Baiting Inside Attackers Using Decoy Documents

Brian M. Bowen, Shlomo Hershkop, Angelos D. Keromytis,
and Salvatore J. Stolfo

Department of Computer Science Columbia University

Abstract. The insider threat remains one of the most vexing problems in computer security. A number of approaches have been proposed to detect nefarious insider actions including user modeling and profiling techniques, policy and access enforcement techniques, and misuse detection. In this work we propose trap-based defense mechanisms and a deployment platform for addressing the problem of insiders attempting to exfiltrate and use sensitive information. The goal is to confuse and confound an adversary requiring more effort to identify real information from bogus information and provide a means of detecting when an attempt to exploit sensitive information has occurred. "Decoy Documents" are automatically generated and stored on a file system by the D^3 System with the aim of enticing a malicious user. We introduce and formalize a number of properties of decoys as a guide to design trap-based defenses to increase the likelihood of detecting an insider attack. The decoy documents contain several different types of bogus credentials that when used, trigger an alert. We also embed "stealthy beacons" inside the documents that cause a signal to be emitted to a server indicating when and where the particular decoy was opened. We evaluate decoy documents on honeypots penetrated by attackers demonstrating the feasibility of the method.

1 Introduction

Much research in computer security has focused on the means of preventing unauthorized and illegitimate access to systems and information. Unfortunately, the most damaging malicious activity is the result of internal misuse within an organization, perhaps since far less attention has been focused inward. Despite classic internal operating system security mechanisms and the body of work on formal specification of security and access control policies, including Bell-LaPadula [1] and the Clark-Wilson models [4], we still have an extensive insider attack problem. Indeed in many cases, formal security policies are incomplete and implicit or they are purposely ignored in order to get business goals accomplished. There seems to be little technology available to address the insider threat problem.

Insider attack has overtaken viruses and worm attacks as the most reported security incident according to a report from the US Computer Security Institute

Y. Chen et al. (Eds.): SecureComm 2009, LNICST 19, pp. 51–70, 2009.

(CSI) [19]. The annual Computer Crime and Security Survey for 2007 surveyed 494 security personnel members from US corporations and government agencies, finding that insider incidents were cited by 59 percent of respondents, while only 52 percent said they had encountered a conventional virus in the previous year. The state-of-the-art seems to be still driven by forensics analysis after an attack, rather than technologies that prevent, detect, and deter insider attack.

We define insider threats by differentiating between Masqueraders (attackers who impersonate another inside user) and Traitors (an inside attacker using their own legitimate credentials). One possible solution for masquerade detection involves anomaly detection [27]. In this approach, users actions are profiled to form a baseline of normal behavior. Subsequent monitoring for abnormal behaviors that exhibit large deviations from this baseline [16] signal a potential insider attack. The common strategy to prevent inside attacks involves policy-based access control techniques to limit the scope of systems and information an insider is authorized to use, and hence, limit the damage the organization may incur when an insider goes awry. Prevention techniques may not always succeed, and thus, monitoring and detection techniques are needed when prevention fails. In this paper, we are focused on different techniques aimed at detecting masqueraders and traitors.

We note that some external attackers can become insiders when an outsider attains internal network access. Many attacks use spyware and rootkits [3], which give outsiders internal access. Such software can easily be installed on systems from physical or digital media (*e.g.,* email, downloads) and allow an attacker administrator or "root" access on a machine along with a means to gather sensitive data. Rootkits have the ability to conceal themselves and elude detection, especially when the rootkit is previously unknown, as is true in zero-day attacks [8]. An external attacker that manages to install rootkits internally in effect becomes an insider, thereby multiplying the ability to inflict harm. Although the techniques described in this paper may have utility for these cases, in this paper our primary focus is on human insiders attempting to exfiltrate sensitive information. By exfiltration we mean unauthorized copying and transmission of information by any means.

The insider attack defense system described in this paper is of an offensive nature, intended to confuse and deceive a traitor by leveraging uncertainty, to reduce the knowledge they ordinarily have of the systems and data they might be authorized to use. This work considers methods to detect insider actions against enterprise systems as well as individual hosts and laptops. We introduce a deception system to distribute potentially large amounts of decoy information with the aim to detect nefarious acts as well as to increase the workload of an attacker to identify real information from bogus information, rather than providing unfettered access as broadly exists today. We developed a system to generate and place *decoy documents* within a file system. Our system generates decoy documents containing decoy credentials that are monitored (*e.g.,* Gmail credential monitoring) for misuse and stealthily embedded beacons that signal an alert when the document is opened.

To achieve the goal of wide spread deception we must consider methods to trap a wide variety of potential insiders with varying levels of sophistication. Toward this goal, we developed a proof-of-concept system we call D^3, the Decoy Document Distributor system. Samples of D^3 generated documents are presented in the Appendix. The contributions of this paper include:

- A novel set of generally applicable properties are proposed to guide the design and deployment of decoys and maximize the deception they induce for different classes of insiders who vary by their level of knowledge and sophistication.
- A large-scale automated creation and management system for deploying decoys that can detect the presence (and, in some cases, "identity") of malicious insiders, or at least indicate malicious insider activity. This provides a means for ordinary users to deploy honey documents without having to setup sophisticated honeypot systems and sensors.
- An offensive trap-based defense system is proposed to detect masqueraders and traitors, and to flood attackers with bogus exfiltrated information that they must analyze in order to find real information of value. Hence, our long term goal is to flood the miscreant marketplace with bogus information devaluing their quarry.
- A design of decoy information that combines a number of methods and monitors, both internal and external, to detect insider exploitation using a common and ubiquitous set of baited targets, ordinary looking documents.
    1. A watermark is embedded in the binary format of the document file to detect when the decoy is loaded in memory, or egressed in the open over a network.
    2. A "beacon" is embedded in the decoy document that signals a remote web site upon opening of the document indicating the malfeasance of an insider illicitly reading bait information.
    3. If these methods fail to detect an insider attack or an exfiltration of baited documents, the content of the documents contain bait and decoy information that is monitored as well. Bogus logins at multiple organizations as well as bogus and realistic bank information is monitored by external means.
- An easy to use system to broadly deploy decoys to ordinary users who are alerted by email when a decoy has been touched on their laptops and personal computers; no such system presently exists.

The reader is encouraged to visit the Decoy Document Distribution (D^3) web site to evaluate our technology developed to date at: `http://www.cs.columbia.edu/ids/RUU/Dcubed`[1].

2 Related Work

The use of deception, or decoys, plays a valuable role in the protection of systems, networks, and information. The first use of decoys (*i.e.,* in the cyber domain)

[1] Some features are restricted for internal use only.

has been credited to Cliff Stoll [29,23] and detailed in his novel “The Cuckoos Egg” [24], where he provides a thorough account of his crusade to catch German hackers breaking into Lawrence Berkeley Laboratory computer systems. Stoll’s methods included the use of bogus networks, systems, and documents to gather intelligence on the German attackers who were apparently seeking state secrets. Among the many techniques waged, he crafted “bait” files, or in his case, bogus classified documents that really contained non-sensitive government information and attached “alarms” to them so that he would know if anyone accessed at them. To Stoll’s credit, a German hacker was eventually caught and it was found that he had been selling secrets to the KGB.

Honeypots are effective tools for profiling attacker behavior. Honeypots are considered to have low false positive rates since they are designed to capture only malicious attackers, except for perhaps an occasional mistake by innocent users. Spitzner described how honeypots can be useful for detecting insider attack [22] and discusses the use of honeytokens [23] such as bogus medical records, credit card numbers, and credentials. In a similar spirit, Webb *et al.* [26] showed how honeypots can be useful for detecting spammers. In current systems, the decoy/honeytoken creation is a laborious and manual process requiring large amounts of administrator intervention. Our work extends these basic ideas to an automated system of managing the creation and deployment of these honeytokens.

Yuill *et al.* [29] extend the notion of honeytokens with a “honeyfile system” to support the creation of bait files, or as they define them, “honeyfiles.” The honeyfile system is implemented as an enhancement to the Network File Server. The system allows for any file within user file space to become a honeyfile through the creation of a record associating a filename to userid. The honeyfile system monitors all file access on the server and alerts users when honeyfiles have been accessed. Their work does not focus on the content or automatic creation of files, but they do elicit some of the challenges of creating deceptive files (with respect to names) that we address in section 4.

In this paper, we introduce a set of properties of decoys to guide their design and maximize the deception they induce for different classes of insiders who vary by their level of knowledge and sophistication. To the best of our knowledge, the synthesis of these properties is indeed novel a contribution. Bell and Whaley [2] have described the structure of deception as a process of hiding the real and showing showing the false. They introduce several methods of hiding that include masking, repackaging, and dazzling, along with three methods of showing that include mimicking, inventing, and decoying. Yuill *et al.* [28] expand upon this work and characterize deceptive hiding in terms of how it defeats an adversary’s discovery process. They describe an adversary’s discovery process as taking three forms: direct observation, investigation based on evidence, and learning from other people or agents. Their work offers a process model for creating deceptive hiding techniques based on how they defeat an adversary’s discovery process.

The decoy documents introduced in this paper utilize similar deception mechanisms as well as beacons to signal a remote detect and alert in real-time time when a decoy has been opened. Web bugs are a class of silent embedded tokens which have been used to track usage habits of web or email users [17]. Unfortunately, they have been most closely associated with unscrupulous operators, such as spammers, virus writers, and spyware authors who have used them to violate users privacy. Typically they will be embedded in the HTML portion of an email message as a non-visible white on white image, but they have also been demonstrated in other forms such as Microsoft Word, Excel, and PowerPoint documents [20]. When rendered as HTML, a web bug triggers a server update which allows the sender to note when and where the web bug was viewed. Animated images allow the senders to monitor how long the message was displayed. The web bugs operate without alerting the user of the tracking mechanisms. The advantage for legitimate advertisers is that this allows them to monitor advertisement effectiveness, while privacy advocates worry that this technology can be misused to spy on users' habits. Our work leverages the same ideas, but extends them to other document classes and is more sophisticated in the methods used to draw attention. In addition, our targets are insiders who should have no expectation of privacy on a system they violate.

3 Threat Model - Level of Sophistication of the Attacker

The insider seeks to identify and avoid the decoys and abscond with "real" information. We broadly define four monotonically increasing levels of insider sophistication and capability. Some will have tools available to assist in deciding what is a decoy and what is real. Others will only have their own observations and thoughts.

- **Low:** Direct observation is the only tool available. The adversary largely depends on what can be gleaned from a first glance. We strive to defeat this level of adversary with our beacon documents, even though decoys with embedded beacons may be distinguished with more advanced tools.
- **Medium:** A more thorough investigation can be performed by the insider; decisions based on other, possibly outside evidence, can be made. For example, if a decoy document contains a decoy account credential for a particular identity, an adversary may verify that the particular identity is real or not by querying an external system (such as www.whitepages.com). Such adversaries will require stronger decoy information possibly corroborated by other sources of evidence.
- **High:** Access to the most sophisticated tools are available to the attacker (*e.g.,* super computers, other informed people who have organizational information). The notion of the "Perfect Decoy" described in the next section may be the only indiscernible decoy by an adversary of such caliber.
- **Highly Privileged:** Probably the most dangerous of all is the privileged and highly sophisticated user. Such attackers might even be aware that the system is baited and will employ sophisticated tools to try to analyze, disable, and avoid decoys entirely. As an example of how defeating this level of threat

might be possible, consider the analogy with someone who knows encryption is used (and which encryption algorithm is used), but still cannot break the system because they do not have knowledge of an easy-to-change operational parameter (the key). Likewise, just because someone knows that decoys are used in the system does not mean they should be able to identify them. This is the principal– coming up with a scheme to satisfy it remains an open problem.

4 Generating and Distributing Bait

In order to create decoys to bait various levels of insiders, one must understand the core properties of a decoy that will successfully bait an insider.

4.1 Decoy Properties

We enumerate various properties and means of measuring these properties that are associated with decoy documents to ensure their use will be likely to snare an inside attacker. We introduce the following notation for these definitions.

Believable[2]: Capable of eliciting belief or trust; capable of being believed; appearing true; seeming to be true or authentic.

A good decoy should make it difficult for an adversary to discern whether they are looking at an authentic document from a legitimate source or if they are indeed looking at a decoy. We conjecture that believability of any particular decoy can be measured by adversary's failure to discern one from the other. We formalize this by defining a decoy believability experiment. The experiment is defined for the document space M with the set of decoys D such that $D \subseteq M$ and $M - D$ is the set of authentic documents.

The Decoy Believability Experiment: $\mathbf{Exp}^{believe}_{A,D,M}$

- For any $d \in D$, choose two documents $m_0, m_1 \in M$ such that $m_0 = d$ or $m_1 = d$, and $m_0 \neq m_1$; that is, one is a decoy we wish to measure the believability of and the second is chosen at random from the set of authentic documents.
- Adversary A obtains m_0, m_1 and attempts to choose $\hat{m} \in \{m_0, m_1\}$ such that $\hat{m} \neq d$, using only information intrinsic to m_0, m_1.
- The output of the experiment is 1 if $\hat{m} \neq d$ and 0 otherwise.

For concreteness, we build upon the definition of "Perfect Secrecy" proposed in the cryptography community [12] and define a "perfect decoy" when:

$$\Pr[\mathrm{Exp}^{believe}_{A,D,M} = 1] = 1/2$$

[2] For clarity, each property is provided with its definition gleaned from online dictionary sources.

The decoy is chosen in a believability experiment with a probability of 1/2 (the outcome that would be achieved if the volunteer decided completely at random). That is, a perfect decoy is one that is completely indistinguishable from one that is not. A benefit of this definition is that the challenge of showing a decoy to be believable, or not, reduces to the problem of creating a "distinguisher" that can decide with probability better than 1/2.

In practice, the construction of a "perfect decoy" might be unachievable, especially through automatic means, but the notion remains important as it provides a goal to strive for in our design and implementation of systems. For many threat models, it might suffice to have less than perfect believable decoys. For our proof-of-concept system described below, we generate receipts and tax documents, and other common form-based documents with decoy credentials, realistic names, addresses and logins, all information that is familiar to all users.

We note that the believable property of a decoy may be less important than other properties defined below since the attacker may have to open the decoy in order to decide whether the document is real or not. The act of opening the document may be all that we need to trap the insider, irrespective of the believability of its content. Hence, enticing an attacker to open a document, say one with a very interesting name, may be a more effective strategy to detect an inside attack than producing a decoy document with believable content.

Enticing: highly attractive and able to arouse hope or desire; "an alluring prospect"; lure.

Herein lies the issue of how does one measure the extent to which a decoy arouses desires, how well is it a lure? One obvious way is to create decoys containing information with monetary value, such as passwords or credit card numbers that have black market value [14,25]. However, enticement depends upon the attacker's intent or preference. We define enticing documents in terms of the likelihood of an adversary's preference; enticing decoys are those decoys that are chosen with the same likelihood. More formally, for the document space M, let P be the set of documents of an adversary's A preference, where $P \subseteq M$. For some value ϵ such that $\epsilon > 1/|M|$, an enticing document is defined by the probability

$$\Pr[m \rightarrow M | m \in P] > \epsilon$$

where $m \rightarrow M$ denotes m is chosen from M. An enticing decoy is then defined for the set of decoys D, where $D \subseteq M$, such that

$$\Pr[m \rightarrow M | m \in P] = \Pr[d \rightarrow M | d \in D]$$

We posit that by defining several general categories of "things" that are of "attacker interest", one may compose decoys using terms or words that correspond to desires of the attacker that are overwhelmingly enticing. For example, if the attacker desires money, any document that mentions or describes information that provides access to money should be highly enticing. We believe we can measure frequently occurring (search) terms associated with major categories of interest (*e.g.*, words or terms drawn from finance, medical information, intellectual property) and use these as the constituent words in decoy documents. To measure the effectiveness of this generative strategy, it should be possible to execute content searches and count the number of times decoys appear in the top 10 list of displayed documents. This is a reasonable approach also, to measuring how conspicuous, defined below, the decoys become based upon the attacker's searches associated with their interest and intent.

Conspicuous: easily visible; easily or clearly visible; obvious to the eye or mind; Attracting attention.

A *conspicuous* decoy should be easily found or observed. Conspicuous is defined similar to enticing, but conspicuous documents are found because they are easily observed, whereas enticing documents are chosen because they are of interest to an attacker. For the document space M, let V be the set of documents defined by the minimum number of user actions required to enable their view. We use a subscript to denote the number of user actions required to view some set of documents. For example, documents that are in view at logon or on the desktop (requiring zero user actions) are labeled V_0, those requiring one user action are V_1, etc. We define a "view", V_i of a set of documents as a function of a number of user actions applied to a prior view, V_{i-1}, hence

$$V_i = \text{Action}(V_{i-1}) \text{ where } V_j \neq V_i, j < i$$

An "Action" may be any command or function that displays files and documents, such as 'ls', 'dir', 'search.' For some value ϵ such that $\epsilon > 0$, a conspicuous document, d, is defined by the probability

$$\prod_{i=0}^{n} \Pr[V_i] > \epsilon$$

where n is the minimum value where $d \in V_n$. Note if d is on the desktop, V_0, $\Pr[V_0] = 1$ (*i.e.*, the documents in full view are highly conspicuous).

When a user first logs in, a conspicuous decoy should either be in full view on the desktop, or viewable after one (targeted) search action. One simple user action is optimal for a highly conspicuous decoy. Thus, a measure of conspicuousness may be a count of the number of search actions needed, on average, for a decoy to appear in full view. The decoy may be stored in the file system anywhere if a simple content-based search locates it in one step. But, this search act depends upon the query executed by the user. The query can either be a location (*e.g.*, search for a directory named "TAX" in which the decoy appears) or a content query (*e.g.*, using Google Desktop Search for documents containing the word "TAX.") In either case, if a decoy document appears after one such search, it is conspicuous. Hence, we may define the set P as all such files that can be found in some number of steps. But, this depends upon what search terms the attacker uses to query! If the decoy never appears because the attacker used the wrong search terms, the decoy is not conspicuous. We posit that the property of *enticing* is likely the most important property, and a formal measure to evaluate enticement will generate better decoys. In summary, an enticing decoy should be conspicuous to be an effective decoy trap.

Detectable; to discover or catch (a person) in the performance of some act: to detect someone cheating.

Decoys must ensure an alert is generated if they are exploited. Formally, this is defined for adversary A, document space M, and the set of decoys D such that $D \subseteq M$. We use $Alert_{A,d} = 1$ to denote an alert for $d \in D$. We say d is detectable with probability ϵ when

$$\Pr[d \rightarrow M : Alert_{A,d} = 1] \geq \epsilon$$

Ideally, ϵ should be 1.

We designed the decoy documents with several techniques to provide a good chance of detecting the malfeasance of an inside attack in real-time.

- At time of application start-up, the decoy document emits a beacon alert to a remote server.
- At the time of memory load, a host-sensor, such as an antivirus scanner, may detect embedded tokens placed in a clandestine location of the document file format.
- At the time of exfiltration, a NIDS such as Snort, or a stream event detection system such as Cayuga [5] may be used to detect these embedded tokens during the egress of the decoy document in network traffic where possible.
- At time of information exploitation and/or credential misuse, monitoring of decoy logins and other credentials embedded in the document content by external systems will generate an alert that is correlated with the decoy document in which the credential was placed.

This extensive set of monitors maximizes ϵ, forcing the attacker to expend considerable effort to avoid detection, and hopefully will serve as a deterrent to reduce internal malfeasance within organizations that deploy such a trap-based defense. In the proof-of-concept implementation reported in this paper, we focus our evaluation on the last item. We utilize monitors at our local IT systems, at Gmail and at an external bank.

Variability: The range of possible outcomes of a given situation; the quality of being subject to variation.

Attackers are humans with insider knowledge, even possibly with the knowledge that decoys are liberally spread throughout an enterprise. Their task is to identify the real documents from the potentially large cache of decoys. One important property of the set of decoys is that they are not easily identifiable due to some common invariant information they all share. A single search or test function would thus easily distinguish the real from the fake. The decoys thus must be highly varied. We define variable in terms of the likelihood of being able to decide the believability of a decoy given *any* known decoy. Formally, we define *perfectly variable* for document space M with the set of decoys D such that $D \subseteq M$ where

$$\Pr[d' \rightarrow D : \mathrm{Exp}^{believe}_{A,D,M,d'} = 1] = 1/2$$

Observe that under this definition an adversary may have access to *all* N previously generated decoys with the knowledge they are bogus, but still lack the ability to discern the N+1st. From a statistical perspective, each decoy is independent and identically distributed. For the case that an adversary can determine the N+1st decoy only after observing the N prior decoys, we define this as an *N-strong Variant.*

Clearly, a good decoy generator should produce an unbounded collection of enticing, conspicuous, but distinct and variable documents. They are distinct with respect to string content. If the same sentence appears in 100 decoys, one would not consider such decoys with repetitive information as highly variable; the common invariant sentence(s) can be used as a "signature" to find the decoys, rendering them distinguishable (and clearly, less enticing).

Non-interference: Something that does not hinder, obstructs, or impede.

Introducing decoys to an operational system has the potential to *interfere* with normal operations in multiple ways. Of primary concern is that decoys may pollute authentic data so that their legitimate usage becomes hindered by corruption or as a result of confusion by legitimate users (*i.e.,* they cannot differentiate real from fake). We define non-interference in terms of the likelihood of legitimate users successfully accessing normal documents after decoys are introduced. We use $\mathrm{Access}_{U,m} = 1$ to denote the success of a legitimate user U accessing a normal document m. More formally, for some value ϵ, the document space M, $\forall m \in M$ we define

$$\Pr[Access_{U,m} = 1] \geq \epsilon$$

on a system without decoys. Non-interference is then defined for the set of decoys D such that $D \subseteq M$ and $\forall m \in M$ we have

$$\Pr[Access_{U,m} = 1] = \Pr[Access_{U,m} = 1|D]$$

Although we seek to create decoys to ensnare an inside attacker, a legitimate user whose data is the subject of an attacker must still be able to identify their own real documents from the planted decoys. The more enticing or believable a decoy document may be, the more likely it would be to lead the user to confuse it with a legitimate document they were looking for. Our goal is to increase believability, conspicuous, and enticingness while keeping interference low; ideally a decoy should be completely non-interfering. The challenge is to devise a simple and easy to use scheme for the user to easily differentiate their own documents, and thus a measure of interference is then possible as a by-product.

Differentiable: to mark or show a difference in; constitute a difference that distinguishes; to develop differential characteristics in; to cause differentiation of in the course of development.

It is important that decoys be "obvious" to the *legitimate user* to avoid interference, but "unobvious" to the insider stealing information. We define this in terms of an inverted believability experiment, in which the adversary is replaced by a legitimate user. We say a decoy is differentiable if the legitimate user always succeeds. Formally, we state this for the document space M with the set of decoys D such that $D \subseteq M$ where

$$\Pr[\mathrm{Exp}^{believe}_{U,D,M} = 1] = 1$$

How might we easily differentiate a decoy for the legitimate user so that we maintain "non-interference" with the user's own actions and legitimate work? The remote thief who exfiltrates all of a user's files onto a remote hard drive may be perplexed by having hundreds of decoys amidst a few real documents; the thief should not be able to easily differentiate between the two cases. If we store a hundred decoys for each real document, the thief's task is daunting; they would need to test embedded information in the documents to decide what is real and what is not, which should complicate their end goals. For clarity, decoys should be easily *differentiable* to the legitimate user, but not to the attacker without significant effort. Thus, the use of "beacons" or other embedded content in the binary file format of a document, must be judiciously designed and deployed to avoid making decoys trivially differentiable for the attacker.

4.2 The Decoy Document Distributor (D^3) System

The D^3 web-based service generates and distributes decoy documents to registered users. The decoy properties guide the design of decoy templates in D^3 that are used to generate specific documents for download. The content of each decoy document includes several types of "bait" information such as online banking logins provided by a collaborating financial institution[3], login accounts for online servers, and web based email accounts. In our deployment we used Columbia University student accounts and Gmail email accounts as bait, but these can be customized to any set of monitored credentials. These decoy credentials are "bait" and are enticing targets for different types of adversaries [14,13].

4.3 Decoy Document Design

The primary goal of the trap based defense is to detect malfeasance. Since no system is foolproof, we propose that multiple overlapping signals be embedded in the decoy documents to ensure *detectability.* Any alert generated by the multiple decoys is an indicator that some insider activity has occurred. Since the attacker may have varying levels of sophistication, a combination of traps are used in

[3] By agreement, the institution request that its name be withheld.

decoy documents to increase the likelihood one will succeed in generating an alert. A sophisticated attacker may, for example, disable the internal beacon, or cut off network connections avoiding communication, disable or kill local host monitoring processes, or they may exfiltrate documents via a web-browser without opening them locally. The documents are designed with several means of detecting their misuse:

- embedded honeytokens, computer login accounts created that provide no access to valuable resources, and that are monitored when (mis)used;
- embedded honeytoken banking login accounts specifically created and monitored for this trap-based technology demonstration specifically to entice financially motivated attackers;
- a network-level egress monitor that alerts whenever a marker, specially planted in the decoy document, is detected (we are collaborating with Cornell to use Cayuga [5] for this purpose. Presently Snort may be used as simple signature detector as a proof-of-concept);
- a host-based monitor that alerts whenever a decoy document is "touched" in the file system such as a copy operation;
- an embedded "beacon" alerts a remote server at a site at Columbia that we call SONAR. The web site emits an email to the registered user who created and downloaded the decoy document.

The implementation of features are described below.

Honeytokens. This layer of defense is made up of "bait" information such as online banking logins provided by a collaborating financial institution, credit card numbers, login accounts for online servers, and web based email accounts. The primary requirement for bait is that it be detectable when (mis)used. For example, one form of bait that we use are usernames and passwords for Gmail accounts. D^3 is integrated with a variety of services to enable monitoring of these credentials once they are deployed as decoys. In the case of the Gmail accounts, custom scripts access *mail.google.com* to parse the bait account pages, gathering account activity information. The information includes the IP addresses for the previous 5 account accesses and the time. If there is any activity from IP addresses other than D^3's monitor, an alert is triggered with the time and IP of the offending host. Alerts are also triggered when the monitor cannot login to the bait account. In this case, we conclude that the account password was stolen (unless monitoring resumes) and the password changed unless other corroborating information (like a network outage) can be used to convince otherwise. In addition, some of our accounts have password monitors, allowing us to produce a seemingly unbounded collection of decoy variants for individual usernames.

In the case of financially motivated bait, we are beginning to use real credit card numbers in addition to banking login credentials. Many credit card providers offer "one-time-credit-card numbers" and other forms of Controlled Payment Numbers [18], which enable the generation of multiple credit card numbers for a single account. In the case of PayPal, single use credit card numbers can be generated with a predetermined balance. The D^3 monitor is being integrated with the PayPal APIs to automatically monitor the activity of the credit card numbers deployed through D^3. As is the case for all of the decoys, the benefit of deployment through D^3 is the automation, enabling their creation, monitoring, and distribution en masse.

Beacon Implementation. The highly sophisticated attacker will likely attempt to differentiate between a real document and a decoy by analyzing the binary file format prior to opening a file. This necessitates a design where beacon code and watermarks in decoy documents are hidden to avoid their easy identification. The attacker would surely avoid the decoys if they could easily identify them by a simple static test for an embedded beacon. The beacon code can be embedded in documents in a number of ways and made to appear statistically equivalent to its surrounding data using a blending technique called "spectrum shaping" (see [21,6]). Such obfuscation techniques are very hard to defeat [15].

Using common techniques developed for malware, beacons attempt to silently contact a centralized server with a unique token embedded within the document at creation time. The token is used to identify the decoy and document, the IP address of the host accessing the decoy document. Depending on the particular document type and the rendering environment used during viewing of the beacon document, some additional data may be collected.

The first proof-of-concept beacons have been implemented in MS Word and PDF and deployed through the D^3 web site. In the case of the MS Word document beacons, the examples rely on a stealthily embedded remote image that is rendered when the document is opened. The request for the remote image is a positive indication the document has been opened. In the case of PDF document beacons, the signaling mechanism relies on the execution of Javascript within the document. The D^3 site includes a tutorial guiding the user on how to generate, download, and enable the decoys' silent communication on hosts. It is important to point out that there are methods for disabling the beacon mechanism. In Section 5.2, we provide an evaluation of beacon robustness.

Embedded Marker Implementation. Beacon documents contain embedded markers that a host or network sensor may detect either when documents are loaded in memory or transmitted in the clear. The markers are constructed as a unique pattern of word tokens uniquely tied to the document creator. The sequence of word tokens is embedded within the beacon document's meta-data area or reformated as comments within the document format structure. Both locations are ideal for embedding markers since most rendering programs ignore these parts of the document. The embedded markers can be used in Snort signatures for detecting exfiltration.

5 Evaluation

5.1 Masquerade Detection Using Decoy Documents as Bait

We have defined the general properties that decoys should have and discussed how we may measure these properties, but here we focus on the most important property: *detectability.* Under ideal testing conditions, decoy efficacy could be shown through deployment on true operational systems either within an enterprise environment, or on personal computers, by the number of attacks they are able to detect or thwart (they have a deterrence effect). However, given reasonable time limits, the infrequency of attacks within the insider threat model makes this approach impractical within a university environment. As we mentioned we are now seeking a larger user population to study and measure decoy generation over time.

Another approach to evaluation is a user study in which users are organized and asked to evaluate decoys based on each of the key decoy properties mentioned earlier. We take human evaluation to be the gold standard of evaluation since the human mind is the ultimate target of our decoys. That is, we wish to show how well our decoys can induce deception on human test subjects. One of the challenges of conducting a traditional user study lies in the logistics of obtaining volunteers. In our methodology, we attempt to reduce this challenge by leveraging external attackers to serve as participants in our study on masquerade detection. To do so, we "invite" attackers (or more accurately, bamboozle them) into our study by attracting them with a set of vulnerable systems on the university network, which also serve as our testing platform.

Our test platform is embedded within a honeynet [9]. It consists of several virtual machines running Linux and configured with Sebek [10] to capture attacker activities including commands and file references. In order to limit potential damage from system compromise and still allow for testing, we configured the honeynet to allow all incoming connections while restricting the number of outgoing connections.

The virtual machine hosts within the honeynet were configured with accounts and home directories for three decoy usernames. To make the environment as real as possible, genuine data from personal accounts on other systems were loaded into each of the home directories. We changed name references within the data to reflect those of the appropriate decoy users. In total, our phony user accounts contained 15 or more directories and 50-100 files. The hosts were then seeded with several of D^3's decoy files using the decoy distributor utility. The decoy files were generated to have conspicuous names such as "stolen passwords", "credit card", "private data", and "Gmail AccountInfo", but were distributed within the polluted home directories of the decoy accounts, making the environment as real as possible.

To lure test subjects into the study, our initial approach was to use attackers that attempt to gain internal access via password scanning. Password scanning attacks are common on the university network, where attempts on a typical machine are in the range of thousands per day. To enable attacker access, we conducted a short study to first determine the most common usernames and passwords (excluding those for root and actual users) used in these attempts. We created accounts with several of these usernames and passwords, to quickly learn that this breed of attacker was not going to suffice for our user study; their sole purpose seemed confined to creating zombies for botnets. While this may be a valid threat to study while evaluating decoys [7], allowing bots to operate on the university network poses too much risk.

In our second and more aggressive approach, we narrowed our recruitment effort to web forums and IRC channels with the expectation and hope that we would get fewer attacks involving botnets. In this approach, we selected several high volume forums to solicit volunteers and posted variations of invitations with messages that included hostnames, usernames, and passwords. The idea was to provide just enough innocent-looking information from a novice to lure people into our machines without providing direct evidence that we were conducting a deception-based experiment. Note that we deliberately omit the names of the forums used and the exact details of the messages, as this is an ongoing study.

While our methodology could, in theory, provide anyone with access to our test platform, by selectively choosing the location of postings and contents postings, we expected to recruit two primary classes of individuals:

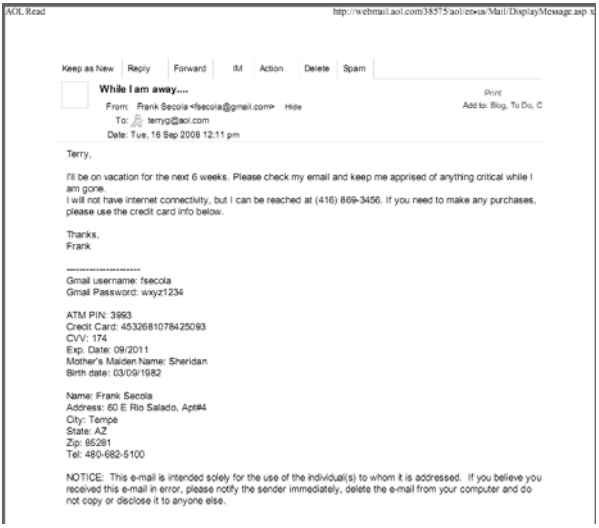
- Legitimate and generally curious computer-savvy individuals. These users have no interest in extending privileges in an unauthorized way, but participate in the study out of curiosity, as there is no other incentive.
- Unscrupulous opportunistic hackers who attempt to extend their network access by whatever means afforded to them. These individuals are enticed by our posting as they see our machines as low “hanging fruit” in their targeting campaign.

In either case, we believe these individuals to be suitable candidates for our study (with one caveat mentioned later). Both classes of individuals can be used in measuring the enticement property of decoys. We measure this by examining the behavior exhibited in file access, both with respect to the particular files a user attempts to read and in the order in which the files are read. For example, if all users consistently read the same file first, we know the file must indeed be enticing.

In regards to indistinguishability of the decoys, we note that the content of these decoys contains bait information in the form of monitored credentials on real systems. Certainly, if our attackers take the time to use the decoy credentials, there is an implication that they must also be believable. More importantly though, if they use the credentials and we detect their use, we have also answered the most important question of – can we *detect* the attacker? Note that the first class of the individuals is by definition, not useful for this part of the study. That is, attempting to use credentials found on our machines is clearly an illicit activity, which they would not partake in. Unfortunately, we do not have a good way of distinguishing the two classes, so we get slightly skewed results.

Over the span of the first week[4], our hosts netted 20 unique users, determined anecdotally by source IP address and reinforced by the unique behaviors exhibited by each. The length of user sessions ranged from minutes to hours. The users that spent hours were focused on activities that included writing code and attempting to install other software, presumably with malicious intent. The

[4] Most attacks occurred within the first 8-hour period after posting.



While I am away....

From: Frank Secola <fsecola@gmail.com>
To: terryg@aol.com
Date: Tue, 16 Sep 2008 12:11 pm

Terry,

I'll be on vacation for the next 6 weeks. Please check my email and keep me apprised of anything critical while I am gone.
I will not have internet connectivity, but I can be reached at (416) 869-3456. If you need to make any purchases, please use the credit card info below.

Thanks,
Frank

Gmail username: fsecola
Gmail Password: wxyz1234

ATM PIN: 3993
Credit Card: 4532681078425093
CVV: 174
Exp. Date: 09/2011
Mother's Maiden Name: Sheridan
Birth date: 03/09/1982

Name: Frank Secola
Address: 60 E Rio Salado, Apt#4
City: Tempe
State: AZ
Zip: 85281
Tel: 480-682-5100

NOTICE: This e-mail is intended solely for the use of the individual(s) to whom it is addressed. If you believe you received this e-mail in error, please notify the sender immediately, delete the e-mail from your computer and do not copy or disclose it to anyone else.

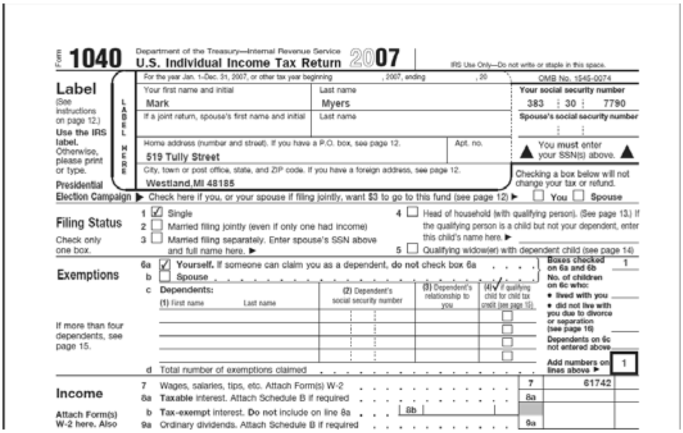
Fig. 1. Decoy email message with embedded Gmail account information

most common activities exhibited for users upon login included activities such as: determining who else was on the machine, checking and deleting the command history, and exploring the system (*i.e.,* reading the most *conspicuous* files). We note that in almost all cases, our decoys were among the first few files viewed by users upon initial login (as noted, viewing the history file was also popular). The most significant observations were made for 6 of the 20 unique users:

1. There were three attempts to use Gmail credentials that were contained in a decoy document, which triggered an alert on SONAR.
2. One attacker changed the password on a bogus Gmail account, which also triggered a SONAR alert.
3. There were at least two attempts to exfiltrate decoy files (with *scp* and *sftp*; one file, named "stolenpasswords", contained credentials to the university systems.
4. There was one attempt to use the university credentials contained in the "stolenpassword" file, which we were alerted to by the monitor that signaled an alert to SONAR.

We take these results as evidence that D^3/SONAR indeed has value as a defense against masqueraders. While only 5 of the 20 users sounded an alarm on SONAR, we emphasize that our methodology did include an unknown proportion of benign users. Furthermore, the focus of study was on masquerade detection; admittedly, we do not yet have a good way of evaluating our system on traitors, but this will be the focus of future work.

One flaw in our evaluation methodology that was revealed during testing was that we allowed users to make changes to the file system. We did this deliberately



Form **1040** Department of the Treasury—Internal Revenue Service **U.S. Individual Income Tax Return 2007** IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2007, or other tax year beginning , 2007, ending , 20 OMB No. 1545-0074

Label (See instructions on page 12.) **Use the IRS label.** Otherwise, please print or type.

LABEL HERE

Your first name and initial: Mark
Last name: Myers
Your social security number: 383 30 7790
If a joint return, spouse's first name and initial
Last name
Spouse's social security number
Home address (number and street). If you have a P.O. box, see page 12.: 519 Tully Street
Apt. no.
▲ You must enter your SSN(s) above. ▲
City, town or post office, state, and ZIP code. If you have a foreign address, see page 12.: Westland,MI 48185

Presidential Election Campaign ▶ Check here if you, or your spouse if filing jointly, want $3 to go to this fund (see page 12) ▶ □ You □ Spouse
Checking a box below will not change your tax or refund.

Filing Status
Check only one box.
1 ☑ Single
2 □ Married filing jointly (even if only one had income)
3 □ Married filing separately. Enter spouse's SSN above and full name here. ▶
4 □ Head of household (with qualifying person). (See page 13.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶
5 □ Qualifying widow(er) with dependent child (see page 14)

Exemptions
6a ☑ **Yourself.** If someone can claim you as a dependent, **do not** check box 6a . . . . .
b □ Spouse . . . .
Boxes checked on 6a and 6b: 1
No. of children on 6c who:
• lived with you
• did not live with you due to divorce or separation (see page 16)
Dependents on 6c not entered above

c Dependents:

(1) First name Last name	(2) Dependent's social security number	(3) Dependent's relationship to you	(4) ✓ if qualifying child for child tax credit (see page 15)
			□
			□
			□
			□

If more than four dependents, see page 15.

d Total number of exemptions claimed . . . . Add numbers on lines above ▶ 1

Income
Attach Form(s) W-2 here. Also

7	Wages, salaries, tips, etc. Attach Form(s) W-2	7	61742
8a	**Taxable** interest. Attach Schedule B if required	8a	
b	**Tax-exempt** interest. **Do not** include on line 8a . . . 8b		
9a	Ordinary dividends. Attach Schedule B if required	9a	

Fig. 2. Decoy tax document with bogus user information

to increase the realism of the environment in the experiments. The problem this created was that it made decoy defense vulnerable to deletion (*e.g.,* several of our visitors executed wholesale deletion of files with "rm -rf *") . This poses a problem in our testing methodology, but not necessarily in practice. That is, the act of deleting files is in itself a detectable behavior that would alert monitors of suspicious behavior.

In this study, we omitted testing decoy documents with embedded beacons. The honeypots set up to attract remote attackers were stripped down Linux machines that had no installed applications necessary to open and render the decoy documents. We believe the value of beacon documents to be self-evident. We encourage the reader to visit and test the D^3 site, and participate in our planned longitudinal study. In the next section we describe tests of the beacon implementation on multiple hosts.

5.2 Beacon Implementation Tests

To test the robustness of the beacon implementations we tested them with the most common configurations of operating systems and document viewers. To this end, we contacted a random group of users across the Internet and sent them each two types of beacon documents along with a request that they open them as part of a benign experiment. The results of tests conducted on PDF and Word beacons are presented in Table 1 and 2 below. These results are a representative sample of real users across multiple hosts accessing the beacon documents. For the most part the beacon technology works well on the windows platform while not as well on Mac and Linux operating systems. The reason is that the default PDF reader is not Adobe's and does not execute Javascript embedded within the documents. Similarly, Word document beacons do not work when applications other than Microsoft Word (*e.g.,* OpenOffice or Google Docs) are used to open

Table 1. PDF Beacon Test Results

OS	Application	#Tests	#Pings
Windows XP	Adobe	6	6
Windows Vista	Adobe	4	4
Mac OS	Preview	1	0
Mac OS	Adobe	1	1
Ubuntu	Evince	1	0

Table 2. Word Beacon Test Results

OS	Application	#Tests	#Pings
Windows XP	Word	5	4
Windows XP	GoogleDocs	1	0
Windows Vista	Adobe	4	4
Mac OS	Word	2	2
Linux	OpenOffice	1	0

them. We are currently researching ways to address these limitations and will focus on them in future work.

6 Conclusions

Our work focuses on the study and creation of bait information with the aim of exposing or thwarting the exploitation of exfiltrated information by malicious insiders. As future work, we intend to explore how this approach might also be applicable in detecting accidental violations of policy, as a means of warning users and organizations about such violations. The benefit of using the proposed decoy document system for this purpose is that it can potentially operate without the privacy repercussions if a mistake is made; such a benefit differentiates the approach from traditional monitoring approaches. Another direction to explore is how to improve the believability of decoys documents. We are planning a series of user studies to help us determine how users treat different attributes of a document in a specific context, such as whether an attacker would find more believable a document purporting to contain tax information that is encrypted/protected with a weak (predictable) passphrase, compared to an unprotected version of the same document.

In conclusion, although the use of bait information and similar trap-based defenses is well known, most of those efforts have focused either on artifacts that are logically separate from the operational systems (*e.g.,* honeypots [22]) or on low-level snippets of information created manually (*e.g.,* fake database records [23]). The D^3 system is a scalable and automated trap-based defensive system that forces attackers to expend considerable effort to identify realistic useful information from purposely planted bogus information intended to deceive. Naturally, the probability of exposing a malicious insider with trap-based defense tactics increases with the amount of decoy information that is generated and disseminated. D^3 offers the novel service of automatically creating and managing decoy documents, enabling the throttling of bait based on the desired protection level or cost (*e.g., interference*) one is willing to pay.

Acknowledgments

This material is based upon work supported in part by the US Department of Homeland Security under Grant Award Number 60NANB1D0127 with the Institute for Information Infrastructure Protection (I3P), the Army Research Office (ARO) Under Grant Award W911NF-06-1-0151 - 49626-CI, and the National Science Foundation (NSF) under Grant CNS-07-14647. The I3P is managed by Dartmouth College. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security, the I3P, ARO, NSF, or Dartmouth College.

We give special thanks to the Sandia National Laboratories Doctorate Study Program for supporting Brian Bowen and to Henner Mohr for his diligent effort and contributions to the development of the D^3 website and decoy document content.

References

1. Bell, D.E., LaPadula, L.J.: Secure Computer Systems: Mathematical Foundations, MITRE Corporation (1973)
2. Bell, J., Whaley, B.: Cheating and Deception. Transaction Publishers, New Brunswick (1982)
3. Butler, J., Sherri, S.: Security: Spyware and Rootkits. In: Login, December 2004, vol. 29(6) (2004)
4. Clark, D.D., Wilson, D.R.: A Comparison of Commercial and Military Computer Security Policies. In: IEEE Symposium on Security and Privacy, pp. 184–194 (1987)
5. Demers, A., Gehrke, J., Hong, M., Panda, B., Riedewald, M., Sharma, V., White, W.: Cayuga: A General Purpose Event Monitoring System. In: CIDR, pp. 412–422 (2007)
6. Detristan, T., Ulenspiegel, T., Malcom, Y., Von Underduk, M.S.: Polymorphic Shellcode Engine Using Spectrum Analysis. Phrack 11, 61–69 (2003)
7. Friess, N., Aycock, J.: Black Market Botnets. Department of Computer Science, University of Calgary, TR 2007-873-25 (July 2007)
8. Hoang, M.: Handling Today's Tough Security Threats. Symantec Security Response (2006)
9. The Honeynet Project, `http://www.honeynet.org`
10. The Honeynet Project, Know Your Enemy: Sebek, A Kernel based data capture tool (November 2003)
11. Honeypot Mailing List, Security Focus, `http://www.securityfocus.com/archive/119`

12. Katz, J., Yehuda, L.: Introduction to Modern Cryptography. Chapman and Hall CRC Press, Boca Raton (2007)
13. Kravets, D.: From Riches to Prison: Hackers Rig Stock Prices. Wired Blog Network (September 2008)
14. Krebs, B.: Web Fraud 2.0: Validating Your Stolen Goods. The Washington Post (August 20, 2008)
15. Li, W., Stolfo, S.J., Stavrou, A., Androulaki, E., Keromytis, A.: A Study of Malcode-Bearing Documents. In: Hämmerli, B.M., Sommer, R. (eds.) DIMVA 2007. LNCS, vol. 4579, pp. 231–250. Springer, Heidelberg (2007)
16. Maloof, M., Stephens, G.D.: ELICIT: A System for Detecting Insiders Who Violate Need-to-know. In: Kruegel, C., Lippmann, R., Clark, A. (eds.) RAID 2007. LNCS, vol. 4637, pp. 146–166. Springer, Heidelberg (2007)
17. McRae, C.M., Vaughn, R.B.: Phighting the Phisher: Using Web Bugs and Honeytokens to Investigate the Source of Phishing Attacks. In: Proceedings of the 40th Hawaii International Conference on System Sciences (2007)
18. Orbiscom, http://www.orbiscom.com/
19. Richardson, R.: CSI/FBI Computer Crime and Security Survey (2007)
20. Smith, R.M.: Microsoft Word Documents that Phone Home. Privacy Foundation (August 2000)
21. Song, Y., Locasto, M.E., Stavrou, A., Keromytis, A.D., Stolfo, S.J.: On the infeasibility of modeling polymorphic shellcode. In: Proceedings of the 14th ACM conference on Computer and communications security (CCS 2007), pp. 541–551 (2007)
22. Spitzner, L.: Honeypots: Catching the Insider Threat. In: Proceedings of ACSAC, Las Vegas (December 2003)
23. Spitzner, L.: Honeytokens: The Other Honeypot. Security Focus (2003)
24. Stoll, C.: The Cuckoo's Egg. Doubleday (1989)
25. Symantec. Global Internet Security Threat Report, Trends for July –December 2007 (April 2008)
26. Webb, S., Caverlee, J., Pu, C.: Social Honeypots: Making Friends with a Spammer Near You. In: Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS 2008), Mountain View, CA (August 2008)
27. Ye, N.: Markov Chain Model of Temporal Behavior for Anomaly Detection. In: Proceedings of the 2000 IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY, June 2000, pp. 171–174 (2000)
28. Yuill, J., Denning, D., Feer, F.: Using Deception to Hide Things from Hackers: Processes, Principles, and Techniques. Journal of Information Warfare 5(3), 26–40 (2006)
29. Yuill, J., Zappe, M., Denning, D., Feer, F.: Honeyfiles: Deceptive Files for Intrusion Detection. In: Proceedings of the 2004 IEEE Workshop on Information Assurance, United States Military Academy, West Point, NY, June 2004, pp. 116–122 (2004)