Lost in Translation: Improving Decoy Documents via Automated Translation

Jonathan Voris, Nathaniel Boggs, and Salvatore J. Stolfo
Department of Computer Science
Columbia University
New York, NY 10027, USA
{*jvoris,boggs,sal*}*@cs.columbia.edu*

***Abstract*—Detecting insider attacks continues to prove to be one of the most difficult challenges in securing sensitive data. Decoy information and documents represent a promising approach to detecting malicious masqueraders; however, false positives can interfere with legitimate work and take up user time. We propose generating foreign language decoy documents that are sprinkled with untranslatable enticing proper nouns such as company names, hot topics, or apparent login information. Our goal is for this type of decoy to serve three main purposes. First, using a language that is not used in normal business practice gives real users a clear signal that the document is fake, so they waste less time examining it. Second, an attacker, if enticed, will need to exfiltrate the document's contents in order to translate it, providing a cleaner signal of malicious activity. Third, we consume significant adversarial resources as they must still read the document and decide if it contains valuable information, which is made more difficult as it will be somewhat scrambled through translation. In this paper, we expand upon the rationale behind using foreign language decoys. We present a preliminary evaluation which shows how they significantly increase the cost to attackers in terms of the amount of time that it takes to determine if a document is real and potentially contains valuable information or is entirely bogus, confounding their goal of exfiltrating important sensitive information.**

***Keywords*-Insider Threat, Decoys, Automated Translation.**

I. Introduction

The insider threat continues to plague both private companies and government agencies. These organizations face the difficult task of differentiating legitimate and illegitimate accesses to data even after users have already been authenticated. A high profile example of an insider attack occurred recently when a software engineer contracted by the Federal Reserve Bank of New York was charged with the theft of ten million dollars worth of source code [1]. This event did not occur in isolation; according to a Cyber Policy Review released by the White House, data theft cost U.S. companies a trillion dollars in 2008 [2]. The amount of damage caused by insider action can only be expected to increase, as the number of companies who experienced such an attack grew by twelve percent from 2008 to 2009 [3].

In addressing this threat, researchers tend to focus on two general populations: "traitors" who abuse their own legitimate access to exfiltrate data and malicious masqueraders who use compromised credentials to gain access to sensitive material. Careful traitors are extremely difficult to deal with from a purely technological standpoint as they can, at the very least, exfiltrate their knowledge without detection. Many promising technologies hope to detect masqueraders and traitors as they change their behavior.

Trap based decoy documents are a particularly promising technique for detecting masqueraders [4]. This approach exploits a knowledge gap that exists between legitimate and unauthorized individuals. Normal users will be familiar with the documents that are present in their system and are therefore capable of differentiating decoys from real files. Masqueraders and even traitors who are less familiar with the particular data sets that are in use, on the other hand, should be more likely to access decoy material due to its believability. By leveraging this difference in data awareness, we can provide a powerful additional layer of security that can serve as one of the last lines of defense against attacks.

A core advantage to the use of decoys to combat insider threats is the way in which they alter adversarial behavior. When decoys are deployed, attackers must execute their attacks flawlessly while being careful not to trip over too many decoys. In this case, malicious entities need only be caught once, which turns the advantage in favor of the system's defense. This stands in contrast to the situation that exists at the initial compromise phase during which attackers must only succeed once while defenders must guard against all potential attacks.

While promising, decoy documents have a number of associated difficulties that must be overcome before they can be widely employed in scale throughout a large organization. Decoy monitoring systems must be difficult to circumvent lest decoys lose their detectability, rendering the entire scheme meaningless. Tamper resistant host sensors are promising in this context. The remaining challenges of effective decoy document usage can be reduced to a cost benefit trade off. The amount of decoys that are deployed and the conspicuousness of the locations in which they are placed must be balanced against the cost of false positives that occur when decoys interfere with the workflow of typical users. The more believable that decoy documents are, the more time that standard users will waste while dealing with them. On the other hand, clever adversaries can find ways to detect and avoid obvious decoys, such as mimicking the

DOI 10.1109/SPW.2012.20

decoy access patterns that are exhibited by normal users.

In this paper we present a novel method of crafting decoys using automated language translation, with the aim of addressing a number of the aforementioned challenges to the use of decoy documents in combating insider threats. We recommend seeding a user's file system with decoy documents that are written in a language that is different than the ones that are typically employed. Though these documents will be written in a language that is not ordinarily in use in the organization, in many contexts it is plausible that such documents would be commonplace. Perhaps an employee speaks a foreign language, for example. It is conceivable that the choice of such a language might be made on the basis of a user's background, making it more plausible that such documents would appear in his or her file system.

By translating some text into a foreign language that is not in business use while leaving some enticing references, such as company names, in English, each decoy provides three important new properties. First, decoys become more differentiable to normal users because they provide a clear signal that these documents are fake. Users will therefore waste less time reviewing them. Second, malicious entities will be more likely to exfiltrate the data that is contained in a foreign language decoy in order to translate the document before discovering that it is a decoy. This improves detectability by providing a clearer signal of malintent, since normal users would have no reason to exfiltrate such decoys. Finally, translating documents means that inside attackers must also invest additional time and computational resources to their task. This is due to the fact that they must first translate decoy documents, then read through a text that is less intelligible in order to discern whether the information is real or fabricated.

In essence, our method expands the knowledge gap between real users and adversaries by utilizing information regarding the languages that are employed by these parties. We intend to leverage this information to increase the accuracy of decoys while raising the bar for adversarial effort. In this work, we conduct a preliminary study of the amount of effort that an attacker must expend in order to decipher a translated decoy.

The remainder of this paper is structured as follows. Section II reviews prior research on the use of decoys to thwart insider attacks. Section III discusses the design and execution of our exploratory study of foreign language decoys. Section IV presents the results of our experiments and discusses their implications for the construction of insider threat defense mechanisms. Section V describes future research we wish to pursue in this area, including user studies and practical deployment work. Finally, Section VI summarizes our conclusions.

II. Related Work

The use of deceptive techniques, such as disinformative propaganda, to thwart one's enemies has long been an established element of warfare. A well known example of this is Operation Bodyguard, which was an Allied plan used during World War II to distract German forces from the invasion of Normandy [5]. The first use of deception in the context of computer security is attributed to Cliff Stoll, who set up and monitored a spurious set of computing resources in order to catch hackers who were attempting to exfiltrate information from Lawrence Berkeley National Laboratory [6].

Decoy files are well suited to the challenge of detecting insider threats because they can be used to issue alerts when attackers start accessing files even after all other defenses are circumvented. Creating, distributing, and managing these files is not a trivial task, however. To help address this, Bowen et al. devised a system which automates the process of creating decoy documents [7]. They identified a set of properties that effective decoys should have, namely:

1) Believability: Decoys should appear legitimate to adversaries.
2) Enticingness: Unauthorized users should find the decoy content to be alluring.
3) Conspicuousness: Decoy material should be easy to find.
4) Detectability: An insider threat defense system must be able to monitor decoy access activity.
5) Variability: There should be no shared properties that set decoys apart from real data.
6) Non-interference: Decoys should not get in the way of the workflow of legitimate users.
7) Differentiability: Real users should be able to easily distinguish between decoys and actual documents.
8) Shelf-life: Decoy material may lose effectiveness after a given time frame.

In addition, Ben Salem and Stolfo investigated deployment techniques to optimize the efficacy of decoy files via a user study [4]. They observed several tradeoffs between the desirable characteristics of decoys. Specifically, they developed guidelines for ways in which such documents can be made more enticing to insider attackers while remaining differentiable by legitimate users and non-interfering with ordinary workflow [4].

Despite this progress in the development of decoy files that can effectively detect insider threats, there remain a variety of dimensions in which decoy documents can be improved. Although techniques for minimizing false positive decoy document accesses during the deployment phase were developed in [4], decoy accesses by benign users are still likely to occur to some degree in practice. These events interfere with normal business practices by wasting users' time or confusing them with fabricated information. In this

paper, we present further ideas for minimizing these false positives.

III. Evaluation

We conduct a preliminary evaluation in order to determine whether translating a decoy file into a foreign language is beneficial. Specifically, we wish to investigate the effect of the translation procedure on a masquerader's ability to differentiate decoys from real files by discerning whether a document's content is authentic or spurious. One indication of a document's authenticity is its intelligibility. A well-written document is more likely to be accepted as legitimate, while a garbled text has a higher probability of arousing suspicion when read. We therefore desire to measure the extent to which an automated translation service mangles its input.

Another method that an adversary may use to expose a file as a decoy is investigating the origin of its subject matter. If an attacker is able to find material on the web that is nearly identical to what is written in a document, he or she may be able to conclude that the file in question is a decoy that was generated using content harvesting techniques. An additional experimental objective is thus to observe how translating a text excerpt complicates the process of searching for its content source on the Internet.

We also seek to measure the impact of adding intermediary languages transitions to a translation path. That is, what effect does converting an excerpt to one language prior to translating it to a final language have on the terminal text? We want to confirm our intuition that this will reduce the clarity of the writing while increasing the difficulty of discovering its source, with an eye towards finding a desirable tradeoff between these factors.

In order to answer these queries, excerpts were selected from five arbitrary articles on the Internet. All are relatively recent, being published within a year, and are readily available in the top results of popular search engines. We deliberately choose technical topics so that there would be terms that would remain in English after the translation process. This is a critical component of foreign language decoys because these English language technical terms will stand out, augmenting the enticingness of the document in which they appear. The topics of the articles that we picked were the Stuxnet worm [8], sanctions against Iran [9], Facebook's initial public offering [10], insider trading arrests [11], and a new email security standard [12].

Two paragraphs were selected from each of these articles as we were particularly interested in measuring whether or not an adversary can easily tell if a large document is a decoy from a small sample. These each went through Google's web based translation service [13]. To ensure that our results were independent from the intricacies of any particular language, different languages were utilized for each hop. As an example, our first article was translated from English to Albanian, then to Galician, followed by Haitian Creole, then Polish, and finally Icelandic. Each text excerpt traveled through a translation path of five arbitrarily chosen languages in this fashion.

We record the resultant text after each additional step of translation and convert each back to English to measure the effect that the translation stack had on the underlying English text up until that point. In order to quantify the distortion that translation introduced, we developed several metrics that we believe are useful for approximating the amount of effort that an adversary would have to expend in order to find the original source text.

The first attribute that we measure is the length of the longest substring that is shared between the original document and its variants that had been translated back in to English. The shorter the shared runs of text are between the original and the manipulated portions of text, the less likely they will be recognized as sharing the same origin. Another characteristic that we analyze is the dictionary of words that comprise each blurb. Having many words in common may serve as an additional indicator that content is harvested from a specific source.

We also search for the first sentence from each translated excerpt and look at how high it appears in Google's search rankings. The first action an adversary may take when determining a document's legitimacy is to consult web search tools. A high search ranking result will send a clear signal that the translated document is nothing more than a disguised article taken from an online source.

While these evaluation measures provide a useful approximation of an adversary's ability to deduce a foreign language decoy's source, they do not indicate anything about the intelligibility of the translations themselves. Since we lacked expertise in the myriad of foreign languages through which our sample excerpts had passed, we sought the assistance of a professor of the Hebrew language from our university. We again send the same five segments of articles from the web through variable length translation paths, but this time we make sure that the terminal language is always Hebrew. To measure how garbled these were, we asked the Hebrew expert to rate their intelligibility on a ten point scale, with 1 representing nonsense and 10 meaning perfectly written Hebrew.

IV. Study Results

Figure 1 depicts the results of our translation analysis in terms of the size of the longest substring that is present in an original document and the English version of its translated counterpart. The measurement is provided as a percentage of the length of the original in order to normalize the length of the various text segments. The average shared substring length starts at 12.17% for the first translation step, then quickly drops to 7.41% before plateauing between 6% and 7%. This shows that multiple layers of translation are an

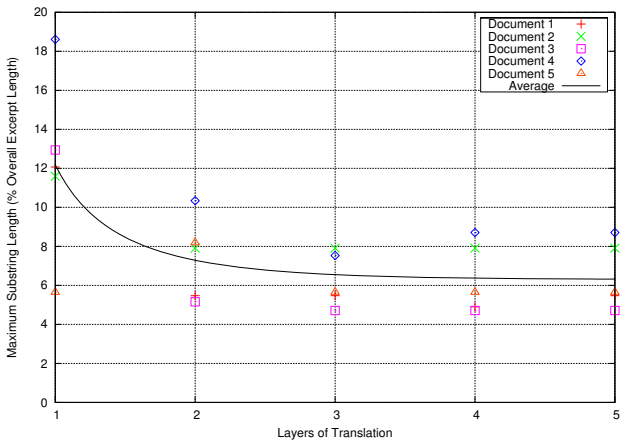

Figure 1. Maximum Shared Substring Length as a Function of Translation Depth

effective way to obfuscate the source of a decoy document, but diminishing returns begin to set in after 2 conversion passes. Note that there are significant variations based on the specifics of each document and language. For instance, Document 5 did not become much more obfuscated after its first translation step, while translating Document 1 from Haitian Creole to Polish between steps 3 and 4 seemed to add more variations than its next hop from Polish to Icelandic between hops 4 and 5.

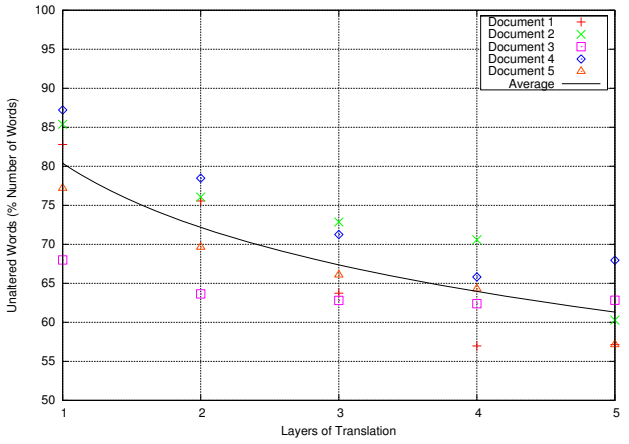

Figure 2. Amount of Shared Words as a Function of Translation Depth

Next, Figure 2 illustrates the impact of translation depth on the number of unique words that are shared between the original excerpt and its translated and decoded variants. This is measured as a percentage of the number of unique words in the altered portions of text to factor in the length changes that occur as a result of the translation process. The effect of language conversion on the number of unaltered words is more linear than on the shared substring size. For translation steps one through four, each document displayed a monotonic decrease in the amount of constant terms. The percentage fell from 80.11% for the first step down to 64.01% on average for the fourth transition. The fifth and final translation link brought an additional decrement of 2.94% on average, but diminishing returns were present here as well since three of the five test excerpts actually showed slight increases in the number of similar vocabulary elements after this step.

Translation Depth	**Google Ranking by Document**				
	1	2	3	4	5
0	2	1	1	1	1
1	4	>100	1	None	None
2	1	7	>100	None	None
3	1	9	>100	None	None
4	3	9	>100	None	None
5	5	None	>100	None	None

Table I
RANKING OF SOURCE DOCUMENT WHEN SEARCHING FOR FIRST SENTENCE OF A TRANSLATED EXCERPT

Table I shows the effect of foreign language translation on the location of the source article in the results of a Google search for the first sentence in each portion of text that was translated and then changed back to English. The relationship between translation depth and search index ranking is not as straightforward as the previous two metrics. Translation quickly removed Documents 4 and 5 from the search results, as these were absent after a single language conversion pass. Document 3 took two steps to no longer appear as the top result, but was not present in Google's top 100 results thereafter.

While translation did an excellent job of obscuring these sources, Documents 1 and 2 proved more difficult to hide. The first translation step brought Document 2 out of the top 100 Google search results, yet it appeared in the top ten results again following hops 2 through 4 before finally disappearing altogether during the last translation step. Translation did the worst job of masking the source of Document 1, which never fell out of the top ten. Counterintuitively, iterations 2 and 3 actually increased the search ranking of Document 1. A possible explanation for this result is that Document 1 concerned the Stuxnet worm [8], a topic which is both very popular and technical.

We believe that the combination of this article's unique terminology and ongoing popularity contributed to its persistence in Google's search results despite repeated rounds of translation. Alternatively, this article may be easy to find because it is fairly new and there is not as much material available about it online as there is for the other topics. While it is difficult to conclude what factors influenced the ranking of this particular document in the results of our translated search queries, the choice of terms and content obviously has a large impact on this. In a realistic insider defense system, these elements should be chosen based

upon the context of the organization that is utilizing decoy documents in their defense.

When asked to rank the intelligibility of our Hebrew excerpts on a ten point scale, our language expert awarded both the baseline single hop translations and the moderate two step translations a 5 on average. She rated the more extreme three iteration translations a mean value of 4. This provides some preliminary evidence that while translating decoy content to a foreign language does degrade the intelligibility of its contents to some extent, repeated rounds of translation do not muddle the text as significantly as the first operation.

V. Future Work

While the results presented in this paper represent a useful first step towards enhancing decoy documents via language manipulation, they do not capture a full picture of their impact on the behavior of users and inside attackers. We intend to perform a much more comprehensive analysis of the effect of language translation on decoy document efficacy as future work. Rather than examining individual examples, we will measure a collection of decoys in order to estimate the overall adversarial workload that is necessary to differentiate a real document that is hidden among a collection of decoys. This will be done via a user study that is performed with decoy documents that have been placed in a realistic environment. We aim to use [4] as a model for this future experiment, but we will use the extent to which documents are translated as an experimental variable rather than the volume of decoys that are deployed.

VI. Conclusion

In conclusion, our work provides preliminary support for a method to increase decoy documents' enticingness to attackers while making them as unobtrusive as possible to legitimate users. Specifically, enhancing decoys through automated language translation can make these documents more enticing to adversaries, leading to an increase in exfiltration attempts. Furthermore, the translation process can potentially make decoys more easily avoidable by typical users while causing adversaries to consume more time and effort. The results of our evaluation provide preliminary evidence in support of our claim that translation between languages provides a useful tool for reducing the error rates associated with decoy based insider threat detection systems.

Acknowledgments

We would like to thank Professor Rina Kreitman for her help with the development and evaluation of our foreign language decoy material. This work was supported as part of the DARPA ADAMS Program, Anomaly Detection at Multiple Scales, No. W911NF-11-1-0140. Professor Stolfo is the founder of Allure Security Technology, Inc.

References

[1] B. Katz, "U.S. charges Chinese man with NY Fed software theft," Available at http://www.reuters.com/article/2012/01/19/us-nyfed-theft-idUSTRE80H27L20120119, 2012.

[2] White House Cyber Policy Review, "Assuring a Trusted and Resilient Information and Communications Infrastructure," Available at http://www.whitehouse.gov/assets/documents/Cyberspace_Policy_Review_final.pdf, 2009.

[3] M. Maxim, "Defending Against Insider Threats to Reduce Your IT Risk," Available at http://www.ca.com/~/media/Files/whitepapers/insider-threat-wp-jan-2011.pdf, 2011.

[4] M. Ben Salem and S. Stolfo, "Decoy Document Deployment for Effective Masquerade Attack Detection," in *Conference on Detection of Intrusions and Malware and Vulnerability Assessment*, 2011.

[5] J. Rubin, "Deception: The other 'D' in D-Day," Available at http://www.msnbc.msn.com/id/5139053/ns/msnbc_tv-the_abrams_report/t/deception-other-d-d-day, 2004.

[6] C. Stoll, "The Cuckoo's Egg," 1989.

[7] B. Bowen and S. Hershkop and A. Keromytis and S. Stolfo, "Baiting Inside Attackers Using Decoy Documents," in *Conference on Security and Privacy in Communication Networks*, 2009.

[8] K. Zetter, "How Digital Detectives Deciphered Stuxnet, the Most Menacing Malware in History," Available at http://www.wired.com/threatlevel/2011/07/how-digital-detectives-deciphered-stuxnet/all/1, 2011.

[9] I. Lakshmanan, "Iran's Oil, Tanker Firms Targeted for Sanctions," Available at http://www.bloomberg.com/news/2012-01-31/iran-s-oil-tanker-companies-targeted-for-sanctions-by-lawmakers.html, 2012.

[10] J. McGregor, "Facebook IPO may be coming Wednesday," Available at http://www.washingtonpost.com/business/economy/facebook-ipo-may-be-coming-wednesday/2012/01/30/gIQAlCPyeQ_story.html, 2012.

[11] J. O'Toole, "Perfect Hedge: 56 found guilty of insider trading," Available at http://money.cnn.com/2012/01/24/news/economy/insider_trading/index.htm?iid=HP_LN, 2012.

[12] K. Higgins, "Google, Facebook, Bank Of America Behind New Email Security Standard," Available at http://www.darkreading.com/authentication/167901072/security/application-security/232500733/google-facebook-bank-of-america-behind-new-email-security-standard.html, 2012.

[13] "Google Translate," Available at http://translate.google.com/, 2012.