From Dockets to Docker: Engineering Realistic Forensic Scenarios for Onion Service Investigations

Pascal Tippe[a,*], Daniel Spiekermann[b], Adrian Tippe[a]

[a] *FernUniversität in Hagen, Germany*
[b] *Fachhochschule Dortmund, Germany*

Abstract

Digital forensics education relies heavily on static datasets, which often fail to replicate the complexity of investigating live, anonymous infrastructures. This paper bridges the gap between theoretical anonymization technologies and isolated forensic artifacts by presenting a novel training environment grounded in historical case data. We first analyzed court documents from 16 high-profile Onion Service takedowns, revealing that successful investigations leverage multiple information sources and rely primarily on linking information and technical misconfigurations rather than protocol exploits. We operationalized these findings into *DarkMarket*, a containerized simulation that requires trainees to navigate a multi-layered investigative chain ranging from identifying information leaks to executing active scanning. A qualitative pilot study with four active-duty law enforcement digital forensic practitioners validated the environment's realism while highlighting a structural gap: current forensic taxonomies proved insufficient for live Onion Service investigations, which instead required adaptive, hypothesis-driven methodologies. Furthermore, we observed a potential *Technical Halo Effect*, where the robust frontend security led experts to heuristically overestimate the adversary's backend competence. These findings underscore that effectively attributing high-value targets requires expanding forensic competencies to include the controlled, active verification strategies found in Network Investigative Techniques.

Keywords: Digital Forensic Education, Cybercrime Investigation, Tor Onion Services, Darknet Markets

Introduction

Digital forensics education and tool validation rely fundamentally on the quality of training data. While real-world datasets offer the highest fidelity, legal and privacy constraints make their use in academic settings nearly impossible. Consequently, the community relies on synthetic datasets, ranging from standard disk images and mobile data found in the Digital Corpora project (DigitalCorpora.org, 2026; Garfinkel, 2012) to the specialized artifacts within the NIST CFReDS portal (National Institute of Standards and Technology (NIST), 2026). However, as noted by Voigt et al. (2025), defining and achieving *realism* in these synthetic environments remains a persistent challenge. While significant progress has been made in generating realistic static artifacts, a critical gap remains: the investigation of complex, distributed cybercrime infrastructures.

Starting from the inception of the first extensive darknet marketplace, Silk Road (The United States Attorney's Office Southern District of New York, 2013), in 2011 and its shutdown in 2013, new marketplaces have continually followed. With each takedown, new platforms emerge, ensuring that the relevance of such investigations remains high. This was evidenced as recently as 2025, when two large-scale operations highlighted the scope of modern investigative efforts: Operation RapTor (Europol, 2025a), announced in May 2025, resulted in 270 global arrests and the seizure of over EUR 184 million in assets, followed by Operation Deep Sentinel Europol (2025b) in June 2025 dismantling the prolific Archetyp Onion Service after it had accumulated over 600,000 users and processed EUR 250 million in transactions. High-profile targets, particularly darknet marketplaces or cryptocurrency tumblers hosted on Tor Onion Services, cannot be adequately represented by single, static forensic images. Investigating these live, adversarial environments requires a shift from passive artifact analysis to *active engagement*, which requires a hybrid skill set merging digital forensics with penetration testing. Currently, training resources for this domain are polarized: they are either abstract Capture-the-Flag (CTF) challenges that prioritize puzzles over procedural context, or static network captures that fail to simulate the interaction required for attribution.

To bridge this gap, we developed a training environment grounded in the failure patterns of actual criminal operations. We analyzed U.S. federal court documents from 16 high-profile Onion Service takedowns. By isolating the specific investigative angles (e.g., server misconfigurations, leaked metadata) that led to these breakthroughs, we engineered a containerized, live simulation named *DarkMarket*. A primary challenge is to validate the realism. While we developed and introduced the training environment as a task for students in a university course, they lack the operational experience to judge the realism of a simulation. To validate our environment, we conducted

*Corresponding author.
Email addresses: pascal.tippe@fernuni-hagen.de (Pascal Tippe), daniel.spiekermann@fh-dortmund.de (Daniel Spiekermann), adrian.tippe@fernuni-hagen.de (Adrian Tippe)

a pilot study with four active-duty forensic experts from a specialized cybercrime unit. While the primary objective was to validate the training environment, the study yielded unexpected insights into the current state of forensic practice. We noted that required techniques to investigate sophisticated criminal infrastructures are not included in forensic knowledge bases like SOLVE-IT (Hargreaves et al., 2025), a gap that correlated with the different approaches that the experienced practitioners took for investigating the live, anonymous environment.

Research Contributions. This paper contributes a novel analysis of the behavioral and technical gaps in the forensics of Tor Onion Services. Specifically, we address the following research questions:

- **RQ1 (Design Validity):** To what extent can public court documents be utilized to construct forensically authentic training environments for complex Onion Service investigations?
- **RQ2 (Expert Assessment):** How do experienced practitioners assess the fidelity and investigative logic of the synthetic environment?
- **RQ3 (Methodological Gap):** What behavioral patterns emerge when forensic experts conduct active investigations of live anonymized infrastructure?

The remainder of this paper outlines the background and related work, details our empirical analysis of historical court cases, and describes the technical design of the *DarkMarket* environment. Finally, we present the expert evaluation results and discuss the implications for modern forensic taxonomies.

Background and Related Work

The development of competent forensic practitioners and advancing forensic research requires exposure to realistic data and complex investigative scenarios. However, the creation and distribution of such training material is constrained by legal, ethical, and technical challenges (Garfinkel et al., 2009).

Shortage of Real-World Data

Access to high-fidelity training data remains a primary bottleneck in digital forensics education. While real-world data offers high educational value, it is inherently encumbered by privacy regulations and non-disclosure agreements. As noted by Breitinger and Jotterand (2023), the sharing of operational datasets involves significant legal risks, often necessitating complex anonymization procedures that can degrade the forensic utility of the evidence. Grajeda et al. (2017) further show that even when researchers create their own datasets, they rarely share them with the community. This lack of accessibility is compounded by the fact that the field relies heavily on experiment-generated datasets (56.4%), which often fail to capture the intricate noise and interactions found in actual criminal operations, as opposed to real-world data (36.7%). This issue is further exacerbated by a critical deficiency in datasets for emerging infrastructures (e.g., IoT), rendering traditional static repositories insufficient for training in modern, live investigative environments like Onion Services.

Gamification and CTFs

To address the engagement gap in security education, the community has long relied on gamification, specifically Capture-the-Flag competitions. Trickel et al. (2017) highlight that CTFs, particularly *Attack/Defense* formats, are powerful educational tools that provide an engaging, interactive environment for learning both offensive and defensive skills. These exercises facilitate faster understanding of complex theoretical concepts compared to static study. However, a significant "technical barrier to adoption" persists for these sophisticated formats. Trickel et al. (2017) note that while most available competitions are "challenge-based" and relatively simple to run, hosting a full Attack/Defense requires an organizer to spend a "substantial amount of time" meticulously building a secure infrastructure and expertise across multiple domains. Consequently, many educators are forced to rely on simpler formats due to the sheer complexity of the required setup. For the forensics domain, the reduction in scope becomes even more pronounced. Forensic categories in CTFs typically target isolated artifacts (e.g., parsing file headers) rather than simulating the full lifecycle of a lengthy investigation. While students may learn to solve individual puzzles, they rarely encounter the architectural complexity or the real-world constraints characteristic of high-stakes investigations. By bridging the gap between static artifact analysis and live infrastructure in a forensic context, our *DarkMarket* platform seeks to provide the interactive "active engagement" described by Trickel et al. (2017).

Realism and Synthetic Datasets

Parallel to the evolution of live CTFs, the forensic research community has focused on generating high-fidelity synthetic datasets to replace sensitive real-world case files. A central tension in this field is the definition and necessity of *realism*. Göbel et al. (2023) argue that the debate over realism is often dispensable. For specific objectives like tool validation, purpose-built data that ensures the presence of specific traces might be more critical than absolute fidelity to human behavior. In contrast, for investigative training that includes behavioral elements, the coherence of user interactions becomes highly relevant. Voigt et al. (2024) demonstrated that Large Language Models can increase this coherence by generating synthetic background activity that accurately mimics the noise of human device usage. Most recently, Voigt et al. (2025) proposed a quantitative approach to quantify this forensic realism, defining it through an indistinguishability game where a synthetic image is successful if an expert cannot distinguish it from a real-world counterpart. However, their analysis of 25 public disk images reveals that they differ significantly in activity and volume (e.g., number of files) compared to personal, real disk images. This focus on static and artificial artifacts highlights a critical gap in forensic education. While established training formats like CTFs competitions perform well at teaching active penetration

techniques (Trickel et al., 2017), they rarely incorporate the procedural rigor and investigative length required for professional forensic practice. Modern investigations, particularly those involving Tor Onion Services, require a hybrid approach: moving beyond stored data to active engagement and live data collection from volatile environments where physical access is (initially) impossible. The current digital forensics knowledge base *SOLVE-IT*, established by Hargreaves et al. (2025), primarily categorizes established techniques like disk imaging (T1002) and memory acquisition (T1003) (SOLVE-IT-DF, 2026). However, these frameworks currently lack specific coverage for the live techniques required for Onion Service investigations. Our *DarkMarket* platform addresses this deficiency by shifting the focus from the realism of static evidence to the realism of system behavior in a distributed, live network.

Empirical Analysis of Tor Onion Service Deanonymization

To ground the design of forensic training scenarios in operational reality rather than theoretical vulnerabilities, this work adopts the framework established by Tippe and Tippe (2024). In this study, the authors shifted the analytical focus from the theoretical anonymity of the Tor protocol to the practical methods employed by Law Enforcement Agencies to bypass it. By analyzing court documents (e.g., indictments, affidavits, and plea agreements) from 136 U.S. court cases, they established a taxonomy of deanonymization vectors that categorizes successful attacks into distinct domains. The findings reveal that despite the robust cryptographic design of the Tor network, the human aspect, manifested as human inattentiveness and failures, remains the primary vector for identification, significantly outnumbering protocol-level exploits. Our study specifically utilizes a subset of these cases involving Onion Service operators to serve as a foundation for designing high-fidelity training simulations. As noted by Tippe and Tippe (2024), the reliance on federal court records introduces a survivorship bias, as the dataset only includes investigations that resulted in successful prosecution. Consequently, the dataset selects for high-profile targets, such as darknet markets, that warranted significant investigative resources. This selection aligns with the educational objective of this paper: training experts in the complex investigation of sophisticated criminal infrastructure.

Active Engagement and the Precedent of Network Investigative Techniques

In the context of anonymized infrastructure, traditional forensic acquisition methods often fail. As noted in the *Freedom Hosting* investigation (U.S. District Court for the District of Maryland, 2013), standard IP identification techniques are rendered ineffective by the Tor network, which masks a target's location by routing traffic through distributed nodes. While NITs are often associated with client-side exploits, they represent a broader shift in forensic practice: the necessity of interacting with a live, remote system to bypass anonymization layers. To bridge this attribution gap, law enforcement agencies have shifted toward *active forensic engagement*, most notably through the deployment of Network Investigative Techniques (NITs). While this shift toward active engagement shares technical similarities with penetration testing, it operates under vastly different legal and evidentiary constraints. Technologically, a NIT functions by delivering specific computer instructions to a target system to trigger the transmission of "certain information to a computer controlled by or known to the government" (U.S. District Court for the District of Maryland, 2013). As analyzed by Ghappour (2017), deploying a NIT is legally classified as a remote search and seizure, effectively executing government code on a computer whose physical location is often unknown. This introduces severe jurisdictional and procedural requirements, tightly regulated under frameworks like Federal Rule of Criminal Procedure 41 (U.S. Government, 2026). Furthermore, because these active techniques inherently alter the target system's state to extract volatile data, they face intense evidentiary scrutiny. Garcha (2018) highlights the constitutional tension during legal discovery: defendants possess the right to examine the technological vulnerabilities and tools used against them to ensure evidence was not fabricated or corrupted. To survive this scrutiny, the data capture must be strictly scoped. For instance, in the *Freedom Hosting* case, the extraction was forensically constrained to five specific volatile artifacts: the actual IP address, the MAC address, the Host Name, the operating system version, and a unique transaction identifier (U.S. District Court for the District of Maryland, 2013). Consequently, active engagement cannot be treated as standard ethical hacking. It requires a highly specialized hybrid skill set, combining technical exploitation with the immutable, cryptographically verifiable logging of all interactions necessary to preserve the chain of custody. For the investigation of Onion Services, these active techniques translate to identifying server-side misconfigurations or weak authentication. Just as a NIT precisely extracts linking evidence from a client, techniques such as targeted port scanning and service-level authentication testing are required to unmask the true location of an Onion Service. Our training task mirrors this approach, requiring students to engage with live infrastructure to link evidence necessary for attribution.

Methodology

Empirical Extraction (Court Case Analysis)

To extract relevant techniques for the training environment and ensure that these reflect the operational reality of high-stakes cybercrime investigations, we grounded our design in a qualitative analysis of historical federal prosecutions. We utilized the case dataset established by Tippe and Tippe (2024), specifically isolating the subset of sixteen cases classified as Onion Service platforms. This specific subset was identified in the prior work via Department of Justice announcements. We selected these targets (comprising darknet marketplaces, cryptocurrency mixers, child sexual abuse material distribution hubs, and hosting providers) because they require complex, multi-layered infrastructures. Unlike transient vendor accounts, these platforms require centralized administration, payment processing, and server maintenance. This complexity in-

creases the attack surface, requiring the convergence of multiple forensic disciplines (e.g., network analysis) that we aim to teach. To translate legal proceedings into training objectives, we employed a qualitative coding process utilizing human researchers, referred to hereafter as *coders*. We utilized two specialized coding pairs (with one overlapping member) for the distinct analysis phases of this study. For the court case analysis, two researchers with expertise in IT security and U.S. court documents conducted the evaluation. The primary coder retrieved the original court dockets (mainly indictments, criminal complaints, and search warrant affidavits) for the selected Onion Services. The primary coder then manually extracted text fragments describing investigative methods, utilizing the established codebook from Tippe and Tippe (2024) as a foundational guide to identify and categorize the quotes. Due to strict page limits, the original codebook has been omitted from the main text but is explicitly provided in our public repository to ensure reproducibility. A list of the 16 analyzed cases can be found in Appendix B. Following the extraction phase, both coders independently clustered these extracted quotes based on a simulatability criterion, specifically, determining whether the described technique relies on a digital artifact that can be simulated without human intervention. After this independent analysis, the coders convened in a live session to merge the clusters and formulate the hypotheses that generated the four artifact categories (subsequently presented in Table 1).

Scenario Engineering and Adaptation

Directly replicating historical case files presents a challenge. Technical developments quickly render technical vulnerabilities obsolete. A specific software exploit used to take down a market in 2013 is likely irrelevant to a student in 2025. To address this, we developed a process of vulnerability abstraction. Instead of reproducing the exact software installations and configurations found in the court documents (e.g., a Captcha leaking an IP address), we extracted the underlying logical failure that enabled the deanonymization. We then re-implemented these logic errors within a modern, containerized technology stack. For example, instead of leaving admin accounts on the website with weak passwords, we configure the SSH server with weak credentials. This ensures that students learn to identify persistent misconfiguration patterns, such as broken access controls, rather than exactly mirroring previously used techniques. We also applied narrative constraints to the extracted vectors. Real-world investigations often involve non-technical antecedents, such as physical surveillance or long-term undercover operations, which cannot be simulated in a compact technical lab. Therefore, we classified vectors like *Physical Operations* and *Active Infiltration* as narrative antecedents. In the training scenario, the outcomes of these phases (e.g., acquired credentials) are provided as the starting state, allowing the trainee to immediately focus on the digital forensic analysis of the live infrastructure. Finally, to mirror the complexity of large-scale platforms serving hundreds of thousands of users, we embedded multiple interdependent vulnerabilities into the environment rather than presenting a collection of isolated puzzles. This design forces the trainee to connect separate artifacts, which simulates the logic of longer investigations. Just as in a real investigation where minor observations monitored over time eventually yield a breakthrough, the simulation requires the user to combine subtle behavioral hints (e.g., dynamic UI changes) with infrastructure flaws to progress through the network.

Evaluation Methodology

The evaluation of the training task was designed as a foundational qualitative pilot study to assess the practical utility and technical realism of the environment from the perspective of active digital forensic practitioners. We collaborated with a specialized cybercrime unit within a German police department. The study involved four forensic experts ($N = 4$) who participated in a controlled, three-hour hands-on session followed by semi-structured interviews. This expert-centric approach was chosen to gain deep qualitative insights into the transferability of the training tasks to real-world investigations, prioritizing professional feedback over a larger, non-specialist sample size.

Expert Evaluation Protocol

To validate that the engineered artifacts accurately reflect the challenges of real-world investigations, the study followed a structured five-phase process:

1. **Initial Assessment:** Participants first completed a pre-study questionnaire to establish their professional background, years of experience in digital forensics, and prior familiarity with Tor and Darknet investigations. The individual baseline established in this survey was used to structure and adapt relevant questions for the final qualitative interview.
2. **Instruction and Briefing:** Participants were presented with the scenario via a start email that provided the legal and operational context of the investigation. They were briefed on the technical scope of the environment but received no specific guidance regarding the integrated vulnerabilities.
3. **Hands-on Investigation:** The experts were given three hours to independently and asynchronously interact with a shared live instance of the environment. This shared setup allowed us to observe how concurrent, uncoordinated actions manifest forensically. The three-hour duration was intentionally designed to simulate the initial triage phase of an investigation, forcing participants to prioritize leads rather than allowing for a comprehensive exhaustion of all forensic possibilities. During this time, they were encouraged to use their standard investigative toolsets to collect court-usable evidence and identify the administrators. Participants were explicitly instructed to maintain individual notes during the investigation to support their memories during the subsequent interview.
4. **Post-Investigation Questionnaire:** Immediately after completing the practical task, experts filled out a questionnaire to capture immediate impressions regarding difficulty and engagement. One coder reviewed these responses to identify specific points of friction and expectations for each participant. Insights from these responses

were used to tailor specific probe questions for the subsequent group discussion, rather than relying solely on a generic script.

5. **Post-Task Interview:** Finally, a semi-structured group interview with all participants was conducted. This collaborative format allowed peer-to-peer discourse, allowing participants to build upon each other's technical experiences and explore insights that might not emerge individually. The interview guide, informed by the preceding surveys, was designed to capture impressions of the task's complexity, the realism of the implemented forensic steps, and specific insights or skills gained.

Data Analysis Method

The qualitative evaluation relied on semi-structured interviews conducted by a single researcher. We utilized a guiding set of open-ended questions to ensure a standardized thematic scope while providing participants the flexibility to share detailed personal experiences and nuanced technical observations. To ensure academic rigor and transparent methodological reporting, the interview transcripts were subjected to an independent coding process by our second coding pair. Both coders possess a background in digital forensics, with one coder contributing several years of hands-on experience in law enforcement forensic investigations. First, both coders independently reviewed the transcripts to identify key topics and mark relevant quotes. Following this independent phase, the coders convened in a live session to compare the identified themes and discuss the marked quotes. During this collaborative process, hypotheses were built for the themes, which directly resulted in our main findings. Because the live reconciliation sessions for both the court case analysis and the interview analysis successfully resolved all analytical disagreements, the use of a third-party arbitrator or the calculation of a quantitative Inter-Rater Reliability (IRR) score was rendered unnecessary.

Analysis of Historical Investigative Angles

Before detailing the technical implementation of our training environment, it is necessary to establish the operational reality it aims to simulate. Deanonymization is rarely achieved by breaking the Tor protocol. Instead, attribution is typically achieved through a convergence of digital footprints and traditional investigative techniques. We categorized these vectors into four primary domains (see Table 1).

The most prevalent failure, observed in 100% of the analyzed cases (16/16), was the leakage of linking information. This category encompasses data that bridges the anonymous infrastructure with the operator's real-world identity. The most critical vector was access timing correlation. In three instances, suspects were identified by correlating connection times to a surveilled server to their home network. Another example is SSH keys that contain metadata (like usernames) or user agent strings that identify a browser combination that can be tracked across platforms. Another classic angle was username correlations across platforms in five cases. In all cases, associated accounts were searched, with email accounts being a primary target. More nuanced technical linkage occurred in cases involving cryptographic artifacts. While the PGP keys themselves were often cryptographically sound, their usage provided additional evidence in five cases. For example, an SSH key found on a seized darknet server contained a comment field with the name of the suspect's computer. Contrary to the perception that deanonymization requires breaking the Tor protocol, our results indicate that protocol exploits are statistical outliers (1 case). Instead, technical attribution was primarily achieved through configuration errors (6 cases). The dominant technical issue was Information Leakage, where the application layer inadvertently exposed backend details. This primarily manifested in either leaking the IP address or being reachable over the IP address. In one case, automated registration emails contained a field populated with the administrator's personal clearnet email address and in another instance an administrative password was insufficiently secured. These findings confirm that the primary threat to Onion Services is not the cryptographic weakness of the network, but the complexity of maintaining perfect isolation between the application layer and the underlying operating system.

Training System Design

To bridge the gap between theoretical vulnerabilities and practical forensic application, we developed a containerized training environment that mimics the complex infrastructure of a high-value Darknet Market named *DarkMarket*. The system architecture, orchestrated via Docker Compose, utilizes a strict network topology: only the Tor gateway container maintains external network access, while all backend services (nginx reverse proxy, Flask web application, and SSH) reside in an isolated internal network reachable only through the gateway. This multi-layered design directly reflects the Onion Service platforms analyzed in our methodology, which typically feature centralized coordination points and distinct separation between frontend and backend assets. By utilizing a modular containerized approach, we were able to precisely inject the specific failure categories identified in our court case analysis (Table 1) into a functional 5-step forensic scenario.

Scenario Initialization and Constraints

The scenario operationalizes the *active infiltration* vector as a narrative antecedent. The trainee assumes the role of an investigator who has received credentials for a market moderation portal from a cooperating lower-level vendor. This setup allows the simulation to bypass the lengthy phase of undercover work and focus immediately on the analysis of the captured digital environment. To enforce forensic methodology over automated exploitation, we implemented aggressive anti-analysis mechanisms. Querying non-existent resources (404 errors) triggers an immediate ban of the user's Tor circuit, and the login panel employs exponential backoff for failed attempts. These constraints mirror our methodological finding that successful investigations rarely rely on noisy protocol exploits, but rather on the patient

Table 1: Prevalence of Investigative Vectors in Analyzed Onion Service Platforms (N=16)

Category	Specific Vector / Artifact	Frequency
Linking Information (OpSec)	*Total Instances*	**16/16**
	Associated Accounts (Email, Hosting, Crypto Exchanges)	16
	Metadata Correlation (IP Logs, User Agents, Access Times)	11
	Public Open Source Intelligence (StackOverflow, Forums)	9
	Cryptographic Keys (PGP, SSH)	5
Active Infiltration	*Total Instances*	**16/16**
	Undercover Buys / Monitoring	16
	Cryptocurrency Tracing (Blockchain Analysis)	11
Physical Operations	*Total Instances*	**13/16**
	Physical Search (Raids/Seizures)	10
	Physical Surveillance (Suspect/Package Tracking)	4
Technical Errors	*Total Instances*	**7/16**
	Misconfigurations (IP Leak, Clearnet Reachability)	6
	Tor Protocol Exploits	1

accumulation of multi-modal evidence. Finally, to ensure visual and immersive fidelity, we visited contemporary darknet marketplaces active in 2025 and selectively incorporated characteristic design choices (e.g., captcha implementations) to mirror the look and feel.

The Forensic Solution Path

The successful attribution of the platform administrator requires the trainee to navigate a five-step investigative chain. This path was designed to force the correlation of technical network artifacts with non-technical behavioral observations.

1. **Network Enumeration and Header Analysis:** Upon accessing the portal (protected by a CAPTCHA to simulate realism), the trainee identifies a list of alternative mirror URLs. One specific mirror is labeled as *under DDoS attack* and configured to load at a restricted rate of 25 bytes per second. While the slow connection mimics a denial-of-service condition, an inspection of the HTTP headers reveals a leaked `X-Load-Balancer` header. This artifact identifies a specific internal hostname, hinting at the existence of a backend nginx server that is not correctly removing internal routing data.
2. **Behavioral Observation (The *Lurking* Phase):** To simulate the lengthy nature of darknet surveillance, we engineered a dynamic background mechanism that updates the interface only after sustained interaction. If the investigator maintains an active session for ten minutes or interacts with the moderator messaging system, the *last login* field on the password panel updates to display a specific, static username. This step tests the investigator's patience and ability to correlate a dynamic interface change (mimicking an administrator checking messages) with the backend infrastructure.
3. **Correlation and Access (The *Hotfix* Error):** The trainee must cross-reference the discovered username with the DDoS mirror found in Step 1. A port scan of this specific mirror reveals that while port 80 is throttled, port 22 (SSH) is open and responsive. This simulates a panic misconfiguration where an administrator, attempting to hotfix the server during an attack, hastily applied firewall rules to the web interface but neglected to secure the administrative port. Using the username discovered in Step 2 and a trivial password, the investigator attempts authentication against the SSH service. The choice of a trivial password mirrors a real-world case of weak credentials found in our dataset. Furthermore, it is a necessary design constraint as the high latency of the Tor network renders complex brute-force attacks impractical for a training timeline. In a forensic context, this step simulates the execution of a Network Investigative Technique based on gathered intelligence rather than arbitrary exploitation.
4. **Internal Pivoting and Tunneling:** Reflecting the multi-layered architecture of real-world platforms, the compromised SSH service is restricted to port forwarding only and has no shell access. Because the backend containers are network-isolated from the public internet, the investigator must establish an SSH tunnel to scan the internal network. This leads to the discovery of an unmapped nginx server listening exclusively on the internal management interface, invisible to the Tor network.
5. **Cryptographic Linking:** The internal server hosts a personal blog belonging to a clearnet persona. While the blog provides soft OSINT leads (e.g., geographic hints), the definitive forensic link is established through a PGP chain of trust. The blog owner's PGP key is found to have signed the Darkmarket administrator's public key (extracted from the moderation panel). Crucially, the blog owner's key is itself signed by the trusted key provided to the participants in the initial briefing email. This final step validates the *linking information* vector, requiring the trainee to synthesize technical access with cryptographic identity analysis

to prove attribution.

By structuring the solution to require these five distinct steps, the system ensures that participants cannot succeed through technical scanning alone. Instead, they must mirror the workflow of a real investigation: obtaining initial access, observing human behavior, identifying a specific configuration error, navigating through the infrastructure, and finally establishing identity through cryptographic hygiene failures.

Evaluation

Participant Demographics and Initial Expectations

The cohort consisted of active-duty professionals from a specialized cybercrime unit, possessing 2 to 4 years of operational experience. Their backgrounds spanned network forensics, malware analysis, and system forensics. While the sample size ($N = 4$) is constrained by the operational availability of active police personnel, the qualitative depth of their expertise offers high-fidelity insights that larger student cohorts cannot provide. Their self-assessed skills, detailed in Table 2, reveal a distinct profile: participants reported high command-line proficiency (Avg=4.5) and web server analysis skills, but lower confidence regarding Onion Service specifics (Avg=2.25). Regarding prior training, all participants had experience with live system analysis, though only one had previously participated in CTF exercises. In the pre-task open-ended survey regarding expectations, two participants emphasized that realistic training must be based on real cases. Others requested "appropriate complexity" and "extensive possibilities to interact with the systems". Notably, two participants specifically requested the "integration of diverse defensive mechanisms" to simulate an authentic environment.

Table 2: Participant Self-Assessment of Forensic Skills ($N = 4$). Shading intensity corresponds to proficiency level (1: None, 3: Medium, 5: Excellent).

Skill Domain	P1	P2	P3	P4	Avg.
Tor & Onion Service Fundamentals	2	2	5	3	3.0
Onion Service Investigation	2	2	3	2	2.25
Web Server Analysis	5	5	3	2	3.75
Web Vulnerability Discovery	3	1	2	2	2.0
Command-line Proficiency	5	5	5	4	4.75
Password Guessing	2	3	3	1	2.25
Forensic Crypto-Artifact Extraction	3	4	3	2	3.0

Note: Data is provided to profile the expert cohort for qualitative context, not for statistical inference.

Immediate Post-Task Perception

The quantitative feedback (Table 3) indicates a high degree of perceived realism (Avg=4.75) and user motivation (Avg=4.5), with the difficulty landing close to the optimal middle ground (Avg=3.5). In the immediate free-text responses, participants provided initial critiques of the investigative constraints. Two participants noted inconsistencies regarding the legal permissions described in the briefing, while another highlighted the lack of collaboration, and also stated that the task was "very focused on pentesting topics". Furthermore, two participants noted "very sensitive security mechanisms", and despite the typical latency of the Tor network, only one expert reported minor irritations regarding the system stability. When comparing this environment to previous training experiences, all participants judged the task as more instructive and sophisticated. Three participants rated it as more time-intensive and more motivating than standard exercises. One participant concluded that a case of this magnitude would realistically require a longer investigation and that in practice "the amount of accumulated information is reflected" in the available material. Regarding personal learning outcomes, two participants identified a need to improve their own penetration testing skills. Another respondent pointed out that Onion Services are technically closer to traditional web services than expected, stating: "don't be deterred".

Table 3: Post-Task Participant Feedback ($N = 4$). For Realism and Motivation, darker shading indicates higher scores. For Difficulty, darker shading indicates the optimal level (3).

Metric	P1	P2	P3	P4	Avg.
Task Difficulty (Target: 3)	2	3	4	5	3.5
Perceived Realism	5	4	5	5	4.75
User Motivation	4	5	5	4	4.5

Note: Task Difficulty follows a non-linear color scale where the target value (3) is shaded darkest to represent an optimal balance between *too easy* and *too demanding*. Data is provided to profile the expert cohort for qualitative context, not for statistical inference.

Qualitative Interview Results

Following the hands-on session, we conducted semi-structured interviews to capture the experts' decision-making processes and their assessment of the environment's fidelity. The analysis of the transcripts yielded four primary themes: investigative strategy, operational realism, defensive friction, and expert bias regarding adversary competence.

Investigative Strategy and Tooling (Addressing RQ3)

Participants initially adopted a passive reconnaissance approach, prioritizing non-invasive observation. They reported starting by "getting an overview" of the platform, systematically viewing all subpages to find "something conspicuous here" and to understand "how the whole [website] functions in the background." This phase involved analyzing deobfuscated source code and manually interacting with the frontend. However, following this initial mainly passive approach, the techniques became increasingly interactive. As the experts shifted to active engagement, their investigative paths diverged significantly. Unlike the standardized workflows typical of disk image analysis, the experts developed distinct hypotheses leading to the investigation of widely disparate components of the environment. This divergence highlights that the complexity of live, distributed environments breaks standardized processes often sufficient for static forensics, requiring a more adaptive, hypothesis-driven methodology.

A significant theme was the ambiguity regarding the rules of engagement. One participant noted uncertainty about whether the objective was "about finding a nickname, or directly attacking the thing." Consequently, the transition to active measures was cautious, reflecting adherence to standard forensic preservation protocols. To interact with the environment, experts employed a mix of standard and specialized tools. Participant 4 reported using `wget` (Free Software Foundation, Inc, 2026) to locally preserve web content, a standard forensic approach, while Participant 1 utilized the `Burp Suite` (PortSwigger Ltd., 2026) to intercept and modify HTTP requests.

Success in the specific identification task varied. Participant 3, who self-reported the highest familiarity with Onion Services, successfully identified the administrator username by observing a dynamic change in the last login field. He validated this finding by attempting a login with the discovered username versus a random string to distinguish the error messages. The participant suspected that a "script ran, that went there every half an hour" to simulate user activity. Another participant also noted the last login field but did not deduce an investigative lead from it. Interestingly, because all experts worked on the same live instance, two participants noted last user attempts in the last login field that hinted at exploitation, correctly deducing these were traces left by their colleagues.

Operational Realism and Collaboration (Addressing RQ2)

The experts rated the narrative plausibility of the scenario highly, with one stating that if "I had gotten that [task] as a student, that would be cool." However, a distinction emerged between *case* realism and *procedural* realism. The critique focused heavily on the lack of organizational context. Participants emphasized that "it is just relatively unrealistic that such a task is done by one of us alone." In a real-world setting, the experts noted they "wouldn't start hacking a page on call without actually having a court order copy in the email". They highlighted that an investigation of this magnitude would involve distributed teamwork, where departments "depending on the importance of the investigative procedure, then try to perhaps get the experts on board." This collaborative aspect was mirrored during the interview itself: when one expert shared a technique (e.g., source code deobfuscation), others often asked for clarification, demonstrating the natural tendency toward knowledge sharing that the solitary training format precluded.

Friction of Defensive Mechanisms (Addressing RQ2)

The implementation of security defenses generated significant discussion regarding the balance between realism and frustration. Opinions on the strictness of the security measures were divided. Participant 3 argued that "if someone runs such a site, then they don't want to tell the attacker, that he has just done something forbidden," suggesting that immediate blocking was unrealistic for a service that needs to retain customers. Participant 2 countered, finding the mechanisms "realistic." Ultimately, a consensus emerged distinguishing between application logic and network logic: blocking a user after "tokens are faked" was deemed valid, whereas an immediate ban for requesting a non-existent resource (404) was considered "too much." This defensive pressure might have influenced the experts' behavior since multiple participants reported hesitation to interact with form fields or send messages (e.g "better not raise too much attention").

Artifact Assessment and Technical Halo Effect (Addressing RQ1, RQ2, and RQ3)

When comparing the training artifacts to those found in CTF challenges, participants provided a granular assessment of technical realism. The PGP linkage, the internal blog, and the dynamic last login leakage were consistently rated as having high to very high realism. In contrast, the presence of an open SSH port was rated low, and the weak SSH password was rated very low. Drawing on their experience, the experts inferred that a sophisticated frontend implied a secured backend. While this heuristic might be efficient in routine investigations, it proved misleading in the context of darknet markets. This assessment revealed a critical cognitive bias that we term the *Technical Halo Effect*: the experts overestimated the competence of the adversary's backend based on the sophistication of the frontend. Participant 3, drawing on "criminalistic experience," stated: "If they run a darknet marketplace..., it is very unlikely that they take a password from the word list of standard password lists." Similarly, Participant 1 noted that because the passwords provided in the briefing "looked relatively good," he did not expect "an open SSH [server] to be this badly secured". This assumption of competence directly impacted the investigation. As Participant 1 admitted: "If the site is already so sensitive to scan attacks, then I wouldn't get very far with the password scan." Consequently, participants did not scan for open ports or expect the SSH server to be this insufficiently secured, despite the technical reality that configuration errors often coexist with sophisticated code.

Discussion

Main Findings

The Gap in Remote Acquisition Taxonomies. While the field of digital forensics has matured to include well-defined processes for *Live Data Collection* and *Network Forensics*, our evaluation suggests these models do not fully address the constraints of Onion Service investigations. Current standardized taxonomies, such as the SOLVE-IT Knowledge Base (Hargreaves et al., 2025), categorize "Live Data Collection" (T1016) primarily as the execution of commands (e.g., `pslist`, `netstat`) on a local, accessible system to capture volatile state. Similarly, data acquisition is largely framed around having physical or logical access, such as Disk Imaging (T1002) or Memory Imaging (T1003). In the context of an anonymous Darknet service, investigators inherently lack the physical access required for these traditional acquisition methods. SOLVE-IT currently offers no dedicated techniques under its live forensics umbrella that capture the remote, active engagement required to deanonymize an Onion Service when physical access is impossible. This limitation stems partly from SOLVE-IT's structural inspiration: operational frameworks like MITRE

ATT&CK (Strom et al., 2020). While ATT&CK comprehensively profiles the technical objectives of an adversary, it models offensive behavior aimed at exploitation against standard enterprise environments, such as generating profit through data theft, ransomware, or denial of service. Adversaries typically do not care about identifying individual operators, nor do they concern themselves with forensic soundness, the chain of custody, or strict evidentiary standards. Information security testing guidelines, which encompass penetration testing workflows (such as NIST SP 800-115 (Scarfone et al., 2008)), emulate these adversary objectives to assess enterprise security. To their credit, NIST SP 800-115 correctly emphasizes the necessity of maintaining an objective record, recommending mechanisms for "automatically logging assessor actions" and recording every test "with a timestamp, test type, tool used, commands" to counter accusations of negative operational impact. While these logging principles are highly applicable, the overarching goal of ethical hacking differs fundamentally from criminal investigations. A successful investigative action against criminal infrastructure might not constitute a traditional corporate security breach (e.g., correlating an SSH banner or discovering an endpoint leaking metadata to identify a clearnet persona), but it is the primary goal for attribution. More critically, standard penetration testing workflows risk unrecorded system state alterations that can render digital evidence legally inadmissible. The deployment of active measures, such as NITs, bridges this acquisition gap. Importantly, these remote techniques are not a replacement for established forensics, but rather the necessary prerequisite. By unmasking a server's true IP address or bypassing an anonymization layer, a NIT enables law enforcement to execute physical search warrants and subsequently perform standard disk and memory imaging. However, this active engagement introduces complexities far beyond traditional vulnerability assessments. Legal and technical analyses by Ghappour (2017) and Garcha (2018) contextualize how these techniques function legally as remote searches and seizures, requiring rigorous warrant scoping and strict disclosure protocols.

Therefore, we argue that the investigation of live, anonymized infrastructure cannot be siloed or delegated to traditional corporate penetration testers. Instead, active remote engagement requires a hybrid skill set. Our findings suggest the necessity of specialized training for select personnel within cybercrime, ensuring investigators can seamlessly integrate the technical exploitation skills of a penetration tester with the criminological rigor and legal domain knowledge required to produce court-admissible evidence.

Active Forensics and Soundness. Our analysis of the Freedom Hosting affidavit (U.S. District Court for the District of Maryland, 2013) highlights a necessary evolution in the definition of forensic soundness. Critics might argue that NITs, which involve injecting code into a client system, violate the traditional forensic principle of non-interference. However, we argue that NITs should be classified under the domain of live forensics. This classification aligns with foundational evidence collection guidelines, such as RFC 3227 (Brezinski and Killalea, 2002), which mandate prioritizing the preservation of highly volatile data, including active routing states and network connections. Just as capturing a RAM image from a running system inevitably alters the system state to preserve volatile evidence, a NIT alters the client's state to preserve the target IP address before the session terminates. The forensic rigor in a NIT lies not in maintaining absolute non-interference, but in the reliability of the process and the minimization of scope. The analyzed warrant demonstrates this rigor by strictly limiting data exfiltration and requiring the evidence to be sent to a controlled environment. Therefore, future forensic training for Onion Service investigations should include these active acquisition workflows. These should not be seen as a hacking attack, but as a court-authorized, distinct forensic acquisition step comparable to executing a remote memory dump. To maintain this legal standard, standard offensive tools are often insufficient. Instead, tools used in this domain must provide transparent protocol interception capable of generating immutable, timestamped, and hashable session logs (e.g., utilizing `mitmproxy` (Cortesi et al., 2010–) configured for strict forensic capture). This ensures that the highly volatile remote evidence actively *generated* during the interaction can be authenticated and reproduced in court.

Heuristics and the Technical Halo Effect. A significant finding regarding expert decision-making was the dissonance between the participants' expectations and the technical reality of the target. Despite empirical data showing that 37.5% of analyzed Onion Service platforms suffered from fundamental technical misconfigurations (Table 1), participants rated the presence of basic vulnerabilities (such as weak SSH credentials) as having low realism. This confirms the Technical Halo Effect identified in our evaluation, where high frontend sophistication effectively masks backend vulnerability. This cognitive bias led experts to omit standard enumeration steps, such as port scanning, despite the potential for configuration errors. This finding underscores the importance of scenarios that force investigators to confront the messy reality of cybercrime, explicitly countering the heuristic assumption of complexity.

Collaborative Forensics and Knowledge Distribution. While the technical realism of the environment was rated highly, the solitary nature of the task was identified as a procedural weakness. As the evaluation showed, knowledge and skill sets are often distributed among team members that are required to deanonymize a platform. In the court cases, this spans from crypto-currency analysis, web application security, and classic Linux forensics which would require extensive specialization in multiple fields. This was evidenced during the interviews, where participants naturally engaged in knowledge sharing, clarifying techniques like source code deobfuscation with one another. This suggests that the single investigator model is misaligned for this domain. Future iterations should support collaborative work, allowing multiple trainees to investigate the same target simultaneously.

Temporal Constraints vs. Operational Reality. The evaluation imposed a strict three-hour time limit on the participants. In operational reality, the investigation of a complex Onion Service is a longitudinal process spanning months of surveillance and correlation. Consequently, the limited success rate in fully attributing the administrator should be viewed through

the lens of this artificial constraint. The study effectively measured the experts' ability to *identify* valid investigative avenues rather than their ability to exhaustively solve the case. Future work could assess whether the observed Technical Halo Effect persists when investigators are given ample time to verify their initial assumptions.

Ethical Considerations

Given the involvement of active law enforcement personnel and the discussion of specific investigative methodologies, this study was conducted under a *framework of cooperation* that replaced standard layperson consent forms with strict institutional protocols regarding data sovereignty and publication vetting. Adhering to the principle of data minimization, data collection was strictly limited to information essential for answering the defined research questions. All survey forms and interview guidelines were explicitly designed to avoid capturing sensitive investigative tactics, focusing instead on generalizable strategies. To protect participant privacy and ensure data hygiene, interviews were recorded solely on an official service device of the police department and locally transcribed immediately following the sessions. These transcripts were reviewed and approved by the department before being released to the researchers, ensuring that no raw audio data left the police department. Access to the transcripts was restricted to the two interview coders. Finally, to ensure the qualitative evidence presented in the evaluation reflects the experts' views without violating the trust necessary for such a partnership, no study results, excerpts, or manuscripts were submitted for publication without a formal review and release process by the participating agency.

Limitations

Evaluation Sample Size and Qualitative Focus. The evaluation was conducted as a foundational qualitative pilot study ($N = 4$). While the expert status of the participants provided deep insights into operational reality, the small sample size limits its statistical generalizability. Future work is required to validate whether the observed behavioral patterns, such as the Technical Halo Effect, persist across a broader forensic cohort.

Dataset Bias and Evolving Strategies. The error taxonomy was derived from a dataset of 16 U.S. court cases. While this number appears small, it represents a significant cross-section of high-value targets such as centralized marketplaces and mixers, which are statistically rare compared to standard vendor accounts. Nevertheless, relying on public court documents introduces two specific biases: (1) **Survivorship Bias:** The dataset reflects only those investigations that resulted in successful prosecution and declassification. It inherently excludes ongoing investigations or cases where attribution was achieved through classified means redacted from the public docket. Access to raw, unfiltered police case files would likely reveal a broader spectrum of unsuccessful attempts and alternative techniques. (2) **Temporal Decay:** The strategies of cybercriminals are dynamic. As specific failures become widely known and software changes, administrators might adapt. Therefore, the specific technical artifacts presented here represent a snapshot of historical failures. The transfer remains valid, but the specific injected vulnerabilities must be continuously updated to match current threat landscapes.

Jurisdictional and Legal Applicability. The training scenario simulates active forensic interaction, requiring techniques that overlap with penetration testing (e.g., active scanning, traffic interception). The legal permissibility of these techniques varies significantly by jurisdiction. The analyzed dataset is grounded in the U.S. judicial system, where specific warrants for Network Investigative Techniques are established. In other jurisdictions, as noted by the participants regarding hacking on demand, the legal threshold for active countermeasures may differ. While the legal permissibility of active countermeasures varies by jurisdiction, the technical reality of the vulnerability is universal. A server misconfiguration creates the same forensic opportunity regardless of the legal framework required to exploit it. The training focuses on the technical capability to recognize these opportunities, leaving the authorization procedures to local agency policy.

Scope of Simulation. Finally, the environment is designed specifically for complex, multi-layered infrastructures. It does not simulate the social engineering or long-term undercover work that often precedes technical analysis. By treating them as a narrative antecedent, the simulation focuses exclusively on the digital forensics phase. This makes the system highly specialized for technical units but less applicable for training general investigators in the psychological aspects of Darknet infiltration.

Conclusion

This work addresses the growing disparity between static forensic training resources and the dynamic reality of investigating anonymized cybercrime infrastructures. By analyzing 16 high-profile Onion Service platforms, we demonstrated that successful attribution relies on multi-modal evidence, frequently hinging on backend misconfigurations rather than protocol exploits. We operationalized these findings into *DarkMarket*, a live training environment that shifts the educational focus from artifact parsing to active engagement. Our pilot evaluation with active-duty practitioners suggests that live investigations require an adaptive, hypothesis-driven methodology. Furthermore, our observations highlighted a potential *Technical Halo Effect*, where experts heuristically overestimated the adversary's backend security based on frontend sophistication. Ultimately, this study highlights that investigating high-value targets requires a paradigm shift. As anonymization layers harden, the definition of a forensically sound investigation must evolve to include controlled, active interaction and collaborative workflows. Future training frameworks must move beyond isolated puzzles to simulate the distributed, volatile nature of modern cybercrime, preparing investigators to actively generate evidence within legal boundaries.

Acknowledgments

The authors used generative AI-based tools to revise the text, improve flow and correct any typos, grammatical errors, and awkward phrasing.

Appendix A. Availability of Data and Materials

To ensure the reproducibility of the presented forensic scenario and to facilitate its adoption in educational curricula, the complete source code and configuration artifacts have been made publicly available. The repository contains the full Docker Compose orchestration for the isolated network environment, the Python/Flask source code for the simulated market application, and the necessary instructor guide.

The artifact is released under the permissive MIT license and is available as open-source software at:

`https://github.com/xNKFTImFop70zeUS1prY/Forensic-Training-DarkMarket-Mock`

Appendix B. Analyzed Court Cases

Table B.4 details the 16 court cases analyzed in this study, categorized chronologically by their primary offense type.

Table B.4: Analyzed Onion Service Court Cases (N = 16).

Case Number	Target Category	Year
2:11-cr-01137 (C.D. Cal.)	Darknet Market	2011
8:13-cr-0010 (D. Neb.)	CSAM Distribution	2012
1:14-cr-00068 (S.D.N.Y.)	Darknet Market	2014
1:14-mj-02427 (S.D.N.Y.)	Darknet Market	2014
5:15-cr-00015 (W.D.N.C.)	CSAM Distribution	2015
1:16-cr-00438 (N.D. Ill.)	Copyright Infringement	2016
1:17-cr-00144 (E.D. Cal.)	Darknet Market	2017
1:18-cr-00243 (D.D.C.)	CSAM Distribution	2018
1:19-cr-00395 (D.D.C.)	Cryptocurrency Tumbler / Mixer	2019
2:19-mj-01843 (C.D. Cal.)	Darknet Market	2019
8:19-cr-00200 (D. Md.)	CSAM Distribution	2019
1:20-cr-00065 (D.D.C.)	CSAM Distribution	2020
1:21-cr-00399 (D.D.C.)	Cryptocurrency Tumbler / Mixer	2021
1:22-cr-00257 (D.D.C.)	Darknet Market	2022
3:22-cr-00143 (N.D. Cal.)	Darknet Market	2022
2:23-mj-00528 (E.D. Pa.)	Cryptocurrency Tumbler / Mixer	2023

References

Breitinger, F., Jotterand, A., 2023. Sharing datasets for digital forensic: A novel taxonomy and legal concerns. Forensic Science International: Digital Investigation 45, 301562. doi:https://doi.org/10.1016/j.fsidi.2023.301562.

Brezinski, D., Killalea, T., 2002. Guidelines for Evidence Collection and Archiving. RFC 3227. doi:10.17487/RFC3227.

Cortesi, A., Hils, M., Kriechbaumer, T., contributors, 2010–. mitmproxy: A free and open source interactive HTTPS proxy. URL: https://mitmproxy.org/. [Version 12.0].

DigitalCorpora.org, 2026. Digital corpora. URL: https://digitalcorpora.org/.

Europol, 2025a. 270 arrested in global dark web crackdown targeting online drug and criminal networks. URL: https://www.europol.europa.eu/media-press/newsroom/news/270-arrested-in-global-dark-web-crackdown-targeting-online-drug-and-criminal-networks.

Europol, 2025b. Europe-wide takedown hits longest-standing dark web drug market. URL: https://www.europol.europa.eu/media-press/newsroom/news/europe-wide-takedown-hits-longest-standing-dark-web-drug-market.

Free Software Foundation, Inc, 2026. Wget 1.25.0. URL: https://www.gnu.org/software/wget/manual/wget.html.

Garcha, R.K., 2018. NITs a No-Go: Disclosing Exploits and Technological Vulnerabilities in Criminal Cases. New York University Law Review 93, 882. URL: https://www.nyulawreview.org/issues/volume-93-number-4/nits-a-no-go-disclosing-exploits-and-technological-vulnerabilities-in-criminal-cases/.

Garfinkel, S., 2012. Lessons learned writing digital forensics tools and managing a 30tb digital evidence corpus. Digital Investigation 9, S80–S89. doi:https://doi.org/10.1016/j.diin.2012.05.002. The Proceedings of the Twelfth Annual DFRWS Conference.

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. Digital Investigation 6, S2–S11. doi:https://doi.org/10.1016/j.diin.2009.06.016. The Proceedings of the Ninth Annual DFRWS Conference.

Ghappour, A., 2017. Searching Places Unknown: Law Enforcement Jurisdiction on the Dark Web. Stanford Law Review 69, 1075. URL: https://scholarship.law.bu.edu/faculty_scholarship/204.

Göbel, T., Baier, H., Breitinger, F., 2023. Data for digital forensics: Why a discussion on "how realistic is synthetic data" is dispensable. Digital Threats 4. doi:10.1145/3609863.

Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics – and what is missing. Digital Investigation 22, S94–S105. doi:https://doi.org/10.1016/j.diin.2017.06.004.

Hargreaves, C., van Beek, H., Casey, E., 2025. Solve-it: A proposed digital forensic knowledge base inspired by mitre att&ck. Forensic Science International: Digital Investigation 52, 301864. doi:https://doi.org/10.1016/j.fsidi.2025.301864. DFRWS EU 2025 - Selected Papers from the 12th Annual Digital Forensics Research Conference Europe.

National Institute of Standards and Technology (NIST), 2026. Cfreds portal. URL: https://www.cfreds.nist.gov/.

PortSwigger Ltd., 2026. Burp suite documentation. URL: https://portswigger.net/burp/documentation.

Scarfone, K., Souppaya, M., Cody, A., Orebaugh, A., 2008. Technical Guide to Information Security Testing and Assessment. NIST Special Publication 800-115. National Institute of Standards and Technology. URL: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-115.pdf.

SOLVE-IT-DF, 2026. Solve-it. URL: https://github.com/SOLVE-IT-DF/solve-it/.

Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G., Thomas, C.B., 2020. MITRE ATT&CK: Design and Philosophy. Technical Report. The MITRE Corporation. URL: https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf.

The United States Attorney's Office Southern District of New York, 2013. Announces seizure of additional $28 million worth of bitcoins belonging to ross william ulbricht, alleged owner and operator of "silk road" website. URL: https://www.justice.gov/archive/usao/nys/pressreleases/October13/SilkRoadSeizurePR.php.

Tippe, P., Tippe, A., 2024. Onion services in the wild: A study of deanonymization attacks. Proc. Priv. Enhancing Technol. 2024, 291–310. doi:10.56553/POPETS-2024-0117.

Trickel, E., Disperati, F., Gustafson, E., Kalantari, F., Mabey, M., Tiwari, N., Safaei, Y., Doupé, A., Vigna, G., 2017. Shell we play a game? CTF-as-a-service for security education, in: 2017 USENIX Workshop on Advances in Security Education (ASE 17), USENIX Association, Vancouver, BC. URL: https://www.usenix.org/conference/ase17/workshop-program/presentation/trickel.

U.S. District Court for the District of Maryland, 2013. Affidavit in support of application for search warrant, case no. 13-1744 gc. Case 8:13-mj-01744-WGC, Document 24, Filed 10/31/16. Details the authorization and constraints of the Network Investigative Technique used against Freedom Hosting.

U.S. Government, 2026. Federal rule of criminal procedure 41: Search and seizure. URL: https://www.law.cornell.edu/rules/frcrmp/rule_41.

Voigt, L.L., Freiling, F., Hargreaves, C., 2025. A metrics-based look at disk

images: Insights and applications. Forensic Science International: Digital Investigation 52, 301874. doi:https://doi.org/10.1016/j.fsidi.2025.301874. DFRWS EU 2025 - Selected Papers from the 12th Annual Digital Forensics Research Conference Europe.

Voigt, L.L., Freiling, F., Hargreaves, C.J., 2024. Re-imagen: Generating coherent background activity in synthetic scenario-based forensic datasets using large language models. Forensic Science International: Digital Investigation 50, 301805. doi:https://doi.org/10.1016/j.fsidi.2024.301805. DFRWS APAC 2024 - Selected Papers from the 4th Annual Digital Forensics Research Conference APAC.