Abstract

This paper discusses the use of active probes to detect insider threats ahead of their manifestation as opposed to the current detection techniques that have generally indicated the presence of a threat post hoc. Users become motivated to engage in insider theft due to a variety of reasons such as greed, disgruntlement, anger, patriotism, and social justice. Once motivated they seek opportunities for data theft, are careful to avoid detection, and often rationalize their behavior which allows them to blur the line between moral and immoral action. Our experimental protocol involves presenting probes to users, which serve as cues that signal the opportunity to steal data (signaled by active probes). We test the effectiveness of the probes by measuring user search and exfiltration behavior before and after the introduction of the probe. The effects of two different probes on student exfiltration behavior were tested in a laboratory setting. Both probes resulted in an increase in curiosity and theft-related behaviors.

Keywords (Required)

Insider threats, active probes, behavioral security, cyber security, information security

Introduction

Data theft has become a key security issue particularly for organizations holding sensitive, propriety information. The risk of a security breach exists both from outside (Goel et al., 2017) and within organizations, with the latter often being more difficult to detect and defend against. The challenge in dealing with the “insider threat” is multifaceted. Individuals engaging in malicious behavior from within an organization readily have access to particularly sensitive information, have privileged access to the inner workings of an organization, and are often granted a level of trust that allows them to engage in data theft in a particularly stealthy and efficient fashion. Additionally, imposing restrictive security protocols run counter to the desire to maintain a cohesive and collegial culture within an organization, making the detection and apprehension of malicious insiders an especially sensitive process.

Information theft is not a new problem, however, the means by which data is stolen have evolved considerably. The proliferation of electronic and media storage options, increasing network connectivity and emerging communication technologies have all increased the potential for data theft, presenting a formidable challenge to those with a vested interest in preventing such actions (Billur, Tayur, & Mahesh, 2015). Unfortunately, as the result of the exponential growth and increased sophistication within the digital realm, many of the traditional surveillance techniques that are no longer adequate. For instance, the small footprint of storage devices can make it easy for infiltrators to evade physical detection. Pvt. Chelsea (formerly Bradley) Manning could exfiltrate all the information she stole by putting it on a compact disk disguised as a music CD (Erbacher, 2011). The surveillance and security strategies of intelligence agencies
Detecting Insider Threats using Active Indicators

are based on existing vectors of attack, however, with emerging technologies these vectors are changing, leading to increased vulnerability of organizations to data exfiltration.

The motivation for data theft varies considerably, with personal disgruntlement, social justice, patriotism, and greed being some of the common motives. For instance, workers who feel that they have been treated unfairly by their organization may steal data to bring to a competitor. Insiders who engage in malicious insider activity (e.g. sabotage, theft, and apathy) often rationalize their behavior as justified, perhaps to correct a perceived transgression (e.g. “the company is not treating me fairly”) or out of a sense of entitlement (“I contributed heavily to this project so should be able to retain ownership”) (Greitzer and Hohimer, 2011). Such defensive cognitive processes may explain how individuals can engage in such malicious behavior without acknowledging the consequences of their actions (Baumeister, Dale, & Sommer, 1998; Mazar, Omir & Ariely, 2008). This mechanism for psychological defense in the face of unethical actions makes the insider threat problem particularly poignant. Even if the probability of an individual engaging in such malicious actions is low, and notwithstanding the fact that most insider data theft acts are minor; the sheer number of potential threats aggregate to a strong, palpable risk to organizations and national security.

The present literature regarding the insider threat phenomenon typically cites most common motives as personal financial gain, disgruntlement, and a sense of entitlement (Moore et al., 2011). Employees may be able to sell information to competitors, hackers, or criminal groups, or those who have taken a job with another company may steal data right before their departure, in the hope that access to the data will provide them leverage for a favorable standing within the new organization. Many inside attackers seem to act out of feelings of resentment over perceived injustices by their employer, either in terms of inequitable distribution of rewards (i.e., distributive injustice) or unjust treatment (e.g., procedural injustice; Willison and Warkentin, 2009). To such individuals, often the theft of an organization’s secrets may be perceived as a way of restoring equity and fairness. Insider threats may also be motivated by feelings of social injustice (rather than personal justice), whereby workers seek to redress what they perceive as immoral acts on the part of their organization. Additionally, insider theft or attacks may also be motivated by employees’ feelings of proprietorship over data created through their own work (Azaria, Richardson, Kraus, & Subrahmanian, 2014).

Detecting and countering insider behavior has traditionally been achieved through forensic data analysis, where sensors are embedded in the network and on individual computers to collect data on user actions (e.g. file transfers, logins, USB usage), which are then analyzed for aberrant, potentially malicious behavior (Raytheon, 2009; Whitty, 2014). Several forensic data analytic techniques have been used to track insider behavior, such as Eldardiry et al.’s (2013) sensor fusion technique to analyze anomalies in user behavior. This method involves flagging anomalous patterns of behavior, such as when a user exhibits behavior that reflects activities of a group to which he/she does not belong, when the user’s behavior differs from that of peers in his/her own group, among other deviations from normative behavior that form a threat detection system. They use logging patterns, device and data access, email metadata, search history, and standard data mining techniques in their threat analysis (Eldardiry, 2013). Legg et al. (2015) similarly use login, USB, email, web and file usage data to determine anomalies in the system, performing a series of analyses based on hourly and daily usage patterns. Additionally, Shultz (2002) suggests a framework for insider behavior that utilizes regression/data mining on multiple data vectors including usage patterns, actions, meaningful errors (e.g. deleting log files), verbal behavior (e.g. hatred in emails, or hostility towards employer), and personality characteristics. A report from the defense company Raytheon (SIFMA, 2014) on insider threats provides general guidelines for protection using a risk-based approach, including valuing assets, profiling individuals, investigating previous incidents, conducting surveillance on activities, and selectively analyzing data based on risk. In most of the research on insider threat based on data analytics, the data collected for analysis is very similar (i.e. file usage/access/transfer patterns, email metadata, search history, etc.) and represents a static state of affairs regarding the paradigm in insider threat detection.

The fundamental challenge with most of these data analytic systems is that they rely on data analysis post-incident, rather than a priori. Detection may take weeks or even months, whereas mitigation requires quickly controlling the damage and attempting to fix the breach to prevent future occurrences. Consequently, such methods often are of limited utility, as their efficacy in detecting insider threats is
largely nullified given that the damage has already been done. Passive post-hoc analysis is not sufficient; identifying individuals who pose risks a priori is of paramount importance to preventing insider theft. Although complex problems rarely are rectified by simple solutions or answers, there are several promising directions in which a priori insider threat detection may be both viable and efficacious. The use of empirically grounded methods that consider the motivations, traits, situational triggers, and nuances that compel an individual to engage in malicious insider behavior in context holds the most promise in developing a robust insider threat detection system that is efficacious a priori.

**Conceptual Framework: Behavioral Probes to Identify Insider Threats**

Insider theft is intentional behavior and can be conceptualized as a function of motivation, capability, and opportunity (Schultz, 2002). Inside attackers are motivated by some means to steal data from their employer, they have (or perceive themselves to have) the technical skills and expertise to successfully complete an attack, and the situation provides the opportunity for theft to occur. In addition, attackers may rationalize their behavior as justified, perhaps to correct a perceived company transgression (“the company is not treating me fairly”) or out of a sense of entitlement (“I contributed heavily to this project so should be able to retain ownership”). Rationalization explains how people who may see themselves as moral and honest engage in minor acts of data exfiltration. Cumulatively, minor acts may result in significant damage to the firm, but the individual acts are easy to self-justify (Mazar, Omir & Ariely, 2008). In this manner, parallels can be drawn between insider theft and other forms of counterproductive work behavior, such as fraud and sabotage. For example, the “fraud diamond” model of corporate fraud identifies the four elements of fraud as incentive (motivation), opportunity, rationalization, and capability (Wolfe & Hermanson, 2004; see also Cressey, 1973). Similarly, we have adopted a four-component model of insider threat behavior: motivation, capability, opportunity, and rationalization. These four elements of threat behavior will be used to guide the development of probes.

Understanding the motivation for inside attacks is important because it provides insight to the specific environmental stimuli that give rise to motivated behavior. For example, people motivated by financial gain are likely to respond to stimuli that signal the availability of valuable assets. The literature on insider threats identifies various motives for data exfiltration, with the most commonly cited motives being personal financial gain, disgruntlement, and a sense of entitlement. Employees may be able to sell information to competitors, hackers, or criminal groups. Employees who have taken a job with another company may steal data right before their departure, in hopes that the data will boost their performance or improve their status in their new organization. Many inside attackers seem to act out of feelings of resentment toward the company over perceived injustice (Willison and Warkentin, 2009). These are disgruntled employees who feel that they have been treated unfairly by their employer, either in terms of inequitable distribution of rewards (i.e., distributive injustice) or unjust treatment (i.e., procedural injustice) and are looking to exact revenge. To them, stealing company secrets may be a way of restoring perceptions of equity and fairness. Insider threat may also stem from feelings of social injustice. Workers may perceive immoral acts on the part of their company or organization and seek to redress the wrongdoing. Finally, insider theft or attacks may also be motivated by feelings of proprietorship over data created by employees through their own work.

Opportunity plays an important role in insider theft. Opportunity arises when the employee has access to information that has value, either through their own work or through access through corporate network. Employees must choose when and how often to attack (Martinez-Mayano, Rich, Conrad, Andersen, & Stewart, 2008), and the decision process involves a benefit-cost analysis. Inside attackers weigh the potential outcomes of their actions against the probability and consequences of getting caught. The likelihood of theft increases when the perceived rewards of action outweigh the cost and consequences.

An employee must also have the capability and skills to carry out an attack. This may include the technical computer skills to download secure files or it may also be related to the position a person in the organization hierarchy. A VP of a company has a greater ability and coverage to influence others and take decisions than a lab technician.
The theory of self-concept maintenance posits that “for certain types of actions and magnitudes of dishonesty, people can categorize their actions into more compatible terms and find rationalizations for their actions” (Mazar et al., 2008, p. 634). For example, people routinely add small expenses to expense reports or claim certain deductions on their tax returns because they feel that there are other expenses that they have not been compensated for or that “everyone does it.” Insider attackers may feel that their behavior is justified because they deserved more reward for their hard work or that their voice was not being heard.

Probes may target any of the four behavioral components of insider threat or combinations thereof. For example, a probe targeting personal gain might involve a notice highlighting the value of particular company documents, thereby attracting the attention of a malicious insider’s to discovering and accessing those documents. Probes that signal opportunity might involve notice that new software is being introduced to detect illegitimate file search behaviors or that monitoring software is suspended. The former signals increased scrutiny and should lead to decreased searching by attackers; the latter signals decreased scrutiny (more favorable opportunity) and should lead to increased suspicious activity. Probes that signal capability might involve notice that additional ways of accessing information, such as Bluetooth, are available, allowing people the chance to use various skills at their disposal. Finally, probes designed to minimize rationalization could take advantage of the literature showing that minor forms of cheating and dishonesty decrease in the presence of morality reminders (Mazar et al., 2008). Email reminders of ethical code of conduct or company loyalty statements may result in decreased activity by potential attackers.

The fraud diamond model suggests that data exfiltration follows a defined cycle from tipping point to trigger point wherein the user first is motivated by the environment to exfiltrate data. The motivations may be internal to the organization, such as a poor review, reassignment to a different task, and company practices, or external to the organization, such as radicalization by propaganda, personal financial difficulties, or national politics. We call this the “trigger point.” Subsequently, the user engages in reconnaissance activity to determine data locations and ways to gather data clandestinely. Once the user finds an opportunity he pulls the trigger and exfiltrates data; we call this the “tipping point.”

In the present research, we design opportunity probes that signal opportunity for data exfiltration and study their effects on malicious actions. Specifically, we introduce probes that may signal to malicious insiders that the chances of detection are very low or that relevant data is accessible, and then observe subsequent changes in their search and exfiltration behavior. According to the fraud model, these “cues” should be the “tipping point” for those with motive and capability.

**Experimental Design**

Our experimental protocol was designed to manipulate motivation and opportunity for data exfiltration and test the effectiveness of active probes designed to detect malicious insider behavior a priori. Within each experiment, probes based on the results of pilot testing were introduced and changes in user behavior measured.

In the experiments students were brought to the lab to perform specific tasks with the expectation that they will be compensated for the quality of their work performance; the better their performance, the higher their reward. The task students were asked to perform was to gather intelligence on hacker groups and record their findings on templates that were provided in a shared drive.

We developed an application through which we sent messages to the students’ computers, which appeared in front of the windows on their screen so that they were seen instantaneously. Experimental manipulations were sent as messages through this application. We introduced two kinds of manipulations: 1) contextual prompts that elicit motives or psychological states related to insider threat, and 2) probes that signaled increased or decreased opportunity for data exfiltration. The objective of the contextual prompt was to increase the number of people motivated to exfiltrate data. Given the performance-based incentive, it was assumed that participants who were performing poorly may be more motivated to exfiltrate data. Hence, half of the participants were sent a message indicated that their performance was subpar.
Based on our fraud diamond model, we expected that probes that signal opportunity for data theft would increase malicious actions. We tested two probes, one indicating that security controls were inoperable and one signaling access to valuable data. We expected that exfiltration behavior would increase after the introduction of the probes, especially for those motivated to improve performance. Accordingly, we have two hypotheses:

**H1:** Insider threat behavior will be influenced by context, such that instances of insider threat behavior will be higher for those who are told that they are performing poorly.

**H2:** Insider threat behavior should increase after introduction of a probe that signals opportunity for exfiltration.

**H3:** Context and probe will interact, such that insider threat behavior will increase most after introduction of a probe for those who are told that they are performing poorly.

We tested the effectiveness of the two different probes, but had no a priori expectations that one was superior to the other.

**Methodology**

*Participants:* Forty-seven undergraduates enrolled in digital forensics and information systems courses participated in the experiment. The experiments were conducted in a dual use lab/classroom that had a capacity of 32 participants, with 22 computers that were configured for the present research. Students were recruited by way of email and faculty notifications. Emails were distributed via listserv to Digital Forensics students (180 students). A faculty member also distributed recruitment emails to students in a sophomore level Introduction to Information Systems class (650 Students). In the experiment, students with a background in cyber-security and digital forensics were given a task that required significant cognitive load, time-pressure, and a significant incentive so that there was pressure to perform and monetary incentive to motivate the use of material from other participants.

*Procedures:* The experimental procedures were reviewed and approved by the University’s Institutional Review Board. The study employed a 2 (Performance Feedback) x 2 (Probe) experimental design. Participants reported to a computer laboratory and were assigned a data intensive task with financial incentives for better performance. Specifically, participants were asked to do research on hacker groups and collect information about their profile and activities as follows:

1. Identify as many global hacker groups in 40 minutes
2. Determine their physical location: Country that they operate out of
3. Find names of their group members
4. List and describe their major attacks
5. Create a profile of their victims (Who do they attack: i.e. individuals, companies, nation-states etc.)

We also provided participants with sample hacker group names and created templates for them to complete with the data that they collected. They were also encouraged to research new groups through web searches using key words such as “hacker groups”, “cyber criminals” and “state sponsored hackers”. It was stipulated that their performance will be evaluated based on the quantity of groups and quality of the information collected. Each student was provided a $10 gift card for their participation with a potential bonus of another ten dollars based on their performance. Additionally, a grand prize of one hundred dollars was announced for the overall best performer in the experiment. A shared drive was created for the experiments such that all machines used in the experiments could access, modify and write to this drive. The participants were asked to save their work to the shared drive in the folder that matched the identification number for their workstation. Each participant was provided a dedicated folder on the shared drive and given instructions for saving the file to their folder on the shared drive.

*Performance Feedback Manipulation:* Once the users started their tasks, 19 of the participants were randomly selected and informed they were performing poorly in comparison to their cohorts, whereas the
rest of the participants were told nothing. Specifically, after 15 minutes, these participants received a message with the following question: How many groups have you fully completed templates for so far? After they responded to the question, these participants received a follow up message: “Based on our analysis of the entries into the templates into the group folder, you’re performing well below the pace of the average performer, you may want to pick up the pace!”

**Probe Manipulation:** Twenty minutes later, participants received one of the two probes, sent as a message that popped up on their computer screen:

1. System Error: Microsoft Group Policy Editor has been corrupted, all security controls have been disabled.
2. Due to group policy misalignment some participants are unable to access their folders. So, we have changed the configuration so that all participants can access the entire shared drive.

**Measures:** Detectors for the experiments were key loggers that could capture precise file movements and manipulations of a user. Two different tools (NetVisor and Veriato 360) were used such that they complemented the functionality missing in the other as well as to help verify the correctness of the tools. Three behavior attributes were derived from data collected via user activity monitoring software:

1. Curiosity: The opening of other participant’s folders on the shared drive (Veriato 360)
2. File Opening: Opening of other participant’s files in the shared drive (Veriato 360)
3. Cheating: Using another participant’s results from documents in the shared drive (SpyTech NetVisor)

Analyses examined changes in these measures from pre to post introduction of the probe, and as a function of performance feedback. ANOVA was used for the omnibus tests. The data were analyzed using SPSS.

**Results**

The performance feedback manipulation did not influence insider threat behavior. Overall, the level of curiosity was similar for those who received performance feedback (11.5% of participants) and those that did not (11.5%), as was the number of files opened (Ms = 6.5% and 7.5%, respectively).

Table 1 presents results pertaining to the performance and probe manipulations. Curiosity and file opening behaviors both increased after introduction of the probe. The incidence of curiosity nearly doubled from pre- to post-probe from .08 to .15, a difference that approached conventional levels of statistical significance, \( p = .09 \). Additionally, the incidence of file opening more than tripled, from .04 to .13, \( p = .04 \). The effect of the probe was slightly stronger for the group receiving negative performance feedback, especially for file opening behavior. This effect was, however, not statistically significant, \( p = .4 \). These results provide support of our hypothesis that insider threat behavior will increase after introduction of probes that signal opportunity for data theft (Hypothesis 2). However, they do not support the hypothesis that the probes will be stronger for those receiving negative performance feedback (Hypothesis 3).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>n</th>
<th>Pre-Probe Curiosity</th>
<th>Post-Probe Curiosity</th>
<th>Opened File(s) Pre-Probe</th>
<th>Opened File(s) Post Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Performance Message</td>
<td>29</td>
<td>0.09</td>
<td>0.14</td>
<td>0.06</td>
<td>0.14</td>
</tr>
<tr>
<td>Performance Evaluation</td>
<td>18</td>
<td>0.06</td>
<td>0.17</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>0.08</td>
<td>0.15</td>
<td>0.04</td>
<td>0.13</td>
</tr>
</tbody>
</table>

**Table 1: Incidence of Curiosity and File Opening Prior to After Introduction of the Probe, by Performance Condition**
Table 2 presents the results by probe type. Results show that the “security down” probe had a stronger effect on behavior than the “facilitate accessibility” probe. The proportion of participants displaying curiosity increased from .08 to .18, \( p = .1 \), after the introduction of the security down probe. The effect on file opening was even more dramatic, increasing from .04 to .18, \( p = .02 \).

<table>
<thead>
<tr>
<th>Probe</th>
<th>n</th>
<th>Pre-Probe Curiosity</th>
<th>Post-Probe Curiosity</th>
<th>Opened File(s) Pre-Probe</th>
<th>Opened File(s) Post Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error: Security Down</td>
<td>28</td>
<td>0.08</td>
<td>0.18</td>
<td>0.04</td>
<td>0.18</td>
</tr>
<tr>
<td>Facilitate Accessibility</td>
<td>19</td>
<td>0.07</td>
<td>0.11</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>0.08</td>
<td>0.15</td>
<td>0.04</td>
<td>0.13</td>
</tr>
</tbody>
</table>

**Table 2: Effects of Probe Type on Curiosity and File Opening**

It is notable that while both probes instigated increases in malicious, and/or “pre-malicious” behavior in curiosity, the security down probe proved to be more efficacious in stimulating malicious insider behavior. Additionally, it is important to note the significance of the curiosity metric, which was qualified by users opening other participant’s folders and/or files. While at face value, devoid of any data theft, this behavior may seem relatively innocuous, it is notable that there was a moderate to strong correlation between pre-probe curiosity and post-probe file opening, \( r = .391, p = .003 \). The significance of this finding is that even though not all participants who engaged in “pre-malicious” curiosity stole data, the present results provide evidence that curiosity indeed was a powerful predictor for subsequent malicious insider behavior.

**Summary and Conclusions**

The present research has demonstrated the feasibility of using active probes for detecting malicious insider activity. The initial survey data obtained in our scenario analyses helped to ascertain the specific probes that could be used to detect insider behavior in employees. Two probes were tested during the experiment (i.e., “Security System Down” and “Access Control Relaxation”), both of which were found to be effective and in inciting curiosity (i.e., pre-malicious insider behavior) and file opening. The “Security System Down” probe demonstrated a larger effect size in spurring such malicious behavior as compared to the “Access Control Relaxation” probe.

Additional research is necessary to continue to validate additional probes across different environmental conditions that to test their robustness, and to evaluate insiders’ reactions to probes that seek to mimic naturalistic events within an organization. The implications for this work are strong wherein intentions of a user can be revealed through innocuous tasks which will help organization in establishing better access control on information based on average threat levels in the organization or for individual employees.

There are several limitations of this work: 1) The computation of certain metrics rely on self-reporting, which can be inaccurate and/or biased; 2) additional data needs to be collected for further testing; perhaps recruit individuals who are cyber security experts in addition to those who are well versed with it; 3) the contextual settings needs to be varied in order to mimic various naturalistic settings of interest; 4) additional probes should be developed and tested to evaluate generalizability of the present experimental paradigm; and 5) additional measures should be taken to allow for greater precision in monitoring user activity during the course of an experimental session. For example, our key loggers captured key strokes but not the cut and paste activity of the users. If participants copied and pasted when stealing data rather than typing in the document, it would be difficult to use key strokes to ascertain whether participants copy and pasted information from other documents.
REFERENCES


