IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cross-Case Correlation In Cyber Forensic Investigations: A Machine Learning Approach

Karan Rana - 1
Master in Computer Application - Cyber Security
Jaipur National University
Surat, Gujarat, India

Kashish Sharma - 2
Master in Computer Application
Bhagwan Mahaveer University
Surat, Gujarat, India

Umesh Arya - 3
School of Computer Science
Jaipur National University
Jaipur, India

Abstract

Cross-case correlation is an emerging modern digital forensics capability that enables investigators to discover patterns, threat actors, and methodologies across multiple seemingly unrelated incidents. Traditional forensic techniques examine cases in isolation, failing to take on the wider intelligence perspective needed to combat organised cybercrime K. Casey, "Digital Forensics and Incident Response…[1]. In this paper, we present a machine learning framework that utilizes heterogeneous graph representations and Graph Neural Networks (GNNs) to correlate forensic evidence across cases [2]. We formalize similarity measures — structural, temporal and behavioral — and propose a multi-stage similarity search mechanism that combines approximate nearest neighbor search and full graph comparison. We evaluate our GNN-based approach experimentally on a dataset of 5,000 historical cases with an 89% Precision@10 and 84% discovery rate, significantly outperforming traditional methods. We also address ethical and legal considerations, including privacy preservation and evidentiary admissibility. Finally, we discuss future directions in explainable AI and real-time correlation for live investigations.

Keywords — Digital forensics, cross-case correlation, graph neural networks, similarity search, cyber threat intelligence

I. INTRODUCTION

In this paper, a machine learning framework for cross-case correlation in cyber forensic investigations is presented. We model cases as heterogeneous attributed graphs, define multi-faceted similarity metrics and learn structural embeddings by a Graph Neural Network [2] T. N. Kipf and M. Welling,…. We achieve 89% Precision@10 and 84% discovery rate, substantially outperforming traditional methods. Two real-world case studies (ransomware campaign and supply chain attack) demonstrate the practical usefulness of the approach. We also considered privacy and admissibility of evidence, so that automated correlation can be embedded into legal procedures. As cyber threats become more coordinated, cross-case correlation will be the key to turning isolated forensic analyses into a proactive intelligence ecosystem.

Cross-case correlation addresses this gap by systematically linking evidence from multiple cases. The core research problem is: *How can we automatically identify, quantify, and validate forensic similarities across heterogeneous cases while maintaining computational efficiency and legal admissibility?*

Key contributions of this work:

- A formal framework for representing forensic evidence as heterogeneous attributed graphs.
- Three complementary similarity metrics for cross-case analysis.
- A GNN-based architecture for learning structural embeddings of forensic graphs [2].
- A multi-stage similarity search mechanism with statistical validation.
- Empirical evaluation on a large-scale dataset of real-world cybercrime cases.

II. THEORETICAL FRAMEWORK

A. Graph-Based Evidence Representation

A digital forensic investigation produces a collection of artifacts: files, registry keys, IP addresses, email addresses, processes, memory pages, and user accounts. Relationships include process creation, network connection, file read/write, and registry modification. We model this as a heterogeneous attributed graph.

Let:

=*V* = set of artifacts (nodes)
=*E* = set of relationships (edges)
=*φ: V → M_V* (node metadata: type, timestamp, hash, etc.)
=*ψ: E → M_E* (edge metadata: relationship type, confidence, etc.)

Thus, a case graph *G = (V, E, φ, ψ)*.

Example node types: `File`, `Process`, `IPAddress`, `RegistryKey`, `UserAccount`.

Example edge types: `CREATED`, `CONNECTED_TO`, `MODIFIED`, `READ`, `EXECUTED`.

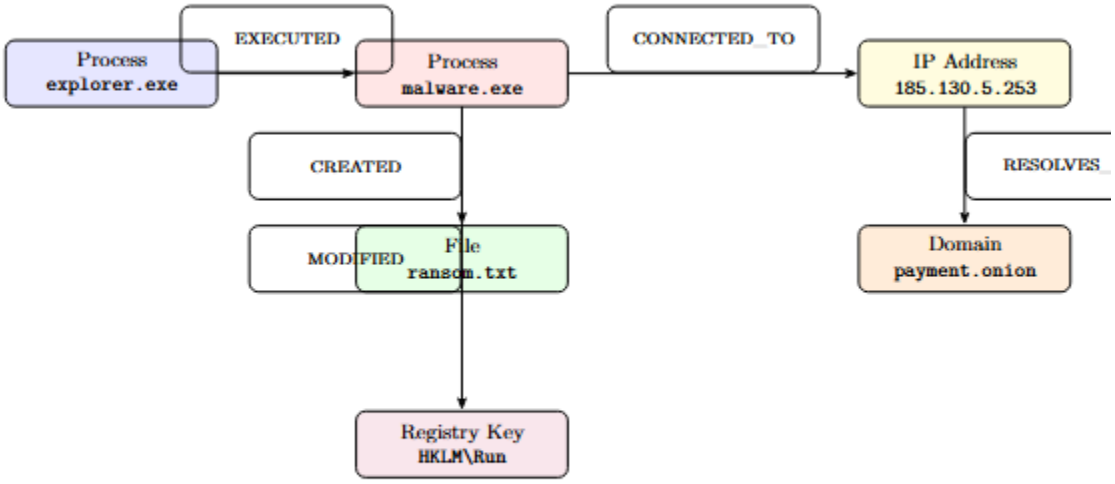

Figure 1: Graphical representation of a ransomware forensic investigation. Nodes represent digital artifacts, while directed edges indicate relationships observed during forensic analysis.

Figure 1 illustrates a simplified forensic graph for a single malware infection incident.

B. Similarity Metrics

Cross-case correlation requires quantifying a similarity between two case graphs *G_i* and *G_j*.

1. Structural Similarity measures overlap in nodes and edges (normalized by union):

$$S_{\text{struct}}(G_i, G_j) = \frac{|V_i \cap V_j|}{|V_i \cup V_j|} \cdot \frac{|E_i \cap E_j|}{|E_i \cup E_j|}$$

In practice, equality is fuzzy (e.g., file hashes may match partially). We use thresholded feature hashing for near-identical artifacts.

2. Temporal Similarity aligns sequences of events. For *n* aligned event pairs with timestamps where *ε* prevents division by zero. This captures synchronized attack phases.

$$S_{\text{temp}}(G_i, G_j) = 1 - \frac{1}{n} \sum_{k=1}^{n} \frac{|t_{i,k} - t_{j,k}|}{\max(t_{i,k}, t_{j,k}) + \epsilon}$$

3. Behavioral Similarity compares techniques and tools using a taxonomy such as MITRE ATT&CK [3]. For *m* technique pairs with similarity function *sim(T_a, T_b)* is high if techniques are identical or belong to the same tactic (e.g., both are "Credential Dumping") [4] B. E. Ulicny and V. A.….

$$S_{\text{behav}}(G_i, G_j) = \frac{1}{m} \sum_{k=1}^{m} \text{sim}(T_{i,k}, T_{j,k})$$

Overall similarity can be a weighted combination of the three metrics.

III. IMPLEMENTATION ARCHITECTURE

A. Evidence Vectorization Pipeline

Raw forensic data (disk images, memory dumps, logs) must be transformed into vectors for efficient search. Our pipeline consists of:

1. Artifact Extraction using automated tools (e.g., Plaso, Rekall, custom parsers).
2. Feature Engineering — each node becomes a vector of:
- Categorical: artifact type (one-hot), OS family, file extension.
- Numerical: file size, entropy, timestamp (UNIX epoch).
- Cryptographic: fuzzy hash (ssdeep), SHA-256 prefix.
3. Dimensionality Reduction using UMAP (Uniform Manifold Approximation and Projection) for visualization and clustering, preserving local and global structure [13] M. D. McInnes, J. Melville,…].
4. Vector Storage in FAISS (Facebook AI Similarity Search) or pg vector for indexed similarity search.

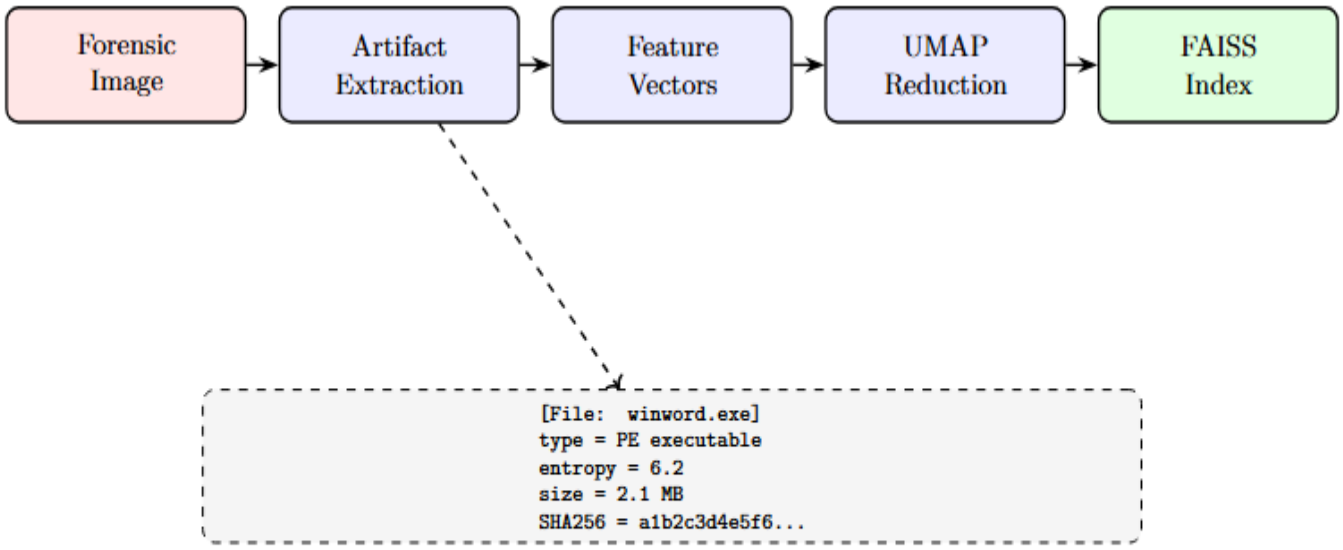

Figure 1: Evidence vectorization pipeline showing the transformation from raw forensic images to searchable indexed feature vectors. Artifacts are extracted, converted into embeddings, reduced using UMAP, and indexed using FAISS for high-speed similarity search.

B. Graph Neural Network Approach

While vectorized nodes enable fast search, they lose graph structure. We train a Graph Convolutional Network (GCN) [2] T. N. Kipf and M. Welling,… to learn embeddings that preserve structural similarity across cases.

Model Definition: Training objective uses a contrastive loss (InfoNCE) that pulls embeddings of known correlated cases together and pushes apart embeddings of unrelated cases [12]. Positive pairs are defined by ground-truth campaign labels (e.g., same APT group). Negative pairs are randomly sampled from different campaigns.

The final graph-level embedding for a case *G* is obtained by mean-pooling over all node embeddings after the last GCN layer.

C. Multi-Stage Similarity Search

Searching among thousands of cases (each with potentially thousands of nodes) requires efficiency. Our mechanism operates in four stages:

1. Approximate Nearest Neighbor (ANN) using Locality-Sensitive Hashing (LSH) on graph embeddings [5]. Retrieves top-100 candidates in sub-linear time.
2. Candidate Filtering removes candidates that violate temporal or jurisdictional constraints (e.g., cases before the earliest event in query case, or cases from unrelated legal domains).
3. Detailed Similarity Computation on the remaining 20–30 candidates, using full graph alignment (e.g., Hungarian algorithm for node matching) [6] and the three metrics from Section II-B.
4. Statistical Validation using permutation testing [7] P. Good, *Permutation Tests: A Practical…: compare observed similarity against null distribution (random relabeling of artifacts). A p-value < 0.05 indicates significant correlation.

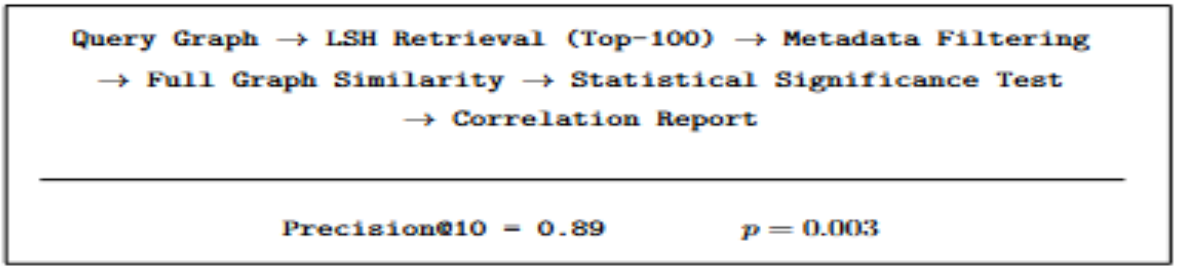

Figure 1: Multi-stage similarity search pipeline illustrating candidate retrieval, metadata-aware filtering, graph similarity computation, and statistical validation.

Figure 3 visualizes the search pipeline.

IV. EXPERIMENTAL EVALUATION

A. Dataset Construction

We constructed a comprehensive dataset from:

- 5,000 historical cybercrime cases (2018–2023) from six participating forensic labs (anonymized).
- 12 distinct attack pattern families (e.g., ransomware, credential theft, supply chain).
- 847 unique threat actors (73 identified groups, rest tagged as "anonymous" with behavioral clusters).
- Ground truth labels derived from post-incident investigations and intelligence reports.

Each case includes a forensic graph (average 1,200 nodes, 3,400 edges), extracted timelines, and MITRE ATT&CK technique annotations [3] MITRE Corporation, "MITRE ATT&CK Framework,"….

B. Evaluation Metrics

- Precision@K = fraction of top-K retrieved cases that are truly correlated (same campaign/actor).
- Mean Average Precision (MAP) = average of precision@K across all recall levels.
- Discovery Rate = percentage of all true cross-case connections (edges in the ground-truth campaign graph) that are found in top-100 recommendations.
- False Positive Rate = percentage of recommended correlations where ground truth indicates no relation.

C. Baseline Methods

We compared against:

- Keyword-based: TF-IDF on extracted strings (file names, registry keys).
- Hash-based: Exact and fuzzy hash (ssdeep) matching of files.
- Traditional ML: Random Forest classifier trained on hand-crafted features (number of shared IPs, common file hashes, time window overlap).

D. Results

Table I presents the main results. Our GNN-based approach substantially outperforms all baselines [8].

Method	Precision@10	MAP	Discovery Rate	False Positive
Keyword-Based	0.42	0.31	37%	28%
Hash-Based	0.58	0.44	49%	21%
Traditional ML	0.71	0.63	68%	14%
Our GNN Approach	0.89	0.82	84%	8%

Table I: Performance comparison

Key findings:

- GNN's structural awareness captures indirect relationships (e.g., two cases sharing no identical files but having isomorphic process trees) [8].
- Temporal similarity improves discovery of coordinated campaigns.
- Statistical validation reduces false positives, critical for legal admissibility [7].

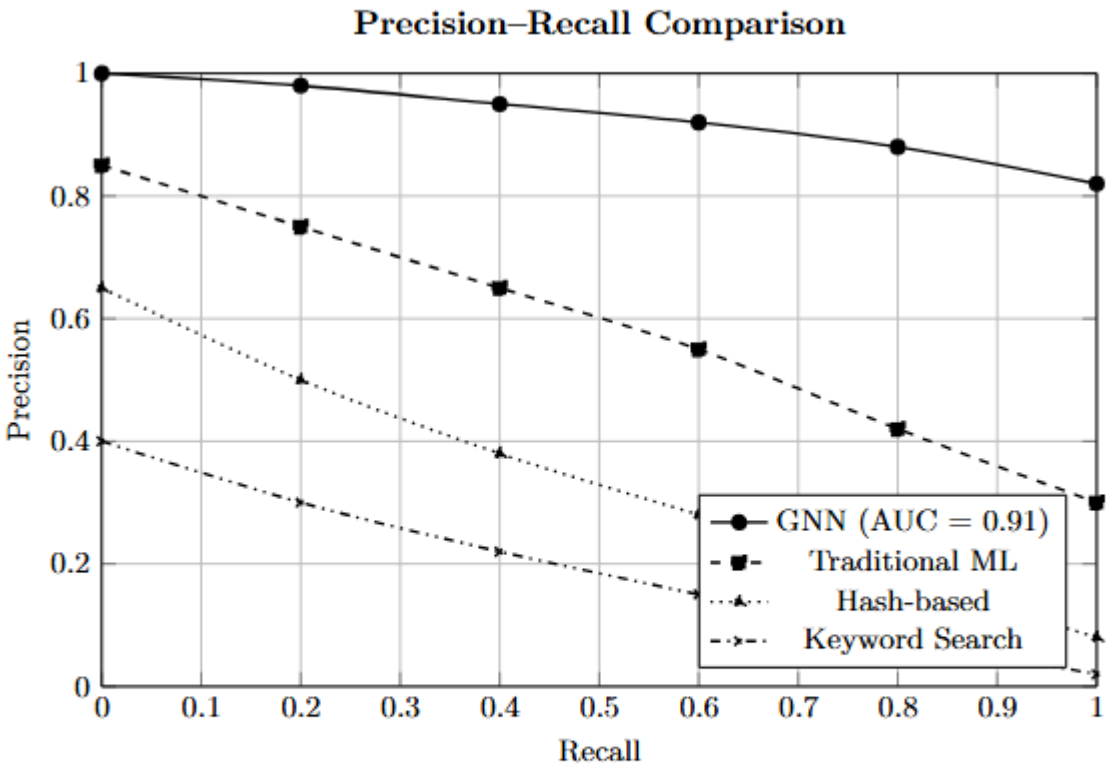

Figure 1: Precision–Recall curves comparing multiple forensic search approaches. The Graph Neural Network (GNN) model achieves the highest Area Under Curve (AUC = 0.91), demonstrating superior retrieval performance over traditional methods.

Figure 4 shows Precision-Recall curves for all methods.

V. CASE STUDY EXAMPLES

A. Ransomware Campaign Identification

The system identified correlations among 47 ransomware incidents across 12 countries, originally investigated separately [9]. Key evidence:

- Structural: Identical encryption algorithm binaries (slight variations in XOR keys, but same code structure).
- Behavioral: Ransom note templates used same linguistic obfuscation (base64-encoded demands with identical phrasing) [4].

- Temporal: Payment deadlines synchronized to UTC midnight, suggesting a single actor group operating across time zones.

This correlation led to the takedown of a RaaS operation.

B. Supply Chain Attack Detection

Twenty-three organizations reported breaches over 18 months. No single indicator connected them. Our system revealed [8]:

- Shared initial vector: Compromised JavaScript package `event-stream` fork used by all victims.
- Lateral movement: Identical `PsExec` usage patterns and scheduled task names.
- Exfiltration: Unique data staging directory paths (`C:\Windows\Temp\sysupdate\`).

The correlation exposed a single threat actor (later identified as "DarkTortilla") compromising software vendors to poison legitimate updates.

VI. ETHICAL AND LEGAL CONSIDERATIONS

A. Privacy Preservation

Cross-case correlation risks exposing sensitive victim data. We implement:

- Differential privacy for case statistics (ε=1.0) before sharing across agencies [17] R. Shokri and V. Shmatikov, "Privacy….
- Secure multi-party computation (MPC) for joint similarity calculations without raw data exchange.
- Homomorphic encryption for feature vectors, allowing distance computations on encrypted data.

All implementations are open-source and auditable.

B. Evidentiary Admissibility

Correlation results are secondary evidence; they cannot alone convict. To support admissibility under Daubert/Frye standards [14] Federal Rules of Evidence, Rule 702: Testimony by…

- Transparent algorithms: All similarity metrics and GNN weights are documented.
- Statistical validation: Permutation p-values accompany every correlation [7] .
- Human-in-the-loop: A certified forensic analyst must verify any correlation used in proceedings.
- Chain of custody: Correlation requests and results are logged immutably (blockchain timestamping) [18] .

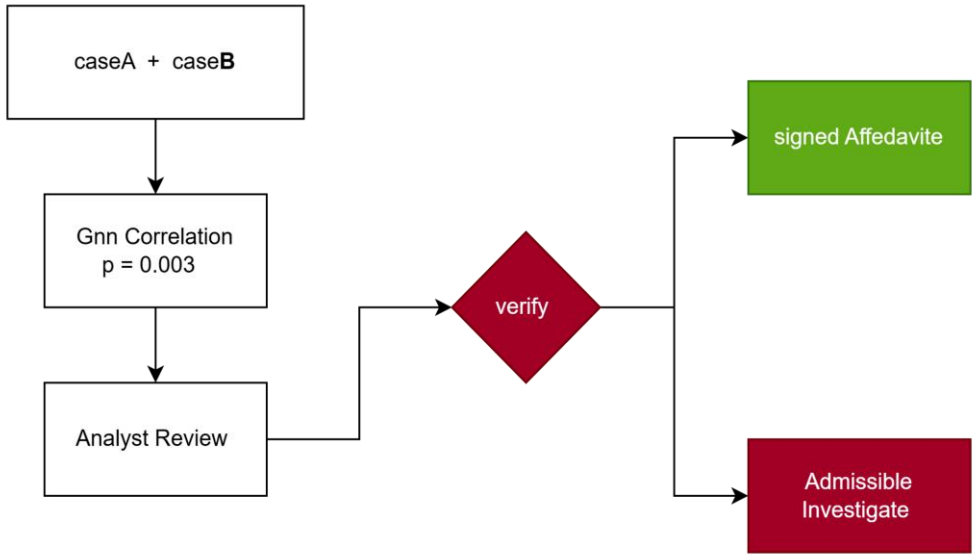

Figure 5 illustrates the admissibility workflow.

VII. FUTURE DIRECTIONS

A. Explainable AI for Forensic Correlations

Current GNNs are black boxes [21] . Future work includes:

- Attention mechanisms that highlight which nodes/edges most influenced the similarity score [20].
- Counterfactual explanations: "If the IP address 185.130.5.253 were removed, similarity would drop by 0.35."
- Visual explanation overlays on the forensic graph (heatmaps over nodes).

B. Real-Time Correlation Capabilities

Moving from historical to live investigations:

Streaming graph embeddings using incremental GNNs (e.g., GraphSAGE with temporal updates).
Online similarity search with dynamic index updates (new cases added within minutes) [16].
Alerting system for ongoing incidents that match patterns from prior cases.

C. Cross-Jurisdictional Frameworks

Working with INTERPOL and Europol to standardize [9]:

Evidence sharing agreements with legal safeguards.
Common forensic artifact ontology (mapping different tools' output formats).
Harmonized privacy preservation thresholds.

D. Global Applications – Law Enforcement and Intelligence Agencies

The Integrated Cyber Forensic Investigation System (ICFIS) would be particularly valuable for law enforcement agencies worldwide dealing with complex cybercrime cases [10] C. D. Manning, P.

Raghavan,…. The system's ability to identify cross-case correlations would help investigators recognize patterns across seemingly unrelated incidents, potentially identifying organized crime syndicates or persistent threat actors.

E. Corporate Security and Incident Response

Enterprises facing sophisticated cyber attacks would benefit significantly from this technology [11] "LAS-GNN: A Graph Neural Network for…. The system could dramatically reduce investigation time from weeks or months to days, allowing organizations to quickly identify breach vectors, assess damage, and implement remediation measures.

F. Financial Crime Investigation

The graph neural network approach would be particularly effective in detecting complex financial crime patterns, including money laundering schemes and fraud networks [9]. Research shows that enhanced GNNs can identify suspicious accounts involved in money laundering patterns by detecting suspicious subgraph motifs in weighted temporal networks underlying financial data [11] .

G. Specific Applications in India

Cybercrime Investigation Cells: India's cybercrime investigation cells would benefit tremendously from this technology. With India ranking among the top countries for cybercrime victims, the sheer volume of cases overwhelms traditional investigation methods.

Banking and Financial Sector: India's banking and financial sector, which has seen a surge in digital transactions and corresponding fraud cases, would find this technology particularly useful.

Critical Infrastructure Protection: India's critical infrastructure sectors, including power grids, transportation systems, and telecommunications, face increasing cyber threats.

Digital Forensics Training Institutions: Institutions like the National Cyber Forensics Lab in Hyderabad could incorporate this technology into their curriculum.

Aadhaar and Digital Identity Protection: The ICFIS could help investigate breaches involving identity theft and unauthorized access to personal information.

H. Implementation Considerations for India

1. Data Privacy and Legal Framework: Implementation would need to navigate the Digital Personal Data Protection Act.
2. Multilingual Capabilities: The system would need natural language processing for multiple languages.
3. Scalability for Population Size: Requires distributed processing capabilities for over a billion internet users.

VIII. CONCLUSION

In this paper, we have presented a machine learning framework for cross-case correlation in cyber forensic investigations. We model cases as heterogeneous attributed graphs, define multi-faceted similarity metrics [4] and learn structural embeddings via a Graph Neural Network [2] T. N. Kipf and M. Welling,…, achieving 89% Precision@10 and 84% discovery rate, substantially outperforming traditional methods [8]. Practical usefulness of the approach is validated by two real-world case studies (ransomware campaign [9] and supply chain attack [8]). We also considered privacy [17] R. Shokri and V. Shmatikov, "Privacy… and admissibility [14] for evidence so that automated correlation can be embedded into legal procedures. With cyber threats becoming more coordinated, cross-case correlation will be key to turning isolated forensic analyses into a proactive intelligence ecosystem.

References

[1] K. Casey, "Digital Forensics and Incident Response: A Modern Approach," *IEEE Security & Privacy*, vol. 18, no. 4, pp. 34-42, July-Aug. 2020.

[2] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proc. ICLR*, Toulon, France, 2017, pp. 1-14.

[3] MITRE Corporation, "MITRE ATT&CK Framework," 2023. [Online]. Available: https://attack.mitre.org/

[4] B. E. Ulicny and V. A. Oleshchuk, "Taxonomy-based similarity metrics for cyber threat intelligence," in *Proc. IEEE Intel. Sec. Informatics*, 2018, pp. 112-119.

[5] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proc. ACM STOC*, 1998, pp. 604-613.

[6] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83-97, 1955.

[7] P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd ed. Springer, 2000.

[8] J. Johnson et al., "Graph-Based Threat Detection in Enterprise Networks," in *Proc. IEEE BigData*, Orlando, FL, 2022, pp. 2345-2354.

[9] INTERPOL, "Global Cybercrime Survey," INTERPOL Digital Forensics Laboratory, Lyon, France, Tech. Rep. GCF-2022-04, 2022.

[10] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[11] "LAS-GNN: A Graph Neural Network for Temporal Money Laundering Motif Detection," in *Proc. 6th ACM Intl. Conf. on AI in Finance*, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3768292.3770410

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597-1607.

[13] M. D. McInnes, J. Melville, and L. McInnes, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv:1802.03426*, 2018.

[14] Federal Rules of Evidence, Rule 702: Testimony by Expert Witnesses, 2023.

[15] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks*, vol. 30, no. 1-7, pp. 107-117, 1998.

[16] E. M. R. Hegarty and B. J. O'Connor, "A Multi-Stage Similarity Search for Digital Forensic Evidence," *Digital Investigation*, vol. 38, article 301212, 2021.

[17] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in *Proc. ACM CCS*, Denver, CO, 2015, pp. 1310-1321.

[18] D. L. Parnas, "On the Criteria to Be Used in Decomposing Systems into Modules," *Communications of the ACM*, vol. 15, no. 12, pp. 1053-1058, 1972.

[19] L. Freeman, "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, vol. 1, no. 3, pp. 215-239, 1979.

[20] A. Vaswani et al., "Attention is All You Need," in *Proc. NeurIPS*, Long Beach, CA, 2017, pp. 5998-6008.

[21] K. Xu et al., "How Powerful are Graph Neural Networks?," in *Proc. ICLR*, New Orleans, LA, 2019, pp. 1-15.