Computers & Security 172 (2027) 105180

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Full Length Article

Challenges in exfiltration, evasion, and behavioral sequences in malicious insider threat: Evidence from a live behavioral experiment

James Doodson [a,*], Deanna D. Caputo [a], James F. Murphy [b], Insider Threat Research & Solutions, The MITRE Corporation

[a] The MITRE Corporation, 7515 Colshire Drive, McLean, VA, 22102-7539, USA
[b] Formerly of The MITRE Corporation, 7515 Colshire Drive, McLean, VA, 22102-7539, USA

ARTICLE INFO

Keywords:
Insider threat
Insider risk
Exfiltration
User activity monitoring (UAM)
Potential risk indicators
Sequences
Behavioral experiment
Evasion

ABSTRACT

Insider risk and threat detection have several well-meaning assumptions that are difficult and rarely tested using scientific methods. For example, exfiltration is often seen as a "smoking gun" of malicious insider activity. This study drew on data from a live behavioral experiment in which 154 employees searched for, collected, and exfiltrated sensitive information from a real corporate network under benign, malicious, and super malicious task conditions. Endpoint- and network-based cyber sensors collected over 7000 separate types of user activity resulting in over 10 terabytes of logs before, during, and after the task, and employee participants completed interviews and questionnaires describing their actions and reasoning. Two analyses identified valuable indicators and/or detectors for insider threat or risk programs. First, qualitative content analysis identified 13 exfiltration techniques with 11 sub-techniques, and 76 evasion techniques categorized in 14 types. Malicious and super malicious intent participants used a broad range of exfiltration and evasion techniques, many relying on ordinary workplace tools and workflows rather than advanced tradecraft. Most techniques were observable in cyber logs in some form, but were difficult to distinguish from benign intent employee activity. Second, sequential analysis of cyber logs produced 5957,005 permutations of malicious cyber activity yet all but 13 had limited practical value. The study's findings indicate that insider risk detection should focus less on employee exfiltration and more on earlier patterns of search and collection behavior, and that sequence-based analytics should not be a core feature required for insider threat detection.

1. Introduction

1.1. Distinguishing insider threats and insider risks

Insider threats include espionage, intellectual property (IP) theft, unauthorized disclosure, sabotage, domestic violent extremism, and workplace violence. These threats cause significant harm to employees themselves, people they work with, their organization, their nation, and international partners—including significant operational, financial, and reputational losses which can take years to recover from and repair. While there are numerous definitions, the U.S. Government defines an insider threat as:

> A threat presented by a person who has, or once had, authorized access to information, a facility, network, a person, or a resource of the [organization] and wittingly or unwittingly commits an act in contravention of law or policy that resulted in or might result in harm through the loss or degradation of government or company information, resources, or capabilities, or a destructive act, which may include physical harm to oneself or another (U.S. Department of Defense, 2024, p. 19).

Insider threat can be differentiated from insider risk: insider *risk* refers to the potential for an employee to cause harm; insider *threat* refers to actions an employee takes resulting (or might result) in harm. The concepts are related as insider risk is a continuum from low to high risk. When insider risk becomes high enough, it becomes an insider threat.

The insider threat community distinguishes between malicious insiders and non-malicious insiders (with non-malicious including outsmarted, mistaken, and negligent insiders; Caputo, 2020). The current study focuses on malicious insiders. Malicious insiders intend to cause

* Corresponding author at: Insider Threat Research & Solutions Capability, The MITRE Corporation, 7515 Colshire Drive, McLean, VA, 22102-7539, USA.
E-mail addresses: doodsonj@mitre.org (J. Doodson), dcaputo@mitre.org (D.D. Caputo), jfmurphy@mitre.org (J.F. Murphy).

https://doi.org/10.1016/j.cose.2026.105180
Received 19 June 2026; Received in revised form 5 August 2026; Accepted 18 September 2026
Available online 18 September 2026
0167-4048/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

harm; they work consciously to steal or damage resources, information, or assets. Importantly, malicious insiders use their legitimate access, existing knowledge, and organizational connections to achieve their malicious goals without having to violate rules. Many cyber activities of malicious employees resemble benign employees doing their job, reflecting the complex challenge Insider Risk Programs (IRPs) face identifying and detecting insider risks and threats. Malicious insiders can blend in using everyday business tools and techniques, and do not need to break rules to achieve malicious goals.

1.2. Building potential risk indicators for insider risk

Insider risk is recognized as an important security challenge (Cappelli et al., 2012) but there remains limited scientific research on potential risk indicators (PRIs) for insider risk using real employees engaging in known malicious activity in real organizations. Instead, PRIs are often developed through expert elicitation and case study analysis, with limited empirical evidence supporting validity (Black, Yeung, and Yeung, 2022; Greitzer and Hohimer, 2011). Case studies are intriguing and useful for socializing and motivating action on insider risk, but are ineffective for developing PRIs. Case studies do not distinguish specifics of an employee's activities or other events relevant to their malicious goals from their other activity, nor do they distinguish an employee's activities from their larger organization (Caputo, 2020). Instead, case studies result in IRPs "chasing the last insider" and not anticipating emerging risks. Effective insider risk management instead relies on systematically combining and analyzing data to characterize known bad behaviors and identify risk patterns (Homoliak et al., 2019).

1.3. Cyber sensors for insider risk detection

Insider risk is inherently a human challenge. Effective IRPs deter, detect, and mitigate insider risk by combining human, organizational, cyber, and physical sensors and approaches (U.S. Office of the Director of National Intelligence, 2024). Some organizations have developed IRPs near-entirely focused on cyber components of managing insider risks, often placing the program in Security Operations Center (SOC), information technology, or cybersecurity functions. IRPs in those functions often lead to disappointment and wasted investment when the program is less effective, as it prioritizes cyber components over all other sensors and approaches. "Cyber-only" or "cyber-first" is not enough for successful insider risk management; IRPs must connect disparate data from human, organizational, physical, and cyber sensors across the organization rather than relying on any one sensor type. However, it would be a mistake to conclude that cyber sensor data are unimportant in isolation. One of the only sensors IRPs often directly own/manage is a User Activity Monitoring (UAM) solution, which is a cyber sensor that collects and records information on employee interaction with organization-furnished endpoints (National Insider Threat Task Force, 2018). UAM solutions collect data on various user activities including application (e.g., opening, closing, or switching between applications), file (e.g., creating, reading, copying, moving, renaming, or deleting files), device (e.g., attaching removable media), network and web (e.g., uploading or downloading files, visiting Intranet and external sites), clipboard (e.g., copy/paste operations), printing (e.g., sending documents to print queues), and login/logoff activities (e.g., successful/-failed logons, session start/stop, workstation lock/unlock). For each user activity, often referred to as an event, UAM solutions collect metadata like timestamp, user/device identifiers, file name and path, URL, data size, and other context. This event-level data provides information needed to implement UAM cyber PRIs.

UAM solutions are distinct from Data Loss Prevention (DLP) tools when looking at insider threat exfiltration. DLP tools focus on information content and predefined rules about information movement (National Institute of Standards and Technology, 2020; Liu and Kuhn, 2010). DLP tools attempt to detect, but are rarely approved to block the transfer of data labeled or recognized as sensitive. Contrastingly, UAM solutions focus on user actions and behavioral patterns, regardless of whether data is pre-tagged or recognizable by content signatures. In practice, DLP helps identify what sensitive data moved and where, whereas UAM helps identify what a user did before, during, and after that movement to inform risk decisions about whether activity was harmful, malicious, or not.

In developing PRIs, there are no "silver bullets" (Cappelli et al., 2012). The key word in PRI is "potential". Generally, no single employee activity or user event proactively demonstrates insider risk or threat beyond a reasonable doubt. Instead, PRIs are activities and events which help IRPs determine whether a risk is worth mitigating or escalating to an inquiry or investigation. Many seemingly logical PRIs struggle to differentiate risk worthwhile for analyst review versus risk that is not; this can lead to limited IRP effectiveness, analyst overload, and wasted analyst effort. IRP analysts with too many unvalidated PRIs triage too many false positive alerts, leaving malicious insiders undetected and putting organizations at risk.

Collecting more cyber data does not inherently improve detection of insider risk. IRPs should only request and collect the specific data required to make decisions. Limiting IRP data collection can demonstrate to important stakeholders that employee data is treated carefully, counters perceptions of a "surveillance" organization, and avoids the "data lake" fallacy, where increasingly more data is thrown at a problem in hope that a data scientist may eventually solve it. To date, that approach has been costly and unsuccessful for insider risk. Instead, the insider threat community needs more effective cyber PRIs that distinguish malicious insider activity from routine employee work—this approach may also reduce the amount of data an IRP needs to generate effective alerts.

1.4. Exfiltration, evasion, and sequence-based detection challenges

Overlap with routine activities and employee job creativity challenge assumptions that malicious insiders can be reliably identified through simple technical cyber exfiltration or evasion signatures. Exfiltration refers to techniques a malicious insider uses to remove data from an organization's network or devices (National Institute of Standards and Technology, 2020). Exfiltration is often treated as a "smoking gun" of malicious insider activity. However, employees remove information from an organization's network as part of their standard job duties using ordinary business tools, routine workflows, and approved systems—which can also be leveraged by a malicious insider. Evasion refers to techniques a malicious insider uses to reduce likelihood of activity being quickly discovered, attributed to malice, or attributed to them. These techniques can be subtle, and often overlap with benign activity of employees doing their regular work. For example, malicious employees can compress files, split files across multiple actions, rename documents, use collaboration platforms, or use removable media to try to hide their malicious activities—these activities also occur routinely by employees for entirely legitimate reasons. In other instances, malicious insiders use creative evasion techniques enabled by their role, knowledge of the environment, and familiarity with organizational processes.

Detection challenges are compounded when organizations apply external threat frameworks to insider risk. Insider risks and threats are not the same as Advanced Persistent Threats (APTs) or compromised accounts. External adversaries who use compromised accounts to act as masqueraders or impersonators do not look the same as actual employees within the organization. For example, external adversaries must gain access and move without authorization, whereas insiders begin with legitimate or indirect access. Insiders already understand internal systems, organizational terminology, social relationships, and routine workflows. Consequently, insider exfiltration and evasion differ meaningfully from external tradecraft, and external cyber threat frameworks and DLP approaches often do not help IRPs as much as expected.

Applying cyber APT kill chains to malicious insider detection has proven equally problematic. The idea of identifying specific sequences of cyber activities and assuming that insider risk has its own version of a kill chain is attractive, particularly to cyber practitioners accustomed to using such approaches for APTs. However, no credible insider risk kill chain has been identified to date. Cybersecurity vendors are increasingly marketing the potential for sequences as a key future feature for insider threat detection. Moreover, even if such a kill chain does exist, there is currently no scientific evidence supporting or refuting it. Sequences could be useful, but that remains an empirical question.

1.5. Research on cyber detection of insider risks and threats

Recent research on cyber-focused insider threat detection has moved beyond purely theoretical discussions, but most still relies on synthetic datasets or simulated user behavior. In published research, many studies evaluate detection performance on Carnegie Mellon University Software Engineering Institute CERT's synthetic datasets rather than observed malicious behavior in deployed enterprise environments (Glasser and Lindauer, 2013). For example, recent machine learning (ML) studies proposed increasingly sophisticated methods, including BiLSTM (Manoharan et al., 2024) and transformer-based sequential models (Elbasheer and Akinfaderin, 2025). Although those techniques often report improved classification performance in studies, development and evaluation using synthetic data rather than operational deployments or ground truth "known" malicious behavior by real employees causes usability challenges. It is difficult for insider risk practitioners and IRPs to secure support from leadership and stakeholders (e.g., Legal, Human Resources) to deploy those findings due to the lack of evidence. From a scientific perspective, synthetic datasets (and methods using the datasets) do not necessarily reflect real human behavior and often fail to capture organizational context, human creativity, and ambiguity that characterize real malicious insider activity.

Although most published or open-source insider threat research uses synthetic datasets or singular case studies, a small number of studies have used datasets based on real human behavior, including competition- or game-based environments (e.g., Harilal et al., 2017; Taylor et al., 2013; see Alzaabi and Mehmood, 2024 for more detailed discussion). These studies advance detection discussions but typically contain few known malicious insiders or occur in settings differing substantially from real work environments, limiting ecological validity, and operational applicability.

Due to the lack of real cyber data, several studies have conducted retrospective analysis for patterns across past insider threat cases including analysis of cyber techniques. For example, Robinson (2025) analyzed 909 publicly reported malicious insider cases from U.S. legal records and documented exfiltration pathways across physical media, cloud services, messaging platforms, and personal devices. Similarly, Whitty et al. (2024) used 86 publicly available IP theft cases drawn from court and legal records. They identified IP theft does not follow a single linear progression but emerges through multiple, recurrent, and non-linear events involving cyber and physical behaviors (e.g., competitor contact, overseas travel, presentations, photographing sensitive information). Despite focusing on exfiltration, there remains little empirical research on malicious insiders deliberately evading detection.

The persistent gap in the use of real data for developing and refining cyber-based insider threat and risk analytics causes challenges transitioning techniques to operational deployment within IRPs. The dearth of empirical scientific research mirrors insider risk and threat PRI research more broadly, where patterns and indicators are often derived from expert frameworks (e.g., Greitzer et al., 2018), retrospective case analysis (e.g., Nurse et al., 2014), or in some cases indirect operational or anecdote-based assumptions.

1.6. Present study

An applied scientific study is needed that moves beyond case studies and untested hypotheses or assumptions. The present study uses data from a live behavioral experiment conducted on a real corporate network with real employees. Specifically, data from that behavioral experiment was analyzed to identify and categorize exfiltration and evasion techniques for insider risk and examine the viability of sequences for insider risk detection. The findings are presented to help IRPs and UAM solutions better understand practical value and limitations of exfiltration, evasion, and sequences in proactive insider risk detection.

The present study addressed two research questions:

1. What exfiltration and evasion techniques do employees with malicious intent use when carrying out malicious search, collection, and exfiltration tasks on a live corporate network, and to what extent are those techniques visible in cyber sensor data?
2. Are there cyber activity sequences that meaningfully distinguish malicious intent insiders from benign intent participants and offer practical value for insider risk detection?

In answering these research questions, the study provides novel empirical evidence and behavioral patterns to help IRPs and UAM solution vendors operationalize exfiltration, evasion, and sequences for defensible, proactive insider risk detection.

2. Methods

2.1. Design

This applied behavioral experiment examined how information search, collection, and exfiltration activities on a live corporate network differed between employees acting with benign intent, malicious intent, and super malicious intent. Employee participants were asked to search for, collect, and exfiltrate sensitive information from the organization's Intranet to an external destination. Three experimental conditions varied on two factors: participant intention and cyber technical expertise. In the benign intent condition, employee participants completed the task of searching, collecting, and removing information from the network in support of a legitimate customer (i.e., developing and sending requested information). In the malicious intent condition, employee participants completed the same task sending the same information but to a competitor (i.e., stealing information). In the super malicious intent condition, employee participants completed the same malicious intent scenario but were selected for their deep cyber defensive and/or offensive expertise, and were asked to behave like an insider threat without getting caught. Other than intention and technical expertise differences, participants completed the same task under the same time and technical constraints.

2.2. Participants

Participants were 154 employees recruited from a medium-sized Defense Industrial Base (DIB) organization in the United States. Employees were recruited through internal email distribution lists and organizational announcements. Of these, 102 participants were recruited from the general employee population and randomly assigned to either a benign intent condition or malicious intent condition. An additional 52 participants with substantial cyber defensive/offensive expertise were recruited into the super malicious intent condition.

Participants were aged 23–74 years, with 64.2% of participants self-reporting being male, 33.1% female, and 2.7% did not answer. Average organizational tenure was 9.9 years (range: 0–43 years) with an average of 19.5 years of expertise in their field (range: 3–45 years). Most participants (87.4%) held a government security clearance and worked in

primarily technical (91.4%), managerial (4.6.%), administrative/support (2.0%), or other roles (2.0%). The employee sample was highly educated, with 29.1% holding a bachelor's degree, 54.3% a master's degree, and 11.3% a doctorate; 0.7% reported an associate degree, 0.7% reported a high school diploma/GED, and 4.0% did not answer. Participants in experimental conditions did not meaningfully differ on major background characteristics, beyond cyber technical expertise.

2.3. Procedure

Before beginning the task, a research team proctor met individually with each employee participant to explain study procedures and administer a pre-task questionnaire collecting demographic, job role, and working environment information. Participants were provided with one of two scenarios describing the reason they were searching for, collecting, and exfiltrating information from their company's Intranet. Both scenarios involved participants being placed in a situation involving difficult financial circumstances. Benign intent participants were instructed to send sensitive information externally as part of legitimate work supporting a customer to gain a promotion. Malicious and super malicious intent participants were instructed to send sensitive information externally to a competitor to get a higher-paying job, resolving their financial issues.

All participants had six hours to complete the task over the course of five business days and could divide the time alongside usual work.[1] Participants could distribute hours however they chose rather than completing the task in a single sitting. For example, some participants used the six hours in multiple sessions over multiple days. Participants were provided a time log to record exact dates and times they worked on the task, which were used by researchers to distinguish on-task activity from normal work activity. All participants searched for information on the same topic, anti-money laundering techniques, and were permitted to use any method to exfiltrate information. There were no restrictions on the devices participants could use to complete the task. However, participants generally searched for and collected information using their organization-furnished computer but some used other devices during exfiltration.

After completing their assigned scenario and task, employee participants completed a post-task questionnaire and deep-dive interview asking about their actions, decision-making, and reasoning during the task. The experiment was reviewed and approved through an Institutional Review Board, and organizational safeguards were implemented to reduce risks associated with employees exfiltrating sensitive information from their corporate live production environment. Safeguards included a clean-up plan for material successfully exfiltrated, and confirmation that no classified or highly sensitive information about the topic was present on the network.

Participants were covered by an information security waiver if they broke organizational security rules doing study activities, although it functioned differently across experimental conditions. In benign and malicious intent conditions, the waiver was an internal employee protection mechanism not disclosed to participants so as not to directly shape their behavior, thus maintaining a realistic sense of risk and constraint. In the super malicious intent condition, the waiver was disclosed to encourage the broadest range of technically creative exfiltration and evasion behaviors. For some participants in the super malicious condition, the waiver was consistent with the types of authorized exceptions encountered in their professional cyber work. Findings from benign and malicious intent conditions better reflect everyday employee behavior under realistic uncertainty, whereas the super malicious condition was an intentionally broader test of technically sophisticated malicious behavior. Despite the waiver, many super malicious participants described apprehension about breaking rules in their post-task questionnaires and interviews, suggesting the waiver only partially reduced perceived risk.

[1] In multiple earlier versions of the study, participants were provided with more time but did not use more than six hours. Task time was set to six hours to minimize disruption to employees' normal work.

2.4. Cyber sensors data collection

Cyber sensors collected data about each employee participant's activity during the week of, week before, and week after the task. This design allowed the research team to compare on-task activity with each participant's own baseline behavior. Data sources included two commercially-available UAM solutions in addition to network and operating system sensors. Sensors collected data and metadata for application, clipboard, device, file, log-on/off, network, printing, registry, web browsing, and window activity. Over 7,000 separate activities from UAM were collected.

Cyber sensor data were extracted into a secure research environment and converted into consistent file formats and data types. The research team developed code distinguishing user-generated activity from system-generated activity and categorizing individual cyber events into meaningful activity types (e.g., file copy/move, clipboard use, search, email send). The team also grouped meaningfully similar activities when raw logs reflected them as multiple low-level events. For example, when a participant moved many files with a single user action, the low-level log entries were grouped into one higher-level file move activity with accompanying metadata on the number of files.

Time log data were used to isolate on-task activity from off-task activity. This distinction was important because participants completed the study alongside normal work and personal use of organization-furnished computers. The resulting dataset therefore contained known bad activity (malicious and super malicious participants during the task), known good activity (benign participants during the task), and presumed good activity (all participants outside on-task periods).

The purpose of including two UAM solution solutions was to reduce coverage gaps and account for differences in what specific products captured, not to compare vendor coverage or analytics (e.g., conduct product evaluation). This study was not designed as a benchmark comparison against User and Entity Behavioral Analytics (UEBA) solutions, UAM solutions, or other solutions. Instead, UAM solutions functioned as cyber sensors for data collection only, not analysis. Consequently, the present findings inform how UAM solutions (and to a smaller extent UEBA solutions) could collect and analyze cyber data, rather than comparatively evaluating existing solutions; the findings enable vendors and IRP leads to enhance existing solutions rather than advocating for new solutions.

2.5. Controls

This study was designed as a live applied behavioral experiment conducted in a real organizational environment rather than a fully naturalistic observation of insider behavior. Use of real employees and work systems on a corporate network together provide important ecological realism uncommon in insider threat research. As in other applied behavioral experiments, participant awareness of being in a study could influence some aspects of behavior or self-report. Several standard control procedures measured and mitigated such effects. First, employee participants in the benign and malicious intent conditions were not told that the study examined insider risk/threat, but instead received a general workplace framing centered on Intranet navigation and information gathering. This control was intended to standardize participants' expectations across conditions and minimize demand characteristics arising from awareness of the study's aims, reducing the likelihood that knowledge of the study's specific purpose would influence their behavior (McCambridge, de Bruin and Witton, 2012; Orne, 1962). Participants in the super malicious intent condition were

informed of the malicious framing in order to maximize the number of feasible exfiltration and evasion behaviors.

Second, all participants were informed that the quality of material they produced would be evaluated; this encouraged participants to take the task seriously and maintain consistent effort throughout completion, consistent with the evaluation apprehension literature (Cottrell, 1972; Henchy and Glass, 1968). Third, researchers reviewed post-task questionnaires, debrief interviews, cyber sensor data, and researcher notes to identify any participants meeting predefined exclusion criteria including whether participants adopted assigned roles as intended. Example questions included: "Did the scenario narrative affect the ways in which you performed the task? If yes, please explain", "During the task, how easy was it to assume your assigned role?", and "If you were to perform the same task without the assigned narrative/role, would there be any discernable differences in behavior?" One participant was excluded under the predefined task validity criteria because post-task questionnaire and debrief materials indicated their task behavior was driven by perceived study purpose rather than the assigned scenario (i.e., participant reported information was gathered for the proctor, not the target described in the scenario).

As an experimental manipulation check, a one-way ANOVA found no significant differences between conditions in the extent participants responded to the question "During the task, to what extent did you feel that you assumed your assigned role?", $F_{(2,\ 146)} = 1.05$, $p = .353$, $\eta^2 = .014$, $R^2 = .014$.[2] All remaining exclusions occurred for reasons outside task validity criteria: incompatible computer or other technical difficulties ($n = 9$), discomfort acting maliciously or against company policy (7), non-response or no-show after initial interest (5), failure to allocate sufficient time to the study (5), family or health obligations (4), inability to consent to monitoring because of nature of normal work (1), and cyber sensor collection issues (10). Together, these controls provide confidence that employee participants engaged with assigned roles as intended and observed behavioral patterns are meaningful for detecting insider activity in a realistic enterprise environment, with direct value for insider risk research and operational analytic development; where other research data often lacks in quantity and quality.

2.6. Data preparation and analysis

Two analytic approaches were used. First, exfiltration and evasion techniques were identified using qualitative content analysis of employee participants' interview and questionnaire data. Two researchers independently reviewed qualitative responses, hand-annotated relevant content, collated annotations into broader codes, and developed a codebook to organize exfiltration and evasion techniques. Roughly 30% of the data was double-coded for reliability and consistency, then the coding structure was refined to reduce overlap. Final coded data was then cross-referenced with cyber sensor logs to determine whether identified techniques were visible and could be reverse engineered to become visible in available cyber data. For the present study, questionnaire and interview data were used to identify, interpret, and contextualize exfiltration and evasion techniques, including behaviors only partially visible in cyber sensors. The development of integrated multi-source risk models combining human, organizational, physical, and cyber sensors was outside the study scope, which focuses specifically on cyber observable behavior and the practical value of cyber-based PRIs.

Second, sequential analysis was conducted by time ordering cyber logs to determine whether participants in the malicious and super malicious intent conditions used different sequences of cyber activities than participants in the benign intent condition. Participants' cyber logs were organized in time order, and relevant activities and metadata extracted then categorized using automated Python scripts (e.g., "file move"). For each participant, the researchers used a fixed-length Python n-gram algorithm to create n-grams from abbreviations. An n-gram is a continuous sequence of n items (e.g., words or characters) extracted from a text corpus, where n is the sequence length (e.g., bigrams for 2 items, trigrams for 3 items; Manning and Schütze, 1999). For this analysis, n-grams were based on sequences of unique abbreviations (e.g., "AA", "BB"), separated from neighboring abbreviations by whitespace to prevent ambiguity (e.g., "AA BB", "AA CC BB"). Each abbreviation referred to a specific cyber activity (e.g., "AA" could refer to "Conduct new Intranet search", "BB" to "Copy text content from Microsoft Teams to clipboard", and "CC" to "Renamed file"). Researchers produced n-grams of cyber activities up to tetragrams (i.e., four cyber activities in length) because the required computational power became unmanageable with limited return-on-investment beyond tetragrams. Analysis did not require cyber activities to occur directly adjacent to one another in the raw cyber logs. Instead, sequences were based on a subset of 100 higher-risk cyber activities, with other activities permitted to occur between them. This approach focused analysis on the order of meaningful cyber activities without sequences being disrupted by less diagnostic cyber activity occurring in between. To reduce overfitting, the dataset was split into training (70%) and testing (30%) subsets, and only sequences meeting threshold criteria across both subsets were retained as potential PRI sequences (threshold described in Results section).

The 100 high-risk cyber activities used to produce the sequences were identified using prior screening analyses from the same study. First, a "snapshot" analysis used Wilcoxon rank-sum tests to identify cyber activities whose frequency significantly differed between benign and malicious intent or benign and super malicious intent participants during the on-task period. Second, a "baseline comparison" analysis used a nonparametric longitudinal repeated-measures model with one between-subjects factor (experimental condition) and one within-subjects factor (task period). One model identified activities showing stronger pre-task to on-task change in malicious than benign intent participants, and a separate model identified activities showing stronger pre-task to on-task change in super malicious than benign intent participants. Follow-up Wilcoxon rank-sum tests and Wilcoxon signed-rank tests were used to interpret statistically significant changes. The resulting 100 activity subset was then used for sequence analysis so that n-gram modeling focused on activities already identified as more relevant to insider risk behavior while keeping candidate permutations computationally manageable.

The researchers used n-gram analysis because the study's primary objective was to determine whether recurring behavioral sequences contain useful information for proactive insider risk and threat detection. As previously discussed, recent insider threat research has increasingly adopted advanced sequential ML techniques (e.g., Elbasheer and Akinfaderin, 2025; Manoharan et al., 2024).

Unlike more advanced sequential ML methods designed primarily to maximize predictive accuracy and only later permit interpretation of which sequence features contributed to a classification, n-gram analysis treats the sequence itself as the unit of analysis. The n-gram approach more directly examined the association between sequences and insider risk, enabling initial assessment of whether cyber sequences have utility for insider risk and threat detection.

The n-gram approach also enhanced the interpretability of behavioral sequences for operational application. Sequential ML models learn complex internal (i.e., latent) representations of behavioral sequences, producing a prediction (e.g., maliciousness score or classification), rather than explicit, human-readable behavioral sequences (Li et al., 2022). For example, a sequential ML model may indicate a user's activity has a high probability of being malicious, with SharePoint access, file copying, and archive creation identified as important ordered contributors to that prediction. In contrast, n-gram analysis directly produces a finite ordered sequence (e.g., Open SharePoint, then Copy File,

[2] Responses ranged from 1=Not at all to 5=Completely. Benign condition: Mean (M) = 3.83, Standard Deviation (SD) = 1.08; Malicious condition: $M = 3.57$, SD=1.06; Super Malicious condition: $M = 3.81$, SD=0.82.

then Compress Archive), which can be interpreted by IRP analysts and examined, validated, and operationalized as a candidate PRI or UAM detection rule.

N-gram analysis also serves a different purpose from qualitative sequence reconstruction approaches (e.g., Whitty et al., 2024). Whereas qualitative approaches reconstruct how malicious insider activity unfolds across broader behavioral, organizational, and contextual pathways providing insight into insider decision-making and activity progression, n-gram analysis focuses on recurring sequences of observable cyber activities drawn directly from empirical user activity allowing specific cyber patterns that recur across users to be examined individually as candidate PRIs.

3. Results

3.1. Exfiltration techniques

Content analysis of interview and questionnaire data identified 13 exfiltration techniques, with four techniques further divided into 11 sub-techniques, yielding 24 exfiltration techniques and sub-techniques (Table 1). Exfiltration techniques were defined as techniques used to remove data from the organization's network or devices.

Several exfiltration techniques were used across benign, malicious, and super malicious intent conditions and have limited value differentiating malicious from benign activity in isolation. These techniques included sending information in email attachments, uploading information to organization-approved secure file transfer service, and uploading information to cloud storage.

Although sending files as attachments to an email was common across all conditions, a Fisher-Freeman-Halton Omnibus test indicated benign employees sent attachments via corporate email significantly more than malicious ($p \leq .001$) and super malicious intent conditions ($p \leq .001$). This reflects that the benign intent condition would only use corporate email, as the condition represented what employees normally do in their job. Even though benign intent employee participants used corporate email more, many malicious (56.5%) and super malicious intent employees (34.0%) also used it, demonstrating employees acting with benign, malicious, and super malicious intent can use the same ordinary business tools to achieve differing goals. The same is true of benign, malicious, and super malicious intent participants using organization-permitted file transfer services albeit these were universally rare.

In contrast, several exfiltration techniques were used only by malicious *and* super malicious intent participants. Some techniques significantly differentiated malicious from benign intent including photographing or videoing the screen, sending files through burner or personal email accounts, transferring files to removable media especially USB devices, and sending email attachments using burner and personal email accounts. Other techniques were also observed only in malicious and super malicious intent conditions, but occurred too infrequently to produce statistically reliable differences including printing hard copies and uploading files to remote machines, unsanctioned file transfer services, and photo storage. Although rare, presence of these techniques may be operationally concerning because they reflect exfiltration methods not observed in the benign intent participants. Exfiltration techniques used by malicious and super malicious participants reflected efforts to move data through channels less directly attributable to routine organizational workflows or provided additional distance between the employee participants and final data destination of the data (albeit likely to generate more opportunities for cyber logging).

Super malicious intent participants used the broadest range of exfiltration techniques. Several techniques were unique to employee participants trained in cyber defensive or offensive capabilities, and not used by either benign or malicious intent participants. Based on pairwise Fisher exact tests, no technique produced statistically significant differences between malicious and super malicious intent participants after

Table 1
Insider risk exfiltration techniques and sub-techniques.

Technique	% of Participants using technique			Omnibus Test[a]
	Benign intent (n = 51)	*Malicious intent (n = 51)*	*Super malicious intent (n = 52)*	
1. Photographed or videoed screen	0.0	12.1	12.6	.003**
2. Printed hard copy	0.0	2.0	1.0	.999 N.S.
3. Saved email in unsent email and retrieved through remote email access session	0.0	0.0	1.0	.999 N.S.
4. Sent as email attachment	100.0	85.9	80.6	.001***
a. Corporate email account	100.0	56.6	34.0	.001***
b. Third-party corporate email account access (e.g., email account from a voluntary position)	0.0	0.0	2.9	.999 N.S.
c. Burner email account+	0.0	16.2	32.0	.001***
d. Personal email account	0.0	19.2	18.4	.001***
5. Sent via collaboration communications platform (e.g., Microsoft Teams, Slack)	0.0	0.0	1.9	.999 N.S.
6. Streamed to screen recording device connected via HDMI	0.0	0.0	1.0	.999 N.S.
7. Transferred to corporate-owned but publicly accessible server	0.0	1.0	0.0	.999 N.S.
8. Transferred to removeable media	0.0	7.1	11.7	.020*
a. SD card	0.0	0.0	1.0	.999 N.S.
b. USB device (e.g., external hard drive, USB "thumb" drive)	0.0	7.1	10.7	.032*
9. Transmitted over Bluetooth	0.0	0.0	1.9	.999 N.S.
10. Uploaded to secure file transfer service	1.1	2.0	2.9	.999N.S.
a. Permitted file transfer service	1.1	1.0	2.9	.999 N.S.
b. Unsanctioned file transfer service	0.0	1.0	0.0	.999 N.S.
11. Uploaded to remote machine (e.g., remote share via SMB, FTP)	0.0	2.0	1.0	.999 N.S.
12. Uploaded to code repository (e.g., Bitbucket, GitHub)	0.0	0.0	1.9	.999 N.S.
13. Uploaded to cloud storage	1.1	3.0	3.9	.999N.S.
a. File cloud drives (e.g., Box, Dropbox, Google Drive, OneDrive)[b]	1.1	2.0	2.9	.999 N.S.
b. Photo storage (e.g., Google Photos)	0.0	1.0	0.0	.999 N.S.
c. AWS Server	0.0	0.0	1.0	.999 N.S.

* $p < .05$, ** $p < .01$, *** $p < .001$, N.S $p > .05$.

[a] Omnibus test was Fisher-Freeman-Halton exact test with a 2 (present, no present) x 3 (experimental conditions) design, reflecting proportion of employee participants using the technique. The test was selected over Pearson chi-square test because many techniques were rare, resulting in sparse cells and zero counts violating chi-square assumptions. The Holm procedure accounted for multiple omnibus tests conducted across techniques and controlled familywise error rate.

[b] Secure file transfer services were subdivided into permitted and unsanctioned categories because the organization maintained a clear distinction between approved and nonapproved tools. Cloud drive services were not similarly subdivided because approval could depend on waiver status, and available data

did not always distinguish corporate from personal account use. Accordingly, cloud drive services were grouped by function rather than sanction status.

correction for multiple comparisons, likely because each occurred much less often. However, these techniques remain important because they were observed only among super malicious intent participants. These techniques included saving an email in draft or unsent form and later retrieving it through remote access, sending files through a third-party corporate email account, sending files via collaboration communications platforms, streaming content to a screen recording device connected via HDMI, transferring files to an SD card, transmitting files over Bluetooth, uploading files to a code repository, and uploading files to cloud storage. Together, these techniques suggest that super malicious intent participants leveraged their technical knowledge for more technically creative exfiltration options than malicious intent participants.

Several of those techniques do not require deep cyber expertise, administrator privileges, or scripting skills; instead, the techniques reflect greater creativity in using ordinary technologies and workflows for exfiltration. Based on interview data, some super malicious intent participants described deliberately selecting exfiltration methods they believed would blend into normal work activity. For example, several participants described using code repositories, cloud infrastructure, or collaboration tools because those services were already common in their cyber roles and perceived as less likely to attract scrutiny; the exfiltration technique also functioned as an evasion technique. Statistically significant differences between malicious and super malicious intent participants emerged only for the type of email account used for attachments: malicious intent participants were more likely to use corporate email ($p = .002$), whereas super malicious intent participants more likely to use burner email ($p = .003$). This pattern suggests email was the dominant choice irrespective of malicious intent, but malicious intent and technical expertise influence how email is used to exfiltrate (i.e., type of email account).

3.1.1. Reverse engineering exfiltration techniques

Attempts to reverse engineer exfiltration techniques in cyber sensor logs yielded limited success. Techniques were often observable in some form, but not in a way that meaningfully distinguished benign from malicious intent use. Three recurring challenges emerged.

First, several exfiltration techniques had both benign and malicious uses. Email attachments, approved file transfer services, and cloud storage were used by participants with benign intent and those with malicious intent. Focusing on exfiltration techniques in isolation identified too many examples of benign activity to be useful as standalone indicators.

Second, metadata associated with observed exfiltration techniques was usually not diagnostic in isolation. Filenames, file sizes, destination, and window titles often lacked sufficient contextual meaning to distinguish malicious from benign intent. Benign intent participants also changed filenames or used common services, reducing the usefulness of metadata fields.

Third, particularly among super malicious intent participants, some techniques were described in detail but were not fully completed because participants ran out of time or chose a lower-effort alternative. Thus, the qualitative data sometimes revealed exfiltration intentions that did not produce a complete trace in cyber sensor logs.

These challenges were especially evident in several commonly observed techniques where techniques were relatively easy to observe but difficult to interpret. Sending files as attachments to emails was one example. UAM solutions could identify that an email was sent from a dedicated email application or webmail interface and, in some cases, could capture recipient information, subject lines, and attachment metadata. However, no clear metadata pattern was found during analysis that reliably distinguished benign, malicious, and super malicious intent use. In addition, sending email to a personal or privacy-focused email service may indicate higher risk, but such activity alone was insufficient to distinguish malicious intent.

Similarly, uploading files to secure file transfer services could be observed through network logs and web browser metadata. Approved services were difficult to use as indicators because benign intent participants also used them. Unsanctioned file transfer services were inherently higher risk, but rarely observed directly on organization-furnished systems because malicious intent participants likely used personal devices instead. The same interpretive problem appeared for code repositories. Uploads to code repositories could be inferred from destination addresses, URLs, or repository-related applications, but super malicious intent participants also used these services during legitimate work, limiting their value as isolated indicators. Printing was also observable, but not useful in isolation. UAM solutions captured printing events and metadata such as document name, page count, and printer device. However, printing behavior did not reliably distinguish benign from malicious intent participants.

One exfiltration technique—Bluetooth file transfer—was more promising as a specific, manageable indicator. It was rare in the dataset and used only by super malicious intent participants whilst on-task. One UAM solution explicitly captured Bluetooth file transfer activity, and another allowed inference through application names and window titles associated with Bluetooth transfer utilities.

Connecting a screen recording device via HDMI was theoretically detectable through device logs, but the specific event could not be reliably identified in the available data. The device may have appeared as a normal peripheral, connected before monitoring began, or not have sufficiently distinctive manufacturer/model information.

3.2. Evasion techniques

Rigorous content analysis of interview and questionnaire data identified 76 evasion techniques grouped by the research team into 14 categories (Fig. 1). Intentionally, analysis did not assess whether evasion attempts were successful, but instead how employee participants attempted to evade detection; this distinction is important because some techniques may reflect participant beliefs about detection, even when beliefs were inaccurate. Additionally, the post-task questionnaire asked participants about evasion motivation and concealment including whether they felt the need to hide or cover up their behavior during the task, what methods they used, and details on attempts to be covert without using a computer. Responses, with debrief interviews, were reviewed to inform qualitative analysis of evasion behavior.

The evasion catalog identified a breadth of techniques and patterns within the study evasion techniques. Concurrently, percentages for many individual evasion techniques were low for several reasons. Evasion behavior was categorized across a large number of techniques, and participants typically needed only one or two techniques rather than many. Furthermore, the task was time-bounded to six hours over five business days so some techniques were considered or partially attempted but not completed. The low evasion percentages across 74 techniques reflect the variability and selectivity of evasion behavior rather than a lack of importance. Therefore, Appendix A provides the percentages for each category of evasion techniques, but publishing percentages individually would lead to insider risk practitioners being pressed to develop detectors of techniques that make less sense for their organization, or low percentages leaving practitioners unable to build detectors for techniques that hold real promise for detection in their organization. The research team considers the evasion techniques catalog itself as a set of promising hypotheses worthy of future insider risk research.

Many evasion techniques did not require deep cyber expertise, specialized tooling, or an ability to script/code. Instead, malicious and super malicious intent participants often relied on ordinary user behaviors like renaming files, creating compressed containers, labeling emails with benign subject lines, avoiding downloads, clearing browser history, and splitting files across multiple emails to evade detection. This

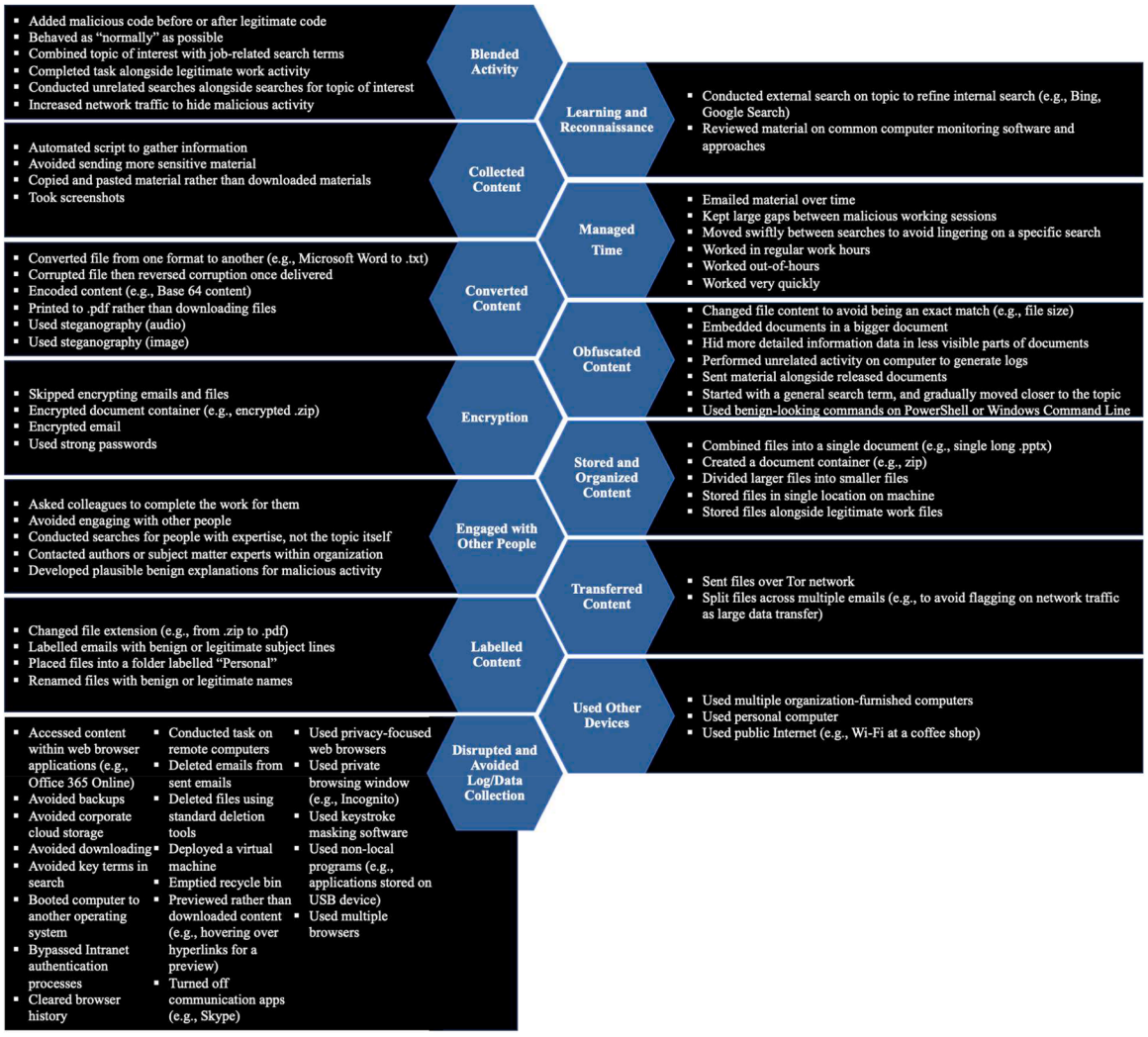


Fig. 1. UAM study evasion catalog.

pattern reinforces that insider evasion is often embedded in routine workplace actions rather than advanced tradecraft. Compared with super malicious intent participants, malicious intent participants were more likely to rely on straightforward techniques including those from Labelled Content, Purged Content, and Stored and Organized Content categories. These included renaming files with benign names, labeling emails with legitimate subject lines, removing document markers, and combining files into a single document. This pattern suggests that malicious participants without deep cyber expertise could still enact meaningful evasion by altering how sensitive material appeared, without needing specialized tools or advanced technical methods.

Many evasion techniques were file-centric involving disguising, reorganizing, or packaging files. Malicious and super malicious intent participants renamed files, changed file extensions, created compressed containers, encrypted document containers, combined files into a single document, divided files into smaller files, and stored files in ways they described as intended to reduce suspicion. This observation suggests that malicious insiders can focus on making sensitive material appear less meaningful, less visible, or more routine within ordinary file workflows. The observation also extended beyond files with malicious and super malicious intent participants describing behaving as "normally" as possible, combining topic-of-interest searches with normal job-related search terms, conducting unrelated searches alongside task-related searches, developing plausible benign explanations, and labeling emails with legitimate subject lines. Thus, malicious insiders can also evade detection by normalizing suspicious activity so that it blends into expected work rather than making it invisible.

A subset of evasion techniques were used by benign intent participants as well as malicious and super malicious intent participants, reinforcing the challenge of operationalizing detection for insider risk. These techniques included contacting subject-matter-experts, conducting external searches to refine internal searching, creating compressed containers, splitting files across multiple emails, printing to PDF, and renaming files with benign names. This overlap reinforces that many evasion techniques are weak proxies of malicious intent in isolation hence the need for PRI models (i.e., combinations of activities).

A final notable pattern was the contrast between participants' evasion approaches when techniques often opposed or were inverse of one another. Some participants encrypted files or email, believing it would prevent inspection of the contents. Other participants explicitly avoided encryption, believing it would appear suspicious. Additionally, some participants attempted to behave as normally as possible, whereas others actively increased network traffic or conducted unrelated activity to create noise. These contrasts indicate that evasion should not be treated as a single, consistent approach.

Overall, analysis demonstrates evasion was varied, selective, and often embedded in routine workplace behavior. Maliciously-tasked employee participants typically relied on a small number of accessible techniques, many centered on changing how content appeared, was organized, or was transferred, rather than advanced technical tradecraft.

Malicious intent participants tended to favor straightforward appearance- and organization-based techniques, while super malicious intent participants drew on a broader mix of methods. As several techniques also appeared in benign activity, no single technique is a strong indicator of malicious intent alone.

3.2.1. Time log analysis and reverse engineering evasion techniques

Malicious and super malicious intent participants described intentionally managing when they worked on the task to avoid detection. However, statistical analyses of participant time logs data did not support meaningful differences between benign, malicious, and super malicious intent participants in how they used their on-task time. There were no statistically significant differences between experimental conditions in the number/average duration of sessions or average time of day sessions took place, $fs < 1.02$, $ps > .35$. Consequently, there was no support for the common assumption that unusual work hours or careful time management are reliable indicators of malicious insider activity in this context.

In contrast, some rare evasion techniques appeared more feasible to reverse engineer and identify in study data. Tor use was theoretically detectable through network activity involving known Tor IP addresses or applications, and steganography tools could be identified in principle through tool names, hashes, or associated activity. However, direct examples were limited because participants often used personal devices for the most sensitive portions of these techniques. Even so, nonapproved Tor use or steganography software may warrant concern in many enterprise environments because there are few routine business uses.

Most evasion techniques, like most exfiltration techniques, were difficult to operationalize as standalone indicators. Often, the technique was visible, but available cyber data could not reliably distinguish malicious from benign use (e.g., "Deleted files using standard deletion tools", "Encrypted document container", "Renamed files benign or legitimate subject lines"). Other times, the most distinctive behavior occurred outside the monitoring boundary of organization-furnished systems. Accordingly, the exfiltration and evasion analyses indicate these behaviors are more useful as contextual amplifiers of elevated risk than as primary detection points on their own.

3.3. Sequences of cyber activity

Sequential analysis was conducted on the ordered cyber logs to determine whether participants in malicious and super malicious intent conditions used different sequences of cyber activities than participants in the benign intent condition. The n-gram analysis produced 5,957,005 different permutations of sequenced cyber activities across data from 154 employee participants. An n-gram[3] is a contiguous sequence of n items (e.g., words or characters) extracted from a text corpus, where n is the sequence length (e.g., bigrams for 2 items, trigrams for 3 items; Manning and Schütze, 1999). Candidate sequences were evaluated by comparing their frequency across experimental conditions and task periods (i.e., week before vs. week after the task). Sequences were retained as PRI sequences based on the following inclusion criteria (1) they occurred for participants in malicious or super malicious intent conditions during the week of the task, (2) did not occur with comparable frequency during the week before the task, (3) did not occur at all for benign intent participants, *and* (4) were observed in both training and testing datasets. A fifth inclusion criterion was also set requiring PRI sequences to occur more than five times during the week of the task to be confirmed as a sequence. With that threshold, no n-grams successfully differentiated benign, malicious, and super malicious intent participants. As the initial threshold produced no retained sequences, an exploratory analysis used a lower threshold requiring sequences to only occur more than once during the week of the task. Under that most lenient threshold, 202 candidate n-grams emerged as PRI sequences for malicious intent relative to benign intent participants. After removing near-equivalent and redundant sequences, five unique PRI sequences remained (Fig. 2). For super malicious intent relative to benign intent participants, 1147 candidate n-grams emerged, of which eight unique PRI sequences remained after removing near-equivalent sequences (Fig. 3). Lastly, comparing super malicious and malicious intent participants produced the same eight unique sequences.

3.3.1. Malicious intent PRI sequences

For malicious intent relative to benign intent participants, four of five PRI sequences involved file activity with one also involving web browsing activity. In Fig. 2, one representative example is that participants opened a file with the sensitive topic in the filename, completed an intervening activity (e.g., copying files from Windows File Manager to clipboard, copying text to clipboard, printing a file, or entering a command in PowerShell command line) and then opened another file with the sensitive topic in the filename. These sequences suggested malicious intent participants sometimes moved quickly between files associated with the sensitive topic while interleaving a small set of collection or handling behaviors.

In practical terms, malicious intent PRI sequences were underwhelming. Each PRI sequence identified a small proportion of malicious intent participants (no more than 7.0%), and reflected patterns already captured by existing PRIs from the broader dataset. For example, file-based PRI sequences primarily reflected interaction with sensitive files alongside already concerning handling activities, and the web browsing PRI sequence reflected attempts to access a shared document on the corporate Intranet and then receipt of access denial messages. An access denial pattern had already emerged as a useful indicator in prior data analyses. Consequently, the sequence analysis added relatively limited new detection value beyond showing a small subset of already known higher-risk cyber activities occurred in a specific order for a small percentage of malicious intent participants.

3.3.2. Super malicious intent PRI sequences

For super malicious intent relative to benign intent participants, eight retained PRI sequences fell into two broad categories: compression sequences and email sequences. In Fig. 3, one representative example is that participants engaged with a file by opening a file, creating a file, and then deleting a file within the compression application. Other retained PRI sequences followed similar logic, suggesting super malicious intent participants sometimes engaged in repeated manipulation of compressed content as part of packaging information prior to exfiltration. Like malicious intent, the super malicious intent PRI sequences were somewhat underwhelming in practical terms. Although they differentiated super malicious intent participants from both benign intent malicious intent participants, each PRI sequence identified only a small proportion of super malicious intent participants (6.3%–12.5%) and largely reflected patterns already evident in exfiltration and evasion findings. For example, super malicious employee participants more often used compressed archives and encrypted containers during exfiltration and evasion. Similarly, the email-based PRI sequences reflected ordinary handling activities occurring around a sent attachment rather than a distinctive insider risk cyber kill chain. Thus, the sequence analysis of UAM data provided limited new detection value.

4. Discussion

This behavioral experiment examined exfiltration, evasion, and cyber behavior sequencing by having employees search for, collect, and exfiltrate sensitive information from a real corporate network under benign, malicious, and super malicious intent conditions. Three overall findings emerged. First, employee participants acting with malicious intent (regardless of technical sophistication) used a wide range of exfiltration and evasion techniques, many being creative and embedded

[3] Fuller description provided in Section 2.6 (Data Preparation and Analysis).

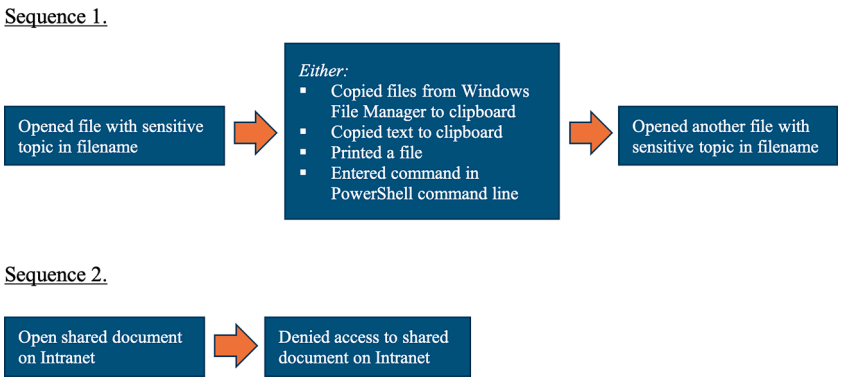


Fig. 2. PRI sequences for malicious intent employee participants.

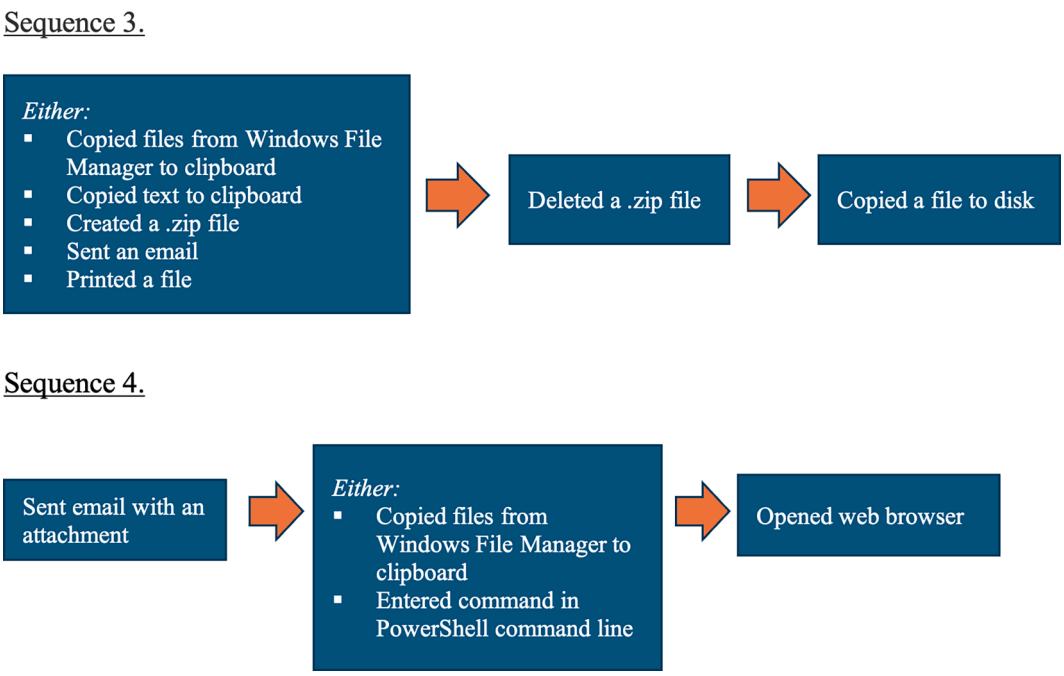


Fig. 3. PRI sequences for super malicious intent participants.

in ordinary workplace activities rather than obviously "advanced" tradecraft. Second, most techniques were difficult to operationalize as standalone PRIs because they overlapped substantially with benign intent activity or occurred outside visibility of enterprise monitoring. Third, behavioral sequence analysis identified a small subset of malicious and super malicious intent participants, but PRI sequences were rare and underwhelming and did not support sequence-based analytics as a primary detection approach. These findings suggest insider risk detection should focus less on exfiltration as a "smoking gun" and more on the broader behavioral context in which search, collection, staging, and removal occur. Findings highlight the practical importance of science-driven insider risk management, particularly in challenging assumptions widely adopted in practice without strong empirical support.

A consistent pattern across findings was that malicious insider behavior often looked operationally ordinary. Employee participants did not need to rely on illicit tools or highly distinctive paths to complete the task. Instead, many identified exfiltration and evasion techniques leveraged everyday business functions like email, file transfer, cloud storage, collaboration platforms, compression, renaming, and deletion. This pattern is important because it reflects a central property of insider risk: malicious insiders begin with legitimate access and can pursue harmful goals by blending into routine workflows. For IRPs, the challenge for detection is not simply identifying whether an activity occurred, but determining whether that activity occurred in a pattern, context, or combination that makes malicious intent more plausible.

The findings also reinforce the distinction between UAM and DLP. DLP focuses on movement of known sensitive content, whereas the most useful observations in the current study involved user behavior before, during, and around that movement. For IRPs, UAM visibility is more suitable for proactive identification of malicious search, collection, and staging, activities, while DLP is more useful to identify the point at which data is exfiltrated from the organizational perimeter.

4.1. *Insider threat exfiltration*

Exfiltration techniques showed employee participants used a broad range of methods to move information outside the organization, including email attachments, removable media, cloud storage, file transfer services, photo capture, Bluetooth, remote machines, collaboration platforms, code repositories, Tor, steganography, and public-facing infrastructure. The same objective—removing information from the organization—was achieved through multiple channels, some overt and some covert, and these channels varied in visibility to cyber sensors. For IRPs, the results indicated that exfiltration among insiders is both

more diverse and more context-dependent than many operational assumptions imply.

A key pattern in the exfiltration findings was that technical sophistication broadened the range of options used rather than fundamentally changing the nature of behavior. Employee participants in the super malicious intent condition used more varied techniques, especially those involving code repositories, cloud infrastructure, Bluetooth, collaboration tools, and other channels that plausibly blend into technical work. Super malicious intent participants did not abandon ordinary mechanisms but expanded the menu of available approaches and selected channels they believed would be less conspicuous in their role. This distinction matters for practice. For IRPs, this suggests "more cyber sophisticated" malicious insiders may not always look more overtly technical; instead, they may look more role-consistent and selective in their use of legitimate services.

Concurrently, reverse engineering results showed exfiltration is often a poor standalone detection point. Many exfiltration techniques were readily observable in cyber logs, but observations were not sufficiently discriminating. Sending email attachments, uploading files to approved transfer services, or using cloud storage are all activities that occur routinely in legitimate work. Even when metadata were available, like filenames, file sizes, or destinations, those fields did not reliably separate benign from malicious use. This demonstrates the problem was not always visibility but interpretability. For IRPs, exfiltration can often be seen, but not confidently understood in isolation.

Quantitatively, only 7 of 24 listed exfiltration techniques and subtechniques significantly differentiated conditions, with 2 reflecting greater benign rather than malicious use. The most common exfiltration channel (email attachment) was heavily used across all three conditions, reinforcing that exfiltration alone is often a weak discriminator of malicious intent. The study's emphasis on earlier detection during search and collection aligns with prior work showing that detection opportunities often arise during precursor behaviors and attack progression, rather than only at terminal outcomes such as exfiltration (e.g., Agrafiotis et al., 2015; Claycomb et al., 2012).

Only a small subset of exfiltration behaviors appeared promising as more specific indicators, particularly when they had few legitimate business uses or were rare in the environment. Examples included Bluetooth file transfer, Tor use, and steganography tool use. Even these findings should be interpreted cautiously. Their practical value lies less in serving as singular proof of malicious activity and more in adding weight to an already elevated risk picture. For IRPs, exfiltration techniques may function best as late-stage contextual amplifiers rather than the starting point of detection.

4.2. Insider risk evasion

Evasion analysis reinforced the same theme: many malicious insider evasion behaviors are subtle, role-adaptive, and difficult to distinguish from benign conduct. Employee participants described and used 76 techniques in 14 categories, including blending with normal work, altering content or file properties, avoiding specific logging mechanisms, using compressed or encrypted containers, renaming files, deleting artifacts, and using personal devices or public networks. For IRPs, this breadth indicates evasion in insider contexts is not limited to a small number of tactics. Rather, it includes a wide range of decisions about how to search, collect, disguise, organize, stage, and transmit information while reducing the likelihood of standing out.

A clear evasion pattern was malicious intent participants relied on ordinary file handling as a form of concealment. Malicious and super malicious participants disproportionately created compressed archives, encrypted document containers, changed file names, changed file extensions, and deleted or cleaned up files. These are not advanced techniques, but are operationally meaningful because they reflect a view of files as movable and modifiable containers rather than only content to be read. For IRPs, evasion was not about defeating monitoring through specialized malware or anti-forensic tooling. It was about using common user behaviors to alter how sensitive material appeared, was stored, and could be interpreted.

A second notable pattern was evasion techniques frequently contradicted one another. Some malicious intent participants encrypted content to reduce the ability of the organization's information security and cybersecurity teams to inspect content, whereas others avoided encryption because they believed it would attract suspicion. Some employee participants tried to behave as normally as possible; others deliberately created noise to hide their actions in larger volumes of activity. This finding underscores evasion should not be conceptualized as a single script. Malicious insiders may adopt opposite strategies based on their beliefs about monitoring, technical expertise, available time, and assumptions about what looks suspicious in their work environment. For IRPs, this means that detection logic built around one narrow model of evasion is likely to miss meaningful variation.

The time management results are especially relevant for practice because they challenge a longstanding assumption in insider risk detection. Although malicious employee participants described intentionally managing when they worked on the task, there were no statistically significant differences in the number of sessions, duration of sessions, or time of day across conditions. In this dataset and others researched by this team, odd work hours and unusual time management did not emerge as useful indicators of malicious intent. This finding suggests that IRPs should be cautious about using out-of-hours work as a generic indicator unless it is extreme, role-inconsistent, or clearly coupled with other concerning behavior. In many environments, employees work irregularly for benign reasons; overemphasizing time-based heuristics is likely to inflate false positives without materially improving detection.

4.3. Insider risk sequence-based detection

Sequential analysis produced a mixed picture. Thirteen retained sequences are potentially useful because they identified a small number of malicious and super malicious intent employee participants. Quantitatively, none of the 5,957,005 candidate permutations met the original confirmation threshold, and only 13 sequences met the most lenient exploratory threshold. The retained sequences identified only a small proportion of malicious or super malicious intent participants (3) reinforcing their limited practical value. The findings do not support the existence of a credible, widely applicable insider risk cyber kill chain. Although a small number of higher-risk activities may sometimes occur in a meaningful order, the variability of insider behavior is too great for sequence-based detection to serve as a primary analytic model.

These findings support a more parsimonious operational approach. Sequence-based analysis requires additional analytic complexity, including preserving event order, defining sequence structure, and maintaining more specialized detection logic. When that added complexity identifies very few risks and provides little value beyond already concerning individual activities or simpler combinations of activities, it is more practical for IRPs to prioritize less cumbersome approaches (e.g., combinations of cyber activities within a specific period).

Sequence types differed between malicious and super malicious intent participants. Malicious participants were characterized primarily by file-to-file movement and a web-based access denial sequence consistent with rapid traversal and handling of sensitive materials. The access denial sequence, by contrast, reflected attempts to reach information beyond current permissions, suggesting opportunistic searching behavior exceeding ordinary work boundaries.

Super malicious intent participants' sequences centered on compression applications, archive manipulation, and email workflows. Practically, super malicious intent participants appeared to engage more repeatedly in packaging, altering, deleting, and restaging archived content. This pattern aligns closely with exfiltration and evasion findings, where technically sophisticated participants used compression and

encrypted containers more frequently. For IRPs, this suggests that super malicious intent participants differed less in the broad goal of exfiltration and more in how deliberately they handled intermediate artifacts during staging and removal. Compared with malicious intent participants' file-to-file movement, the super malicious intent participants' behavior appeared more oriented toward preparation, transformation, and post-action file handling.

Another important pattern was results which did not emerge. Although employee participants clearly engaged in search, collection, and information handling, those behaviors did not consolidate into stable, reusable ordered chains across individuals. Instead, the most diagnostic sequences tended to remain within narrower activity families, like file handling, compression, or email. This suggests that real insider behavior may be too variable to support wide-ranging sequence models, even when smaller intrafamily sequences retain value.

Overall, for IRPs the insider risk sequencing findings support a restrained conclusion: explicit ordered combinations of behaviors can identify a very small number of risks but sequence-based analytics are unlikely to function as a scalable primary approach. They are better conceptualized as supplementary detectors that may add confidence in a layered analytic program. It is important to highlight, the n-gram analysis focused on sequences within the course of a week. The findings do not represent extended efforts over the course of months, nor ML approaches with the goal of optimizing predictive accuracy. Consequently, future research should examine sequences over longer time periods to determine whether sequences can offer additional value beyond the limited signal observed in this study.

4.4. Overall implications for practice

Taken together, the analysis findings suggest IRPs should place less weight on exfiltration as a primary proactive detection point and more weight on the broader behavioral context in which search, collection, staging, and removal occur. Often, exfiltration was either difficult to distinguish from benign work or occurred partly outside enterprise visibility. More operationally useful signals were often earlier and surrounding behaviors, particularly unusual search, sensitive file traversal, repeated access denials, archive creation or manipulation, and other forms of local file handling that helped clarify intent of the cyber activity. In this study, employee participants were primarily in technical roles, with most having government security clearances as contractors. Consequently, findings are most applicable for IRPs operating in technically oriented, "knowledge worker" organizations with enterprise monitoring, including DIB, government, intelligence, and parts of most critical infrastructure organizations (e.g., financial sector). More caution is required when interpreting these findings for organizations with less frequent cyber monitoring, in less security-centric settings, and with workforces whose frequent activities do not center on search and collection of information on computers.

The findings also suggest that insider risk detection should be cumulative rather than singular. Ordinary workplace actions such as sending attachments, using cloud services, compressing files, renaming files, or deleting artifacts rarely function well as standalone indicators. Their value increases when they occur in combination, within a similar timeframe, or in a pattern that is inconsistent with an employee's role or baseline (i.e., recent behavior). For practice, the findings support detection analytics that prioritize specific combinations of higher-risk activities identified through empirical behavioral and data science, rather than broad anomaly detection or any single exfiltration or evasion technique in isolation.

The study highlights an important visibility gap for IRPs. Some participant behavior, particularly during evasion and exfiltration, shifted to personal devices or otherwise occurred outside enterprise-managed system visibility. These blind spots are not only UAM solution limitations but reflect the operational reality that no single cyber sensor fully captures insider behavior across all devices, networks, and contexts (hence why this study did not rely on UAM alone but collected post-task questionnaire, interview, and network-level data beyond the endpoint). For practice, this challenge reinforces the need for IRPs to negotiate access to a range of cyber data and alerts beyond traditional UAM, to include supplementary data from network, cloud, and mobile sensors. It also highlights why IRPs must interpret cyber data within a broader insider risk framework combining human, organizational, physical, and cyber sensors rather than relying on cyber monitoring alone.

Finally, the study reinforces the need for insider-specific visibility and analytic models. Endpoint-level observation of routine user activity was often more informative than network-centric visibility alone, and external threat frameworks did not map cleanly onto insider behavior involving authorized access and ordinary workflows. Organizations are therefore likely to gain greater value from detection approaches tailored to role context, legitimate service use, and subtle behavioral change than from models assuming a clear insider kill chain or relying heavily on generic heuristics.

Importantly, the research design and findings should not be interpreted as suggesting cyber data alone are sufficient for insider risk or threat detection. Instead, the study focuses on cyber observable behavior because UAM is one of the few sensors IRPs often directly manage. In practice, more effective and proactive insider risk detection combines human, organizational, physical, and cyber sensors rather than cyber sensors alone.

4.5. Future research

Several directions for future research follow from this behavioral experiment. First, the exfiltration techniques and evasion catalog should be tested in additional organizational environments, especially those with different technical restrictions, workforce compositions, and monitoring boundaries. The usefulness of specific techniques may vary substantially depending on whether removable media are allowed, collaboration platforms are common, technical staff regularly use code repositories, or employees work in highly controlled environments.

Second, future work should evaluate how exfiltration and evasion behaviors interact with earlier search and collection indicators in integrated risk models. Results strongly suggest that later-stage techniques are more informative when interpreted in context. Formal models that combine earlier and later indicators may provide a better balance of recall and precision than either approach alone.

Third, additional research is needed on the role of technical expertise. Super malicious intent employee participants did not simply become "more malicious" versions of the malicious intent employee participants; rather, they exhibited different behavioral emphases, particularly around compression, staging, and role-consistent exfiltration paths. Future work should examine how occupational role, technical specialization, and familiarity with enterprise monitoring shape insider decision-making and observable behavior.

Fourth, sequence analysis warrants continued but targeted development. The current findings suggest broad, highly repeated activity chains are unlikely to offer scalable solutions. However, narrower sequences within activity families, individually baselined sequence changes over time, or sequences tied to the journey of a file or data object may prove more useful. Future research should also examine whether sequence-based approaches perform better in environments where employee behavior is more constrained and task variability is lower.

Fifth, a critical next step for the insider risk community is the continued development of potential risk models combining human, organizational, physical and cyber sensors. The present study does not report the results of such analyses because they contain sensitive operational details not appropriate for public dissemination. Furthermore, those broader modeling efforts require a different analytic framework out of scope for the current study. Those broader models are most

valuable and should be developed and tuned within operational IRPs using their own real data, local context, and risk priorities.

Sixth, in this study, participants were intentionally provided with six hours of task time to distribute across five business days as they wished. Although this six-hour window was tested in advance, additional or different forms of malicious insider behavior could unfold over much longer periods (e.g., so-called "low and slow" approaches). To help interpret the impact of this design decision, participants were also asked in the post-task questionnaire whether they had enough time to complete the task and what they would have done differently if given more time. Only 23.3% of participants indicated that they would have benefited from more time. Consequently, the findings are most applicable for malicious insider activity occurring over shorter operational windows. Future research should examine whether longer timeframes yield different exfiltration, evasion, or sequencing patterns.

Finally, future studies should address visibility gaps created by personal devices and off-network activity. Several more distinctive or sensitive exfiltration and evasion behaviors were determined primarily through employee participant self-report rather than cyber logs. This limitation is not unique to the behavioral experiment; it reflects an operational reality for many organizations. Better integration of cyber, human, and organizational data therefore remains essential to maturing insider risk detection.

5. Conclusion

This behavioral experiment provides empirical examination of exfiltration, evasion, and behavioral sequences in malicious insider activity using ground-truth data from a live real organization. Rigorous analyses show malicious insiders use a broad and often ordinary-looking set of techniques to remove information and reduce detection, most of those techniques are difficult to operationalize as standalone indicators, and explicit activity sequences can identify a small number of risks with relatively high confidence but limited coverage. Overall, results argue for a more proactive and context-sensitive model of insider risk detection emphasizing earlier behavioral indicators, interprets exfiltration and evasion within broader patterns of activity, and uses sequence-based analytics selectively rather than as a primary line of defense.

Funding

This work was supported by internal corporate research funding from The MITRE Corporation.

Data statement

The data that support the findings of this study are not publicly available because they contain potentially identifiable participant information, sensitive organizational information, and confidential details relating to insider risk/threat detection protocols. In addition, participants in the malicious and super malicious conditions were de-identified under Institutional Review Board approval to participate in activities that could otherwise violate organizational information security rules without the Study's approved waiver.

CRediT authorship contribution statement

James Doodson: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Deanna D. Caputo:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **James F. Murphy:** Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Dr. Paul Seymer and Mr. Brian Shaw for early contributions to the hand-annotation and reverse engineering of exfiltration and evasion techniques; and Dr. Nathaniel Ratcliff for early contribution to hand-annotation of exfiltration and evasion techniques.

Appendix

Appendix A
Catalog of Evasion Techniques and Percentages by Condition.

	Condition		
Evasion Technique	Benign	Malicious	Super Malicious
Blended Activity			
1. Added malicious code before or after legitimate code	0.0	0.0	3.9
2. Behaved as "normally" as possible	0.0	2.0	1.0
3. Combined topic of interest with job-related search terms	0.0	0.0	1.9
4. Completed task alongside legitimate work activity	0.0	1.0	0.0
5. Conducted unrelated searches alongside searches for topic of interest	0.0	2.0	1.9
6. Increased network traffic to hide malicious activity	0.0	1.0	0.0
Converted Content			
7. Converted file from one format to another (e.g., Microsoft Word to .txt)	0.0	0.0	1.0
8. Corrupted file then reversed corruption once delivered	0.0	0.0	1.0
9. Encoded content (e.g., Base 64 content)	0.0	0.0	1.0
10. Printed to .pdf rather than downloading files	2.2	0.0	1.9
11. Used steganography (audio)	0.0	0.0	1.0
12. Used steganography (image)	0.0	0.0	3.9
Collected Content			
13. Automated script to gather information	1.1	0.0	1.9
14. Avoided sending more sensitive material	1.1	1.0	0.0
15. Copied and pasted material rather than downloaded materials	1.1	1.0	1.9
16. Took screenshots	1.1	3.0	1.0

(*continued on next page*)

Appendix A (*continued*)

	Condition		
Evasion Technique	Benign	Malicious	Super Malicious
Disrupted and Avoided Log/Data Collection			
17. Accessed content within web browser applications (e.g., Office 365 Online)	0.0	0.0	1.0
18. Avoided backups	0.0	1.0	0.0
19. Avoided corporate cloud storage	0.0	0.0	1.0
20. Avoided downloading	0.0	4.0	4.9
21. Avoided key terms in search	0.0	2.0	1.0
22. Booted computer to another operating system	0.0	0.0	1.0
23. Bypassed Intranet authentication processes	0.0	0.0	1.9
24. Cleared browser history	0.0	3.0	1.0
25. Conducted task on remote computers	0.0	0.0	1.0
26. Deleted emails from sent emails	0.0	0.0	1.9
27. Deleted files using standard deletion tools	1.1	3.0	4.9
28. Deployed a virtual machine	0.0	0.0	1.9
29. Emptied recycle bin	0.0	2.0	4.9
30. Previewed rather than downloaded content (e.g., hovering over hyperlinks for a preview)	0.0	2.0	1.0
31. Turned off communication apps (e.g., Skype)	0.0	1.0	0.0
32. Used privacy-focused web browsers	0.0	0.0	1.0
33. Used private browsing window (e.g., Incognito)	0.0	0.0	1.9
34. Used keystroke masking software	0.0	1.0	0.0
35. Used non-local programs (e.g., applications stored on USB device)	0.0	0.0	1.0
36. Used multiple browsers	0.0	1.0	0.0
Encryption			
37. Skipped encrypting emails and files	0.0	0.0	1.0
38. Encrypted document container (e.g., encrypted .zip)	0.0	13.1	21.4
39. Encrypted email	0.0	1.0	2.9
40. Used strong passwords	0.0	0.0	1.0
Engaged with Other People			
41. Asked colleagues to complete the work for them	1.1	0.0	0.0
42. Avoided engaging with other people	0.0	1.0	1.9
43. Conducted searches for people with expertise, not the topic itself	1.1	1.0	3.9
44. Contacted authors or subject matter experts within organization	5.4	2.0	3.9
45. Developed plausible benign explanations for malicious activity	0.0	3.0	3.9
Labelled Content			
46. Changed file extension (e.g., from .zip to .pdf)	0.0	2.0	4.9
47. Labelled emails with benign or legitimate subject lines	0.0	5.1	1.0
48. Placed files into a folder labelled "Personal"	0.0	0.0	1.0
49. Renamed files with benign or legitimate names	2.2	30.3	22.3
Learning and Reconnaissance			
50. Conducted external search on topic to refine internal search (e.g., Bing, Google Search)	4.3	1.0	5.8
51. Reviewed material on common computer monitoring software and approaches	0.0	1.0	0.0
Managed Time			
52. Emailed material over time	0.0	0.0	1.0
53. Kept large gaps between malicious working sessions	0.0	0.0	1.0
54. Moved swiftly between searches to avoid lingering on a specific search	0.0	0.0	1.0
55. Worked in regular work hours	0.0	1.0	0.0
56. Worked out-of-hours	0.0	2.0	1.0
57. Worked very quickly	0.0	1.0	1.0
Obfuscated Content			
58. Changed file content to avoid being an exact match (e.g., file size)	0.0	0.0	1.9
59. Embedded documents in a bigger document	0.0	1.0	0.0
60. Hid more detailed information data in less visible parts of documents	0.0	0.0	1.0
61. Performed unrelated activity on computer to generate logs	0.0	0.0	1.0
62. Sent material alongside released documents	0.0	0.0	1.0
63. Started with a general search term, and gradually moved closer to the topic	0.0	2.0	1.0
64. Used benign-looking commands on PowerShell or Windows Command Line	0.0	0.0	1.0
Purged Content			
65. Changed document properties (e.g., "last saved by")	0.0	0.0	3.9
66. Removed document markers (e.g., "For Internal Use Only")	0.0	3.0	0.0
Stored and Organized Content			
67. Combined files into a single document (e.g., single long .pptx)	1.1	5.1	0.0
68. Created a document container (e.g., zip)	6.5	19.2	24.3
69. Divided larger files into smaller files	1.1	1.0	0.0
70. Stored files in single location on machine	0.0	0.0	1.0
71. Stored files alongside legitimate work files	0.0	0.0	1.0
Transferred Content			
72. Sent files over Tor network	0.0	0.0	1.0
73. Split files across multiple emails (e.g., to avoid flagging on network traffic as large data transfer)	2.2	2.0	4.9
Used Other Devices			
74. Used multiple organization-furnished computers	0.0	0.0	1.0
75. Used personal computer	0.0	1.0	2.9
76. Used public Internet (e.g., Wi-Fi at a coffee shop)	0.0	2.0	1.9

Data availability

The data that has been used is confidential.

References

Agrafiotis, I., Nurse, J.R.C., Buckley, O., Legg, P., Creese, S., Goldsmith, M., 2015. Identifying attack patterns for insider threat detection. Comput. Fraud Secur. 2015 (7), 9–17. https://doi.org/10.1016/S1361-3723(15)30066-X.

Alzaabi, F.R., Mehmood, A., 2024. A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods. IEEE Access 12, 30907–30927. https://doi.org/10.1109/ACCESS.2024.3369906.

Black, M., Yeung, J., Yeung, D., 2022. Insider Threat and White-Collar Crime in Non-Government Organisations and Industries. RAND Corporation. Retrieved online from: https://www.rand.org/content/dam/rand/pubs/research_reports/RRA1500/RRA1507-1/RAND_RRA1507-1.pdf [Accessed June 18th, 2026].

Cappelli, D.M., Moore, A.P., Trzeciak, R.F., 2012. *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes* (Theft, Sabotage, Fraud). Addison-Wesley.

Caputo, D.D., 2020. Insider threats: leveraging the benefits of behavioral science research. In: Meloy, J.R., Hoffmann, J. (Eds.), International Handbook of Threat Assessment, 2 edition. Oxford University Press, New York, NY, USA, pp. 301–321.

Claycomb, W.R., Huth, C.L., Flynn, L., McIntire, D., Lewellen, T., 2012. *Chronological Examination of Insider Threat Sabotage: Preliminary Observations* (CMU/SEI-2012-TN-026). Carnegie Mellon University, Software Engineering Institute. Retrieved online from: https://www.sei.cmu.edu/library/chronological-examination-of-insider-threat-sabotage-preliminary-observations.

Cottrell, N.B., 1972. Social facilitation. In: McClintock, C. (Ed.), Experimental Social Psychology. Holt, Rinehart and Winston, New York, NY, USA, pp. 185–236.

Elbasheer, M., Akinfaderin, A., 2025. User-based sequential modeling with transformer encoders for insider threat detection. arXiv. https://doi.org/10.48550/arXiv.2506.23446.

Glasser, J., Lindauer, B., 2013. Bridging the gap: a pragmatic approach to generating insider threat data. In: 2013 IEEE Security and Privacy Workshops, pp. 98–104. https://doi.org/10.1109/SPW.2013.37.

Greitzer, F.L., Hohimer, R.E., 2011. Modeling human behavior to anticipate insider attacks. J. Strateg. Secur. 4 (2), 25–48. https://doi.org/10.5038/1944-0472.4.2.2.

Greitzer, F.L., Purl, J., Leong, Y.M., Becker, D.E.S, 2018. SOFIT: sociotechnical and organizational factors for insider threat. In: 2018 IEEE Security and Privacy Workshops (SPW). IEEE, pp. 45–52. https://doi.org/10.1109/SPW.2018.00035.

Harilal, A., Toffalini, F., Castellanos, J., Guarnizo, J., Homoliak, I., Ochoa, M., 2017. TWOS: a dataset of malicious insider threat behavior based on a gamified competition. In: Proceedings of the 2017 International Workshop on Managing Insider Security Threats, pp. 45–56. https://doi.org/10.1145/3139923.3139929.

Henchy, T., Glass, D.C., 1968. Evaluation apprehension and the social facilitation of dominant and subordinate responses. J. Pers. Soc. Psychol. 10 (4), 446–454. https://doi.org/10.1037/h0026814.

Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., Ochoa, M., 2020. Insight into insiders and IT: a survey of insider threat taxonomies, analysis, modeling, and countermeasures. ACM Comput. Surv. (CSUR) 52 (2), 1–40. https://doi.org/10.1145/3303771.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., et al., 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl. Inf. Syst. 64 (12), 3197–3234. https://doi.org/10.1007/s10115-022-01756-8.

Liu, S., Kuhn, D., 2010. Data loss prevention. IT Prof. (IEEE) 12 (2), 10–13. https://doi.org/10.1109/MITP.2010.52.

Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA.

Manoharan, P., Hong, W., Yin, J., Wang, H., Zhang, Y., Ye, W., 2024. Optimising insider threat prediction: exploring BILSTM networks and sequential features. Data Sci. Eng. 9 (4), 393–408. https://doi.org/10.1007/s41019-024-00260-z.

McCambridge, J., De Bruin, M., Witton, J., 2012. The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. PloS one 7 (6), e39116. https://doi.org/10.1371/journal.pone.0039116.

National Insider Threat Task Force, 2018. NITTF tech bulletin 20180527: how committee on national security systems directive (CNSSD) 504 defines UAM. Retrieved online from: https://www.dni.gov/files/NCSC/documents/nittf/20181022-UAM-Definition.pdf.

National Institute of Standards and Technology, 2020. *Data loss prevention* (*DLP*). Retrieved online from. https://csrc.nist.gov/glossary/term/data_loss_prevention.

Nurse, J.R., Buckley, O., Legg, P.A., Goldsmith, M., Creese, S., Wright, G.R., Whitty, M., 2014. Understanding insider threat: a framework for characterising attacks. In: 2014 IEEE Security and Privacy Workshops, pp. 214–228. https://doi.org/10.1109/SPW.2014.38.

Orne, M.T., 1962. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. Am. Psychol. 17 (11), 776–783. https://doi.org/10.1037/h0043424.

Robinson, M.K. (2025). *When trust fails: evolving data exfiltration methods in insider threat cases.* Retrieved online from: http://doi.org/10.2139/ssrn.5784170.

Taylor, P.J., Dando, C.J., Ormerod, T.C., Ball, L.J., Jenkins, M.C., Sandham, A., Menacere, T., 2013. Detecting insider threats through language change. Law. Hum. Behav. 37 (4), 267–275. https://doi.org/10.1037/lhb0000032.

U.S. Department of Defense, 2024. DoD Instruction 5205.16: The DoD Insider Threat Program. December 20. Office of the Under Secretary of Defense for Intelligence and Security. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/520516p.pdf.

U.S. Office of the Director of National Intelligence, 2024. Insider threat mitigation for U.S. critical infrastructure entities. Retrieved online from: https://www.dni.gov/files/NCSC/documents/nittf/20240926_Insider-Threat-Mitigation-for-US-Critical-Infrastructure.pdf.

Whitty, M.T., Ruddy, C., Keatley, D., Butavicius, M., Grobler, M., 2024. The prince of insiders: a multiple pathway approach to understanding IP theft insider attacks. Inf. Comput. Secur. 32 (4), 509–522. https://doi.org/10.1108/ics-11-2023-0210.