Insider Data Theft Prevention System

Mr. Kunal Madhukar Shirkande [*], **Prof. Prajakta A. Satarkar [MTech CSE]** [**]

Computer Science & Engineering Department, SVERI's College Of Engineering, Gopalpur-Ranjani Road, Gopalpur, Pandharpur 413304, Dist. Solapur.

Abstract- Insider data theft attacks are characterized by an adversary stealing a legitimate user's credentials and using them to impersonate the authenticate user and to perform malicious activities. Prior works also combines a user behavior profiling technique with a baiting technique, but profiling user behavior using single modeling technique suffers from a considerable number of false positives. Also decoy documents are placed at conspicuous locations rather than using automatically generated decoys which may not give significant accuracy to the detection system. Proposed system will extend prior work and presents an integrated detection approach where behavior profiling will be done by combining more than one classifier, each uses different modeling algorithm to reduce false positive rate. Along with this proposed system will include a baiting approach based on automated generation of demand decoy documents on the user's file system and user authentication by challenge questions, to provide more accuracy. Proposed system could provide a strong defense mechanism against malicious insider data theft attacks.

Index Terms- Behavior profiling technique, Anomaly detection, Decoy documents, User authentication by challenge questions

I. Introduction

In many business organizations, the most destructive form of attacks against a computer network is not from the outside hacker, but from trusted entities that belong to the internal organization. Data theft attacks are augmented if the attacker is a malicious insider. This is considered as one of the top threats to cloud computing by the Cloud Security Alliance.Insiders may get the credentials of authorized user by sniffing password for accessing system illegitimately. The insider can be considered extremely dangerous, due to an in-depth knowledge of the target and trusted access to sensitive information on the system.

The Twitter incident is one of the popular examples of an insider data theft attack from the Cloud. Several Twitter corporate and personal documents were illegitimately copied to technological website TechCrunch. Customer's accounts, including the account of U.S. President Barack Obama, were illegitimately accessed. The attacker used a Twitter administrator's password to get access over Twitter's corporate documents, hosted on Google's infrastructure as Google Docs.

Much research in security has paying attention on ways of preventing insider data theft by developing sophisticated access control and encryption mechanisms. However these mechanisms have failed to prevent data damage caused by insider data theft.
A possible solution proposed here is, an integrated approach for masquerade attack detection which uses behavior profiling (using ensemble of classifiers with different modeling algorithm) with generating on demand decoy documents to bait attackers along with user authentication by challenge questions. The approach could improve detection accuracy over prior techniques and will be less vulnerable to impersonation attacks.

II. MOTIVATION

Existing algorithms used for modeling user behavior uses statistical features, such as the sequence of user commands or co-occurrence of multiple events combined through logical operators. The anomaly detectors built using these algorithms are used then to determine deviations from normal user behavior. Anomaly detectors suffer from low accuracy and particularly from high false positive rates. One way to overcome this shortcoming is by combining several base classifiers to create one ensemble of classifiers. Each classifier uses a different modeling algorithm to profile user behavior.

If the different classifiers have highly overlapping anomaly spaces, then when evading detection by one classifier by impersonating real user behavior, an attacker is likely to get away from detection of the other classifiers. Combining different classifiers in this case does not comprise a good defense mechanism against mimicry attacks. To overcome the limitations of model diversity, proposed system will merge the two different detection techniques .The first is a modeling technique based on profiling behavior profiling. The second is a baiting technique where decoy technology is used.

In prior work, Decoys files are strategically placed at conspicuous locations by the legitimate user in their own file system, but there is no guarantee that fake user will surely touch these files. So to overcome this limitation proposed system will generate on demand decoy documents, if user is suspected as masquerader by behavior profiling. Any access to these decoy documents is then considered as indicative of malicious insider activity.

III. LITERATURE REVIEW

Stolfo et al proposed a combined approach for detecting masquerade attacks [1]. The authors focused on modeling user search behavior with a baiting technique to reveal an attacker's malicious intent. They hypothesized and showed that a masquerader would engage in search activities different from those of the legitimate user in terms of their volume and frequency.

Maloof et al applied a user behavior profiling technique to detect malicious insider activities which violated `Need-to-Know' policy [2]. In order to identify bad insider behavior, they defined the malicious user scenarios and had to combine results different sensors through a Bayesian net. Although the few attack scenarios tested were detected, there was no real evaluation of the false positive rate associated with the overall classifier.

Hershkop et al surveyed that most of the prior user behavior profiling work focused on auditing and modeling sequences of user commands including work on enriching command sequences with information about command arguments. A thorough review of these machine learning techniques can be found in this survey [3]. The detection rates of these anomaly detection techniques ranged between 75.8% and 26.8%, with false positive rates ranging between 1% and 7%. These results are obviously far from satisfactory.

Chawla et al presented a novel approach to distributed learning using fuzzy clustering [4]. This intelligent method of partitioning a dataset is compared to simpler, random methods of partitioning. The results presented in this paper suggest that for very large datasets, the creation of ensembles of classifiers can perform reasonably well.

Dzeroski et al empirically evaluated several state-of-the art methods for constructing ensembles of heterogeneous classifiers with stacking and shown that they perform comparably to selecting the best classifier from the ensemble by cross validation [5]. They had proposed a new method for stacking which uses multi-response model trees at the meta-level.

Bowen et al concluded as masquerade attacks pose a grave security problem and detecting masqueraders is very hard [6]. In this paper, the author has investigated the use of such trap-based mechanisms for the detection of masquerade attacks. They evaluated the desirable properties of decoys deployed within a user's file space for detection.

IV. PROBLEM STATEMENT

Proposed system presents an integrated detection approach where profiling user search behavior will be created by combining more than one classifier to reduce false positive rate. Along with this proposed system will use a baiting approach based on the crafting on demand decoy documents on the user's file system and user authentication by challenge questions. The proposed system will be designed to prevent unauthorized and illegitimate access to the system and to provide security to the user's data by combining behavior profiling and decoy documents.

V. PROPOSED WORK

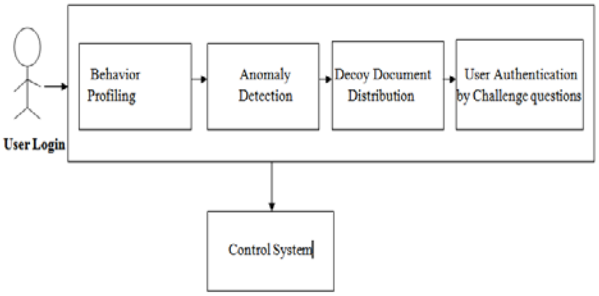
A. SCOPE

Profiling user behavior is an anomaly detection technique. Anomaly detection has the potential to identify new and unknown attacks. While profiling user behavior, any abnormal behavior is indicative of masquerade activities. However, it may produce high false positive rates, particularly if the user model does not simplify well because of lack of model training data. An anomaly detector suffers from low accuracy, and mainly from high false positive rates. One way to overcome this shortcoming is by combining several base classifiers into ensemble classifier. Each classifier will use a different modeling algorithm to profile user behavior.

The ensemble methods output collectively one classification label which reacts the meta-learning from these models or the consensus amongst them. The objective of using such ensemble methods is to improve robustness and classification accuracy over single-model methods.An adversary may know how a user behaves and execute a `mimicry attack'. However, they are improbable to know what the sufferer knows. If the sufferer baits the system with automatically generated decoy documents may trap the adversary later. Even a sophisticated adversary who mimics the sufferer or target user may still get trapped, as they do not know that the decoys are being accessed.

If the anomaly detector is subject to insider attack, then the decoy file monitoring is likely to catch the masquerade activity. If decoy documents are wrongly accessed by user then user authentication by challenge questions is likely to catch masquerade activity. Hence dependency and interlinking between above techniques could increase the overall efficiency of proposed system by reducing false positive rates.

Following Figure 1 shows the Architecture of the insider data theft detection system:


User Login
Behavior Profiling
Anomaly Detection
Decoy Document Distribution
User Authentication by Challenge questions
Control System

Profiling search behavior detects anomalous user search behavior forming a baseline of standard search behavior. Then it monitors for abnormal search behaviors that reveal large deviations from the baseline. The system builds a normal user model C that models the user's search behavior by extracting features. System measures the deviation between actual user behavior and the historical user behavior as defined by the normal user model C. The distance D is compared in order to determine whether there is enough proof for masquerade activity.

If there is difference between current behavior and standard behavior, that user will be flooded by automatically generated decoy documents. The true user has an idea about decoy files. That why true user will not supposed to access to these files. On the other hand fake user does not know that files are fake files. So he will access that files.

The insider data theft detection system generates alert when decoy documents are being accessed, copied or read. As soon as the decoy document is loaded into memory by any application, system verifies whether that file is an original file or a fake decoy file by computing a Hypertext Message Authentication Code (HMAC) which is embedded in that file and comparing it with previously embedded document. If the two HMACs match, the document seems to be a decoy otherwise, the document seems to be normal. If the decoy documents are not being accessed then user activity is not malicious enough.

Following figure describes the flowchart of overall decision process related to masquerade alert generation.

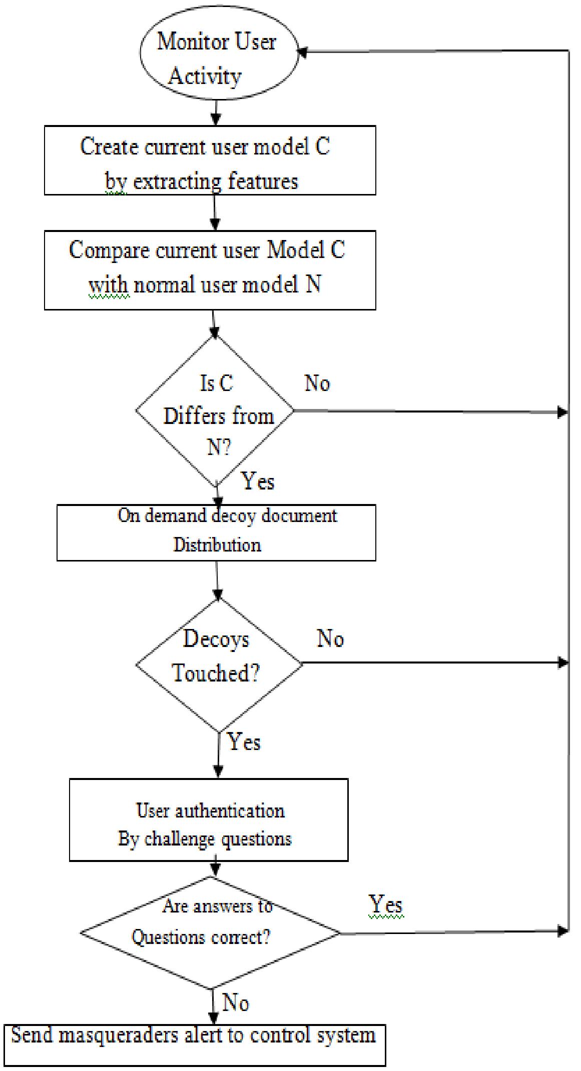

Monitor User Activity
Create current user model C by extracting features
Compare current user Model C with normal user model N
Is C Differs from N?
No
Yes
On demand decoy document Distribution
Decoys Touched?
No
Yes
User authentication By challenge questions
Are answers to Questions correct?
Yes
No
Send masqueraders alert to control system

Fig.2. Alert Generation Process

B .METHODOLOGY

Insider data theft attack detection system will be developed using following modules:

Module 1: Validating user logins

The application will be deployed on a system. The Application will be used to validate the insider data theft attack detection system. User logins are the basic inputs for detection system. Application will include the following options:

1. It will store user name, password, confirm password, contact number and at least ten secrete questions at the time of account creation.

2. It will allow strict entry time checking by querying the user with randomly selected secret questions on each login.
Choice of selecting questions will be done by user at the time of account creation.

Module 2: User access behavior profiling

Anomalous user behavior can vary from normal user behavior baseline. Regular user has an idea about the file structure and contents of the system. So his search for files is limited and specific. According to this assumption standard user model will be prepared by extracting features such as search pattern, number of touches to file, keystrokes, mouse movements etc. My contribution to existing work will be to reduce feature set by selecting minimum distinguishing features among them.

Also I wish to contribute by creating ensemble of classifier for reducing low accuracy of anomaly detection. Some of the following classifiers from following will be combined in my work.

1. One class support vector machines: They are Linear classifiers used for classification and regression. It maps input data into a high-dimensional feature space using a kernel function.

2. Naïve bayes classifier: It uses only data from a single user when training a classifier to profile a distinct user. It calculates the likelihood that a command block belongs to masquerader or not.

3. Multivariate bernoulli event model: A vector of binary attributes is used to represent a document indicating whether the command occurs or doesn't occur in the document.

4. Multinomial model: It uses the number of command occurrences to represent a document, which is called "bag-of-words", capturing the word frequency information in documents.

Module 3: Anomaly detection

The system will be developed where current user behavior will be modeled. It will be compared with the standard behavior model of that user. If the difference is exceeding the limit, then that user is suspected to be masquerader. It will be the first suspecting alert of my detection system. User will be exposed to next module only if this alert is generated. If the current user behavior is same as the past behavior, there is no need to traverse the next modules and the user is allowed to continue his work on original data.

Module 4: Decoy document distribution

Whenever alert by anomaly detection is generated, decoy information such as tax returns, bank receipts, policy documents etc may be supplied immediately on demand. My contribution to existing work will be to supply decoy documents automatically on generation of first alert rather than keeping decoy traps. The system will maintain same directory and file structure for the decoy file system and the original file system rather than providing irrelevant bogus data in the decoy files for confusing the attacker. The information contained in the decoy file is delivered in such a way as to appear completely normal to the attacker and he will not get any doubt of fake or worthless data is being served to him.

Module 5: User authentication by challenge questions

If a decoy document is loaded into memory then current user's behavior seems abnormal. If decoy documents will be accessed by authenticate user by mistake, even authenticate user also will be treated as masquerader by detection system. Hence it will increase false positive rate. To overcome this limitation, a set of challenge questions will be asked to the user whose answers are only known to the real user. This will improve accuracy of overall detection system.

The answers to these questions will be given by the administrator of the system during the installation of the application. Choice of these questions will not be done by the user; administrator will choose the questions and will provide answers which are totally confidential between the user and him. Questions can be like, what is the secret password key of the company members? This module can be shifted upwards (before the decoy document distribution) on the basis of experimental results.

Module 6: Control system

The control system will represent an interface to view the malicious insider accesses. It will allow the administrators to enforce give/reject policies for the remote users. Administrator can block the masquerader by denying access to the systems data. When attacker will be caught then administrator will notify true user to set his credentials again. It will maintain logs of anomaly detection system.

VI. FACILITIES AVAILABLE AND REQUIREMENTS

Hardware required	Few systems with interconnected network Intel® Pentium IV & above, Min 512MB RAM.
software required	jdk 1.7 .

Proposed system will require any heterogeneous network and will work on any platform.

REFERENCES

[1] Salvatore J. Stolfo,Malek Ben Salem, Angelos D. Keromytis,“ Fog Computing: Mitigating Insider Data Theft Attacks in the Cloud”, IEEE Symposium on Security and Privacy Workshops,July 2012.

[2] Maloof M.A.and Stephens et al,“ Detecting Insider Data Theft of Trade Secrets” ,published by the ieee computer and reliability societies ,november/december 2009 .

[3] Shlomo Hershkop et al “A survey of insider attack detection research”, In Insider Attack and Cyber Security: Beyond the Hacker, Springer(2008).

[4] Chawla, N. V., Eschrich S. and Hall L. O., “ Creating ensembles of classifiers.” In Proceedings of the 2001 IEEE International Conference on Data Mining (Washington, DC, USA, 2001),IEEE Computer Society, pp. 580-581.

[5] Dzeroski S., and Zenko B. “Is combining classifiers better than selecting the best one” In Proceedings of the Nineteenth International Conference on Machine Learning (San Francisco, CA, USA, 2002), ICML '02, Morgan Kaufmann Publishers Inc, pp. 123-130.

[6] Bowen B. M., Hershkop S., Keromytis A. D., and Stolfo S. J. , “ Baiting inside attackers using decoy documents.” In SecureComm'09: Proceedings of the 5th International ICST Conference on Security and Privacy in Communication Networks (2009).

[7] Ben-Salem, M., and Stolfo, S. J., “ Detecting masqueraders: A comparison of one class bag-of-words user behavior modeling techniques.” In MIST '10: Proceedings of the Second International Workshop on Managing Insider Security Threats, Japan (June 2010), pp. 3-13.

[8] M. Ben-Salem and S. J. Stolfo, “Combining a baiting and a user search profiling techniques for masquerade detection,” In Columbia University Computer Science Department, Technical Report#cucs01811,2011.

[9] Lingaswami, G. Avinash Reddy, “Offensive Decoy Technology For Cloud Data Attacks.”, International Journal of P2P Network Trends and Technology(IJPTT)–Vol.3 Issue 10-Nov 2013.

[10] Cloud Security Alliance, “Top Threat to Cloud Computing V1.0,” March 2010. [Online]. Available: https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf.

[11] M. Arrington, “In our inbox: Hundreds of confidential twitter documents,” July 2009. [Online].Available: http://techcrunch.com.

AUTHORS

First Author – Mr. Kunal Madhukar Shirkande, Computer Science & Engineering Department, SVERI’s College Of Engineering, Gopalpur-Ranjani Road, Gopalpur, Pandharpur 413304, Dist. Solapur., E-Mail: kmscsecoep@gmail.com

Second Author – Prof. Prajakta A. Satarkar [MTech CSE], Computer Science & Engineering Department, SVERI’s College Of Engineering, Gopalpur-Ranjani Road, Gopalpur, Pandharpur 413304, Dist. Solapur.