Towards Reproducible Cyber Threat Attribution: A Bayesian Scoring Framework over STIX 2.1 Knowledge Representations

Aykhan Huseynli[1], Jamal Hamzazade[2]
[1]Azerbaijan Technical University, Baku, Azerbaijan
[2]University College Dublin, Dublin, Ireland
[1]*aykhan.huseynli@cert.az*, [2]*jamalhamzazade@gmail.com*

***Abstract*— Cyber threat attribution remains non-reproducible and poorly calibrated in uncertainty, limiting its operational utility. This paper addresses both deficiencies through a Bayesian attribution pipeline grounded in STIX 2.1 knowledge representation. Standardized threat intelligence reports are transformed into structured bundles, from which features are extracted and aggregated across multiple per-actor campaigns. Attribution scoring combines campaign-frequency and IDF-weighted Jaccard similarity with feature-class-specific Laplace smoothing, correctly distinguishing durable behavioral evidence from perishable technical indicators. Proposed pipeline is evaluated against 17 known actors from 57 reports and yields realistic, analyst-interpretable probability distributions.**

Keywords*— *cyber threat attribution; STIX 2.1; threat intelligence; Bayesian inference.

I. INTRODUCTİON

Attributing a cyberattack to a specific threat actor is foundational to strategic cyber defense. Accurate attribution enables proportionate countermeasures, informs diplomatic and legal responses, and supports long-term defensive posture. Yet the technical process remains largely artisanal: analysts correlate indicators of compromise (IoCs), behavioral patterns, and contextual clues across heterogeneous sources without a reproducible methodological framework.

Three principal deficiencies characterize existing attribution practice. First, results are non-reproducible – they depend on implicit analyst judgment that cannot be audited, challenged, or transferred across organizations. Second, intelligence is rarely represented in a common schema, making cross-organizational comparison prohibitively difficult. Third, probabilistic uncertainty is seldom quantified. Outputs are typically binary (attributed or not attributed) which misrepresents the epistemic state of the evidence and can lead to misattribution with serious geopolitical consequences [1].

This paper addresses these gaps with a fully automated attribution pipeline built on two open standards. STIX 2.1 provides a rigorously defined ontology for representing threat actors, malware, tools, infrastructure, vulnerabilities, and the semantic relationships between them [3]. Bayesian inference provides a principled probabilistic framework for updating belief about actor identity as evidence accumulates.

The central contributions of this work are: (1) a JSON-to-STIX 2.1 transformation layer enforcing a fixed nine-object schema suitable for attribution; (2) a union-based multi-report aggregation mechanism that builds complete actor profiles from multiple campaign reports while preserving source provenance and confidence filtering; (3) campaign frequency (CF) scoring that rewards behaviorally consistent features observed across multiple campaigns; (4) Inverse Document Frequency (IDF) weighted Jaccard similarity that automatically down-weights observables common across many actors; (5) feature-class-specific Laplace smoothing that correctly distinguishes behavioral evidence from perishable network indicators; (6) min-max normalization replacing softmax to prevent probability collapse; and (7) a dark figure advisory mechanism acknowledging the reporting-gap problem inherent to open-source intelligence.

The remainder of this paper is organized as follows. Section II surveys related work. Section III details the attribution methodology. Section IV presents case study results. Section V discusses limitations. Section VI concludes.

II. RELATED WORK

A. *Threat Attribution Frameworks*

Early attribution research established conceptual frameworks that remain influential. The Diamond Model decomposed intrusions into four interconnected features (adversary, capability, infrastructure, and victim) providing a structured vocabulary for analyst reasoning [4]. Later that approach was extended into a kill-chain-based taxonomy mapping adversary TTPs to observable behavioral signatures [5]. Association analysis over threat intelligence based on statistic, extension, behavior pattern, and probability similarity to identify actor linkages through shared technical artifacts was introduced later [6]. Threat playbook extensions for Cyber Threat Intelligence (CTI) that incorporate behavioral attribution signals at the strategic level was used to perform Advanced Persistent Threats (APTs) was performed in [7]. These works establish the conceptual vocabulary of attribution evidence but do not produce quantified probability estimates or operationalize uncertainty.

B. *Probabilistic and Machine Learning Attribution*

Reference [8] demonstrated machine learning-based

DOI: 10.25045/forensic.gov.2026.19

attribution using high-level IoCs for FinTech actors, achieving reasonable accuracy on a closed-world benchmark but without addressing uncertainty quantification. Mutiple ML classifiers were applied to cyber-attribution problems and found that feature selection significantly affected accuracy, presaging the IDF weighting approach developed here [9]. A modular framework using opinion pools to combine probabilistic attribution signals from multiple sources were proposed in [10]. Attribution over unstructured CTI reports using NLP-derived features and their context-aware extension further improved accuracy through hybrid feature fusion were proposed in [11], [12]. Deep learning over behavioral signatures was leveraged for soft attribution, achieving strong performance on larger datasets [13]. In addiiton to that, a comprehensive recent survey of APT attribution methods identified reproducibility and uncertainty quantification as persistent open problems, a gap which this work addresses [14].

C. *Knowledge Graphs and STIX-Based Intelligence*

Structured representation of threat intelligence has gained considerable attention. Mavroeidis and Bromander [3] evaluated CTI taxonomies, sharing standards, and ontologies, establishing STIX as the dominant interoperability standard. Bromander proposed automation-enabling extensions to CTI data models for sharing and exchange [15]. Reference [16] demonstrated automated extraction of STIX objects from narrative CTI reports, enabling scalable pipeline construction. A knowledge graph from open-source APT reports for attribution and applied graph-based reasoning to IoC repositories for APT identification was implemented in [17], [18]. Deep neural networks to APT attribution at the nation-state level was applied, achieving high closed-world accuracy but without transparency or uncertainty communication [12].

D. *False Flag and Reporting-Gap Challenges*

Pahi and Skopik demonstrated that false flag operations – (actors deliberately mimicking other groups' tooling) represent a fundamental limit on technical attribution [2]. Furthermore, [2] catalogued the types of technical artifacts used in attribution and their susceptibility to manipulation. In addiiton to that, Delphi and AHP methods were applied to rank artifact types by attribution evidential value [19]. These works motivate both the dark figure advisory mechanism introduced in this paper and the explicit uncertainty representation in the output reports.

III. METHODOLOGY

The pipeline is organized into five sequential stages, illustrated in Fig. 1: (A) STIX 2.1 transformation, (B) feature extraction, (C) multi-report aggregation with campaign frequency scoring, (D) CF-IDF-weighted Bayesian scoring, and (E) min-max normalization with saturation detection.

A. *STIX 2.1 Knowledge Representation*

Each input is transformed into a STIX 2.1 Bundle containing a fixed set of STIX Domain Objects (SDOs) and Relationship Objects (SROs). The schema is deliberately constrained to ensure consistent feature extraction across all actors and reports.

The nine SDO types produced are: Campaign, ThreatActor, Malware, Tool, Identity, Location, Vulnerability, and Indicator (IP address, domain, URL, email address, SHA-256, SHA-1, and MD5 hash). Each Indicator carries a STIX pattern derived from its type.

SDO and SRO relations are depicted in Fig. 1.

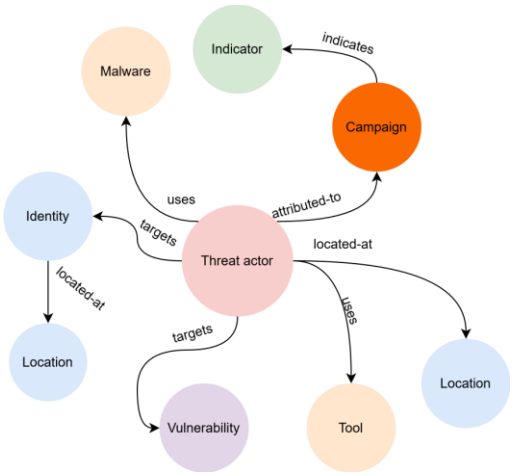

Figure 1. SDO and SRO relations.

B. *Feature Extraction*

Features are extracted from each STIX bundle by resolving SRO source and target references against an identifier-to-object lookup dictionary. The per-report feature vector *F(r)* contains ten keyed sets: malware names, tool names, targeted sector labels, actor origin country, target country names, CVE identifiers, IPv4 addresses, domain names, URLs, and file hashes.

C. *Multi-Report Aggregation and Campaign Frequency*

In operational practice, multiple reports document different campaigns attributed to the same actor. Treating each report as a separate known actor splits probability mass and degrades attribution performance [14]. The aggregation layer collapses per-report feature vectors into a single profile per actor using set union:

$$F_{agg}(a) = \bigcup_{r \in R(a)} F(r) \tag{1}$$

where $R(a)$ is the set of reports for actor a passing the source confidence filter (confidence ∈ {high, medium, low}).

Union aggregation is preferred over intersection, which penalizes well-documented actors whose campaigns exhibit diverse observable sets, and over per-report averaging, which dilutes strong individual matches.

A key limitation of naive union aggregation is that it destroys information about how consistently a feature value appeared across campaigns. To capture this behavioural repetition, a Campaign Frequency (CF) score is computed for each feature value before the profile is collapsed,

$$CF(a, f, v) = \frac{|\{r \in R(a) : v \in F_r(f)\}|}{|R(a)|} \tag{2}$$

where $F_r(f)$ is the set of values for feature f in report r. A CF value of 1.0 indicates that the observable appeared in every known campaign of that actor which is a strong behavioral fingerprint. CF scores are stored alongside the union profile and carried into the scoring function.

D. IDF Weighting

Not all feature values are equally discriminative for attribution. Observables common across many actors (e.g., widely-used tools, frequently-exploited vulnerabilities) provide weaker attribution signal than values unique to specific actors. We adopt a smoothed Inverse Document Frequency (IDF) weight computed over aggregated actor profiles. Let N be the number of known actors and $df(f, v)$ the count of actors whose aggregated profile contains value v for feature f. The smoothed IDF weight is,

$$IDF(f, v) = \log(\frac{N+1}{df(f,v)+1}) + 1, \quad (3)$$

Values appearing across many actors receive IDF approaching 1.0; values unique to one actor receive IDF approaching $\log(N + 1) + 1$. IDF is computed over aggregated profiles rather than individual reports, so a value that union aggregation introduces from one outlier report still receives the correct cross-actor rarity weight.

E. CF-IDF-Weighted Jaccard Similarity

Standard Jaccard similarity treats all feature values as equally discriminative and all occurrences as equally consistent. The proposed CF-IDF-weighted Jaccard integrates both the global rarity of a value (IDF) and its behavioural consistency within the known actor's campaigns (CF). Let U_f and A_f denote the feature sets of the unknown and known actor, respectively. The similarity for feature f is,

$$J(U_f, A_f, f) = \frac{\sum_{v \in U_f \cap A_f} \mathrm{IDF}(f,v) \cdot CF(a,f,v)}{\sum_{v \in U_f \cup A_f} IDF(f,v) \cdot \max(CF(a,f,v), 1)}, \quad (4)$$

where the denominator treats the unknown actor's single-report features as maximally consistent (CF = 1.0), ensuring the union weight is never reduced below the IDF contribution alone. The numerator rewards matched values that are simultaneously rare in the global actor corpus and behaviourally persistent across the known actor's documented campaigns. The function returns 0.0 when both sets are empty and 1.0 when the sets are identical under uniform weights.

F. Log-Posterior Scoring

We adopt a Naive Bayes framework with a uniform prior over actors. The log-posterior score for actor a_i given evidence E is:

$$\log S(a_i) = \sum_f w_f \cdot \log\left(J(U_f, A_f^i, f) + \varepsilon_f\right), \quad (5)$$

where w_f is the base feature weight (Table I) and ε_f is a feature-class-specific Laplace smoothing floor. A single global floor fails to account for the fundamentally different evidential meaning of absence across feature types. Network indicators such as IP addresses and domains are rotated frequently; their absence is expected and uninformative. Behavioral features such as malware families and targeted sectors are more stable; their absence is meaningful. Table II defines the class-specific floors used in this work.

TABLE I. FEATURE-CLASS-SPECIFIC LAPLACE SMOOTHING (ε)

Feature Class	ε	Rationale
Malware, Tools, Sectors, CVEs, Countries	0.05	Absence is informative – behavioral
File Hashes	0.30	Semi-stable; moderate absence penalty
IP Addresses	0.90	Perishable; near-zero absence penalty
Domains	0.85	Perishable; near-zero absence penalty
URLs	0.95	Very perishable; minimal absence penalty

TABLE II. FEATURE WEIGHTS

Feature	Weight	Rationale
Malware	3.0	Highly actor-specific
CVEs	2.5	Rare exploitation patterns
Tools	2.0	Behavioral fingerprint
Target Sectors	2.0	Targeting intent
File Hashes	2.0	Binary-level evidence
IP Addresses	1.5	Infrastructure overlap
Domains	1.5	Infrastructure overlap
Target Countries	1.5	Geographic intent
Actor Country	1.0	Origin corroboration
URLs	1.0	Perishable indicator

G. Min-Max Normalization and Saturation Detection

Softmax normalization applied to log-scores produces catastrophic probability collapse: when one actor scores above the floor on high-weight features and all others do not, exponential amplification assigns probability ≈1.0 to the winner regardless of actual score differences. This is a mathematical artifact, not an evidential finding. We replace softmax with min-max normalization:

$$S' = \frac{\log s(a_i) - S_{\min}}{S_{max} - S_{min}}, \quad (6)$$

where S_{min} and S_{max} are the minimum and maximum log-scores across all known actors. Normalized scores are then re-scaled to sum to unity to yield relative probabilities,

$$P(a_i \mid E) = \frac{S'(a_i)}{\sum_j S'(a_j)}. \quad (7)$$

This preserves rank ordering, distributes probability mass across the full actor set, and does not amplify small score differences. When $S_{max} = S_{min}$, (all actors score identically), scores are set to $1/N$. A saturation detector flags outputs for mandatory analyst review when $P(a_1 \mid E) > 0.85$ or the gap $\Delta = P(a_1 \mid E) - P(a_2 \mid E) > 0.60$, signalling that a single feature is likely dominating the posterior.

IV. CASE STUDY: CAMPAİGN TARGETİNG AZERBAİJAN

A. Dataset and Unknown Actor

The evaluation dataset comprises 17 known threat actors profiled from 57 threat intelligence reports. The corpus includes nation-state groups (APT28, APT29, APT34, Lazarus Group, MuddyWater, Sandworm, OilRig), regionally focused actors (TGR-STA-1030, Batavia, UNC2970, Gamaredon), and other documented groups. Reports per actor range from 3 to 5. All reports carried medium or high source confidence.

The unknown actor is derived from a campaign targeting government, energy, transportation, and financial sector organizations in Azerbaijan and Turkey. Technical artifacts include a specific malware family, CVE-2017-11882 exploitation, seven IPv4 addresses, three domains, fourteen SHA-256 file hashes, and two URLs. The actor's origin country could not be determined from available reporting.

B. Attribution Results

Table III presents the full attribution ranking. In this case no saturation condition was triggered ($\Delta = 0.203$). The top attribution is TA558 at 31.99%, followed by MuddyWater at 11.70% and Sandworm at 8.04%.

TABLE III. ATTRIBUTION RANKING

Rank	Actor	Reports	Score	Probability
1	TA558	3	-31.55	31.99%
2	MuddyWater	5	-36.16	11.70%
3	Sandworm	3	-36.99	8.04%
4	OilRig	3	-37.04	7.79%
5	APT34	3	-37.14	7.37%

C. Evidence Analysis

Table IV summarizes the shared features for the top two candidates with IDF and CF scores.

TA558's attribution rests on two high-quality matches. The malware family match carries IDF = 3.197 and CF = 1.000, indicating the tool is rare across the 17-actor dataset and appeared in every TA558 campaign report – the maximum possible behavioral consistency signal. The CVE-2017-11882 match carries IDF = 3.197 with CF = 0.333, indicating the exploitation pattern is rare but was observed in only one of three TA558 reports.

The second-ranked candidate shares no malware or CVE overlap with the unknown actor. Its case rests entirely on geographic alignment: Azerbaijan (IDF = 3.197, CF = 0.400) and Turkey (IDF = 2.792, CF = 0.600) are both rare targets in the dataset and fall squarely within MuddyWater's established operational footprint across the MENA and South Caucasus regions. The combined geographic signal is substantive but is outweighed by TA558's high-IDF, high-CF malware match under the current base weight assignments ($w_{malware} = 30$ versus $w_{target\_countries} = 1.5$).

TABLE IV. CONTRIBUTING FEATURES

Actor	Feature	Value	IDF	CF
TA558	Malware	Remcos RAT	3.197	1.000
TA558	CVE	CVE-2017-11882	3.197	0.333
TA558	Sector	Financial	1.693	0.333
MuddyWater	Country	Azerbaijan	3.197	0.400
MuddyWater	Country	Turkey	2.792	0.600
MuddyWater	Sector	Transportation	2.504	0.200
MuddyWater	Sector	Financial	1.693	0.200

The diverging signals for TA558 – no geographic overlap with Azerbaijan or Turkey are explicitly labeled in the pipeline output as potentially reflecting reporting gaps rather than behavioral absence. TA558's profile is constructed from 3 reports from a single vendor, a coverage tier that carries elevated dark figure risk: the actor may operate in regions not reflected in public reporting.

V. DİSCUSSİON

A. Limitations

Independence assumption. The Naive Bayes framework treats all features as statistically uncorrelated. In practice, geographic targeting, sector targeting, and actor origin are structurally correlated – nation-state actors operate within geopolitically defined mandates. A Bayesian network or joint probability model would capture these dependencies but requires substantially more training data than available in open-source corpora.

Dark figure problem. The pipeline operates exclusively on public CTI. The true behavioral profile of any actor extends beyond what vendor reports document. The absence of a feature in an actor's profile may reflect unreported activity rather than genuine behavioral absence – the dark figure problem [2]. This limitation is irreducible within any open-source attribution pipeline. The dark figure advisory mechanism in the output report makes this uncertainty explicit rather than suppressing it, but cannot quantify it.

Closed-world IDF. IDF weights are computed over the 17-actor dataset. Values appearing in only one actor receive high IDF not because they are globally rare but because the dataset is small. As the dataset grows toward more actors, IDF weights will converge toward their true discriminative values.

B. Operational Implications

The pipeline's primary operational value is to support attribution efforts. An analyst should interpretate the output as: "prioritize TA558 and MuddyWater as primary investigation hypotheses, with the following evidence for and against each."

The pipeline's transparency property is its most operationally significant characteristic since every score is traceable to specific feature matches with explicit IDF and CF values. An analyst with classified knowledge that TA558 has conducted unreported operations in Azerbaijan can rationally override the geographic divergence signal. This is not possible with black-box attribution systems discussed earlier.

CONCLUSİON

This paper presented a structured Bayesian attribution pipeline for cyber threat actor identification grounded in STIX 2.1 intelligence representations. The pipeline addresses three core deficiencies of existing approaches – non-reproducibility, absence of a common intelligence schema, and failure to quantify probabilistic uncertainty – through five methodological contributions: CF-IDF-weighted Jaccard similarity, feature-class-specific Laplace smoothing, min-max normalization, multi-report aggregation with union profiles, and a dark figure advisory mechanism.

Case study evaluation against a 17-actor, 57-report dataset produced realistic probability distributions with meaningful differentiation between candidates, and surfaced a genuine analytical ambiguity between toolchain evidence (favoring TA558) and geographic behavioral evidence (favoring MuddyWater) that binary attribution methods would have suppressed.

Future work will address the independence assumption through Bayesian network modeling, incorporate temporal decay for perishable indicators, and develop a confidence calibration layer mapping model scores to empirically validated posterior probabilities.

REFERENCES

[1] T. Pahi and F. Skopik, "Cyber Attribution 2.0: Capture the False Flag".

[2] F. Skopik and T. Pahi, "Under false flag: using technical artifacts for cyber attack attribution," Cybersecur, vol. 3, no. 1, p. 8, Dec. 2020, doi: 10.1186/s42400-020-00048-4.

[3] V. Mavroeidis and S. Bromander, "Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence," in 2017 European Intelligence and Security Informatics Conference (EISIC), Athens: IEEE, Sep. 2017, pp. 91–98. doi: 10.1109/EISIC.2017.20.

[4] S. Caltagirone, A. Pendergast, and C. Betz, "The Diamond Model of Intrusion Analysis," 2013, doi: 10.13140/RG.2.2.31143.56481.

[5] P. N. Bahrami, A. Dehghantanha, T. Dargahi, R. M. Parizi, K.-K. R. Choo, and H. H. S. Javadi, "Cyber Kill Chain-Based Taxonomy of Advanced Persistent Threat Actors: Analogy of Tactics, Techniques, and Procedures," Journal of Information Processing Systems, vol. 15, no. 4, pp. 865–889, Aug. 2019, doi: 10.3745/JIPS.03.0126.

[6] Q. Li, Z. Yang, Z. Jiang, B. Liu, and Y. Fu, "Association Analysis Of Cyber-Attack Attribution Based On Threat Intelligence," in Proceedings of the 2017 2nd Joint International Information Technology, Mechanical and Electronic Engineering Conference (JIMEC 2017), Chongqing, China: Atlantis Press, 2017. doi: 10.2991/jimec-17.2017.49.

[7] K. Edie, C. Mckee, and A. Duby, "Extending Threat Playbooks for Cyber Threat Intelligence: A Novel Approach for APT Attribution," in 2023 11th International Symposium on Digital Forensics and Security (ISDFS), Chattanooga, TN, USA: IEEE, May 2023, pp. 1–6. doi: 10.1109/ISDFS58141.2023.10131867.

[8] U. Noor, Z. Anwar, T. Amjad, and K.-K. R. Choo, "A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise," Future Generation Computer Systems, vol. 96, pp. 227–242, Jul. 2019, doi: 10.1016/j.future.2019.02.013.

[9] M. Alkaabi and S. O. Olatunji, "Modeling Cyber-Attribution Using Machine Learning Techniques," in 2020 30th International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt: IEEE, Dec. 2020, pp. 10–15. doi: 10.1109/ICCTA52020.2020.9477672.

[10] K. T. W. Teuwen, "A Modular Approach to Automatic Cyber Threat Attribution using Opinion Pools," in 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy: IEEE, Dec. 2023, pp. 3089–3098. doi: 10.1109/BigData59044.2023.10386708.

[11] E. Irshad and A. Basit Siddiqui, "Cyber threat attribution using unstructured reports in cyber threat intelligence," Egyptian Informatics Journal, vol. 24, no. 1, pp. 43–59, Mar. 2023, doi: 10.1016/j.eij.2022.11.001.

[12] I. Rosenberg, G. Sicard, and E. David, "DeepAPT: Nation-State APT Attribution Using End-to-End Deep Neural Networks," vol. 10614, 2017, pp. 91–99. doi: 10.1007/978-3-319-68612-7_11.

[13] E. Böge, M. B. Ertan, H. Alptekin, and O. Çetin, "Unveiling Cyber Threat Actors: A Hybrid Deep Learning Approach for Behavior-Based Attribution," Digital Threats, vol. 6, no. 1, pp. 1–20, Mar. 2025, doi: 10.1145/3676284.

[14] N. Rani, B. Saha, and S. K. Shukla, "A comprehensive survey of automated Advanced Persistent Threat attribution: Taxonomy, methods, challenges and open research problems," Journal of Information Security and Applications, vol. 92, p. 104076, Jul. 2025, doi: 10.1016/j.jisa.2025.104076.

[15] S. Bromander et al., "Investigating Sharing of Cyber Threat Intelligence and Proposing A New Data Model for Enabling Automation in Knowledge Representation and Exchange," Digital Threats, vol. 3, no. 1, pp. 1–22, Mar. 2022, doi: 10.1145/3458027.

[16] F. Marchiori, M. Conti, and N. V. Verde, "STIXnet: A Novel and Modular Solution for Extracting All STIX Objects in CTI Reports," in Proceedings of the 18th International Conference on Availability, Reliability and Security, Benevento Italy: ACM, Aug. 2023, pp. 1–11. doi: 10.1145/3600160.3600182.

[17] Y. Ren, Y. Xiao, Y. Zhou, Z. Zhang, and Z. Tian, "CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution," IEEE Trans. Knowl. Data Eng., vol. 35, no. 6, pp. 5695–5709, Jun. 2023, doi: 10.1109/TKDE.2022.3175719.

[18] I. J. King, R. Ramirez, B. Bowman, and H. H. Huang, "Trail: A Knowledge Graph-Based Approach for Attributing Advanced Persistent Threats," in 2025 IEEE 41st International Conference on Data Engineering (ICDE), Hong Kong, Hong Kong: IEEE, May 2025, pp. 1207–1220. doi: 10.1109/ICDE65448.2025.00095.

[19] S. Hwang and T.-S. Kim, "An Exploratory Study on Artifacts for Cyber Attack Attribution Considering False Flag: Using Delphi and AHP Methods," IEEE Access, vol. 11, pp. 74533–74544, 2023, doi: 10.1109/ACCESS.2023.3295427.