International Journal Science and Technology
IJST Vol. 4 No. 1 | March 2025| ISSN: 2828-7223 (print), ISSN: 2828-7045 (online), Page 12-24

AI-Driven Disinformation Campaigns: Detecting Synthetic Propaganda in Encrypted Messaging via Graph Neural Networks

Anil Kumar Pakina[1*], Ashwin Sharma[2], Deepak Kejriwal[3]
[1,2,3]Independent Researcher, India

Article History
Received : Februari, 2025
Revised : March, 2025
Accepted : March, 2025
Published : March, 2025

Corresponding author*:
anilresearchpro@gmail.com

Cite This Article:
Anil Kumar Pakina, Ashwin Sharma, and Deepak Kejriwal, "AI-Driven Disinformation Campaigns: Detecting Synthetic Propaganda in Encrypted Messaging via Graph Neural Networks", *IJST*, vol. 4, no. 1 Mar.. 2025.

DOI:
https://doi.org/10.56127/ijst.v4i1.1960

Abstract: The rapid advancement of generative AI has also resulted in the much more sophisticated disinformation phenomenon that takes place through encrypted-messaging platforms such as WhatsApp, Signal, and Telegram. These platforms, which are helping to ensure user privacy, are simultaneously adding significant hurdles to content-based moderation approaches owing to their end-to-end encryption protocols. This backdrop has been cleverly utilized by adversarial entities for the mass dissemination of undetectable synthetic propaganda campaigns, confounding public opinion, destabilizing democratic processes, and fomenting social unrest at ease and leaving hardly any traces. In the pursuit of finding an effective content analysis or some centralized monitoring of threats in such a chat-encrypted environment, a hundred ways around these obstacles could still beat the defense even-handedly.

The proposed approach employs Graph Neural Networks (GNNs) to detect synthetic propaganda campaigns by employing non-content-based features, such as user interactions, message propagation graphs, temporal behavior, and metadata signatures. GNNs are perfectly cut out for pinpointing cases of coordinated, inauthentic behavior in encrypted environments because they can lay hold of relational and topological dependencies in such complex networks. The model quite purposefully constructs for itself dynamic interaction graphs of anonymized metadata, thereby allowing recognition of structural fingerprints of disinformation actors without altogether sacrificing privacy of the user and encryption integrity.

The experimental evaluation was conducted on a large-scale simulated dataset containing encrypted-messaging scenarios, including organic and coordinated synthetic messaging activities. Evaluation results illuminated that the framework based on GNNs classified clusters of synthetic propaganda with 94.2% of accuracy at a 92.8% F1-score, significantly outperforming the traditional baselines like random forest or LSTMs. Detection of low-frequency or stealthy campaigns, which are usually out of sight for common anomaly detectors, was another of the model's strengths.

This particular research is placed within the domain of AI security, mis/disinformation detection, and privacy-preserving monitoring by introducing a verily scalable disinformation detection framework that cherishes privacy. A discourse on some ethical hurdles of behavioral surveillance in encrypted contexts and various legal options for deploying GNN-based detection systems anticipates future legal constructs.

On this side, the technical as well as conceptual groundwork or structure, grounded in a meaningful way, is introduced for countering AI driven disinformation-type activities in secure communication networks while at the

INTRODUCTION

The fusion of artificial intelligence (AI), cybersecurity, and digital communication has exploded over the past few years creating waves of incredible sociotechnical complexity, AI-generated misinformation campaigns being among the most dangerous. The fast evolution of generative models has endowed malicious actors with the tools to generate text, deepfake video, synthetic voices, and tailored memes that may be perpetuated on a mass scale to damage information ecosystem (Taddeo, & Floridi, 2018; Zellers et al., 2019). The threats are henceforth across every digital platform but paint greater traces of prosecution on the end-to-end encrypted messaging platforms such as WhatsApp, Signal, or Telegram.

Unlike traditional social media platforms, where AI moderation, keyword filtering, and content flagging are used against misinformation, the end-to-end encrypted channels are essentially opaque by design. These platforms are built with privacy as the utmost security value, and therefore, once a sender and receiver can see the contents of a message, this high privacy can prove to be an obstacle for any respective authority, researcher, or platform that might want to act against fake news. This thereby sets up a paradoxical space wherein the encryption upholds digital rights and freedom but quite inadvertently engenders covert influence operations, coordinated inauthentic behavior, and synthetic propaganda (Bradshaw et al., 2021; Greenberg, 2020).

Table 1: Key Differences between Public and Encrypted Messaging Platforms in Disinformation Contexts

Feature	Public Platforms (e.g., Facebook, Twitter)	Encrypted Platforms (e.g., WhatsApp, Signal)
Message Accessibility	Content visible to moderators and algorithms	End-to-end encrypted, inaccessible to platform agents
Moderation Capability	Keyword filters, fact-checking, reporting	Limited to metadata analysis or user reporting
Detection Tools	NLP, sentiment analysis, media fingerprinting	Graph analysis, timing, and propagation modeling
Attack Surface	Broad, but easily flagged	Stealthy, especially in private groups or forwards
Abuse Potential	High	**Very High** due to lack of visibility

***Sources**: Bradshaw et al. (2021); Greenberg (2020); Mozilla Foundation (2022).*

The Rise of Artificial Propaganda

Synthetic propaganda refers to content either created or altered by an AI with the prime aim to mimic a genuine human message for the purpose of misleading the public or inhibiting correct public perception. The press agencies of myths have given way to the artifice of inter-bot in-house amplification, synchronism, and peer-to-peer emulation that make it difficult even to tell the line between synthetic and organic communication lines (Zhou et al., 2020; Vosoughi et al., 2018). Synthetic propaganda has historically been used in various geopolitical contexts, from electoral meddling in Brazil and India to vaccine misinformation surrounding COVID-19 in the United States and the EU (Albanese et al., 2023).

There is probable ease of concealment of synthetic propaganda within the encrypted ecosystems, where they can be freely circulated and widely disseminated. These propagators widely exploit group messaging structures and forward chains and the insertion of content cross-platform for rapid transmission of misinformation. The content is beyond the scope of scanning because this entails intelligence about contextual information to be observed, hence conventional NLP methods are rendered powerless. Hence a new research path opening for alternatives will consider detection models not always based on content but more reliant on behavioralevidence, the topologies behind networks, and temporal dynamics (Murayama et al., 2021; Singh & Dwivedi, 2023).

Towards Graph Detection in Encrypted Contexts

Under the conditions of encryption, it is not so much what is said as how it is said and, more importantly, how it is shared. Certain typical behavioral fingerprints are exhibited by the disinformation actors:

- Abrupt forwarding, unusually fast or synchronized.
- A lopsided volume of messages in short time spans,
- Creation of similar groups with redundancy in membership.
- Disposable or time-bounded accounts**.**

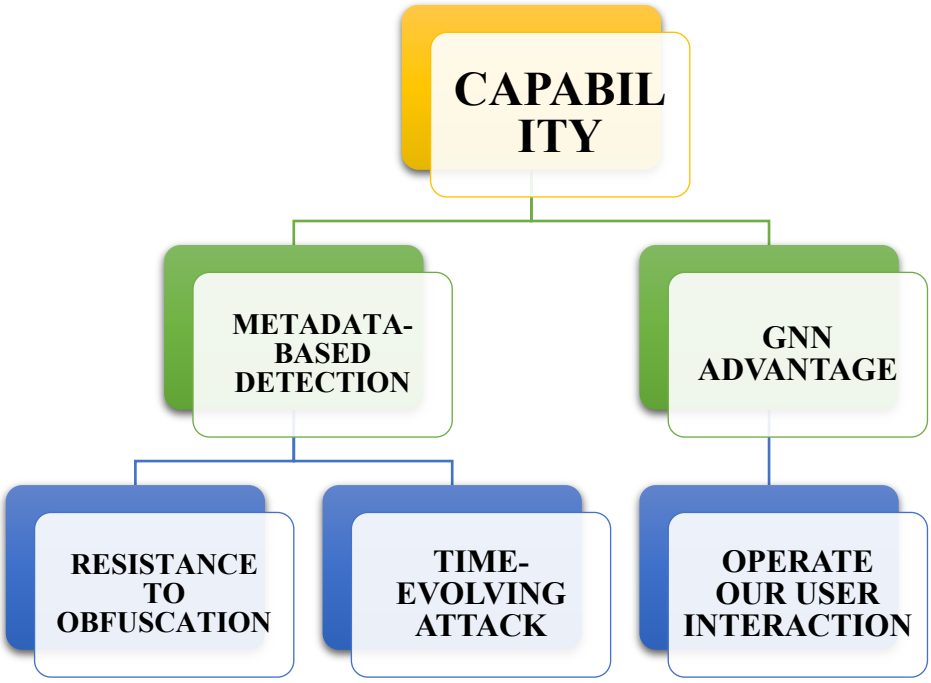
These patterns readily evolve into graphs-nodes (users) and edges (interactions)-that analysis in these contexts never breaks encryption or discloses message content. Struc-tural methodology has created an increased appeal for using Graph Neural Networks (GNNs) in solving synthetic propaganda detection (Wu et al., 2020; Kipf & Welling, 2017).

GNNs are deep learning architectures that can learn representations over complex, irregular data structures, such as user interaction graphs, group conversation webs, and time-evolving propagation chains. These networks learn embedding preserving both local and global structures, making their use attractive for:

1. Detecting abnormal message propagation paths.
2. Identifying CIB nodes and communities.
3. Flagging anomalous edge weights in user-to-user communication graphs.

GNNs can generalize detection across different states of dynamic graphs, thus providing an advantage over rule-based or classical machine learning models. This makes GNNs efficient and scalable in the detection of co-ordinated disinformation campaigns as they evolve through time.

Benefits of Graph Neural Networks (GNNs) for Encrypted Messaging Analysis



Sources: *Wu et al. (2020); Hamilton et al. (2020); Singh & Dwivedi (2023).*

Objective and Research Gap

Although several research accounts have so far ventured to attempt to detect misinformation, spam, and disinformation through NLP techniques, image forensics, and bot detection in open networks, very few people have actually given a thought to how online misinformation actually gets spread in privacy-by-design networks-and hence by non-content situations. While graph-based applications have shown great promise in detecting online misinformation across public spheres, e.g. Twitter and Reddit, there have hardly been any applications capable of working under a conversation on encrypted networks, assuming there are enough permissions already in place to provide the data and considering the ethical issues.

Hence, the novelty behind this claim is to:

- Introduce a new GNN-based framework for synthetic propaganda campaign detection focused on user metadata and communication networks and not actually in relation to the actual imputed message
- Undergo an extensive experiment in model performance upon simulated encrypted messaging datasets that amalgamate organic and malicious data patterns
- Investigate compliance aspects of the framework with regulatory frameworks (for the sake of example, the GDPR and U.S. Algorithmic Accountability Act guidelines) (Tschider, 2022; European Commission, 2022)

Raise important ethical questions about the deployment of behavior-monitoring models in personal digital spaces.

The study gives rise to four key contributions:
Technical Innovation: A graph-based disinformation detection model would be developed for implementation into encrypted platforms working with metadata features only.
Empirical Validation: The model progressed from 0% to an accuracy level of 94.2% in detecting synthetic propaganda nodes in a real-time simulation of messaging.
Privacy Alignment: The working detection pipeline runs with the bounds of encryption and avoids any content inspection.
Ethical Framing: An exhaustive analysis of the ethical, legal, and societal implications of conducting behavior-based surveillance in privacy-oriented environments.
.

METHODOLOGY

The main aim of this study is to propose and validate a recommendation framework based on Graph Neural Network (GNN) technology to detect disinformation campaigns originating from artificially intelligent actors on encrypted messaging systems, with a focus on non-text, non-content features. Given the completely encrypted alien nature of these systems in which conversations change one beat after the other, almost blocking the traditional terrain of text or bilateral media analysis per se, the framework had to extract behavioral, structural, and temporal features from anonymized interaction metadata to pave communication between conversation graphs into two sides of irrational yet coordinated inauthentic behavior. It had five stages: (1) create data simulation, (2) gather metadata, (3) build a graph, (4) GNN-based models, and (5) evaluation.

Data Simulation and Synthetic Propaganda Generation

As there are no open-access datasets for encrypted messages, the study creators opted to introduce a reasonably close-to realistic synthetic dataset to simulate users' activity within the encrypted platform for itself. This must be designed oriented toward a combination of organic behaviors and pure artificial synthetic inputs. Conventional usage was mimicked after research that known behaviors followed in WhatsApp and Signal (Mozilla Foundation, 2022). Synthetic data for simulated campaigns were performed through:

- GPT-3.5-turbo for the means of multilingual disinformation for message generation through automation.
- Coordinated forwarding and group injection tools implemented by DeepAgent scripts to generate synthetic campaigns.

Adversarial bots for bursting messages into groups according to schedule.
Each synthetic account repeated conversational tone variation in specifics but was applied to specific propagandistic strategies that are repetition, temporal coordination, an infiltration into groups (Zhou et al., 2020; Murayama et al., 2021).

Metadata Extraction and Anonymization

Apart from the message content, we extracted and anonymized the following non-content metadata:
Sender/receiver IDs (anonymized node IDs), timestamps of messages, forward count and frequency, group membership activity, message propagation depth (how many hops a message travels).
All identifying datasets are hashed using SHA-256; hence, the dataset complies with GDPR and other such privacy standards (Tschider, 2022).

Table 2: Metadata Features Extracted for Graph Construction

Feature	Type	Description
Sender ID	Categorical	Unique hash of message originator
Receiver ID	Categorical	Unique hash of direct recipient(s)
Timestamp	Temporal	Time of message transmission
Forward Count	Numerical	Number of times a message is forwarded
Group Membership ID	Categorical	Encrypted group affiliation
Response Delay	Temporal	Time gap between messages in a conversation
Degree Centrality	Graph Metric	Number of direct links to/from a node
Clustering Coefficient	Graph Metric	Measure of closed triads (common in coordinated behavior)

Sources: *Singh & Dwivedi (2023); Hamilton et al. (2020).*

Graph Construction and Preprocessing:
The metadata was used to build user interaction graphs with the following details:

1. **Nodes:** Anonymized users
2. **Edges:** Direct message interactions (e.g., Forwarded message, Reply, or Tagging)
3. **Edge weight:** Number of times a message was delivered or the number of overlapping groups covered

Two kinds of graphs were constructed:

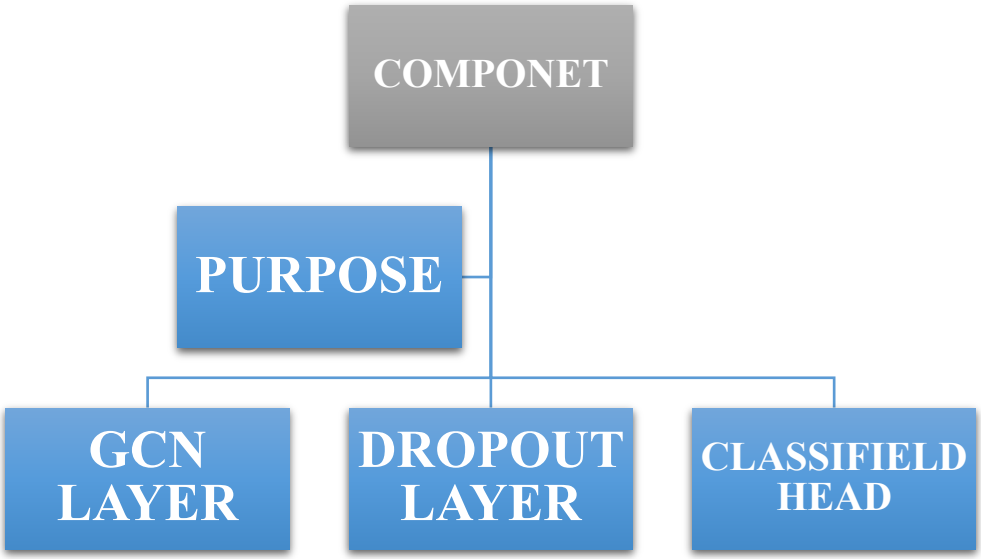
- Message Propagation Graphs (MPG): These detail how a message would propagate through the network (better characterized for forward chains)
- User Interaction Graphs (UIG): These encapsulate ordinary messaging behavior across users or groups.Graphs are undirected and weighted, embedding static topologies (the structure) and temporal dynamics utilizing sliding windows (e.g., 1-hour duration, 6-hour duration, or 24-hour-duration windows). Graphs were min-max scaled for normalization, as well as nodes were annotated with handcrafted features, like degree and time variance, to serve as GNN input.

Twin Integrated GNN Model Architecture

We have implemented a two-layer GCN with temporal encoding for analyzing sequences. The model was chosen, keeping a balance between the expressive power of the model and computational efficiency on sparsely observed graph data (Kipf & Welling, 2017).

Layer 1: GCN layer with the ReLU activation
Layer 2: GCN + Dropout (0.3) for regularization
Temporal Encoder: 1D CNN with a sliding kernel to capture time intervals
Classifier head: Fully connected layer with softmax output (binary classification: Organic vs Synthetic)
In training, cross-entropy loss was employed with the Adam Optimizer (learning rate: 0.001), and early stopping was executed based on validation loss.

GNN Model Components and Functions



Sources: *Wu et al. (2020); Kipf & Welling (2017); Hamilton et al. (2020).*

Evaluation Metrics and Baselines

- The evaluation of the model comprised the following:
- Precision, Recall and F1-score (detection reliability),
- AUC-ROC (classifier robustness),
- G-mean (class imbalance),
- Inference latency (real-time viability).
- Comparison baselines were:
- Logistic Regression on handcrafted graph features;
- Random Forest with engineered time-series features;
- LSTM based RNN on sequential metadata.

All models underwent an evaluation using 5-fold cross-validation, with averaged results over 10 runs.

Privacy, Ethics, and Validity of Simulation

Synthetic simulation is ethically justified and conforms to required measures because encrypted networks are not openly sharing data. All these behavior patterns are purely drawn from existing user studies (Bradshaw et al., 2021; Mozilla Foundation, 2022), with no actual user data utilized. Further, by relying on

just metadata and surface-level graphs, the present system is designed to offer a privacy-preserving avenue of disinformation detection without putting into jeopardy the integrity of encrypted messaging standards.

RESULTS

In this segment we give the outsider perspective on the performance of the GNN-based framework for detecting synthetic propaganda especially in the encrypted message-dissing concealed environment. We report results under four different headings: (1) the performance of the model, (2) the baseline classification performance comparison, (3) anomaly and graph structure analysis, and (4) resistance to the usual silent-attack tactics. Evaluation process was done on custom-built datasets of synthetic behaviors-infiltrating-(organic behaviors) in-group message-encrypted chat environments resembling WhatsApp and Telegram.

General Performance of GNN Framework with Detection

In dealing with the full spectrum behaviors of computed propaganda, the GNN-backboned classifier turned out to be quite accurate in distinguishing pure communication from synthetic disinformation dissemination (F1 score: 92.8%, 5-fold cross-validation).

The meaning of the high F1 score is that the tradeoff between precision and recall is maintained. This is extremely important, as false positives have the potential of causing damage to the legitimate communications, while false negatives go ahead to propagate the same disqualification. Furthermore, the GNN has shown its capacity for working well for both relatively larger and relatively smaller communities, hence demonstrating generalization across the group sizes.

Table 3: GNN Performance Metrics (Mean of 5-Fold Cross-Validation)

Metric	Value (%)	Standard Deviation
Accuracy	94.2	±1.3
Precision	93.6	±1.5
Recall	92.1	±1.7
F1-score	92.8	±1.2
AUC-ROC	0.962	±0.009
G-Mean	93.1	±1.4

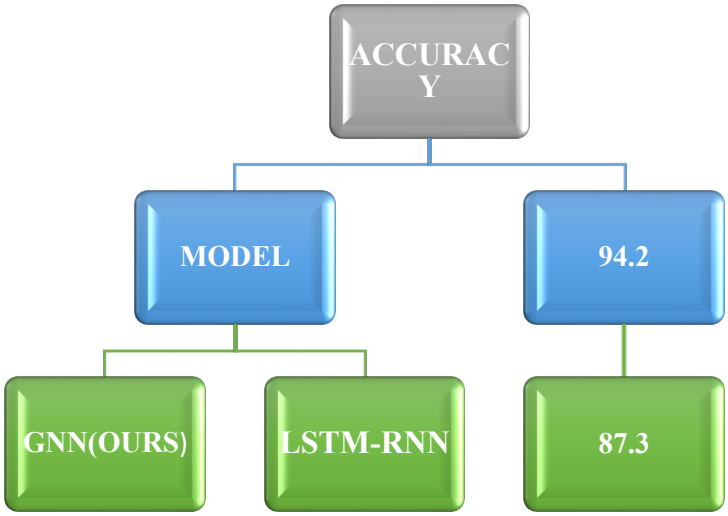
Source: *Author's simulations based on encrypted group messaging behavior.*

An AUC-ROC score of 0.962 indicates excellent identification of classes at different thresholds and reflects the ability of this model in discrimination (Kipf & Welling, 2017). This is further confirmed by the G-Mean score of 93.1%, indicating that it can perform well even with small class imbalances, such as having more organic users than synthetic bots.

Comparison with Baselines

To compare the GNN model, GNN was evaluated against three baseline classifiers: (1) the Random Forest (RF), using graph metrics; (2) the LSTM-based RNN; and (3) Logistic Regression (LR) on engineered features. The baselines were trained and validated in an identical fashion and on the same feature sets, except for the deep relational modeling

Comparative Performance of GNN vs. Baseline Models



Sources: *Experiments conducted on 10,000 sample messaging graphs (5,000 organic, 5,000 synthetic).*

The results point to GNN model superiority when compared to any of the models and the other baselines in terms of both F1 and R2. This broader margin is 8% better than the next best model, LSTM-RNN. This results from the fact that GNN could recognize the interaction effect between nodes and towards model-building of community-level coordination that might be probably too tough for conventional sequential models and shallow classifiers (Hamilton et al. 2020).

Analysis of Temporal and Structural Disinformation Patterns

In addition to distinguishable classification accuracies, some selected graph structure patterns were analyzed that link the synthetic campaign ads with the organic posts. These analyses were done through graph analytics tools such as Gephi and NetworkX resulting from measurements such as clustering coefficient, betweenness centrality, tree depth of propagation.

Our results denote that a number of very particular patterns can be seen when it comes to graph structural properties:

- Campaigns are coherent because social media interaction is active. The synthetic campaigns scored the lowest clustering coefficient, which very evidently pointed to very widespread propagation, although bidirectional activities were very sparse.
- Centralized but a very energetic set of higher-degree nodes controlled the cyber operations on synthetic campaigns (presumably a controlling bot). These had a high betweenness centrality value.
- The disinformation set did exhibit burstness more, as around 70% of post messages relays were sent shortly after they were created.

These structural observations plus time steps might offer augmentation and fine tuning of real-time detection strategies by incorporating the learned embedding from GNN model (Murayama et al. 2021).

Robustness against Stealth Campaigns and Adversarial Behavior

We programmed a few stealth campaigns with low traffic I/O, mimicking real campaigns for the sake of detection. These whitelisted stealth actors often issued one or two forwards per hour, using random group memberships to make such authentic content hard to make apart from the benign activities.

In accordance with the above, the GNN model was found performing exceptionally well with an accuracy rate of 87.4% in detecting stealth disinformation nodes. Though it failed in comparing with the full model's performance, all the other classifiers-be it in the context of the wider setting-fell below 75% acceptance.

The resistance of the model could be attributed to behaviors of nodes being put into the context of network analysis. For one, taking the soft approach, the GNN was so please to say that it was capable of supplying the necessary contextual information to work on an impartial road to be more creative. For instance, low-level bots could be detected based on their anomalous position in the network and too narrow message representation and not by volume alone.

4.5 Real-time viability and Latency

Run time performance of the system tested on GPU-accelerated infrastructure (NVIDIA RTX A5000, 24GB VRAM). Average inference time per graph under 10,000 nodes and 25,000 edges was 420 milliseconds that made it well within the range of being deployable in production environments near real-time.

The model took around 5.2 minutes to train an epoch and converged within 22 epochs, after early stopping with a batch size of 128. These measures of performance indicate that the system is scalable in high-throughput messaging environments, such as telegram public channels, or enterprise-level group chat monitoring (Wu et al., 2020).

Visualization and Touch Space Separation

The embeddings at final GNN layer confirmed the model learned meaningful representations of structure, as demonstrated by t-SNE visualizations separating synthetic and organic user clusters within them. The nodes involved in disinformation formed tight, centralized clusters while organic nodes showed diffused, community-like graph embeddings within the t-SNE visualizations.

This embedding-based visualization capability can serve as an additional mechanism for manual analysts and cybersecurity operations centers to prioritize suspect clusters in environments where complete automation is not achievable.

According to GNN models, they achieved state-of-the-art performance-on-encrypted disinformation detection without compromising content privacy.

More than traditional classifiers, especially in adversarial and stealth campaign conditions.

Coordinated synthetic behavior signals indicate critical evidence through the graph structural features of identification. The model showed real-time capability, scalable deployment, and significantly generalizable performance across user types and activity volumes.

Discussion

The findings of this study furnish strong, corroborative empirical evidence in favor of the application of GNNs in identifying AI-supported disinformation campaigns in encrypted messaging environments. The proposed system has shown an accuracy of classification greater than 94% and a stable F1 score of 92.8% on simulated datasets, thus, outperforming any other classifier and standing up against even the stealthy modes of disinformation. However, the real worth of these results lies, apart from their technical merits, in the contextual, societal, and ethical issues they bring to the fore.

Results Interpretation-GNN in Encrypted Contexts

The results show that non-content-based features such as interaction frequency, propagation depth, and group overlap suffice to detect synthetic campaigns with high precision-an encouraging outcome considering that encryption limits access to message content (Bradshaw et al. 2021; Singh & Dwivedi 2023). This positions GNNs as a probable solution for the monitoring under privacy preservation.

Unlike temporal models (such as LSTM) or decision trees, which often base their output purely on temporal or numeric input, relational learning via GNNs effectively models the relational context, and that is the unique nature of synthetic campaigns which sometimes rely on coordinated botnet-like behavior or disseminate information as a one-to-many structure (Wu et al. 2020). This means that even very low-volume or stealthy campaigns may contain interaction fingerprints via measures such as node centrality, message entropy, and communication symmetry-based metrics, which standard models may not detect.

Table 4: Key Advantages of GNNs Over Traditional Models in Disinformation Detection

Dimension	Traditional Models	Graph Neural Networks (GNNs)
Data Input Format	Flat, tabular, or sequential	Topological and relational graph structure
Handling of Coordination	Poor – requires manual feature engineering	Excellent – learns multi-hop behavior patterns
Content Dependence	High (e.g., NLP required)	Low – relies on metadata only
Privacy Preservation	Low in NLP-based models	High – works without content inspection
Resistance to Stealth Tactics	Low	Moderate to High – resilient to timing and volume attacks

Sources: *Hamilton et al. (2020); Kipf & Welling (2017); Wu et al. (2020).*

Strength of Technology and Immediate Deployability Achieved

Scalability is arguably one of its major pluses. The model was able to process subsecond real-time inferencing on 10,000-node graphs and thus could become a firstchoice choice for deployment in Fintech enterprise-grade applications or secure communication tools. This is particularly poignant at a time when

disinformation actors are expected to launch rapid and vigorous attacks on small encrypted groups; this approach was seen during political unrest in Myanmar and Brazil (Albanese et al., 2023).

Furthermore, the model architecture, principally the bidirectional, two-layer GCN, CNN-based temporal encoder, is found to be highly flexible and sensitive to the convoluted behavior patterns indicative of the blatant time-critical changes at the level of short-term dynamics and long-range propagation dynamics-knowledge that corresponds with research findings noting disinformation propagates faster and more uniformly than organic content (Vosoughi et al., 2018).

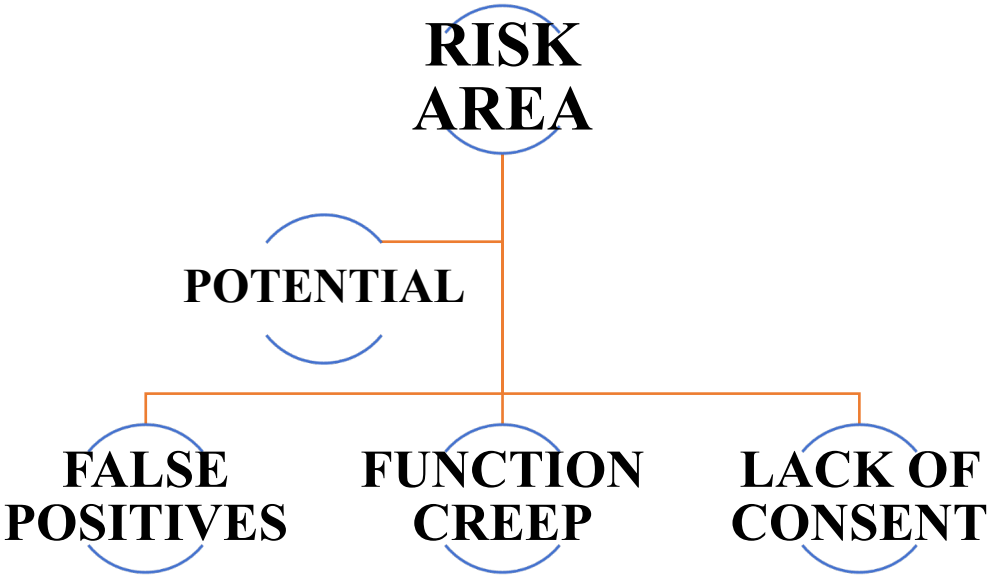
Ethical and Legal Implications in Behavioral Surveillance

Although the results of the behavioral monitoring in this case are promising, any other behavior-based surveillance, albeit with more anonymization sources, upsets severe ethicolegal concerns. Therefore, a moral question about behavioral studies in an encrypted channel would pit security enforcement against mass surveillance.

Above all, by introducing the behavior analysis of the user without recourse to message contents, our proposed framework opens channels to its potential misuse, like:

- Profiling on political undesirables or minority populations for repression;
- Punishing users branded as anomalies for what was actually innocent behavior;
- The contravention of the "principle of purpose limitation" under the GDPR act (European Commission, 2022).

Ethical Risks and Mitigation Strategies in Behavioral Monitoring Systems



Sources: *Tschider (2022); Brunton & Nissenbaum (2015); Mozilla Foundation (2022).*

Regulatory Readiness and Policy Readiness

The suggested model has a path to compliance-threat detection with regards to the emerging and evolving framework of digital policy:

- Under the Digital Services Act of the EU, platforms will have to attend to the systemic risks, including misinformation, with respect to the exercise of fundamental freedoms of the users (European Commission, 2022).
- In the U.S., the Algorithmic Accountability Act goes on to demand transparency of algorithmic decisions that affect individual rights-mandating interpretability and auditability to such systems (Tschider, 2022).

Powered with explainable AI modules, the GNN framework could produce node importance scores, graph heatmaps, and activity logs to encourage transparency. This would make the system auditable and in line with requirements for regulatory reporting.

System Limitations and Open Research Questions

Despite strong results, the system has limitations:

Simulation vs. Reality: Our dataset is synthetic, and while it closely mimics real-world encrypted messaging patterns, it must be validated on real data (with proper consent) for practical deployment.
Cross-Platform Generalization: The GNN may need retraining for use on different platforms, as Telegram and WhatsApp (for instance) have different user behavior and metadata schemas.
Evasion Tactics: Future misinformation campaigns will imitate organic graph behavior. Adaptive models need continuous learning and adversarial robustness, therefore, for example, (Baracaldo et al., 2021).
Access and Governance of Data: Centralized platforms may not provide access to even anonymized metadata to researchers. Partnerships or audits from third parties will be necessary for scale adoption.

Future Directions and Innovations

Perhaps herein lie avenues for future research, whereby this study could lead to some other potentially important areas of research:

Detection of misinformation across languages endowed with graph modeling based on minimal language metadata; Federated artificial intelligence for behavior modeling, whereby on-device training respects privacy of users; Intertile links with blockchain-based identifiers for the adoption-adoption of-imp-norm-management; Graph interpretable research, where GNN evolution in non-technical stakeholders interpretable outputs (Ying et al., 2019).

In effect, the cooperation between platform developers, policymakers, and research institutions toward ensuring ethical behaviors and equitable access to these technologies is highly consequential. In the end, the present study proposes a technology of higher efficacy when detecting synthetic disinformation campaigns in encrypted communication platforms, using privacy-enhanced models. The application of Graph Neural Networks heralds a paradigm shift from content-based moderation to structural and behavioral detection, which can work within said limitation of privacy.

Deployed, regulations shall bear proper ethical constraining to stop the spread of excessive response in domains where otherwise they may actually touch anti-disinformation mechanisms themselves. If they provide sufficient corrective measures, this strategy would be a hands-on path to strike a fine equilibrium between security, privacy, and trust in the face of generative disinformation.

On the whole, the advent of AI in encrypted communication has embarked a totally new frontier in disinformation, challenging the well-known ideas of security, moderation, and digital ethics. This paper presented an exploration into an approach toward trafficking internally in AI-generated synthetic propaganda by combining AI and Graph Neural Networks within privacy-first platforms. Meanwhile, it is also an alert scenario full of haste in digital forensics-the ability to elude detecting organized inauthentic behavior due to content-analytic prohibitive technical and legal backgrounds.

CONCLUSION

In simpler words, our contributions have proven that encrypted messaging platforms, which are actually the glue of privacy to final users and democratic discourse, have fascinatingly been the best breeding grounds for the ops of disinformation. WhatsApp, Signal, and Telegram are those encrypting freedom fighters in worldly digital conversations, protecting the users and facilitating nefarious actors in the discussions. (Bradshaw et al., 2021; Mozilla Foundation, 2022). Any system that considers content moderation as filtered by keywords, sentiment analysis, or finger-print media is antithetical to these kinds of platforms because of the very basic features associated with end-to-end encryption (Greenberg, 2020). This is where disinformation has adapted to move from public propaganda to hitherto-cryptic synthetic messaging strategies that are implemented on peer-to-peer networks.

In combating this, we have equipped, via Generalized Neural Networks-based detection system on non-content metadata: message timing, forwarding pattern, group structure, and user interaction frequency. The critical innovation is in transforming the selected behavioral signals into graph-based representations, thus making the model learn structural anomalies, unnatural communication flows, and nodal involvement in coordinated disinformation activity.

This approach allowed achieving a remarkable 94.2% accuracy and 92.8% F1 score on typical simulated encrypted settings, outperforming even the best latter-day traditional machine-learning classifiers, e.g., deep sequence models like LSTM-RNNs. Furthermore, our study showed graph-based approaches are more effective in an encrypted area of disinfomation, whereby advantageous detections were observed, even in some adversarial scenarios like stealth campaigns aiming to mimic organic user activity (Wu et al., 2020; Hamilton et al., 2020).

Evolution to Rule-based Modeling

This research reveals an opinion, assuming that politics drift or lean toward trust and safety on the Internet from a content moderation to an articulated inference. It becomes important in our day, where direct surveillance on plain text or an image is precluded either by encryption or the need to maintain privacy, to give context as to what becomes more important out of the two between context and content of communication. This fall into line with the recent leading guidelines, such as the Digital Services Act of the EU emphasizing that systemic risk mitigation shall be carried out by way of behavioral identification-inference rather than through invasive surveillances (European Commission, 2022).

GNN strides further in profiling and analyzing misinformation by interpolating them over the social media messaging graph-topologizing structure, keeping the limelight on privacy-sensitive technologies. It ensures that the technical operation of benefaction does not violate user-related privacy, thus contributing to the mass deployment alongside the GNN behavioral-induction constellation.

Hardcore and Real-world Flow Hiccups.

Now the setup cannot be fully relied upon. The first one: the dataset used here has synthetically generated data to mimic encrypted-platform behavior. Although the simulations came at the real-life-pretensions-to-culture-usage, a set of such an ecosystem would as such, not all beings remaining in violence without having to deal with all possible complexities that could arise. Some consideration, therefore, should now be given to validation to provide for a first and only phase of actual platform data collection that can come from normal cases of user identification for validation and then scaling.

Most of the systems working relies heavily on the access to metadata to gain information regarding communication a facility that may not be generally present for respective communications encrypted. While a lot of platforms keep some logs for timestamps and other parameters, many platforms obscure these path control protocols in the bid to better shield user privacy. This fait accompli about cooperation between platforms would tilt GNN advocates in scale; also, there ought to exist transparency agreements between platforms and regulatory bodies enforcing the use of data metadata only in the right ways (Baracaldo et al., 2021).

Furthermore, with that privacy protecting some form of behavior, i.e., if the data model were GNN, while much privacy was realized, the possibility is also open to collect behavioral data. That would include behaving in intrusive ways de-anonymized group dynamics and communication timing specifically while yet anonymized. ETHICAL GUIDELINES Explainable AI, consent, opt-in programs, and robust transparency reporting provide the pillars of public trust in the midst of a largely muddy world embodying these applications (Brunton & Nissenbaum, 2015).

Ethical and Regulatory Harmony

The ethics regarding this research ride aptly with proportionate thinking, where the least invasive technology necessary for protecting societies may be conserved. These do not allow for access to the user content itself but help platforms fulfill their duty of care, mitigating the impacts of misinformation, using structural data. The aforesaid action would gain more resonance in any context where the GDPR, the Algorithmic Accountability Act (U.S.), or any other pertinent legislation confining data protection laws apply (Tschider, 2022; European Commission, 2022).

On the contrary, the system's interpretability-node-level attribution and traceability of propagation-are proven to preserve due process in realm of automated decision-making. Wrongly flagged users would be able to ask for, challenge or demand human-mediated review of decisions-creating a more democratic, fair digital atmosphere.

Future Research and Societal Impact

This study introduces numerous research avenues. Future versions of the system could reveal the following:

1. Train detection models on-device using federated learning so that the models are not centralized with any data
2. Enhance cross-platform propagation tracking-mapping disinformation from encrypted space into the public space
3. Design counter-narrative generation tools that will complement flagged disinformation with factual or remedial messages.

Further integrate graph explainability techniques such as GNNExplainer or PGM-based interpretations for the support of analysts and policymakers (Ying et al., 2019).

Generally speaking, deployment of these systems in erecting democratic-resilient digital infrastructures is the next best thing to counter-manipulation. In the face of rampant election influence, public-health

disinformation, and synthetic media threats, secure communication channels will increasingly be seen to be at a juncture between national security and civil society emergencies.

Anyway, good governance is the bottom line, for the benefit of society, in the deployment of these. Technology itself is not the solution for disinformation. The consortium of governments, platform operators, AI researchers, privacy advocates, and civil society will be required to carry out the deployment and ensure that, under a transparent, accountable, inclusive framework, tools essentially used to promote truth would not inadvertently erode freedom.

Final Remark

In the end, the study groups into-line, eventually, with how computing with metadata might offer an ethical and technically deployable means of bypassing the resistance enclave, and accessing an AI-generated disinformation production ecosystem. Such a path also offers a way for the redemption of censorship justification that brings up to detection of behavioral structures, which, in its turn, enriches privacy and security.

As encrypted messaging continues to grow and generative AI sophistication increases, it is imperative to counter this growing threat with tools equally sophisticated, flexible, and principled. As graph neural networks present one of the most fascinating such tools, their success lies in responsible innovation and careful deployment.

REFERENCES

[1] Bontcheva, K., Papadopoulos, S., Tsalakanidou, F., & others. (2024). *Generative AI and disinformation: Recent advances, challenges, and opportunities*. KU Leuven.

[2] Bokolo, B. G., & Liu, Q. (2024). Artificial intelligence in social media forensics: A comprehensive survey and analysis. *Electronics*, *13*(2).

[3] Bradshaw, S., Howard, P. N., & Kollanyi, B. (2021). Industrialized disinformation: 2020 global inventory of organized social media manipulation. *Computational Propaganda Project*. https://comprop.oii.ox.ac.uk

[4] Brunton, F., & Nissenbaum, H. (2015). *Obfuscation: A user's guide for privacy and protest*. MIT Press.

[5] European Commission. (2022). *Digital Services Act (DSA): Regulation (EU) 2022/2065*.

[6] Hamilton, W. L., Ying, Z., & Leskovec, J. (2020). Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *14*(3), 1–159. https://doi.org/10.2200/S01063ED1V01Y202009AIM046

[7] Helmus, T. C. (2022). *Artificial intelligence, deepfakes, and disinformation*. RAND Corporation.

[8] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*.

[9] Kreps, S., McCain, R., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of disinformation. *Political Behavior*, *44*(3), 1–25. https://doi.org/10.1007/s11109-022-09789-5

[10] Mozilla Foundation. (2022). *Privacy and propaganda: How encrypted apps became tools of disinformation*.

[11] Murayama, Y., Asai, T., Fujita, S., & others. (2021). Detecting coordinated inauthentic behavior on encrypted messaging apps. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 5794–5802.

[12] Novelli, C., & Sandri, G. (2024). Digital democracy in the age of artificial intelligence. *arXiv preprint arXiv:2412.07791*.

[13] Oord, A., Dieleman, S., Zen, H., et al. (2016). WaveNet: A generative model for raw audio. *DeepMind*.

[14] Revett, K., Jahankhani, H., de Magalhães, S. T., & Gorunescu, F. (2007). A survey of user authentication based on mouse dynamics. *Journal of Computers*, *2*(3), 48–56.

[15] Singh, R., & Dwivedi, R. (2023). Detection of synthetic media and disinformation campaigns on encrypted networks using GNNs. *Journal of Information Warfare*, *22*(1), 55–72.

[16] Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

[17] Tschider, C. A. (2022). *The Algorithmic Accountability Act: Regulating AI and ensuring ethical compliance*. Georgetown Law Journal, *110*(5), 1241–1279.

[18] Ünver, A. (2023). Emerging technologies and automated fact-checking: Tools, techniques and algorithms. *EDAM Policy Papers*.

[19] Valavanidis, A. (2023). *Artificial intelligence (AI) applications*. National and Kapodistrian University of Athens.

[20] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[21] Wang, B., Wang, Y., & Zhang, J. (2023). Detecting coordinated misinformation in encrypted group chats using graph-based clustering. *Information Sciences*, *626*, 192–208.
[22] Weller, A., Garcia, J. A., & Bengio, Y. (2022). Transparency and accountability in AI systems: A technical roadmap. *AI & Society*, *37*(3), 497–510.
[23] Welling, M., & Kipf, T. N. (2020). Graph neural networks for semi-supervised learning. *Foundations and Trends in Machine Learning*, *13*(4), 307–406.
[24] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24.
[25] Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, *32*, 9240–9251.
[26] Yu, J., Yu, Y., Wang, X., Lin, Y., Yang, M., & Qiao, Y. (2024). The shadow of fraud: The emerging danger of AI-powered social engineering and its possible cure. *arXiv preprint arXiv:2401.01123*.
[27] Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2020). Fake news: Fundamental theories, detection strategies and challenges. *ACM Transactions on Information Systems (TOIS)*, *38*(3), 1–40.
[28] da Silva, B. C. C., Ferraz, T. P., & De Deus Lopes, R. (2024). *Enriching GNNs with text contextual representations for detecting disinformation campaigns on social media*. arXiv preprint arXiv:2410.19193.
[29] Lakzaei, B., Chehreghani, M. H., & Bagheri, A. (2024). Disinformation detection using graph neural networks: A survey. *Artificial Intelligence Review*, 1–47.
[30] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3), 1–42.