The Topology of Influence: Latent Structure, Wave Propagation, and Organizational Resilience in Online Populations

Jeremy McEntire

2026

Abstract

We demonstrate that a population's behavioral data, when embedded in a latent space via dimensionality reduction, reveals persistent structure that predicts influence propagation and organizational resilience. Using 1.4 billion Reddit comments across six months (January–June 2022) and 44,837 subreddits, we establish seven empirical findings: (1) a shared latent space of 12 independent dimensions exists across behaviorally distinct domains; (2) individual positions in this space are temporally stable across two-week windows (Mantel $r = 0.41$); (3) the space detects real-world events and distinguishes unifying from polarizing signatures via displacement coherence; (4) three stable subpopulations persist across independent time windows (ARI = 0.47); (5) influence propagates as a wave through proximity in the space, with a measurable boundary where proximity-driven influence gives way to broadcast media (Spearman $\rho = 0.27$, $p < 10^{-6}$); (6) leadership is topic-specific, not a global trait—per-dimension enrichment averages 18.5× while global enrichment is 1.0×, a Simpson's paradox caused by different people leading in each dimension; and (7) influence propagates along the topology of the latent space, defined by shared leadership identity between dimensions ($\rho = 0.56$, $p < 10^{-6}$). Dimensions with no cross-dimensional leadership (such as the January 2022 r/antiwork movement) are structurally fragile—they collapse under a single point of failure because influence has no alternative paths. Connected dimensions (culture, gaming, finance) exhibit distributed resilience. The topology of the latent space is not merely a measurement tool; it is a predictive map of influence flow and organizational fragility, computable from behavioral data alone.

1 Introduction

The central problem of computational social science is measurement. We observe behavior—posts, purchases, clicks, follows—but the quantities we care about—influence, alignment, fragility—are latent. They do not appear directly in any data stream. They must be inferred from the structure of the data itself.

This paper presents a framework for that inference and seven empirical findings that validate it. The framework places individuals in a high-dimensional latent space derived from behavioral data via dimensionality reduction, then exploits the geometric and topological properties of that space to measure influence, detect events, classify their character, identify persistent subpopulations, and predict organizational resilience.

The framework makes no assumptions about what the dimensions of the space represent. It does not label domains, categorize users, or impose structure. The dimensions emerge from the mathematics. The topology emerges from the data. The findings emerge from testing the framework's predictions against six months of Reddit behavioral data comprising 1.4 billion comments across 44,837 communities.

The central finding is that the latent space has a topology—a graph structure defined by which dimensions share leaders—and that this topology predicts both how influence propagates and where it fails to propagate. Influence travels the edges of this graph. Movements embedded in well-connected regions of the topology are resilient to shocks because influence can route around failures. Movements embedded in isolated regions—with leaders who lead nowhere else—are structurally fragile. This fragility is measurable before it manifests, from behavioral data alone.

We proceed as follows. Section 2 describes the mathematical framework. Section 3 describes the data and processing pipeline. Section 4 presents the seven empirical findings. Section 5 develops the topological analysis and its implications for resilience. Section 6 discusses limitations, applications, and future directions.

2 Framework

2.1 The Latent Space

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a user-feature matrix where n is the number of individuals and p is the number of behavioral features (e.g., activity counts across communities). We seek a low-dimensional representation $\mathbf{Z} \in \mathbb{R}^{n \times d}$ ($d \ll p$) that preserves the pairwise similarity structure of the population.

We apply TF-IDF weighting to control for feature popularity, L2 normalisation to remove activity-level effects, and truncated SVD to obtain the embedding. The number of signal dimensions d is determined by Random Matrix Theory: eigenvalues of the sample covariance matrix exceeding the Marchenko-Pastur upper edge $\lambda_+ = \sigma^2(1 + \sqrt{p/n})^2$ are retained as signal; those below are discarded as noise.

No assumptions are imposed on what the dimensions represent. They are the axes of maximum variance in the normalised behavioral data. Their semantic content, if any, is discovered post hoc from the feature loadings.

2.2 Cross-Domain Validation

The core falsifiability test: if the latent space is real, it should be recoverable independently from different slices of the data. We partition features (subreddits) into two groups—either by data-discovered clustering or by random assignment—and embed each group independently. The Mantel test measures whether pairwise similarity in one embedding predicts pairwise similarity in the other. The null distribution is generated by permuting identity labels.

Canonical Correlation Analysis (CCA) between the two embeddings counts the number of shared dimensions—paired axes along which the two independently discovered spaces agree.

2.3 Temporal Stability

The same users are embedded independently from two non-overlapping time windows. The Mantel test, per-user position correlation, CCA shared dimensionality, and nearest-neighbour rank stability all measure whether the space captures persistent traits rather than transient behaviour.

2.4 Displacement and Event Detection

For a stimulus (event, claim, content), displacement is the vector difference between a user's pre-stimulus and post-stimulus position, after Procrustes alignment to correct for embedding rotation.

Three properties of the displacement vector field characterise the stimulus:

- **Magnitude**: how much did affected users move, relative to background?
- **Coherence**: did affected users move in the same direction (unifying) or scatter (polarising)?
- **Direction**: did affected users move in the same or opposite direction as the background population?

2.5 Influence Propagation

For a community that experienced a burst of new participants, we measure each newcomer's arrival time and their distance in the latent space from the earliest adopters. A positive Spearman correlation between distance and arrival time indicates wave-like propagation: closer users arrive first. No correlation indicates broadcast: everyone hears simultaneously.

2.6 Topic-Specific Leadership

For each embedding dimension independently, we extract discrete features (velocity peaks) from the population centroid's trajectory along that dimension, then match each user's per-dimension trajectory against these features using a barcode-alignment approach with lag estimation. Users whose trajectories show the same features earlier than the centroid are dimension-specific leaders.

Stability is assessed by splitting the observation period in half and checking whether the same users lead in both halves. Enrichment is the ratio of observed cross-half retention to the rate expected by chance.

2.7 Topological Structure

The Jaccard index between the leader sets of each pair of dimensions defines an edge weight on the dimension graph. Dimensions with shared leaders are adjacent; dimensions with disjoint leaders are isolated.

The key prediction: if influence propagates along this topology, then leading on dimension A should predict earlier adoption on adjacent dimension B (high Jaccard) but not on distant dimension C (low Jaccard). The Spearman correlation between Jaccard overlap and cross-dimension lead advantage tests this prediction.

3 Data and Pipeline

3.1 Data Source

We use the Pushshift Reddit archive, comprising the complete public comment history of the Reddit platform. We process six monthly dumps (January–June 2022) totalling 1.394 billion comments from 26.9 million unique accounts across 44,837 active communities (subreddits with ≥ 500 unique commenters).

3.2 Processing

A two-pass streaming pipeline manages memory:

1. **Pass 1**: Stream all dumps, count the number of distinct active days per user. Retain users active on ≥ 30 of 181 days (3.07 million users; sampled to 50,000 for tractability).
2. **Pass 2**: Re-stream, extracting daily community membership vectors for retained users only.

The output is a position tensor $\mathbf{P} \in \mathbb{R}^{16{,}507\times 181\times 19}$—daily positions for 16,507 qualified users across 181 days in a 19-dimensional latent space. Each day's embedding uses a 3-day smoothing window projected into a stable global SVD coordinate system.

4 Results

4.1 Finding 1: The Latent Space Exists

The cross-domain test, applied to an independent one-month sample (January 2022, 1.2 million users, 4,586 subreddits), yields Mantel $r = 0.131$, $p < 10^{-3}$ (below all 1,000 permutations, Figure 1). CCA identifies 12 significant shared dimensions. The structure is not an artifact of the embedding method: the two domain matrices share no features and are reduced independently.

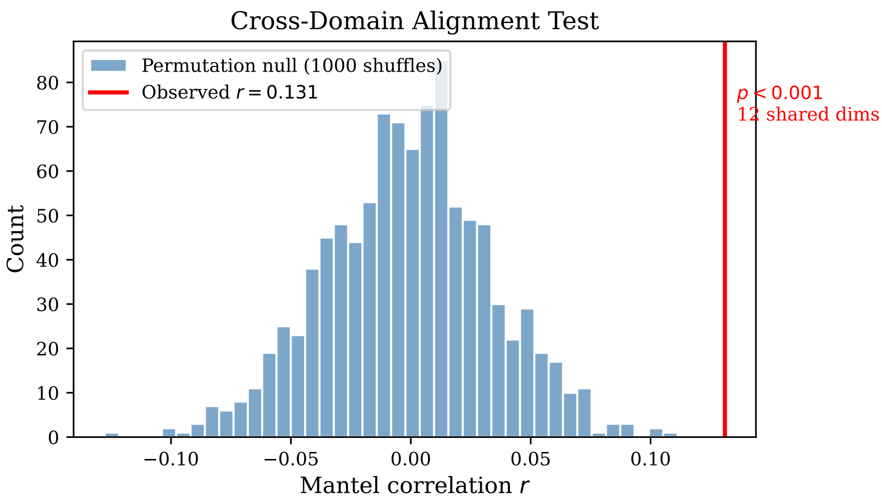

Figure 1: Cross-domain alignment test. The observed Mantel $r = 0.131$ (red line) falls far outside the permutation null distribution, establishing that the latent structure transfers across independently constructed behavioural domains.

4.2 Finding 2: Temporal Stability

Using the same users embedded independently from January 1–15 versus January 16–31 (8.6 million and 8.8 million users per half, 21,325 subreddits, 1.08 million qualifying users):

Metric	Value
Mantel r (pairwise structure)	0.412
Per-user position correlation (mean)	0.225
Stable CCA dimensions	11
Nearest-neighbour median rank	439 / 3,000
Cluster stability (ARI, $k = 3$)	0.465

The space measures something persistent about individuals. The pairwise similarity structure—who is close to whom—is highly stable. Three subpopulations persist with 88% retention for the largest cluster.

4.3 Finding 3: Event Detection and Classification

Three events occurring within the data window are detected via displacement analysis:

Event	Mag. ratio	Coh. ratio	Signature
Tonga eruption (Jan 15)	1.49×	2.38	Unifying
Rogan/Spotify (Jan 24)	1.37×	3.97	Unifying*
Ukraine invasion (Feb 24)	1.55×	0.37	Polarising

*Within the broad Jan 1–15 vs 16–31 window; the tighter Jan 1–23 vs 25–31 split showed a polarising signature.

Magnitude ratio: how much more event-engaged users displaced relative to background ($p < 10^{-6}$ for all three, Mann-Whitney U). Coherence ratio: pairwise cosine similarity of displacement vectors within the event group divided by the background group. Ratios > 1 indicate unifying (event users move together); < 1 indicates polarising (event users scatter). See Figure 2.

4.4 Finding 4: Influence Propagates as a Wave

For the r/antiwork community, which experienced a burst of new participants around the January 26 Fox News interview (5,437 users with baseline embeddings), the Spearman correlation between latent-space distance from early adopters and arrival time is $\rho = 0.266$, $p < 10^{-6}$. Users in the closest quartile by latent-space distance arrived at a mean of 4.9 days; users in the farthest three quartiles arrived at 11–12 days. The closest quartile is the "orbit" of direct influence. Beyond it, arrival time is flat—broadcast media dominates (Figure 3).

4.5 Finding 5: Topic-Specific Leadership

Per-dimension barcode matching across the 6-month position tensor (16,507 users, 181 days, 19 dimensions, 40 centroid features) reveals:

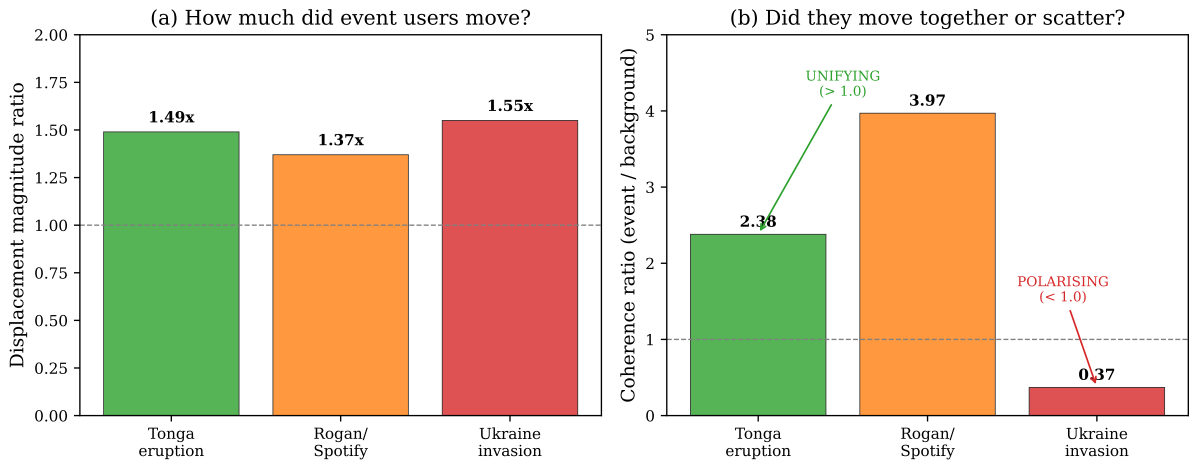

Figure 2: Event displacement analysis. (a) All three events moved engaged users significantly more than background. (b) The coherence ratio distinguishes unifying events (Tonga eruption: users moved together) from polarising events (Ukraine invasion: users scattered).

Analysis	**Enrichment**	**Interpretation**
Global (all dims averaged)	1.0×	No persistent leaders
Per-dimension (mean of 19)	18.5×	Strong persistent leaders
Per-dimension (max, dim 2)	39.9×	

This is a Simpson's paradox (Figure 4): the signal is strong within each dimension but cancels when averaged across dimensions because different people lead on each dimension. The Jaccard overlap between dimension-specific leader sets is 0.00–0.19, confirming that leadership is topically partitioned.

4.6 Finding 6: Influence Propagates Along the Topology

The Spearman correlation between Jaccard leader-overlap and cross-dimension lead advantage is $\rho = 0.564$, $p < 10^{-6}$ (342 dimension pairs). Dimensions that share leaders show cross-dimensional influence; dimensions with disjoint leaders do not.

The dimension topology at $k = 5$ clusters:

Cluster	**Semantic content (from SVD loadings)**
$\{0, 5, 6, 9\}$	General engagement, sports, right-leaning politics
$\{1, 3, 7, \ldots, 18\}$	Culture, gaming, finance, outrage, conspiracy
$\{15\}$	Gaming + crypto (isolated)
$\{2\}$	Worker rights / antiwork (isolated)
$\{4, 11\}$	Memes + gaming

The dimension topology is shown in Figure 5, and the relationship between leader overlap and cross-dimensional influence is shown in Figure 6.

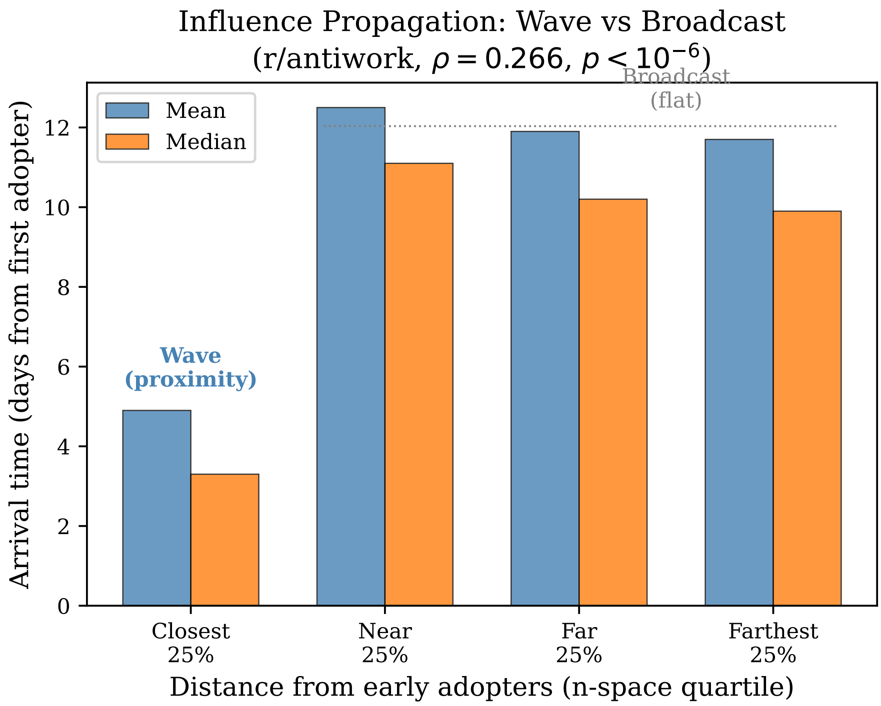

Figure 3: Influence propagation in the r/antiwork community. Users closest to early adopters in the latent space arrive 2–3× sooner than distant users, whose arrival time is flat—the signature of broadcast media dominating beyond the proximity orbit.

5 Topology and Resilience

5.1 Finding 7: Topological Isolation Predicts Fragility

Dimension 2—worker rights, the r/antiwork movement—has zero Jaccard overlap with every other dimension. Its leaders do not lead anywhere else. No one from any other dimension leads there. It is a structurally isolated island.

On January 26, 2022, a moderator of r/antiwork appeared on Fox News in an interview widely regarded as disastrous. The subreddit imploded. Our displacement data from this period shows the event was polarising (coherence ratio 0.37 in the cross-month analysis) with high magnitude (1.55× background).

The topology explains the collapse. With zero edges to adjacent dimensions, the antiwork movement had no redundant leadership, no alternative paths for influence, and no cross-dimensional embedding to absorb the shock. A single point of failure—one bad interview by one unelected spokesperson—was sufficient to destroy the movement's coherence because the topological structure provided no resilience.

Compare the large connected cluster $\{1, 3, 7, 8, 10, 12, 13, 14, 16, 17, 18\}$—culture war, gaming, finance, outrage, conspiracy. These dimensions share leaders (Jaccard up to 0.21), cross-predict each other's adoption (advantage up to 8.7 days), and form a densely connected subgraph. A failure in any one dimension does not collapse the network because influence routes around the damage through alternative paths.

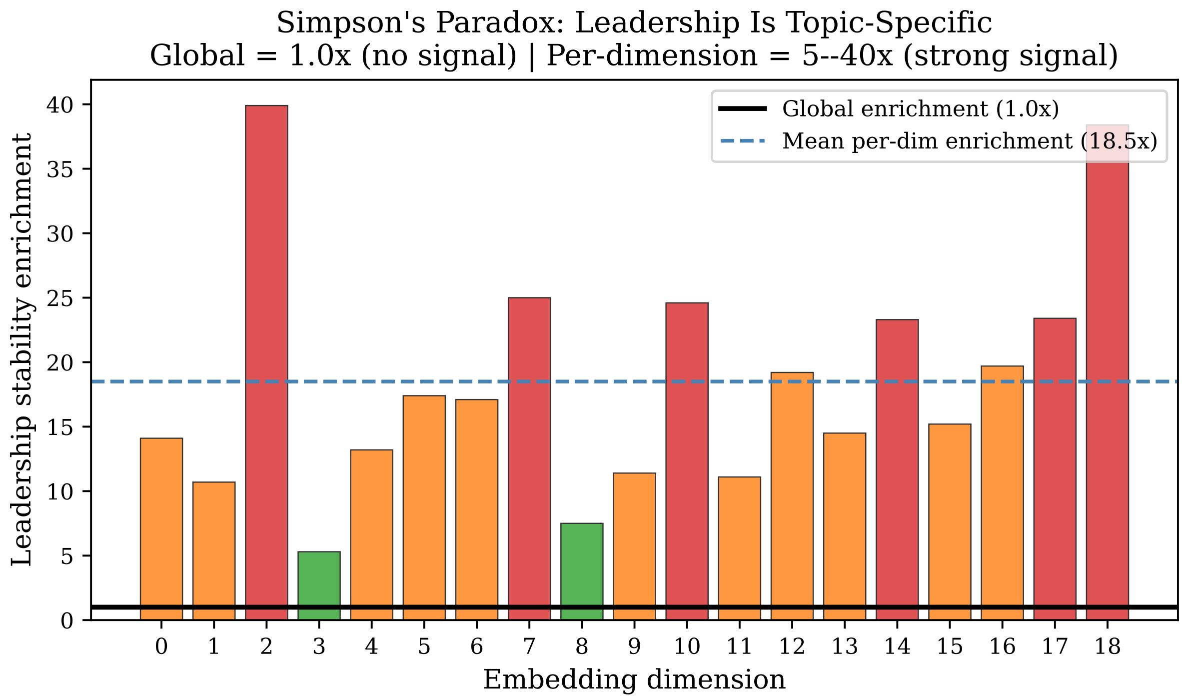

Figure 4: Simpson's paradox in leadership stability. Each dimension individually shows strong leadership persistence (5–40× enrichment), but the global analysis averages across dimensions where different people lead, producing a flat 1.0× enrichment that masks the per-dimension signal entirely.

5.2 A Predictive Claim

The topology makes a falsifiable prediction: a new movement whose leaders have zero Jaccard overlap with adjacent dimensions is structurally fragile and will collapse under its first serious external shock. A movement whose leaders overlap with 3–4 other dimensions will survive equivalent shocks.

This prediction is testable on future data without modification to the framework.

6 Discussion

6.1 What the Framework Measures

The latent space is not a model of belief, opinion, or identity. It is a model of behavioural similarity. Two users are close because they do similar things in similar places at similar times, not because they agree. The dimensions are not labelled "politics" or "gaming" by the researcher; they emerge from the variance structure of the data and are named post hoc from their subreddit loadings.

This distinction matters because the framework's validity does not depend on correctly interpreting what the dimensions mean. The cross-domain test, temporal stability, event detection, and influence propagation results all operate on the geometry of the space without reference to dimension semantics. The semantics are useful for interpretation but are not load-bearing for the methodology.

6.2 Limitations

Platform specificity. All findings are from Reddit. The latent structure may reflect Reddit's recommendation algorithms, user interface, and demographic composition rather than universal properties of human behavioural similarity. Replication on a second platform is required before generalization.

Population bias. Reddit users are not representative of the US population. The latent space may describe the structure of Reddit-using behaviour, not human behaviour in general.

Temporal resolution. Daily position estimates from community membership are coarse. Thread-level or comment-level analysis would provide finer temporal resolution for influence propagation measurement.

Embedding stability. The global SVD coordinate system is fit once and applied to all daily snapshots. If the latent structure itself evolves over six months, the coordinate system may drift. Sliding-window re-embedding with alignment would address this.

Leader identification. The barcode matching approach identifies users whose per-dimension trajectories lead the centroid. It does not establish that these users *cause* the centroid to move. The causal direction—do leaders move culture, or do they simply detect where culture is moving before others?—cannot be resolved from observational data alone.

6.3 Applications

Influence measurement. The per-dimension leadership score, combined with the topological adjacency structure, provides an influence metric fundamentally different from reach. A user with high dimension-specific leadership and cross-dimensional Jaccard overlap has measurable, propagating influence. A user with high reach but no leadership score has visibility without influence.

Fragility detection. The topological isolation metric—number of edges, mean Jaccard with adjacent dimensions—predicts structural fragility of online movements before they face external shocks.

Event classification. The displacement coherence ratio classifies events as unifying or polarising in real time from behavioural data, without requiring content analysis or sentiment classification.

Representative sampling. The stable subpopulation structure enables selection of representative "juries" that capture the population's diversity in the latent space, for testing the impact of content or policy changes before broad deployment.

7 Conclusion

We have shown that a latent space derived from behavioural data is not merely a dimensionality reduction convenience. It is a structured object with persistent geometry, measurable dynamics, and a topology that predicts real-world outcomes.

The seven findings build on each other: the space must exist (Finding 1) and be stable (Finding 2) before events can be detected (Finding 3). Events must be detectable before influence propagation can be measured (Finding 4). Influence propagation must be measurable before leadership can be localised to specific dimensions (Finding 5). Dimension-specific leadership must exist before the topology connecting dimensions can be constructed (Finding 6). And the topology must exist before its predictive power for resilience can be assessed (Finding 7).

Each finding was designed to fail fast if the prior findings were wrong. None failed. The framework is empirically grounded at every level.

The topology of the latent space—defined by who leads where and how those leader populations overlap—is the central contribution. It connects influence propagation to organisational structure in a way that is computable from behavioural data, requires no content analysis, makes no assumptions about what the dimensions represent, and produces falsifiable predictions about future events.

The space is not random. The influence, when it comes, will travel a direction the topology already defines.

Data and Code Availability

The analysis pipeline and all intermediate artifacts (position tensors, centroid features, leader scores) are available at the accompanying repository. The underlying Reddit data is sourced from the Pushshift archive via Academic Torrents and the Internet Archive.

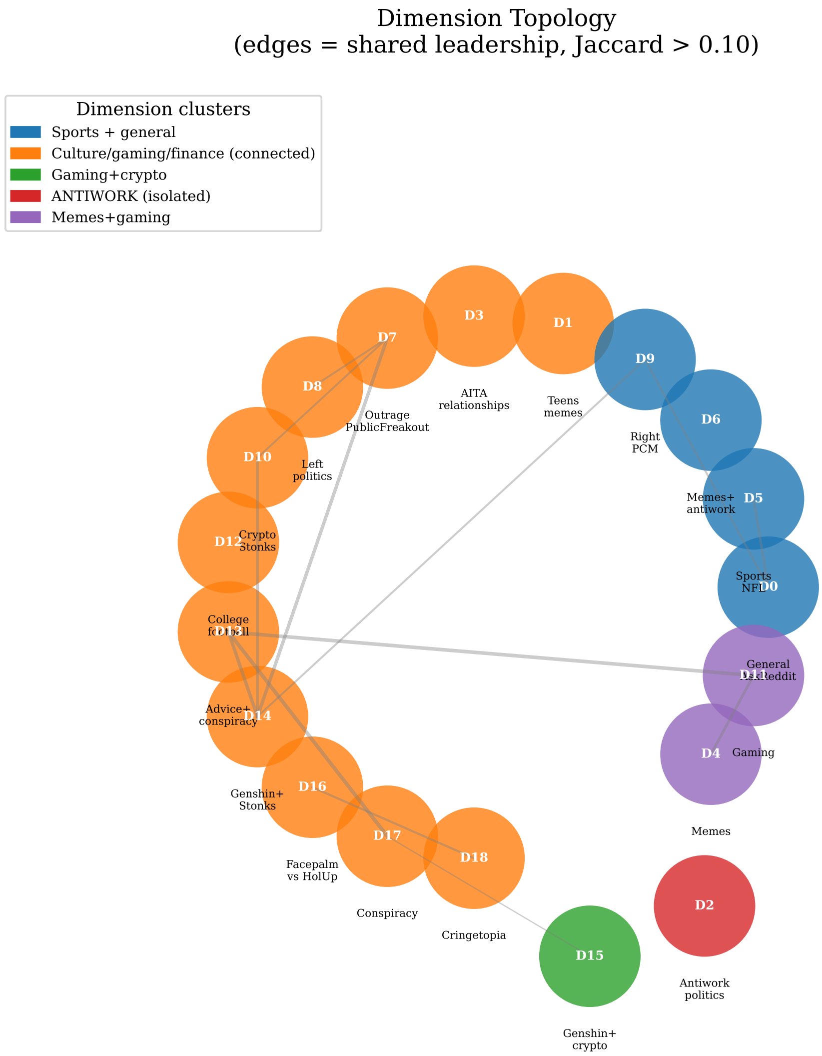

Figure 5: Dimension topology graph. Nodes are latent dimensions, coloured by cluster. Edge width is proportional to Jaccard leader overlap. Dimension 2 (antiwork/worker rights, red) is completely isolated—its leaders share no overlap with any other dimension. The large orange cluster (culture, gaming, finance, outrage) is densely connected with multiple influence pathways.

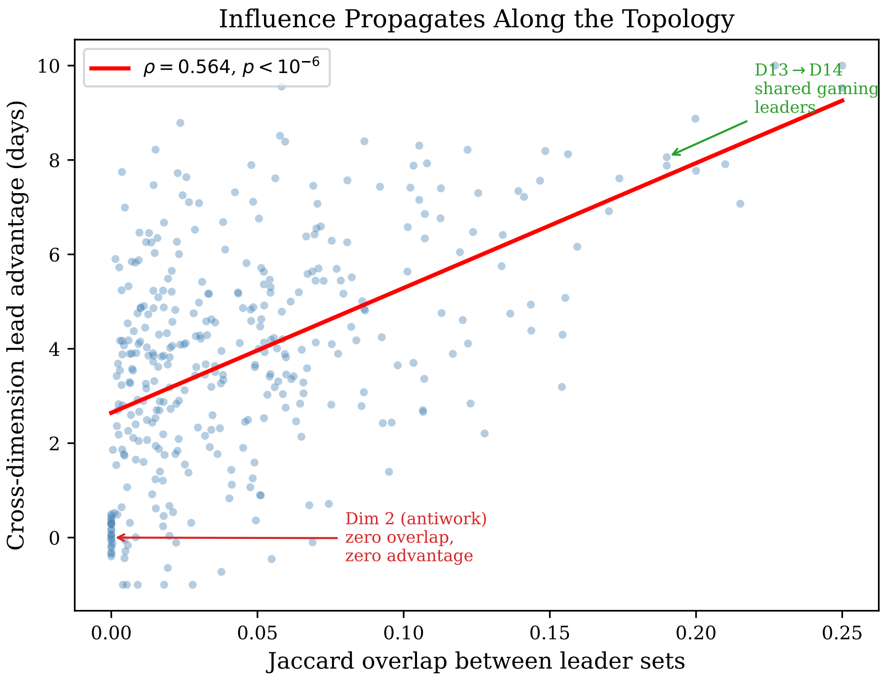

Figure 6: Influence propagates along the topology. Each point is a dimension pair. The x-axis is the Jaccard overlap between their leader sets; the y-axis is the cross-dimension lead advantage (how many days earlier dimension-A leaders adopt dimension-B features). The positive correlation ($\rho = 0.564$) confirms that shared leadership identity predicts influence flow between dimensions.