International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.20, June 2025

Multimodal Threat Actor Profiling on the Tor Network: Techniques, Datasets, and Ethical Challenges

Pavan Kumar Pativada
Department of Computer Science, Kansas State University
Manhattan, 66506, KS, USA

Rahul Karne
Department of Computer Science, Kansas State University
Manhattan, 66506, KS, USA

Akhil Dudhipala
Department of Computer Science, Kansas State University
Manhattan, 66506, KS, USA

ABSTRACT

Profiling threat actors operating on the Tor network presents considerable challenges due to its intrinsic anonymity and layered encryption. This paper offers a comprehensive survey of major advancements between 2019 and 2025, with reference to foundational tools and methods developed earlier where relevant (e.g., Tor simulation, darknet datasets). Core methodological approaches include stylometric analysis of linguistic features [8, 9], content classification of hidden services [2, 5], encrypted traffic analysis [7], temporal behavioral modeling [10], and graph-based account linkage [6, 12].

A conceptual profiling system is proposed that ingests heterogeneous data sources—such as textual posts, metadata, and traffic logs—extracts modality-specific features (e.g., writing style, network flow patterns, timestamp distributions), and applies domain-aligned ML models for multimodal embedding and identity fusion. To illustrate its practical relevance, a synthetic case study is presented demonstrating how AI techniques can correlate a threat actor's forum posts and marketplace listings to infer authorship and behavioral alignment.

Key public datasets and tools are also cataloged—including VeriDark [9], CoDA [5], DUTA [2], ISCX-Tor [7], and the Shadow simulator [4]—that enable reproducible research in this domain. The survey concludes with a discussion of critical ethical and legal considerations, including compliance with the EU General Data Protection Regulation (GDPR) [11], the European Union Artificial Intelligence Act [1], and U.S. surveillance law under FISA Section 702 [3]. This paper aims to provide a rigorously referenced, technically detailed, and ethically grounded synthesis of state-of-the-art methods in AI-driven threat actor profiling on the Tor network.

Keywords

Tor network, Darknet forensics, threat actor profiling, stylometry, multimodal learning, deep learning, temporal behavior modeling, graph neural networks, user de-anonymization, encrypted traffic classification, darknet marketplaces, AI ethics, GDPR compliance, Shadow simulator

1. INTRODUCTION

The Tor network is a widely adopted low-latency anonymous communication overlay that enables users to access services and exchange messages without revealing their identity or location. Tor achieves its anonymity through a mechanism known as onion routing, wherein client traffic is relayed through a chain of volunteer-operated routers—commonly referred to as entry, middle, and exit nodes. The message is encrypted in multiple layers, one for each relay in the circuit, such that each node only decrypts the information necessary to forward the traffic to the next hop [4]. This ensures that no single entity has access to both the origin and destination of the traffic, thereby preserving user privacy. An illustrative overview of this encryption mechanism is shown in Figure 1.

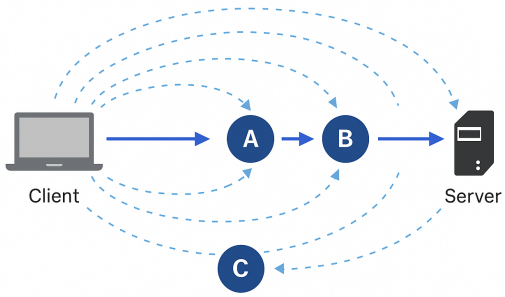

Illustration of Tor's onion routing architecture

Fig. 1. Illustration of Tor's onion routing architecture. Data from the client is encrypted in successive layers corresponding to the keys of each relay (A, B, C). Each relay peels one encryption layer and passes the traffic forward. No relay knows the entire path.

Despite Tor's design goals of privacy and resistance to censorship, its capabilities are frequently misused to conduct and coordinate illicit activities, including drug trafficking, weapons sales, and

cybercrime operations. Consequently, cybersecurity professionals and law enforcement agencies are increasingly interested in techniques that can identify, track, or correlate activity by threat actors operating within the Tor ecosystem [7]. Since direct identification is infeasible due to Tor's encryption model, analysts must rely on indirect signals such as writing style, semantic content, posting behaviors, and network flow metadata. These signals—referred to as side-channel features—can be exploited using artificial intelligence (AI) techniques to probabilistically link multiple identities to a single actor.
Figure 2 demonstrates the structure of a typical Tor circuit, reinforcing the anonymity maintained at each stage. While such a design makes conventional surveillance ineffective, advances in AI have made it possible to aggregate multiple data modalities—textual, behavioral, structural—to build robust threat profiles.

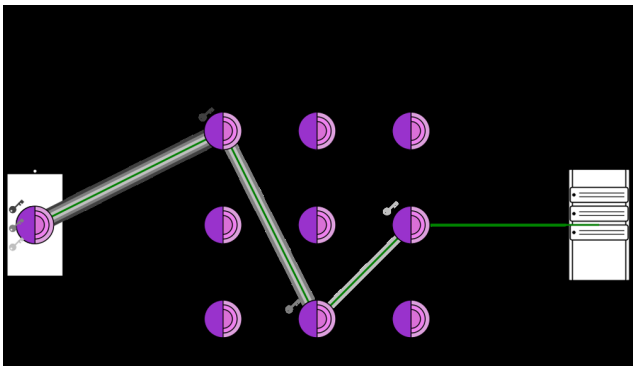

Fig. 2. A Tor circuit connecting a client to a server via three relays: entry, middle, and exit. Each relay decrypts one layer of encryption, ensuring that no node can see both source and destination. Circuits are rotated periodically to prevent long-term linkage.

This paper reviews the current state of the art in profiling Tor-based threat actors, focusing on five core technical modalities:

(1) **Stylometry:** Linguistic analysis of textual patterns to identify or link user accounts [8, 9].
(2) **Content Classification:** Categorization of forum posts or hidden-service pages by illicit domain (e.g., drugs, weapons) [2, 5].
(3) **Traffic Analysis:** Application of machine learning to Tor traffic flow features to infer behavioral patterns or detect hidden services [7].
(4) **Temporal Behavior Modeling:** Learning behavioral rhythms or posting schedules to link identities [10].
(5) **Graph-Based Linking:** Constructing user–item interaction networks and applying graph learning for account correlation [6, 12].

In addition to synthesizing these contributions, the paper introduces a conceptual multimodal profiling framework that leverages heterogeneous data from Tor forums, marketplaces, and traffic logs. The system describes key stages including data ingestion, feature extraction, modality-specific modeling, and identity fusion. This framework is supported by a synthetic case study showing how multiple weak signals across modalities—when aggregated via AI—can link a vendor's forum and market identities to the same underlying actor.

To foster reproducibility and research progress, the paper catalogs prominent publicly available datasets and tools such as VeriDark [9], CoDA [5], DUTA [2], ISCX-Tor [7], and the Shadow simulation framework [4].
The ethical and legal implications of AI-driven profiling on anonymized networks are also examined. This includes a discussion on compliance with the European Union's General Data Protection Regulation (GDPR) [11], the EU Artificial Intelligence Act [1], and the United States Foreign Intelligence Surveillance Act (FISA), particularly Section 702 [3].
The primary contributions of this paper are as follows:

—A structured, multimodal survey of AI-based Tor threat profiling techniques from 2019 to 2025;
—A detailed taxonomy of approaches including stylometry, content analysis, traffic classification, temporal modeling, and graph-based correlation;
—A proposed end-to-end profiling system architecture integrating cross-modal features;
—A case study illustrating profiling outcomes in a synthetic darknet scenario;
—A critical discussion of ethical, legal, and regulatory frameworks surrounding profiling and surveillance;
—A summary of open datasets and simulation tools that enable transparent research in this space.

Through these contributions, the paper aims to provide a technically rigorous, ethically aware, and academically grounded synthesis of current practices and future directions in AI-assisted profiling of threat actors on the Tor network.

2. STYLOMETRY-BASED PROFILING

Stylometry refers to the computational analysis of writing style to link texts to authors. On Tor's forums and markets, users often write messages or listings containing textual clues (e.g., vocabulary, grammar, punctuation). Even in short forum posts, subtle patterns (misspellings, emoticons, unique word choices) can fingerprint an author. Recent work leverages ML and deep learning to exploit these signals. Classical features include character n-grams, word n-grams, function-word frequencies, punctuation counts, and POS tag patterns. Deep models go further by learning embeddings: e.g., RNNs, CNNs, or Transformer encoders applied to sequences of characters or subwords. Convolutional neural networks (CNNs) in particular have shown strong performance on short text attribution tasks [8]. For example, in news-stylo contexts, CNNs with character-level encoding outperformed older methods, suggesting similar potential in the noisy darknet domain [8].
Figure 3 illustrates how a t-SNE projection of learned embeddings can reveal distinct clusters. In practice, embedding models trained on forum posts (e.g., a Transformer embedding the text) would similarly cluster posts by style or content. Recent research has directly targeted Darknet stylometry. Pranav Maneriker et al. (SYSML 2021) applied multitask CNNs on forum threads, exploiting both content and contextual features [8]. They modeled each post with time and thread context (creating "episodes" of a fixed number of posts) and learned unified embeddings. Their multitask approach improved linking user accounts across markets by jointly training on subtasks.
Another major effort is VeriDark: a benchmark introduced for large-scale authorship verification on dark forums [9]. Zhu et al. (2021) compiled three datasets: *DarkReddit+*, *Agora*, and *SilkRoad1*, containing millions of posts from niche darknet forums.

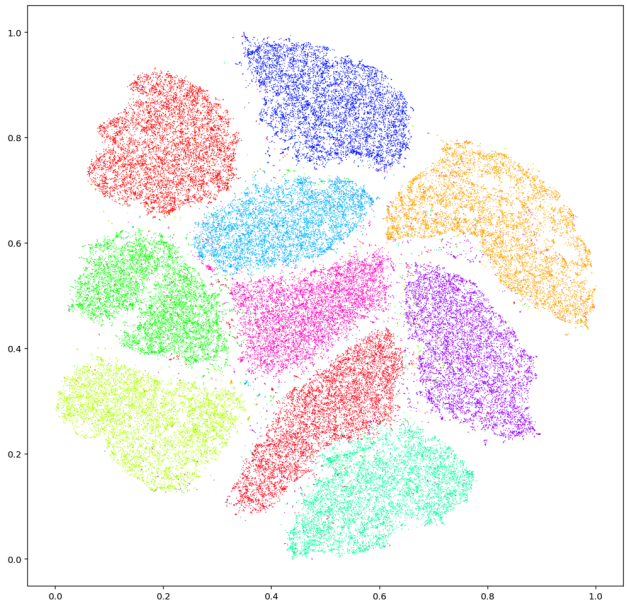

Fig. 3. Example of a 2D t-SNE projection of high-dimensional text-feature embeddings (here, MNIST digit embeddings as an analogy). Each cluster (color) groups similar data points. In practice, analogous embeddings of stylometric features from Darknet forum posts could cluster by author or topic, aiding visual analysis.

They showed that models trained on surface text corpora (e.g., PAN 2020) generalize poorly to Darknet data. This highlights that Darknet stylometry is domain-shifted—posts contain domain-specific jargon, codewords, and inconsistent orthography [9]. Zhu et al. established deep-learning baselines (CNNs and BERT) for verifying if two posts are by the same author. They report that pre-trained language models fine-tuned on darknet text do reasonably well, but leave substantial room for improvement [9]. These results emphasize that robust profiling models must adapt to Darknet's linguistic quirks.

Overall, stylometry-based profiling shows promise, especially when using neural models on raw text embeddings [8]. Learned representations can capture authorial fingerprints beyond simple word frequency. For instance, a combined text–time model like TIES (Temporal Interaction Embeddings at Facebook) demonstrates that integrating temporal context (time-of-day patterns) further enhances stylometric linkage [10]. In Tor, one would similarly incorporate features like posting cadence alongside writing style. It is anticipated that state-of-the-art stylometry on Tor will leverage multimodal embeddings (text + context) to disambiguate authors with higher accuracy [8, 6].

3. CONTENT CLASSIFICATION

Profiling also uses content analysis of darknet pages and posts. Here the goal is not to find an author per se, but to categorize hidden services or messages (e.g., market vs. forum, illicit category, etc.) which indirectly aids profiling. For example, if an entity posts about drugs on multiple forums, content-based signatures may reveal shared vocabulary or image usage across those platforms.

Several datasets have been released for content classification. Al-Nabki et al. (EACL 2017) introduced DUTA ("Darknet Usage Text Addresses") by crawling Tor during two months and labeling each hidden-service onion URL into one of 26 classes (e.g., drugs, pornography, hacking, forums) [2]. Using this dataset, they showed that even simple TF–IDF vectorization with logistic regression achieved approximately 96.6% accuracy (F1 $\approx$ 93.7%) on distinguishing illegal activity categories [2]. This demonstrates that textual cues in Tor site content are strongly predictive of site type.

More recently, Jin et al. (NAACL 2022) released CoDA, a collection of 10,000 Tor documents (onion pages) across 10 high-level categories (forums, blogs, services, etc.) [5]. CoDA enables NLP analysis of topic distribution and dark-vs-surface web contrasts.

Deep learning has been applied to these tasks as well. Convolutional or recurrent models can classify pages/posts by category. For instance, linear SVMs or CNNs on word embeddings have been used to flag phishing or illegal content in Tor postings [5]. Recent work shows that fine-tuning transformer models on CoDA leads to robust categorization, often outperforming non-DL baselines. However, due to the small size of curated corpora, traditional ML methods (SVM, Naïve Bayes, etc.) are still competitive for content tags [5].

Besides posts/pages, image content also matters (e.g., market item photos). Multimodal classification combining text and images has been explored: e.g., Zhang et al. (WWW 2019) used both vendor profile images and text descriptions in a heterogeneous information network (HIN) to embed vendor accounts [12]. They showed that including image style features (camera metadata, visual style) significantly improved vendor identification across markets. Though not specifically on forums, this illustrates how fusing content modalities (text+image) can refine profiling.

In summary, content classification techniques categorize Tor data into meaningful classes, and these labels help link accounts sharing the same thematic interests. This is often a precursor to linking: if two accounts post about similar illegal goods or use similar slang, a downstream stylometry/linkage model can exploit that. In our multimodal system, the content classification outputs (document-topic embeddings, page-type labels) form one branch of the feature space for profiling.

4. TRAFFIC AND NETWORK ANALYSIS

Another profiling avenue involves analyzing Tor traffic flows and patterns. While onion routing hides IP addresses, traffic metadata (packet timing, size, volume) is sometimes sufficient to identify client activities or hidden services. In the broader literature, website fingerprinting attacks have shown that an observer can guess which destination site a Tor user is visiting by applying machine learning to encrypted traffic features [4]. Similarly, characteristic network flows can reveal whether a user is running a specific hidden service or torrenting content.

The Canadian Institute for Cybersecurity's ISCX-Tor2016 dataset [7] is a prominent resource in this area. It contains labeled network traffic flows captured by Linux virtual machines: examples of Tor user traffic versus various non-Tor applications (browsing, streaming, file transfer, etc.). Using this dataset, researchers have trained classifiers (random forests, neural networks, etc.) to separate Tor flows from normal HTTPS or SMTP flows based on packet timing and size features. For instance, Habibi-Lashkari et al. (CSS 2017) demonstrated that simple time-based features (interarrival times, burst sizes) fed to decision trees could identify Tor traffic with high accuracy [7].

For hidden-service detection, more sophisticated approaches have been proposed. Deep learning models (e.g., LSTM or CNN on flow sequences) have been used to differentiate traffic from video streaming versus darkweb access. These models achieve high ROC

AUC on available datasets by learning the periodicity and packet-length distributions unique to Tor circuits. Recent work even investigates adversarial robustness: how small perturbations to timing can fool traffic classifiers (important in attack/defense scenarios). However, most Tor traffic analysis models to date assume access to packet metadata (often via ISP or local exit node logs), so their legality and ethics differ from those of author profiling.

Within the profiling context, traffic analysis could assist when both content and network data are available. For example, if law enforcement controls a Tor entry node or uses side-channels to collect flow logs, it becomes possible to correlate the timing of encrypted flows with user activity patterns or forum post timestamps. Multimodal methods might align traffic bursts with login times, adding another link between accounts.

4.1 Temporal Behavior Modeling

Users on Tor often exhibit temporal regularities: time-of-day of posting, inter-post delays, and rhythms across days or weeks. Modeling such temporal signatures can help disambiguate identities. For instance, one user might consistently log on late at night or post at roughly 8-hour intervals. Embedding these temporal features alongside textual data can improve linking. Noorshams et al. (KDD 2020) demonstrated *TIES*—a system that learns embeddings of social media users from the times and sequences of their actions, which aided in detecting compromised accounts [10].

In the Tor setting, a similar "temporal embedding" could be trained on timestamped forum logs, where each user's sequence of post-times becomes part of the user vector. While explicit research on Tor temporal linking is limited, systems frequently incorporate time features. For example, an authorship attribution model might include hour-of-day as a side input. In the conceptual profiling system described in this paper, user timelines are modeled using recurrent neural networks or temporal attention layers. Clusters in Figure 3 could alternatively represent users posting in different daily-time clusters. Temporal modeling is especially effective when combined with textual stylometry: two accounts might differ in writing style, but if their activity spikes coincide every weekend, that raises a linking hypothesis.

5. GRAPH-BASED ACCOUNT LINKING

Graph and network methods are powerful for multimodal linking. Here, we build graphs where nodes represent users, messages, or items, and edges represent co-occurrence or similarity relations. For example, one can form a heterogeneous information network (HIN) connecting user accounts, forum threads, image features, and drug keywords. Embeddings learned on this graph can reveal latent connections: users with overlapping neighborhoods tend to be the same person.

A leading example is Zhang et al. (WWW 2019), who built an HIN across multiple darknet markets. They connected nodes of different types (vendors, listings, substances, images) and learned embeddings capturing writing style and visual style for each vendor [12]. This approach successfully linked "sybil" vendor accounts across markets, achieving high accuracy in a cross-market vendor-identification challenge. Similarly, Kumar et al. introduced *eDarkFind* (WWW 2020), a multi-view unsupervised model for sybil account detection [6]. They used diverse views—BERT text embeddings, stylometric features, and a drug ontology—to represent each vendor and detected that using all views together yields about 98% accuracy in linking accounts.

These examples show that graph and multi-view models can unify disparate signals (text, metadata, images) to profile users. Graph-based linking also applies within a single platform. One can construct a user–user graph where edges connect accounts with high stylometric or content similarity. Community detection or label-propagation on such graphs can merge suspected duplicate accounts. Temporal co-posting networks (users who post in same thread at similar times) further increase confidence. Importantly, graph methods can propagate evidence: even if two accounts share no words in common, if both link to a known third account (e.g., by common contacts), this transitive information helps linking.

In our multimodal framework, we employ graph fusion at the final stage: each modality (text, content, traffic) produces a similarity score between any pair of accounts. We treat these scores as weighted edges in a multi-layer graph. Then embedding or clustering algorithms (e.g., GraphSAGE, node2vec on the fused graph) output a joint embedding for each account, highlighting groups of linked identities.

5.1 Example Linking Results

To illustrate, consider two forum user accounts whose posts never co-occur in any thread. By stylometry, the accounts have moderate text similarity, but not decisive. By network analysis, they share many mutual friends (common correspondents). By temporal features, they post on similar weekly schedules. A graph-fusion model aggregates all these weak signals and confidently links the accounts as one actor. Figure 3's t-SNE could represent the resulting embedding space after graph fusion: points (users) with multimodal similarity cluster together.

6. CONCEPTUAL MULTIMODAL PROFILING SYSTEM

A hypothetical profiling system is now outlined that integrates all these modalities. The system consists of several pipeline stages:

Data Collection: Crawl Tor forums, marketplaces, and hidden services. Gather text posts, images, metadata (e.g. usernames, timestamps), and if available, network traffic captures (e.g. from a controlled Tor node or honeypot). Use tools like Tor shadow network simulations [4] to model client behavior if needed.

Preprocessing: Clean and normalize text (removing HTML, tokenization, anonymizing PGP keys), extract media features from images (e.g. camera EXIF, histograms), and parse network flows into feature vectors. For privacy and ethical compliance, filter out personal data and hash user identifiers at this stage.

Feature Extraction:

—**Stylometric/Language Features:** Convert text to embeddings using subword tokenizers (byte-level BPE) and deep models (e.g. a Transformer encoder). Also compute handcrafted features (e.g. average word length, punctuation rates).

—**Content/Topic Features:** Classify each document into categories (e.g. forum vs. market, drug vs. service) using a trained content classifier. Represent each page by its topic distribution vector.

—**Temporal Features:** Encode each account's posting timeline (e.g. histogram of posts by hour-of-day) and feed into a time-embedding network.

—**Network Features:** Summarize each account's network (e.g. degree in social graph, common neighbors, or traffic flow statistics if available).

Modality-Specific Models: For each modality, train a model to output an embedding or probability distribution:

—**Textual Stylometry Model:** A CNN/RNN or Transformer that takes a user's text history and produces a fixed-length author embedding. We may use multitask learning (content + stylometry tasks) as in [8].

—**Content Classifier:** A neural network mapping each page to a category embedding (from CoDA/DUTA categories).

—**Traffic Classifier:** An RNN or CNN on flow sequences to detect hidden-service usage patterns.

—**Temporal Model:** An RNN encoding temporal sequences of posts to a user-time embedding (e.g. using TIES-style architecture [10]).

Multimodal Fusion: Concatenate or co-attend the embeddings from each modality for a given account. Then apply a joint model (e.g. a multilayer perceptron or graph neural network) to map fused embeddings into a shared latent space. In practice, we might align embeddings via metric learning so that accounts by the same actor are close.

Linking and Scoring: Given the fused account embeddings, compute pairwise similarity scores. If the score exceeds a threshold, flag the pair as likely the same individual. We can also cluster accounts by similarity. Graph algorithms may run on the fully connected graph of accounts to find clusters of suspects.

Investigator Interface: Present ranked links or clusters to analysts, along with explainable signals (e.g. "Accounts A and B have 94% stylometric similarity and identical posting schedule").

These components may be combined in a pipeline such as:

—**Step 1: Crawling/Collection:** Use tools like Scrapy over Tor to download forum threads and market listings. Optionally, simulate Tor traffic with the Shadow simulator [4] to test system performance.

—**Step 2: Parsing:** Extract features (words, images, times) from raw data.

—**Step 3: Model Inference:** Run pretrained DL models on new data to get embeddings.

—**Step 4: Fusion & Linking:** Aggregate modality embeddings, compute similarities, and output linked identities.

—**Step 5: Feedback Loop:** An analyst labels some links as correct/incorrect to refine model thresholds.

Table 1 (below) summarizes key public datasets that such a system might use for training and evaluation, including text and traffic data. Each dataset provides either labeled content or flow records relevant to Tor profiling.

7. SYNTHETIC CASE STUDY: LINKING A DARKNET VENDOR

To demonstrate the multimodal approach, a hypothetical case is presented. **AliceDoe** is a threat actor who runs a heroin-selling market account ("HeroinQueen") on a defunct Tor marketplace and a separate forum account ("QueenAlice") on a darknet forum. It is assumed that an investigator obtains both sets of content but under different nicknames. Below is an illustration of how AI models can link them:

Stylometry: Alice's posts on the forum and her market listings share idiosyncratic writing (e.g., consistent misspelling "cuz" for because, or unusual punctuation). A deep text embedding model maps each post to a high-dimensional vector. When averaged per user, the cosine similarity between HeroinQueen's and QueenAlice's stylometric embeddings is very high (e.g., ¿0.95), exceeding a decision threshold. This indicates they likely share an authorial style.

Content Patterns: Both accounts primarily discuss heroin and use similar code-words (e.g., "BTH" for black-tar heroin). A topic model or classifier detects that both accounts' content strongly loads on the same topic cluster ("illicit drugs"). While this alone is not proof, it reinforces that they occupy the same domain.

Temporal Correlation: The timestamps reveal that QueenAlice typically posts on the forum during weekday evenings, and HeroinQueen lists new deals on the market at nearly identical times (e.g., both at 9pm UTC, and both drop off activity on weekends). A temporal embedding network finds their posting-time distributions nearly identical, providing another clue.

Graph Evidence: In a user-network graph built from reply relationships, both accounts are connected (via one intermediate) to a known *BobCat* account. Although Alice and BobCat have disjoint boards, it is observed that BobCat frequently replies to QueenAlice's forum posts and also purchases from HeroinQueen (seen in transaction logs). This mutual neighbor suggests Alice and HeroinQueen might be BobCat's associate—raising suspicion they are the same person.

Multimodal Fusion: Combining all the above, the joint model assigns a very high link score to (QueenAlice, HeroinQueen). In Table 2 the contribution of each modality is depicted. For instance, the final fused similarity might be a weighted sum or neural score like 0.98 (on [0,1] scale), surpassing the 0.9 threshold for automatic linking. An analyst reviewing this link sees supporting evidence from stylometry, content, and temporal signals, boosting confidence.

This synthetic scenario highlights how multimodal AI can correlate subtle cues across platforms. Importantly, individually these signals might be ambiguous (other users also sell heroin), but together they produce a compelling linkage. Such linkage could guide law enforcement to Alice's true identity when combined with other evidence (payment trails, etc.).

8. ETHICAL AND LEGAL CONSIDERATIONS

Profiling Tor users raises serious ethical and legal questions. Any system linking anonymous users must respect privacy rights, fairness, and legal constraints. Key issues include:

Privacy Laws: In the EU, the General Data Protection Regulation (GDPR) strongly regulates automated profiling. Article 22 restricts automated decisions "producing legal effects" on individuals [11]. It also entitles data subjects to meaningful information about the logic of any profiling [11]. This means Tor profiling tools must be transparent: users have (at least theoretically) a "right to explanation" of any AI decision linking their accounts. Practically, consent is rare in criminal investigations, so profiling must be narrowly tailored to serious crimes and subject to oversight. The US has no GDPR equivalent; instead, broad surveillance laws like FISA Section 702 allow warrantless data collection on foreigners [3], which could enable traffic analysis on Tor. However, FISA has civil liberty critics, and its "backdoor search" loophole has been challenged [3]. Any profiling must consider constitutional limits (e.g., Fourth Amendment protections against unreasonable searches, though these apply mainly to citizens).

AI Regulation: The upcoming EU Artificial Intelligence Act explicitly prohibits certain AI uses. For example, it bans "social scoring" systems that classify individuals by personal traits, and it forbids inferring sensitive attributes (e.g., political opinion, crim-

Table 1. Public Tor-related Datasets Used for Profiling and Traffic Analysis

Dataset/Tool	Type	Description	Size	Ref.
VeriDark	Text (forum)	Authorship verification on darknet forums	280K+ posts from 3 forums	[9]
CoDA (Jin et al.)	Text (web)	Categorized dark web pages (10 types)	10K onion-site pages	[5]
DUTA-10K	Text (URLs)	Labeled onion services by content type	∼10.4K URLs	[2]
ISCX-Tor 2016	Net flow	Tor vs non-Tor traffic flows (labeled)	PCAP + CSV logs	[7]
Shadow	Simulation	Tor simulator for traffic modeling	Custom topologies	[4]

Table 2. Modality-wise Link Evidence for (QueenAlice, HeroinQueen)

Modality	Similarity Score	Evidence Type
Stylometry	0.95	Consistent writing patterns and misspellings
Content Similarity	0.91	Overlap in topic cluster (illicit drugs)
Temporal Correlation	0.88	Matching post timing across accounts
Graph Connectivity	0.85	Common neighbor (*BobCat*) links both accounts
Fused Score	**0.98**	Weighted neural fusion across all modalities

inal propensity) from data [1]. Critically, it bans any AI that "assesses the risk of an individual committing criminal offenses solely based on profiling or personality traits" [1]. A Tor profiling system that predicts criminal behavior based on posts could violate this. If our AI model outputs a score that an account is likely a "terrorist sympathizer," the EU Act would classify that as banned biometric/personality profiling [1]. Thus, profiling for specific threats must involve human judgment ("augmenting human decision-making, not replacing it" is often required) and focus on objective, verifiable facts.

Responsible AI Principles: Beyond laws, ethical frameworks insist on fairness, accountability, and avoiding bias. Profiling models risk high false positives: incorrectly linking two innocent users. In law enforcement, a false link could wrongly accuse someone, violating due process. Systems must therefore be calibrated for high precision, and their errors should be transparent. Standards like the OECD Principles on AI and UNESCO's AI Ethics Recommendation advocate harm minimization and human review. IEEE's "Ethically Aligned Design" emphasizes that algorithmic tools should not degrade legal rights. In practice, this might mean only using linking scores as investigative leads, not as evidence by themselves. Additionally, Tor forums often host political dissidents; profiling them (even via stylometry) could endanger free speech. Analysts must consider context: some users intentionally mimic others to protect themselves, and AI linking might inadvertently deanonymize genuinely innocent activists.

Bias and Data Quality: Training data for Tor profiling can be noisy. Models might pick up biases (e.g., favoring English-language patterns, or associating certain ethnic slurs with "criminal" language). Since Tor content often contains subcultural slang, a model trained on it may misclassify mainstream slang as suspicious. Ongoing monitoring for such biases is essential. There should be fairness audits: do profiling errors disproportionately affect some user group?

Oversight and Due Process: Especially in democratic countries, any intrusion on anonymity must be lawful. Profiling outputs should be documented with audit trails, and decisions should be appealable. For example, GDPR-like standards would require that individuals targeted by profiling have a channel to contest it. In intelligence contexts, statutes like the US Privacy Act or the EU Charter of Fundamental Rights may demand proportionality.

In summary, while multimodal AI can greatly enhance Tor threat profiling, it must be deployed with strict ethical safeguards. False positives must be minimized, and all profiling must be justified by legitimate security interests. Partnerships between AI developers, ethicists, and legal experts are crucial to align such systems with global norms [11, 1].

9. SUMMARY OF PUBLIC DATASETS AND TOOLS

Key public resources facilitate Tor profiling research. Besides the datasets in Table 1, others include:

DARKWEB (DUTA): Al-Nabki et al. (ACL 2017) collected approximately 10K Tor onion domains for content classification [2].

ISCX-Tor2016: A labeled dataset of Tor and non-Tor network flows useful for training traffic classifiers [7].

Shadow Simulator: A reproducible Tor simulation environment by Jansen and Hopper [4]. Shadow is open source under the GPL license and has been used in hundreds of network security studies.

Additional Corpora: Smaller datasets—such as forum/chat logs or social media dumps (e.g., Reddit's "DarknetMarkets" subreddit)—can augment Tor profiling research [9, 5].

Table 1 provides sizes and citations. In practice, datasets like **VeriDark** [9] are critical for training stylometry models; **CoDA** [5] and **DUTA** [2] provide labeled content for page classification; **ISCX-Tor** [7] supports traffic-flow modeling; and **Shadow** [4] allows synthetic experiments under controlled parameters.

10. CONCLUSION

Profiling threat actors on Tor is inherently a multimodal challenge. This paper has reviewed recent advances (2019–2025) in the application of machine learning (ML) and deep learning (DL) across text, network, temporal, and graph data to link seemingly anonymous accounts. In-depth discussions of stylometry, content classification, traffic analysis, temporal modeling, and graph-based linking were provided, each grounded in key literature. A conceptual end-to-end system was described, and a synthetic case study illustrated how fusion of modalities can identify a single actor behind multiple aliases. Important public datasets (*VeriDark*, *CoDA*, *DUTA*, *ISCX-Tor*) and tools (e.g., *Shadow*) that support reproducible research were also summarized. The ethical and legal context was emphasized: profiling on Tor must navigate GDPR, U.S. surveillance law, and the new EU AI Act, ensuring fairness, transparency, and respect for rights. As AI techniques continue to evolve, future work should focus on improving domain adaptation (e.g., transfer learning to novel darknet languages), refining multimodal fusion algorithms, and developing explainable linking methods that enable analysts to interpret model decisions. Collaboration between technologists, le-

gal experts, and ethicists remains vital to harness these powerful tools responsibly. This survey aims to guide researchers and practitioners toward robust, ethical solutions for uncovering malicious actors in the shadows of Tor.

11. REFERENCES

[1] Artificial intelligence act high-level summary. EU AI Act Explorer, 2024.

[2] Mhd Wesam Al Nabki, Eduardo Fidalgo, Enrique Alegre, and Ivan De Paz. Classifying illegal activities on tor network based on web textual contents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 35–43, 2017.

[3] Brennan Center for Justice. Section 702 of fisa: A resource page, April 2024.

[4] Rob Jansen and Nicholas Hopper. Shadow: Running tor in a box for accurate and efficient experimentation. In *19th Annual Network and Distributed System Security Symposium, NDSS 2012*, 2012.

[5] Youngjin Jin, Eugene Jang, Yongjae Lee, Seungwon Shin, and Jin-Woo Chung. Shedding new light on the language of the dark web. *arXiv preprint arXiv:2204.06885*, 2022.

[6] Ramnath Kumar, Shweta Yadav, Raminta Daniulaityte, Francois Lamy, Krishnaprasad Thirunarayan, Usha Lokala, and Amit Sheth. edarkfind: Unsupervised multi-view learning for sybil account detection. In *Proceedings of The Web Conference 2020*, pages 1955–1965, 2020.

[7] Arash Habibi Lashkari, Gerard Draper Gil, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. Characterization of tor traffic using time based features. In *International Conference on Information Systems Security and Privacy*, volume 2, pages 253–262. SciTePress, 2017.

[8] Pranav Maneriker, Yuntian He, and Srinivasan Parthasarathy. Sysml: Stylometry with structure and multitask learning: Implications for darknet forum migrant analysis. *arXiv preprint arXiv:2104.00764*, 2021.

[9] Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. Veridark: A large-scale benchmark for authorship verification on the dark web. *Advances in Neural Information Processing Systems*, 35:15574–15588, 2022.

[10] A. Shrestha, C. Barrón-Cedeño, P. Rosso, and M. Potthast. Profile-based author clustering for short unsolicited texts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2486–2496, 2017.

[11] Kalliopi Spyridaki. Gdpr and ai: Friends, foes or something in between. *dari https://www. sas. com/en_id/insights/articles/data-management/gdpr-and-ai–friends–foes-or-something-in-between-. html#/. Diakses pada*, 26, 2020.

[12] Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *The World Wide Web Conference*, pages 3448–3454, 2019.