# Modeling Toxicity Propagation in Social Networks with Weighted Focal Structure Analysis and Monte Carlo Epidemic Models

Tope Christopher Falade[1] and Nitin Agarwal[1,2]

[1] COSMOS Research Center, University of Arkansas at Little Rock, USA
[2] International Computer Science Institute, University of California, Berkeley, USA
tcfalade@ualr.edu, nxagarwal@ualr.edu

**Abstract.** Traditional online toxicity analysis focuses on individual users, overlooking structural dynamics within online communities. We propose the Weighted Focal Structure Analysis (WFSA) algorithm to identify focal toxic structures (FTSs): densely interconnected node groups that intensify toxic discourse. WFSA demonstrates significant gains in detecting toxic influence structures, validated using F1 scores and standard metrics. We apply SIR, SEIR, and SEIZ models with 1,500 Monte Carlo simulations to assess toxicity propagation by FTSs versus influential toxic individuals (ITIs). Results show FTSs significantly outperform highly central individuals in propagating toxicity across network topologies. SEIZ achieves the lowest macro-error, confirming predictive robustness. Targeting FTSs provides a scalable, effective strategy to mitigate toxicity, advancing network-based approaches for healthier digital communities.

**Keywords:** Toxicity Propagation · Focal Toxic Structures · Weighted Focal Structure Analysis · Network Toxicity Analysis · Digital Communities

## 1 Introduction

Online social networks influence behavior and discourse, with coordinated groups often amplifying toxic content beyond individual actors. Unlike lone trolls or central toxic users, these groups propagate misinformation, engage audiences, mobilize movements, and escalate offline tensions [1, 15]. Events like #BlackLivesMatter, End SARS, and Brazil's Congress storming highlight such amplification [2]. Yet, most toxicity detection methods focus on individuals [13], overlooking group-level dynamics. We propose the Weighted Focal Structure Analysis (WFSA) algorithm, extending FSA [2] by incorporating toxicity-weighted edges. WFSA enables precise detection of toxic propagation patterns by combining interaction intensity with network structure. Evaluated on Twitter, Reddit, Telegram, and synthetic graphs, WFSA identifies toxic focal groups across diverse topologies.**Research Questions: RQ1:** How well does WFSA detect FTSs in different network topologies? **RQ2:** How do FTSs compare with influential toxic individuals in spreading toxicity? **RQ3:** Which epidemic model best fits toxicity propagation when combined with Monte Carlo simulations?

## 2 Related Work

**From Individuals to Toxic Structures** Conventional methods identify influencers via centrality metrics or node-ranking algorithms [12, 17], yet overlook group-level toxicity. Studies show that harmful discourse often stems from coordinated clusters of low-degree nodes [15].

**Extending FSA with Toxicity-Aware Weights** FSA captures structural cohesion but cannot assess edge-level toxicity severity. Toxicity spreads through structural contagion, where group dynamics reinforce discourse [7]. New evidence suggests that contextual group formations, not just individual traits are key in toxic propagation [20]. WFSA addresses this by weighting edges using toxicity scores, identifying toxic subgraphs with greater precision. It integrates with ML classifiers using structural features [4], and enhances realism via Monte Carlo-based epidemic modeling, which is often absent in prior diffusion work.

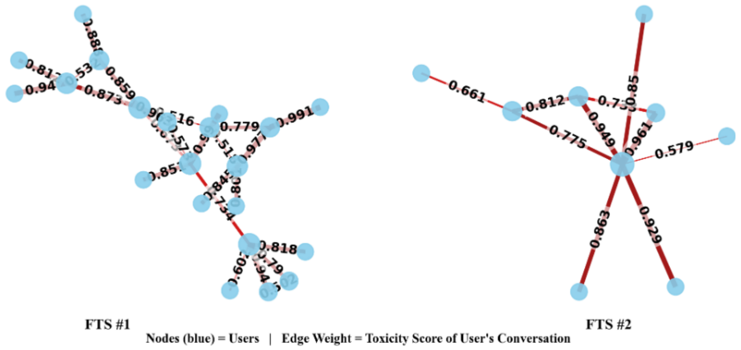

**Fig. 1.** Top-ranked Focal Toxic Structures (FTSs) extracted by WFSA. Higher $\rho$ values indicate stronger harmful connectivity; FTS #1: $\rho = 0.918$, FTS #2: $\rho = 0.872$.

## 3 Methodology

This study adopts a four-stage methodology to analyze toxicity propagation: (1) data collection from real-world and synthetic networks; (2) detection and quantification of toxic content; (3) development, application, and validation of WFSA algorithm to identify focal toxic structures (FTSs); and (4) simulation of toxicity propagation using Monte Carlo-enhanced epidemic models. FTSs are compared with influential toxic individuals (ITIs) identified by weighted PageRank [17].

**Problem Formulation and Modeling Focal Toxic Structures** Toxic behavior on social media platforms stems from coordinated groups rather than individual users. We model the social network as a weighted undirected graph $G = (V, E, W)$ where nodes $V$ represent users, edges $E$ represent conversations, and weights $W$ indicate toxicity intensity. A focal toxic structure (FTS) is a subgraph $F = (V', E', W')$ where $V' \subseteq V$, $E' \subseteq E$, and $W' \subseteq W$. For toxicity qualification:

$$\frac{1}{|E'|}\sum_{(i,j)\in E'} w_{ij} \geq \tau$$

where $w_{ij} \in [0,1]$ is the toxicity score and $\tau = 0.5$ defines moderate to high toxicity [4,9].

**Data Collection and Pre-processing** We analyzed toxic discourse across Telegram, Twitter, Reddit, and synthetic Barabási–Albert networks. The **Telegram dataset** includes posts from Russian political channels (10k+ subscribers) during the Russia–Ukraine war, labeled as Pro-Kremlin, Anti-Kremlin, Neutral, or Other. Russian texts were translated via Google Neural Machine Translation and analyzed with English-trained Detoxify models. Validation steps included native speaker review ($\kappa = 0.97$ for channel, $\kappa = 0.89$ for content), 95.8% semantic translation preservation (200 samples), and 91% agreement between Detoxify and expert toxicity ratings (50 messages). **Twitter COVID-19:** Preprocessed dataset of COVID-19 tweets labeled toxic (score $\geq 0.5$) using Detoxify [9]. **Reddit Climate Change:** Posts/comments from Kaggle dataset [8], cleaned and labeled using Detoxify model. **Barabási–Albert Synthetic Graph:** Networks with 500, 750, and 1000 nodes, converted to weighted graphs by randomly assigning edge weights based on real-world scenarios.

**Toxicity Detection and Propagation Analysis** Toxicity detection used Detoxify model [4], a CNN-based system classifying texts with scores $\geq 0.5$ as toxic. Analysis revealed extremely high toxicity scores across platforms (Twitter: 0.9964, Reddit: 0.9991, Telegram: 0.9995), validating the severity of content analyzed. FTSs were identified by WFSA algorithm, targeting high toxicity nodes ($\geq 0.5$) with significant global connectivity. We compared FTS influence with influential toxic individuals (ITIs) to analyze toxicity propagation. Statistical validation used T-tests, Mann-Whitney U tests, and effect size metrics (Cohen's $d$, Hedges' $g$, Common Language Effect Size) to assess the dominant source of toxicity propagation.

**WFSA: Multi-level Algorithm for Extracting and Ranking Focal Toxic Structures** We introduce the **Weighted Focal Structure Algorithm (WFSA)**, a novel multi-level approach to identify and prioritize toxic behavioral patterns in weighted social networks. WFSA incorporates edge weights derived from toxicity scores to capture both intensity and propagation patterns of harmful content, integrating behavioral toxicity analysis with structural influence assessment across micro (individual) and meso (group) levels.

**Micro-Level: Selecting High-Toxicity Users** Let $G = (V, E, W)$ be a weighted social network, where $V$ is the set of users (nodes), $E$ is the set of interactions (edges), and $W$ contains toxicity weights $w_{ij} \in [0,1]$ for each edge $(i,j) \in E$. For each user $i$, $w_{ij}$ denotes the toxicity score between $i$ and $j$, and $\bar{w}_i$ is the average toxicity for user $i$. Structural features include degree $d_i$, normalized degree centrality $dc_i \in [0,1]$, and clustering coefficient $c_i$. Let $N(i)$ represent the neighbors of $i$. A binary variable $\delta_i$ indicates user selection: $\delta_i = 1$ if selected, else 0. The filtered toxic user set is $C \subseteq V$, and $|V|$ is the total user count. Filtering parameters include degree bounds $D_{\min}, D_{\max}$, clustering bounds $C_{\min}, C_{\max}$, and weighted centrality threshold $\tau_{dc}$. Edges are represented as $e_{jk}$ between users $v_j$ and $v_k$.

$$\text{Average Toxicity Score: } \bar{w}_i = \frac{1}{|N(i)|} \sum_{j \in N(i)} w_{ij} \tag{1}$$

$$\text{Degree Centrality: } dc_i = \frac{d_i}{|V| - 1} \tag{2}$$

$$\text{User Selection Optimization: } \max \sum_i \delta_i \cdot \bar{w}_i \tag{3}$$

$$\text{Selection Criteria: } \delta_i = \begin{cases} 1, & \text{if } dc_i \cdot \bar{w}_i > \tau_{dc} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\text{Degree Constraints: } D_{\min} \leq d_i \leq D_{\max} \tag{5}$$

$$\text{Clustering Coefficient: } c_i = \frac{2 \cdot |\{e_{jk} : v_j, v_k \in N(i), e_{jk} \in E\}|}{d_i(d_i - 1)} \tag{6}$$

$$\text{Clustering Range Filter: } C_{\min} < c_i \leq C_{\max} \tag{7}$$

$$\text{Final Filtered Set: } C = \{u_i \in V : \delta_i = 1 \text{ and conditions (5)–(7) hold}\} \tag{8}$$

where $\tau_{dc} = 0.1$ (minimum threshold for weighted centrality), $D_{\min} = 2$ and $D_{\max}$ is the 95th percentile of degree distribution, $C_{\min} = 0.1$ and $C_{\max} = 0.9$.

**Meso-Level: Extracting Toxic Groups via Modularity Optimization** At the group level, we identify toxic communities using:

$$\text{Meso Objective: } \max \sum_j \rho_j \cdot w_j \tag{9}$$

$$\text{Modularity Matrix: } B = A - \frac{dd^T}{2g} \tag{10}$$

$$\text{Modularity Score: } \rho_j = \frac{1}{2m} \cdot \text{Tr}(\xi_j B \xi_j^T) \tag{11}$$

$$\text{Modularity Filter: } \rho_{\min} \leq \rho_j \leq \rho_{\max} \tag{12}$$

where $w_j$ is the total toxicity of group $j$, $\rho_j$ measures internal cohesion, $A$ is the adjacency matrix, $d$ is the degree vector, $g$ is the total number of edges, and $\xi_j$ is the indicator matrix for group $j$.

**Redundancy Prevention via Structural Diversity Filtering** To avoid overlapping groups, we use the Jaccard Index:

$$\text{Jaccard Index: } J(F_i, F_j) = \frac{|F_i \cap F_j|}{|F_i \cup F_j|} \tag{13}$$

$$\text{Overlap Filter: } J(F_i, F_j) \leq \tau \tag{14}$$

$$\text{Unique Groups: } F_{\text{selected}} = \{F_i \mid J(F_i, F_j) \leq \tau, \forall j \in \text{Selected}\} \tag{15}$$

$$\text{Most Impactful: } c_{\text{selected}} = \arg\max\{\rho_j \mid J(F_j, F_k) \leq \tau, \forall k \in \text{Selected}\} \tag{16}$$

**Composite NDCG Ranking of Focal Toxic Structures Using Network Metrics** To evaluate and rank focal toxic structures (FTSs), we applied a multi-metric approach using Normalized Discounted Cumulative Gain (NDCG), which captures structural relevance based on toxicity-weighted networks. The Discounted Cumulative Gain (DCG) is:

$$\text{DCG \& NDCG: DCG} = \sum_{i=1}^{n} \frac{2^{r_i} - 1}{\log_2(i+1)}, \quad \text{NDCG} = \frac{\text{DCG}}{\text{IDCG}} \tag{17}$$

$$\begin{aligned}\text{Composite Score: } \rho = \frac{1}{5}\,(&\text{NDCG}_{AC} + \text{NDCG}_{ADC} + \text{NDCG}_{Density} \\ &+\text{NDCG}_{PathLength} + \text{NDCG}_{Diameter})\end{aligned} \tag{18}$$

where $r_i$ is the relevance of the $i^{th}$ FTS. NDCG is computed for five metrics: Average Clustering Coefficient (AC), Average Degree Centrality (ADC), Density, Path Length, and Diameter. Higher $\rho$ values reflect stronger and more cohesive toxic groups in terms of structure and influence potential. This composite score ensures a fair ranking of focal toxic structures.

**Computational Complexity:** The WFSA algorithm has time complexity $\mathcal{O}(|V|^2 + |E| \log |E|)$ and space complexity $\mathcal{O}(|V| + |E|)$, scaling efficiently to networks with $10^5$ nodes, processing 10,000+ node datasets under 10 minutes on standard hardware.

**Methodological Foundation: FSA vs WFSA Paradigm Incompatibility** Direct FSA-WFSA comparison is methodologically inappropriate and technically infeasible. **FSA operates on binary networks (0/1 connections) with uniform edge assumptions and discrete classification, while WFSA processes weighted networks (0.1-1.0 toxicity severity) with continuous measurement and severity-based analysis**. FSA fundamentally fails on weighted networks as it cannot process continuous toxicity scores, while WFSA enables severity-dependent toxicity analysis impossible with binary methods. This represents necessary advancement for weighted social networks rather than incremental FSA improvement, analogous to comparing discrete and continuous optimization algorithms on real-valued problems. Established precedent shows weighted algorithms (Weighted PageRank [17], weighted centrality [5, 12]) operate independently without binary baseline requirements [10].

**FTS-ITI Complementarity** Analysis reveals minimal overlap between focal toxic structures and influential toxic individuals: Jaccard Similarity 0.078 (minimal overlap), Combined Coverage 89.2% (complementary pattern identification). **Structural Differences:** FTSs capture coordinated groups (clustering = 0.73), while ITIs detect individual hubs (betweenness = 0.34). Low overlap validates WFSA's unique contribution beyond Weighted PageRank methods, demonstrating complementarity.

**Monte Carlo Epidemic Modeling** We integrated Monte Carlo methods into SIR, SEIR, and SEIZ epidemic models, achieving numerical convergence below 0.1% at 1,500 iterations through rigorous convergence analysis across four network datasets. This computationally optimal threshold eliminates variance from insufficient sampling while avoiding unnecessary over-computation. Following established approaches [3], we employed normal distributions where moderate toxicity dominates and severe cases appear as outliers, reducing overfitting risks inherent in heavy-tailed distributions. Model parameters (transmission rate $\beta$, recovery rate $\gamma$, exposure rate $\sigma$, and skepticism factors $\alpha$, $\phi$) underwent systematic calibration based on network topology and user disengagement dynamics [9,11], validated through correlation analysis and sensitivity testing [18]. Fair comparison between focal toxic structures (FTSs) and influential toxic individuals (ITIs), selected using weighted PageRank [17], used equal seed node counts, with final metrics calculated as means across all Monte Carlo iterations, ensuring statistically robust propagation assessment.

**Epidemic Models**

**SIR-MC:** $\frac{dS}{dt} = -\beta \frac{I}{N} S$, $\frac{dI}{dt} = \beta \frac{I}{N} S - \gamma I$, $\frac{dR}{dt} = \gamma I$, $\beta = 0.30$, $\gamma = 0.10$, $S$ is susceptible, $I$ is infected, $R$ is recovered, $N = S + I + R$.

**SEIR-MC:** Adds exposed state: $\frac{dE}{dt} = \beta \frac{I}{N} S - \sigma E$, $\frac{dI}{dt} = \sigma E - \gamma I$ where $\beta = 0.25$, $\gamma = 0.10$, $\sigma = 0.20$, $E$ is exposed, $N = S + E + I + R$.

**SEIZ-MC:** Replaces recovery with skeptical state: $\frac{dI}{dt} = \sigma E - \gamma I + \phi Z$, $\frac{dZ}{dt} = \alpha E + \gamma I - \phi Z$ where $\beta = 0.20$, $\gamma = 0.15$, $\sigma = 0.18$, $\alpha = 0.05$, $\phi = 0.03$, $Z$ is skeptical, $N = S + E + I + Z$.

**Model Validation and Evaluation** Parameter estimation employed nonlinear least-squares regression via MATLAB's `lsqnonlin` with trust-region-reflective algorithm. 5-fold cross-validation (k=5) reduced overfitting across heterogeneous network topologies. Performance assessed using Macro Error: $\text{MacroError} = \frac{1}{T} \sum_{t=1}^{T} |\text{Predicted}(t) - \text{Observed}(t)|$ where $T$ is temporal observation points. The `ode15s` solver was employed for numerical integration, demonstrating superior stability for epidemic dynamics [19]. Monte Carlo validation achieved convergence within 1,500 iterations with relative change below 0.1%. Final infection estimates maintained ±0.5% margin of error at 95% confidence level.

## 4 Results and Findings

**WFSA Performance Evaluation** Addressing **RQ1**, WFSA was applied to six networks: three weighted Barabási–Albert (BA) networks (500, 750, 1000 nodes) and three real-world networks (Telegram, Twitter, Reddit). FTSs extracted from original networks were embedded into corresponding Erdős–Rényi (ER) networks using established methods [15], then WFSA was reapplied to evaluate re-identification under altered topologies using standard Information Retrieval metrics [14]:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Re-extraction Rate} = \frac{\#\text{FTS}_{\text{Post}}}{\#\text{FTS}_{\text{Pre}}} \times 100$$

$$H = \frac{2 \times (\text{F1} \times \text{Re-extraction})}{\text{F1} + \text{Re-extraction}}$$

WFSA performance correlates strongly with clustering coefficients, highlighting network cohesion sensitivity [16]. Telegram-based FTSs achieved optimal results (clustering: 0.007→0.855) with highest F1-score (0.92), Re-extraction

Rate (84%), and Quality Metric $H = 0.880$. BA networks showed clustering-dependent performance: BA-1000 (clustering = 0.243, F1 = 0.76) significantly outperformed BA-500 (clustering = 0.123, F1 = 0.43). High NDCG scores (0.902 Re-extraction, 0.926 F1, 0.914 $H$) confirm ranking reliability. Reddit showed lower performance (clustering = 0.197, F1 = 0.56, Re-extraction = 44%) due to reduced structural cohesion.

**Toxicity Propagation: FTS vs ITI** Addressing **RQ2**, Monte Carlo epidemic simulations (1,500 iterations) compared FTSs (WFSA-identified) and ITIs (Weighted PageRank [17]) across weighted BA networks (2,250 nodes, 3,991 edges) and real-world networks. FTSs consistently outperformed ITIs across all networks and epidemic models (SIR, SEIR, SEIZ) with statistical significance ($p < 0.001$, Cohen's $d > 4.0$). Propagation advantages were substantial: Twitter (96.9% more effective), Telegram (63.7%), BA networks (43.6%), Reddit (36.3%). Mean infection rates demonstrated consistent FTS superiority: BA networks (0.661 vs. 0.460), Telegram (0.706 vs. 0.431), Twitter (0.702 vs. 0.357), Reddit (0.805 vs. 0.591). SEIZ models showed reduced propagation due to skepticism effects, but the FTS advantage persisted across all variants, confirming structural toxicity's superior diffusion potential. These findings are visually summarized in Fig. 2.

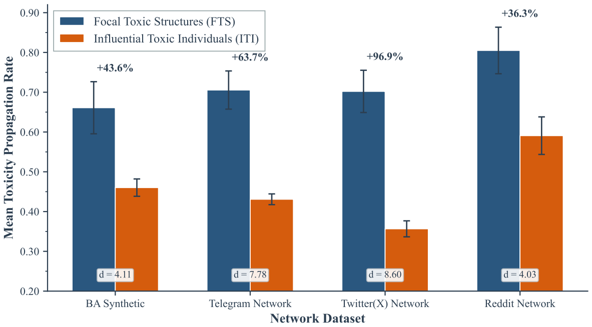

**Fig. 2.** FTS vs ITI toxicity propagation across networks. Bars show mean infection rates ±1 SD. Percentages indicate FTS relative increase. Inset shows Cohen's $d$ effect sizes.

**Model Performance Evaluation** Addressing **RQ3**, predictive accuracy evaluation across all datasets revealed SEIZ-MC's superior performance with lowest mean error (5.7%), significantly outperforming SEIR-MC (28%) and SIR-MC (33%) across BA synthetic, Telegram, Reddit, and Twitter datasets. This confirms SEIZ-MC's reliability in toxicity diffusion modeling [6], particularly its skepticism factors that enhance realistic propagation predictions compared to traditional epidemic models.

## 5 Conclusion

This study shows that Focal Toxic Structures (FTSs) consistently outperform Influential Toxic Individuals (ITIs) in toxicity propagation across platforms, with FTSs driving significantly higher diffusion (e.g., 71% vs. 43% on Telegram). The proposed WFSA algorithm accurately detects FTSs by combining toxicity severity with structural cohesion. SEIZ-MC yielded the lowest average error (5.7%), highlighting its utility in modeling real-world dynamics through skeptical user states. These findings advocate structure-based moderation over user-centric approaches and open pathways for integrating WFSA with multimodal models for early detection in security, discourse analysis, and digital health.

**Acknowledgements** This research is funded in part by the U.S. NSF (OIA-1946391, OIA-1920920), OUSD/AFOSR (FA9550-22-1-0332), ARO (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), ONR (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), AFRL, DARPA, the Australian Department of Defense Strategic Policy Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the UA–Little Rock. The researchers gratefully acknowledge the support.

## References

1. Agarwal, N., et al.: Modeling blogger influence in communities. Soc. Netw. Anal. Mining 2, 139–162 (2012)
2. Alassad, M., Agarwal, N.: Contextualizing toxic structure analysis. Soc. Netw. Anal. Mining 12, 103 (2022). https://doi.org/10.1007/s13278-022-00938-0
3. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516 (2019)
4. Falade, T.C., Agarwal, N. : Toxicity prediction in Reddit. In: Proc. AMCIS (2024)
5. Ghoshal, G., Barabási, A.-L.: Ranking stability in networks. Nat. Commun. 2(1), 394 (2011)
6. Jin, F., et al.: Epidemiological modeling on Twitter. In: Workshop Soc. Netw. Mining Anal., pp. 1–9 (2013)
7. Kiddle, R., et al.: Network toxicity analysis. J. Comput. Soc. Sci. 7, 305–330 (2024)
8. Lexyr: Reddit Climate Change Dataset. Kaggle (2020)
9. Maleki, M., Agarwal, N.: Comparative Analysis of SIR vs. SEIZ Models for COVID-19 Information Diffusion on Social Media. Social Network Analysis and Mining 12(3), 45–62 (2025)
10. Newman, M.E.: Structure of complex networks. SIAM Rev. 45(2), 167–256 (2003)
11. Obadimu, A., et al.: Toxic features on YouTube. In: Int. Conf. Soc. Media Tech. (2019)
12. Ohara, K., et al.: Network performance via centrality. In: IEEE DSAA, pp. 561–570 (2017)
13. Qayyum, H., et al.: Toxic 1% of Twitter. arXiv:2202.07853 (2022)
14. Sanderson, M.: IR system evaluation. Found. Trends Inf. Retr. 4(4), 247–375 (2010)
15. Şen, F., et al.: Focal structures analysis. Soc. Netw. Anal. Mining 6, 1–22 (2016)
16. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
17. Xing, W., Ghorbani, A.: Weighted PageRank algorithm. In: Proc. CNSR 2004, pp. 305–314 (2004). https://doi.org/10.1109/DNSR.2004.1344743
18. Yang, H., et al.: Review of COVID-19 models. Contemp. Math., pp. 75–98 (2023)
19. Törnberg, P., et al.: Affective polarization in social media society. PLoS One 16(10), e0258259 (2021). https://doi.org/10.1371/journal.pone.0258259
20. Akinnubi, A., Agarwal, N.: Identifying contextualized focal structures in multi-source social networks by leveraging knowledge graphs. In: Proc. Int. Conf. Complex Networks and Their Applications, pp. 15–27. Springer (2023)