): Crisis and competition

- Einstein proposes relativity: new component $c_{\text{relativity}}$ emerges
- Initially high variance: some accept, many reject
- Debate corresponds to non-zero λ_k mode activating

Era 3 (1920+): New consensus

- Relativity accumulates evidence, variance decreases
- New consensus: $C_{\text{modern}} = (0.9, 0.1, 0, \ldots)$ (Newton as limiting case)
- Returns to low-variance (consensus) state

5.8 The Laplacian as Mediator: Individual ↔ Collective

The graph Laplacian provides the bridge between individual and collective inference:

Theorem 5.5 (Laplacian Mediation Between Scales)**.** *The graph Laplacian L determines how individual beliefs become collective knowledge:*

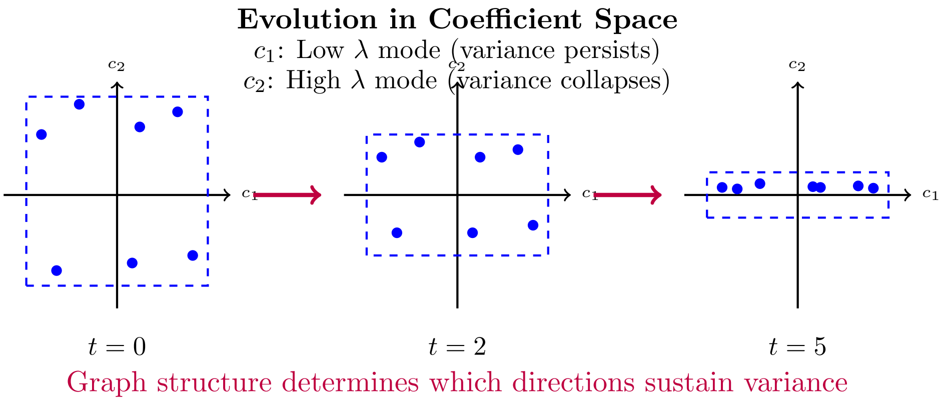

Evolution in Coefficient Space
c_1: Low λ mode (variance persists)
c_2: High λ mode (variance collapses)
c_2
c_1
c_2
c_1
c_2
c_1
$t = 0$
$t = 2$
$t = 5$
Graph structure determines which directions sustain variance

Figure 12: Evolution in 2D coefficient space. Direction c_1 (aligned with low-λ mode) maintains variance—stable disagreement. Direction c_2 (aligned with high-λ mode) collapses to consensus. The graph Laplacian determines which belief components are ”sticky” vs ”spreading.”

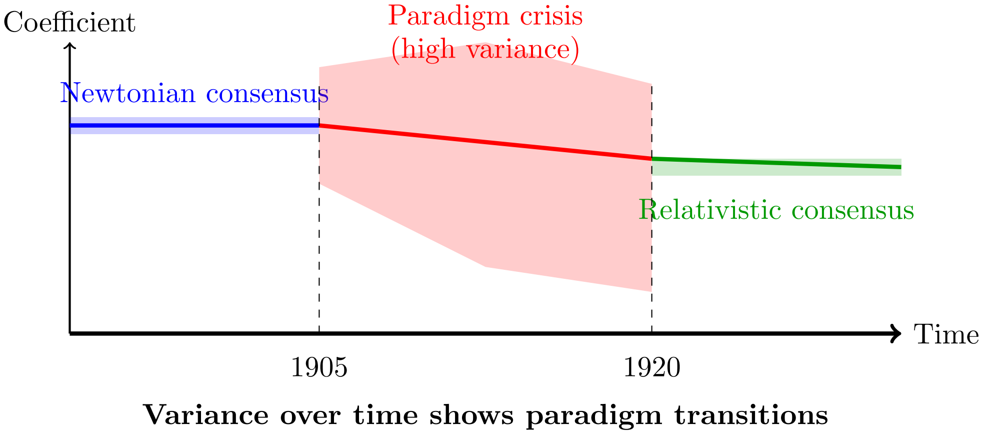


Figure 13: Scientific consensus as a time series. Newtonian era: low variance (consensus). Crisis period (1905-1920): high variance (competing theories). Modern era: new consensus with low variance. Graph Laplacian determines which beliefs can sustain variance.

1. ***Individual updates:*** *$\frac{dc_i}{dt} = -(Lc)_i$ depends only on local neighbors*
2. ***Variance dynamics:*** *$\frac{d\,Var(c)}{dt} = -2\langle c, Lc\rangle$ always decreases*
3. ***Stable patterns:*** *Low-λ eigenmodes sustain population variance (disagreement)*
4. ***Collective emergence:*** *Consensus $\lim_{t\to\infty} c(t) = \langle c(0)\rangle v_1$ determined by initial average*

The eigenmodes reveal which belief configurations can persist collectively: - $\lambda_1 = 0$: Perfect consensus (no variance) - Small $\lambda_2, \lambda_3, \ldots$: Stable clusters or gradients (sustainable disagreement) - Large λ_k: Unstable fluctuations (cannot persist)

This transforms collective intelligence from averaging to *selective stabilization of graph-resonant patterns.*

5.9 When Simple Diffusion Fails: The Need for Extended Models

The elegant spectral theory we've developed has a critical limitation: it assumes all agents update beliefs purely through diffusion, with no resistance to change and no attachment to prior beliefs. This fails to capture crucial real-world phenomena.

[Conspiracy Theory Communities] Consider a network with two communities:

- **Mainstream community (M)**: Scientists, journalists, general public
- **Conspiracy community (C)**: Small group with strong alternative narrative

Under pure diffusion, both communities should eventually converge to consensus. But in reality:

- Community C maintains beliefs despite contrary evidence
- Members are highly resistant to outside information
- The beliefs persist for years or decades

Why does our model fail here?

The pure diffusion model assumes belief change is "free"—agents costlessly average with neighbors. But real belief updating involves:

1. **Cognitive costs**: Mental effort required to revise beliefs [**?**]
2. **Epistemic confidence**: Strong prior beliefs resist change [**?**]
3. **Social costs**: Changing beliefs may threaten group membership [**?**]
4. **Pragmatic considerations**: Some beliefs serve non-epistemic functions [**?**]

Recent work on the variational costs of belief change [**?**] provides a principled framework for incorporating these factors.

5.10 Extended Model: Variational Belief Updating

Following Hyland and Albarracin (2025), we extend our framework to include costs:

Definition 5.6 (Variational Belief Dynamics)**.** *Agent i updates beliefs by optimizing:*

$$\mathcal{F}_i[q_i] = U_i[q_i, o] + \alpha_i \mathbb{E}_{q_i}[\log p(o|s)] - \lambda_i D_{KL}[q_i \| p_i] \tag{25}$$

where:

- $U_i[q_i, o]$*: Affective utility of holding belief* q_i *given observation* o
- $\alpha_i \geq 0$*: Weight agent* i *assigns to evidence*
- $\lambda_i \geq 0$*: Cost of deviating from prior* p_i

In continuous-time multi-agent dynamics, this gives:

$$\frac{dc_i}{dt} = -\alpha_i \sum_{j \in N_i} w_{ij}(c_i - c_j) - \beta_i(c_i - c_i^{\text{prior}}) \tag{26}$$

The first term represents averaging with neighbors (weighted by openness α_i). The second term represents attachment to prior beliefs (weighted by stubbornness β_i).

Remark 5.7 (Interpretation of Parameters).

- $\alpha_i = 1, \beta_i = 0$: *Standard diffusion (costless belief change)*
- α_i *small: Agent resists neighbor influence (confirmation bias)*
- β_i *large: Agent strongly attached to prior (epistemic confidence)*
- c_i^{prior}: *Agent's "preferred" or "default" belief*

In matrix form:

$$\frac{dc}{dt} = -D_\alpha Lc - D_\beta(c - c^{\text{prior}}) \tag{27}$$

where $D_\alpha = \text{diag}(\alpha_1, \ldots, \alpha_n)$ and $D_\beta = \text{diag}(\beta_1, \ldots, \beta_n)$.

This creates an *effective Laplacian*:

$$L_{\text{eff}} = D_\alpha L + D_\beta \tag{28}$$

The spectral structure of L_{eff} determines convergence dynamics, but now includes the costs of belief change.

5.11 Three Detailed Examples

We now demonstrate how this extended framework captures phenomena the simple diffusion model cannot.

5.11.1 Example 1: Pure Diffusion (Baseline)

Setup: Five agents connected in a line, each with equal weight edges ($w = 1$). All agents are fully open to neighbor influence ($\alpha_i = 1$) with no prior attachment ($\beta_i = 0$).

Dynamics: Standard diffusion equation:

$$\frac{dc_i}{dt} = -(c_i - c_{i-1}) - (c_i - c_{i+1}) \tag{29}$$

Spectral Analysis: The graph Laplacian has eigenvalues:

$$\lambda_k = 2\left(1 - \cos\frac{\pi k}{6}\right), \quad k = 1, \ldots, 5 \tag{30}$$

giving $\lambda_1 = 0, \lambda_2 \approx 0.27, \lambda_3 \approx 1.0, \lambda_4 \approx 1.73, \lambda_5 = 2.0$.

Convergence: The spectral gap $\lambda_2 \approx 0.27$ determines mixing time:

$$t_{\text{mix}} \approx \frac{1}{\lambda_2} \approx 3.7 \text{ time steps} \tag{31}$$

Result: All agents converge to consensus at the mean $\bar{c} = 0.5$ within ∼50 time steps. This is the baseline behavior.

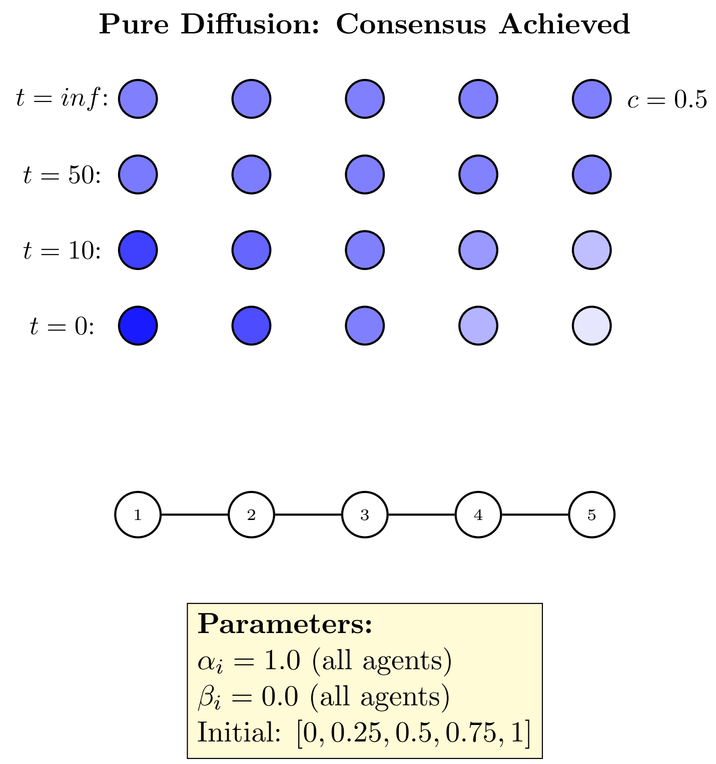

Pure Diffusion: Consensus Achieved
$t = inf$:
$c = 0.5$
$t = 50$:
$t = 10$:
$t = 0$:
1
2
3
4
5
Parameters:
$\alpha_i = 1.0$ (all agents)
$\beta_i = 0.0$ (all agents)
Initial: $[0, 0.25, 0.5, 0.75, 1]$

Figure 14: Five agents in a line with pure diffusion dynamics. Initial beliefs range from 0 to 1. All agents converge to the mean (0.5) within ∼50 time steps. Color intensity represents belief value.

5.11.2 Example 2: Echo Chambers with Weak Connections

Setup: Six agents in two triangular communities (M and C), each internally fully connected with $w = 1$. A single weak edge ($w = 0.01$) connects the communities. All agents have $\alpha_i = 1, \beta_i = 0$.

Dynamics: Still pure diffusion, but graph structure creates bottleneck:

$$\frac{dc_i}{dt} = -\sum_{j \in M}(c_i - c_j) \quad \text{(for } i \in M) \tag{32}$$

$$\frac{dc_i}{dt} = -\sum_{j \in C}(c_i - c_j) - 0.01(c_i - c_{\text{bridge}}) \quad \text{(for bridge node)} \tag{33}$$

Spectral Analysis: The modified Laplacian has:

$$\lambda_1 = 0, \quad \lambda_2 \approx 0.002, \quad \lambda_3, \ldots, \lambda_6 \sim O(1) \tag{34}$$

The critical spectral gap $\lambda_2 \approx 0.002$ is **100× smaller** than the baseline!

Convergence Time:

$$t_{\text{mix}} \approx \frac{1}{\lambda_2} \approx 500 \text{ time steps} \tag{35}$$

Result: Even with perfect diffusion, structural isolation creates persistent disagreement. The communities maintain distinct beliefs for hundreds of time steps, appearing as ”echo chambers” despite no individual resistance to updating.

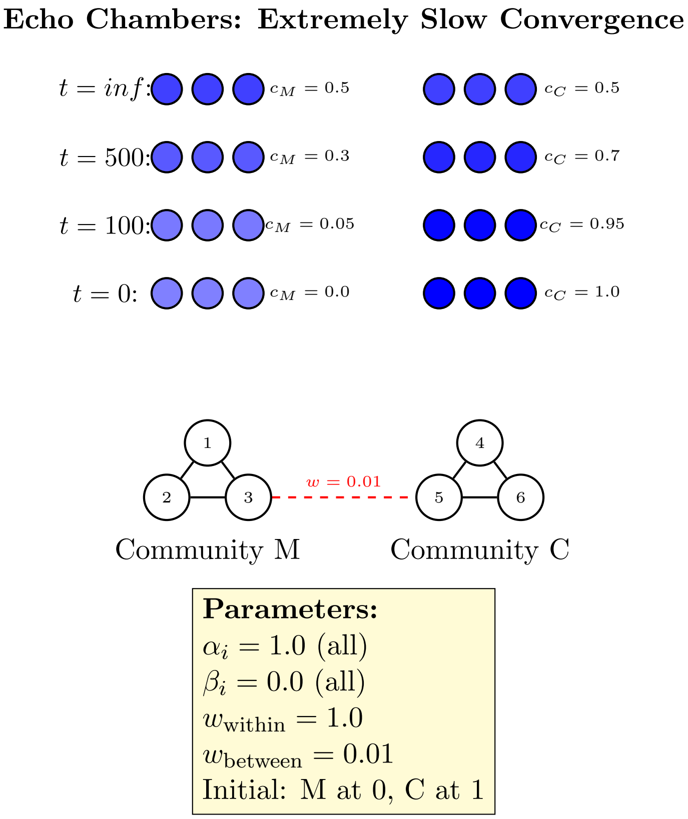


Figure 15: Two communities with strong internal ties ($w = 1$) but weak inter-community connection ($w = 0.01$). Despite using pure diffusion, convergence takes ∼500 time steps due to structural bottleneck.

5.11.3 Example 3: Stubborn Agents with Prior Attachment

Setup: Same graph as Example 2, but now Community M (conspiracy theorists) has:

- Low openness: $\alpha_M = 0.1$ (only 10% weight on neighbors)
- High prior attachment: $\beta_M = 0.5$
- Strong prior: $c_M^{\text{prior}} = 0$

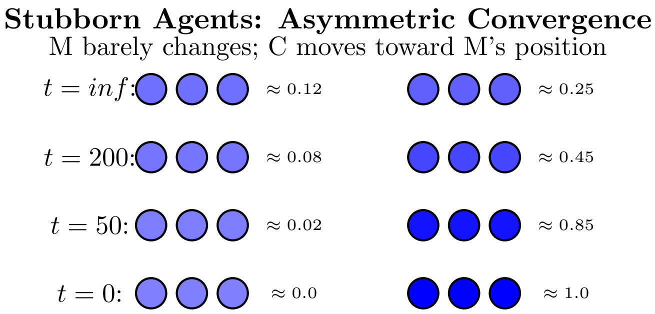
Community C (mainstream) remains open: $\alpha_C = 1.0, \beta_C = 0$.
Dynamics: Asymmetric updating:

$$\frac{dc_i}{dt} = -0.1 \sum_{j \in N_i} w_{ij}(c_i - c_j) - 0.5(c_i - 0) \quad (i \in M) \tag{36}$$

$$\frac{dc_j}{dt} = -1.0 \sum_{k \in N_j} w_{jk}(c_j - c_k) \quad (j \in C) \tag{37}$$

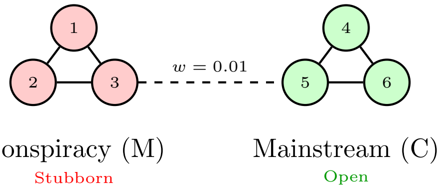
Effective Laplacian: The system no longer converges to the global mean! The effective dynamics are:

$$L_{\text{eff}} = \left(D_\alpha L + D_\beta\right) \tag{38}$$



Stubborn Agents: Asymmetric Convergence

M barely changes; C moves toward M's position



Community M (conspiracy):
$\alpha_M = 0.1$, $\beta_M = 0.5$
$c_M^{\text{prior}} = 0.0$

Community C (mainstream):
$\alpha_C = 1.0$, $\beta_C = 0.0$
Initial: M at 0, C at 1

Figure 16: Identical graph structure to Example 2, but Community M has high stubbornness ($\beta = 0.5$) and low openness ($\alpha = 0.1$). Community M maintains beliefs near prior (0) indefinitely, while Community C gradually shifts toward M's position due to asymmetric updating.

where D_α and D_β are highly non-uniform.

Steady State: Solving $(D_\alpha L + D_\beta)c^* = D_\beta c^{\text{prior}}$:

$$c_M^* \approx 0.12 \quad \text{(barely moved from 0)} \tag{39}$$

$$c_C^* \approx 0.25 \quad \text{(moved significantly from 1)} \tag{40}$$

Key Insight: The **stubborn minority pulls the open-minded majority** toward their position! This is the opposite of what pure diffusion predicts, where everyone converges to the mean (0.5).

Real-World Analog: This captures how conspiracy communities can resist mainstream evidence while gradually shifting public discourse in their direction through persistent messaging, despite being numerically smaller.

5.12 Summary of Examples

These examples demonstrate that understanding collective intelligence requires both:

1. **Structural analysis**: Graph topology (captured by Laplacian spectrum)

Property	**Example 1**	**Example 2**	**Example 3**
Graph structure	Line	Two clusters	Two clusters
Parameters	$\alpha = 1, \beta = 0$	$\alpha = 1, \beta = 0$	α mixed, β mixed
Spectral gap λ_2	0.27	0.002	Effective ~ 0.0005
Convergence time	Fast (∼50 steps)	Slow (∼500 steps)	Never (asymmetric)
Final state	Consensus (0.5)	Consensus (0.5)	Persistent disagreement
Mechanism	Diffusion	Structural bottleneck	Behavioral resistance

Table 1: Comparison of three examples showing how graph structure and individual parameters interact to produce qualitatively different collective dynamics.

2. **Behavioral modeling**: Individual costs and biases (captured by $\alpha, \beta, c^{\text{prior}}$)

The power of this framework is that it unifies both perspectives, showing how individual variational costs and network structure jointly determine collective outcomes.

6 Conclusion

We have developed a mathematical framework for collective intelligence that bridges individual belief updating with emergent collective patterns. Starting from individual agents maintaining beliefs in coefficient space, we showed how the graph Laplacian naturally emerges as the operator governing collective dynamics.

The spectral decomposition of the Laplacian reveals the fundamental modes of collective variation—patterns of belief distribution across the population. Low eigenvalue modes represent sustainable patterns of disagreement; high eigenvalue modes represent unstable fluctuations that quickly dissipate.

However, pure diffusion dynamics assume costless belief change, which fails to capture real-world phenomena like echo chambers, conspiracy theories, and political polarization. By incorporating variational costs of belief updating—following recent work on motivated reasoning and resource-rational cognition—we extended the framework to include individual stubbornness, prior attachment, and selective information processing.

The extended model reveals that persistent collective disagreement emerges from the interplay of:

- **Structural isolation**: Weak connections create small spectral gaps (slow mixing)
- **Behavioral resistance**: High belief change costs prevent convergence
- **Asymmetric updating**: Stubborn minorities can pull open-minded majorities

This framework provides both explanatory power (understanding why certain collective patterns persist) and predictive capability (forecasting how interventions will affect collective beliefs). It suggests that improving collective intelligence requires operating on multiple levels:

- **Network level**: Strengthen cross-cutting ties to increase λ_2
- **Individual level**: Reduce costs of belief revision (lower λ, increase α)
- **Cultural level**: Create norms that value evidence over identity (modify c^{prior})

The graph Laplacian serves as the mathematical bridge between individual active inference and collective intelligence, mediating how local belief updates produce global patterns of agreement and disagreement. Understanding this bridge is essential for designing hybrid human-AI systems that maintain beneficial collective properties as artificial agents become increasingly integrated into human coordination mechanisms.