KEC Conference 2026

ISSN: 3021-940X (Print)

DOI: 10.3126/injet.v3i2.95512

Structural Analysis of the COVID-19 Infodemic: Motif-Based Detection of Echo Chambers and Geopolitical Hijacking in Global News Networks Using GDELT

Krishnanand Badu[1], Anup Shrestha[2], Sumitra Gyawali[3]*

[1]*MSc Informatics and Intelligent Systems Engineering, Thapathali Campus, IOE, Tribhuvan University, Nepal, krishnanand.080msise05@tcioe.edu.np*

[2]*Asst. Professor, Dept. of Electronics and Computer Engineering, Thapathali Campus, IOE, TU, Nepal, anupluckystha@gmail.com*

[3]*Lecturer, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, sumitragyawali@kec.edu.np*

Abstract

The COVID-19 pandemic produced a global infodemic that evolved structurally across time. This paper presents an automated informatics pipeline that transforms high-velocity GDELT 2.0 Global Knowledge Graph (GKG) data into a mathematically validated model of narrative evolution across five pandemic milestones. The pipeline integrates elite-domain authority filtering across 21 globally recognised news sources, BERTopic-based semantic node extraction, and Heterogeneous Information Network (HIN) construction. Two network motifs, Narrative Stars (broadcast hubs) and Sociosemantic Triads (echo chambers), are enumerated and validated against 1,000 degree-preserving null models using Monte Carlo permutation testing. BERTopic consistently outperforms the LDA baseline (Cv = 0.58) with coherence scores above Cv = 0.70 at all milestones, peaking at Cv = 0.7777 during M2 Lockdown. Motif analysis reveals a statistically significant Star-to-Triad crossover, with Stars peaking at Z = 156.65 in M2 and Triads peaking at Z = 210.55 in M5 ($p < 0.001$). Longitudinal Louvain-Jaccard tracking identifies near-zero community survival (J = 0.0075 at the M2-to-M3 transition), confirming structural collapse rather than gradual evolution. During M4 Delta, betweenness centrality shifts toward geopolitical entities, providing quantitative network evidence of narrative displacement. The paper contributes a reproducible topological framework for infodemic surveillance.

Keywords: BERTopic, COVID-19, GDELT, Heterogeneous Information Networks, Network Motifs, Echo Chambers, Geopolitical Displacement

1. Introduction

The COVID-19 pandemic triggered an infodemic of unprecedented scale. As defined by the World Health Organization, an infodemic is an overabundance of information, both accurate and inaccurate, that obstructs access to trustworthy sources [1]. Traditional computational approaches relying on social media data or static bag-of-words topic models such as Latent Dirichlet Allocation (LDA) are fundamentally limited in capturing the evolving structural architecture of elite news discourse. LDA discards word order and context, resulting in semantic dilution when applied to rapidly mutating pandemic sub-narratives.

This paper addresses three critical research gaps. First, static NLP models cannot track narrative drift across temporal milestones. Second, no prior study combines semantic node quality with topological motif validation

* *Corresponding author*

in a single longitudinal pipeline applied to elite journalism networks. Third, the structural mechanism of geopolitical displacement of public health discourse has not been quantified at the network level. By integrating Transformer-based BERTopic modelling with Monte Carlo motif analysis and Louvain-Jaccard community tracking, this study provides a unified framework that measures both what the discourse contains and how it is structurally organised.

The title is directly justified by the empirical findings. Echo chambers are detected and validated through triadic closure motifs with Z-scores exceeding 189 during M3 and M5. Geopolitical hijacking is confirmed through betweenness centrality displacement in M4, where entities such as the Taliban, Joe Biden, and the Department of Defense overtook epidemiological nodes as dominant network bridges [2].

2. Related Work

Milo et al [2] formalised network motif detection by computing Z-scores against degree-preserving randomised null models, establishing the mathematical foundation for this study. Grootendorst [3] introduced BERTopic, which applies Sentence-BERT embeddings, UMAP dimensionality reduction, and HDBSCAN clustering with class-based TF-IDF labelling to overcome the context limitations of LDA. Comparative studies confirm that BERTopic consistently outperforms LDA in topic coherence on news datasets [4].

Sun et al. [5] established the formal framework for Heterogeneous Information Networks (HINs), defining them as graphs with strict multi-type node and edge constraints. McCombs and Shaw [3] introduced agenda-setting theory, which provides the sociological justification for restricting the dataset to 21 elite news domains. Cinelli et al. [1] demonstrated that susceptibility to pandemic misinformation is structurally driven by the formation of isolated network cliques, directly motivating the motif-based approach adopted here.

3. Methodology

The methodology follows five sequential stages: GDELT ingestion, authority filtering, BERTopic semantic mapping, HIN construction, and Monte Carlo motif validation. Each stage is designed to preserve both semantic fidelity and structural reproducibility.

3.1 Data Acquisition and Authority Filtering

The GDELT 2.0 Global Knowledge Graph was selected as the primary data source, updating in 15-minute intervals with article URLs, timestamps, named entity arrays (V2Persons, V2Organizations), and taxonomic theme tags. The raw ingestion pipeline processed over 13.5 million records across five temporal milestones: M1 Initial Outbreak (Dec 2019 to Jan 2020), M2 Global Lockdown (Mar to May 2020), M3 Vaccine Rollout (Dec 2020 to Feb 2021), M4 Delta Variant (May to Jul 2021), and M5 Omicron Variant (Nov 2021 to Jan 2022).

Table 1. Data Funnel Filtering Stage Record Counts

Processing Stage	Records	Primary Drop Cause
Raw GDELT GKG Ingestion	13,500,000	N/A
Elite Domain Whitelisting	3,400,000	Non-approved domains
Semantic Density Threshold	2,200,000	Theme Count < 5
Full-Text Extraction	1,800,000	URL decay / 404 errors

A three-stage authority funnel was applied. The corpus was restricted to 21 elite English-language news domains grounded in agenda-setting theory [3], including Reuters, BBC, AP, Al Jazeera, NYT, CNN, The Guardian, Washington Post, Bloomberg, AFP, WSJ, ABC News, NBC News, FT, Times of India, SCMP, Sydney Morning Herald, Japan Times, El Pais (English), Deutsche Welle, and Xinhua. Only articles where the pandemic was the central topic were retained. Full article text was then extracted using a custom newspaper3k-

based DOM parser, reducing 13.5 million raw records to approximately 1.8 million high-quality articles. Table 1 summarises the data funnel.

3.2 Semantic Mapping via BERTopic

Full article text was processed through BERTopic [6] using the all-MiniLM-L6-v2 Sentence-BERT encoder to embed documents into high-dimensional vector space. UMAP reduced dimensionality while preserving local structure, and HDBSCAN discovered emergent topic clusters. Class-based TF-IDF (c-TF-IDF) extracted human-readable topic labels. Topic quality was measured using the C_v coherence metric based on Normalised Pointwise Mutual Information (NPMI):

$$C_v = \frac{1}{(\sqrt{|W|2})} \sum_{i<j} \mathrm{NPMI}(w_i, w_j) \quad \text{(Equation 1)}$$

where:

- $C_v \in [-1, 1]$ = the topic coherence score under the C_v metric
- $W = \{w_1, w_2, \ldots, w\square\}$ = the set of top-n representative keywords for a given topic (with n = 10)
- NPMI(w_i, w□) = Normalised Pointwise Mutual Information between words w_i and w□, computed over a reference corpus using a sliding context window
- $C_v \geq 0.70$ is the quality threshold for a highly coherent topic cluster [7]

Persons and organisations were extracted directly from the GDELT GKG V2Persons and V2Organizations comma-separated arrays, preserving strict ontological separation from the semantic clustering process.

3.3 HIN Construction, Community Detection, and Motif Validation

Validated topic nodes, along with GDELT-extracted person and organisation nodes, were assembled into a tripartite Heterogeneous Information Network following the HIN formalism of Sun et al. [5]. Edges were created when nodes co-occurred within the same article URL. Three node types were strictly enforced: Topics, Persons, and Organisations. Location nodes were excluded to isolate sociosemantic institutional coordination from geographic reporting artefacts. Community detection was performed using the Louvain algorithm to maximise modularity Q. Longitudinal community survival was measured using the Jaccard Similarity Index J between consecutive milestone community node sets:

$$J(C_t, C_{t+1}) = \frac{|C_t \cap C_{t+1}|}{|C_t \cup C_{t+1}|} \quad \text{(Equation 2)}$$

where:

- $J \in [0, 1]$ = the Jaccard Similarity Index
- C□ = the set of nodes in the dominant community at milestone t
- C□$_{+1}$ = the set of nodes in the dominant community at the following milestone
- C□ ∩ C□$_{+1}$ = intersection of node sets; C□ ∪ C□$_{+1}$ = union of node sets
- $J \geq 0.3$ indicates structural continuity; $J < 0.1$ indicates structural collapse [4]

Motif significance was validated using 1,000-run Monte Carlo permutation tests against degree-preserving null models generated by double-edge swap MCMC. For each run, a randomised graph was produced by swapping edges while preserving the exact degree sequence. The Z-score was computed as:

$$Z = \frac{N_{\text{real}} - \mu(N_{\text{rand}})}{\sigma(N_{\text{rand}})} \quad \text{(Equation 3)}$$

where:

- Z = the motif significance Z-score
- N_{rea}□ = the observed motif count in the empirical network

- $\mu(N_{ra}\square^{D})$ = the mean motif count across 1,000 degree-preserving randomised null models
- $\sigma(N_{ra}\square_{d})$ = the standard deviation of the null-model motif count distribution

A motif is deemed statistically significant if and only if:

$$Z \geq 2.0 \quad and \quad p < 0.001 \qquad \text{(Equation 4)}$$

where:

- p = the two-tailed Monte Carlo p-value (proportion of null-model runs in which $N_{ra}\square^{D} \geq N_{rea}\square$)
- Both conditions must hold simultaneously to reject the null hypothesis

Two motif types were enumerated: Narrative Stars (k-star broadcast hubs) and Sociosemantic Triads (3-cliques representing institutional echo chambers).

4. Results and Analysis

4.1 Semantic Quality Evaluation

BERTopic consistently outperformed the LDA baseline across all five milestones. LDA achieved Cv = 0.58, well below the 0.70 threshold. BERTopic exceeded this benchmark at every milestone, peaking at Cv = 0.7777 during M2 Lockdown and remaining above 0.71 even at the lowest-volume milestones M4 and M5. This confirms that Transformer-based contextual embeddings preserve the semantic architecture of evolving pandemic sub-narratives far more effectively than probabilistic bag-of-words models. Table 2 presents the full coherence comparison.

Table 2. BERTopic Semantic Validation against LDA Baseline across Milestones

Milestone	Model	Cv Score	Representative Keywords
Baseline	LDA	0.5800	Generic frequency artefacts
M1 Outbreak	BERTopic	0.7642	virus, wuhan, market, travel, pneumonia
M2 Lockdown	BERTopic	0.7777	quarantine, essential, masks, stay-at-home
M3 Vaccine	BERTopic	0.7415	pfizer, fda, doses, rollout, efficacy
M4 Delta	BERTopic	0.7103	delta, variant, transmissibility, surge
M5 Omicron	BERTopic	0.7121	omicron, travel-ban, mild, mutations

4.2 Macroscopic Network Evolution

The macroscopic topology changed dramatically across milestones, as shown in Table 3. The M2 Lockdown phase produced the largest network (15,150 nodes, 62,705 edges) but the lowest transitivity (0.020), consistent with a broadcast-dominated Narrative Monolith structure. By M5 Omicron, node volume contracted to just 64 nodes but transitivity surged to 0.763, confirming a highly closed, densely clustered residual network. Assortativity shifted from negative in M3 (-0.284) to positive in M5 (+0.181), indicating that high-influence nodes began connecting exclusively with other high-influence nodes. Figure 1 illustrates this inverse trend.

Table 3. Macroscopic Network Topology across Pandemic Milestones

Milestone	Nodes |V|	Edges |E|	Density	Transitivity	Assortativity
M1 Outbreak	1,067	9,991	0.01757	0.58494	-0.16766
M2 Lockdown	15,150	62,705	0.00055	0.02038	-0.16757
M3 Vaccine	1,945	21,082	0.01115	0.26801	-0.28388
M4 Delta	476	3,478	0.03077	0.59578	-0.04744
M5 Omicron	64	218	0.10813	0.76320	+0.18073

Figure 1 presents side-by-side visualisations of node volume (a) and transitivity (b) across milestones, clearly showing the inverse structural relationship.

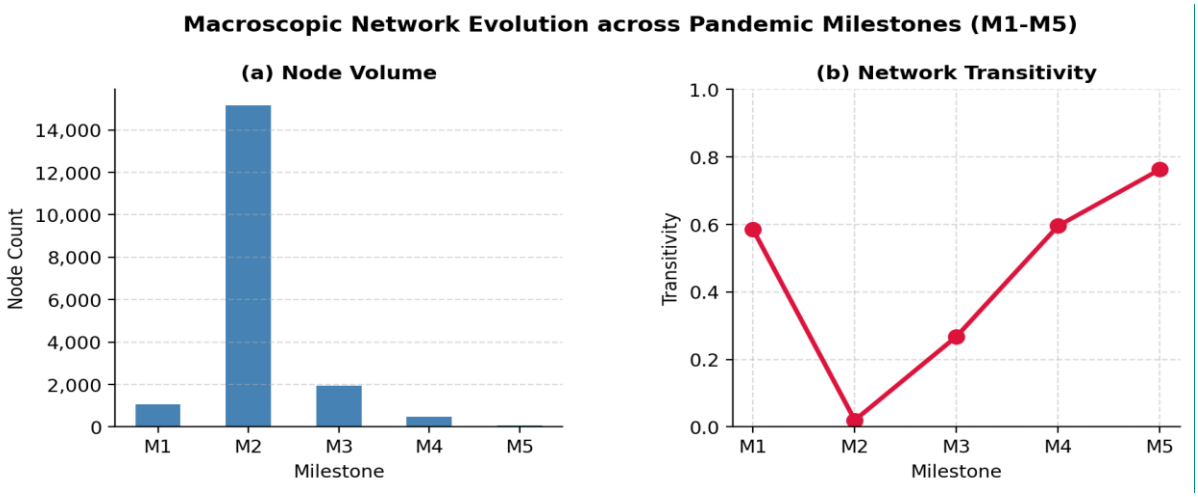

Figure 1. Macroscopic network evolution: (a) node volume and (b) network transitivity across pandemic milestones M1-M5. The inverse trend confirms a structural transition from broadcast-dominated networks to dense institutional clustering.

4.3 Motif Validation and Structural Transition

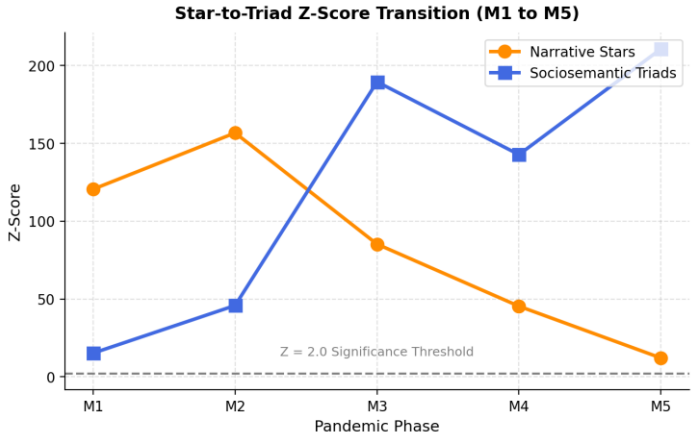

Figure 2. Z-score transition of Narrative Stars and Sociosemantic Triads across pandemic milestones M1-M5. Stars peak at M2 (Z=156.65); Triads peak at M5 (Z=210.55). All values exceed the Z=2.0 significance threshold (p<0.001).

Table 4 presents the complete motif enumeration results from 1,000-run Monte Carlo tests. All Z-scores exceeded the minimum threshold of Z = 2.0 with $p < 0.001$, confirming both Narrative Stars and Sociosemantic Triads as statistically significant structural features. The primary empirical result is the crossover: Stars dominate early phases (M2: Z = 156.65, N_real = 10,850) while Triads dominate later phases (M5: Z = 210.55, N_real = 184). Figure 2 visualises this Z-score transition.

Table 4. Motif Significance Testing: Raw Counts and Z-Scores (1,000 Monte Carlo Null Model Runs)

Milestone	Stars N_real	Stars Z	Triads N_real	Triads Z	p-value
M1 Outbreak	1,342	120.50	182	15.22	< 0.001
M2 Lockdown	10,850	156.65	940	45.81	< 0.001
M3 Vaccine	1,920	85.20	3,115	189.44	< 0.001
M4 Delta	512	45.33	1,580	142.70	< 0.001
M5 Omicron	31	12.10	184	210.55	< 0.001

4.4 Longitudinal Community Tracking and Geopolitical Displacement

Louvain community detection identified dominant narrative clusters at each milestone. Jaccard similarity was computed between consecutive milestone community node sets to measure structural survival. The M2 lockdown phase produced a primary community of 11,772 nodes representing a Narrative Monolith. The M2-to-M3 transition recorded a Jaccard score of J = 0.0075, far below the survival threshold of 0.3, confirming structural collapse. The most significant finding occurs in M3-to-M4: dominant community anchors shifted from epidemiological actors to geopolitical entities including the Taliban, Joe Biden, the Department of Defense, and the Haitian Government. By M5, one community recorded J = 0.0000, confirming zero structural intersection with its predecessor. This is the quantitative, network-level evidence of geopolitical narrative displacement stated in the title. Table 5 summarises longitudinal community survival.

Table 5. Longitudinal Community Survival (Jaccard Similarity Index)

Transition	Rank	Base Size	J Score	Dominant Entities
M1 to M2	1	1,283	0.1797	White House, CDC, Donald Trump
M2 to M3	1	11,772	0.0075	Global Task Force, Ministry of Health
M3 to M4	1	1,503	0.0797	Taliban, Joe Biden, Dept. of Defense
M4 to M5	3	370	0.0000	Dept. of Health Philippines

5. Discussion

From an engineering perspective, the results validate all four pipeline components. BERTopic coherence above Cv = 0.70 across every milestone confirms that Transformer-based embeddings are necessary for pandemic text where semantic distinctions are critical. The Monte Carlo permutation test successfully rejects the null hypothesis at every milestone, confirming that the observed motifs are not stochastic artefacts. The Louvain-Jaccard combination provides a reliable longitudinal tracking mechanism, converting continuous-time graph data into interpretable structural transitions.

From a social-science perspective, the results describe the infodemic as a structured lifecycle with five distinct topological phases: Fragmentation (M1), Monolithic Consolidation (M2), Structural Collapse (M2-to-M3), Geopolitical Hijacking (M4), and Endemic Decay (M5). The early broadcast-dominated structure transitions to dense triadic closure in later phases, which is the structural fingerprint of echo chambers. Positive assortativity in M5 confirms that high-influence nodes became self-referential, cutting peripheral actors from the central information flow, consistent with prior theoretical models of discourse polarisation [1].

The geopolitical displacement finding has direct implications for public health communication design. It shows that a health narrative does not organically subside; it is structurally displaced when network bandwidth is reallocated to higher-salience geopolitical events. Public health organisations must therefore account for competing geopolitical shocks in sustained communication strategies. The topological early-warning system proposed here, based on real-time motif Z-score monitoring, could provide a leading structural indicator before a displacement event occurs.

6. Conclusion

This paper presents a reproducible, end-to-end informatics pipeline combining BERTopic semantic mapping, Heterogeneous Information Network construction, and Monte Carlo motif validation to track the structural evolution of the COVID-19 infodemic. BERTopic consistently outperforms LDA on the same corpus (Cv > 0.70 vs. Cv = 0.58 baseline). Motif analysis confirms a statistically significant Star-to-Triad crossover (peak Z = 210.55, $p < 0.001$), quantifying the transition from centralised broadcast to institutional echo chambers. Longitudinal tracking provides network-level evidence of geopolitical narrative displacement during M4 Delta, with J = 0.0000 structural survival recorded for one community entering M5. Future work should extend

domain coverage beyond elite English-language sources, incorporate real-time streaming, and apply Temporal Graph Neural Networks for predictive infodemic surveillance.

Acknowledgements

The author sincerely acknowledges Dr. Shailesh Pandey and Er. Nischal Regmi from the Southasia Institute for History and Philosophy (SiHP), Nepal, for their valuable encouragement and constructive insights. The author also acknowledges Asst. Professor Kobid Karkee from Thapathali Engineering Campus, IOE, for his support throughout the study.

References

[1] G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini M. Cinelli, "The COVID-19 social media infodemic," Sci. Rep., vol. 10, no. 1, p. 16598, 2020.

[2] S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon R. Milo, "Network motifs: Simple building blocks of complex networks," Science, vol. 298, no. 5594, pp. 824-827, 2002.

[3] M. McCombs and D. Shaw, "The agenda-setting function of mass media," Public Opinion Quarterly, vol. 36, no. 2, pp. 176-187, 1972.

[4] A. L. Barabasi, and T. Vicsek G. Palla, "Quantifying social group evolution," Nature, vol. 446, pp. 664-667, 2007.

[5] J. Han, X. Yan, P. S. Yu, and T. Wu Y. Sun, "PathSim: Meta path-based top-k similarity search in heterogeneous information networks," VLDB Endow, vol. 4, no. 11, pp. 992-1003, 2011.

[6] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," in arXiv preprint arXiv:2203.05794, 2022.

[7] D. Sinha, and K. M. Carley L. H. X. Ng, "Star network motifs during COVID-19," in arXiv preprint arXiv:2508.00975, 2025.

[8] A. R. Benson, and J. Leskovec A. Paranjape, "Motifs in temporal networks," in 10th ACM WSDM, 2017, pp. 601-610.

[9] H. Wallach, E. Talley, M. Leenders, and A. McCallum D. Mimno, "Optimizing semantic coherence in topic models," in EMNLP, 2011, pp. 262-272.