Modeling User Search Behavior for Masquerade Detection

Malek Ben Salem and Salvatore J. Stolfo
Computer Science Department
Columbia University
New York, New York, USA
{malek,sal}@cs.columbia.edu

***Abstract*—Masquerade attacks are a common security problem that is a consequence of identity theft. Masquerade detection may serve as a means of building more secure and dependable systems that authenticate legitimate users by their behavior. Prior work has focused on user command modeling to identify abnormal behavior indicative of impersonation. This paper extends prior work by modeling user search behavior to detect deviations indicating a masquerade attack. We hypothesize that each individual user knows their own file system well enough to search in a limited, targeted and unique fashion in order to find information germane to their current task. Masqueraders, on the other hand, will likely not know the file system and layout of another user's desktop, and would likely search more extensively and broadly in a manner that is different than the victim user being impersonated. We devise a taxonomy of Windows applications and user commands that are used to abstract sequences of user actions and identify actions linked to search activities. The experimental results show that modeling search behavior reliably detects all masqueraders with a very low false positive rate of 1.1%, far better than prior published results. The limited set of features used for search behavior modeling also results in large performance gains over the same modeling techniques that use larger sets of features.**

***Index Terms*—masquerade detection; user behavior profiling; intrusion detection; search behavior; one-class support vector machines;**

I. Introduction

The *masquerade attack* is a class of attacks, in which a user of a system illegitimately poses as, or assumes the identity of another legitimate user. Identity theft in financial transaction systems is perhaps the best known example of this type of attack. Masquerade attacks are extremely serious, especially in the case of an insider who can cause considerable damage to an organization. The masquerade and insider attack detection problem remains one of the more important research areas requiring new insights to mitigate against this threat. Masquerade detection may also serve as a means of building more secure and dependable systems that authenticate legitimate users by their behavior, rather than exclusively by their (possibly stolen) credentials.

A common approach to counter this type of attack, which has been the subject of prior research, is to apply machine learning algorithms that produce classifiers which can identify suspicious behaviors that may indicate malfeasance of an impostor. We do not focus on whether an access by some user is authorized since we assume that the masquerader does not attempt to escalate the privileges of the stolen identity, rather the masquerader simply accesses whatever the victim can access. However, we conjecture that the masquerader is unlikely to know the victim's search behavior when using their own system which complicates their task to mimic the user. It is this key assumption that we rely upon in order to detect a masquerader. The conjecture is backed up with real user studies. Eighteen users were monitored for seven days on average to produce more than 10 GBytes of data that we analyzed and modeled. The results show that indeed normal users display different search behavior, and that that behavior is an effective tool to detect masqueraders. After all, a user will search within an environment they have created. We assume the attacker has little to no knowledge of that environment and that lack of knowledge will be revealed by the masquerader's abnormal search behavior. Thus, our focus in this paper is on monitoring a user's behavior in real time to determine whether current user actions are consistent with the user's historical behavior, primarily focused on their unique search behavior. The far more challenging problems of thwarting mimicry attacks and other obfuscation techniques are beyond the scope of this paper.

Masquerade attacks can occur in several different ways. In general terms, a masquerader may get access to a legitimate user's account either by stealing a victim's credentials, or through a break in and installation of a rootkit or key logger. In either case, the user's identity is illegitimately acquired. Another perhaps more common case is laziness and misplaced trust by a user, such as the case when a user leaves his or her terminal or client open and logged in allowing any nearby co-worker to pose as a masquerader. In the first two cases, the identity thief must log in with the victim's credentials and begin issuing commands within the bounds of one user session. We conjecture that legitimate users initiate the same repeated commands each time they log in to set their environment before using it, initiate some set of applications (read email, open a browser, or start a chat session) and similarly, clean up and shut down applications when they log off. Such repeated behaviors constitute a profile that can be modeled and used to check the authenticity of a user session early before significant damage is done. The case of hijacking a user's session is perhaps a bit more complicated. In either case, a monitoring system ought to detect any significant deviations

from a user's typical profiled behaviors in order to detect a likely masquerade attack. Ideally, we seek to detect a possible masquerader at any time during a session.

In this paper we extend prior work on modeling user command sequences for masquerade detection. Previous work has focused on auditing and modeling sequences of user commands including work on enriching command sequences with information about arguments of commands [1], [2], [3]. We propose an approach to profile a user's behavior based on a 'taxonomy' of Windows applications, Dynamic Link Libraries (DLLs), and Microsoft Disk operating System (MS-DOS) commands. The taxonomy abstracts the audit data and enriches the meaning of a user's profile, thus helping reveal the **user's intent** and focus on specific user actions of interest such as search-related activities. Hence, commands or applications that perform similar types of actions are grouped together in one category making profiled sequences more abstract and meaningful. Most importantly, the use of the taxonomy significantly **reduces the dimensionality of the feature space**, and thereby reducing the complexity of the computation. Commands are thus assigned a type, and the sequence of command types is modeled rather than individual commands.

One particular type of commands is *information gathering* commands, i.e. *search* commands. We conjecture that a masquerader is unlikely to have the depth of knowledge of the victim's machine (files, locations of important directories, available applications, etc.), and hence, a masquerader would likely first engage in information gathering and search activities before initiating specific actions. To this extent, we conduct a set of experiments using a Windows data set that we have gathered in our department. We model search behavior in Windows and test our modeling approach using our own data, which we claim is more suitable for evaluating masquerade attack detection methods.

A. Contributions

The contributions of this work are:

- A **taxonomy of Windows applications and DLLs**: The taxonomy is used to abstract and enrich the meaning of user activities performed on the host system. This abstraction enables the reduction of features used for user behavior profiling, and therefore a significant decrease in computational complexity.
- A **small set of search-related features** used for effective masquerade attack detection: The limited number of these features reduces the amount of sampling required to collect training data. Reducing the high-dimensional modeling space to a low-dimensional one allows for the improvement of both accuracy and performance over prior approaches. We shall use standard machine learning techniques to evaluate the performance of a system composed of these features. Other work has evaluated alternative algorithms. Our focus in this work is on the features that are modeled. The best masquerade attack detection accuracy was achieved using a modern ML algorithm, Support Vector Machines (SVMs). SVM models are easy to update, providing an efficient deployable host monitoring system. We shall use one-class SVM (ocSVM) models in this work.
- A **Windows data set** [4] collected specifically **to study the masquerade attack detection problem** as opposed to the author identification problem: The data set consists of normal user data collected from a homogeneous user group of 18 individuals as well as simulated masquerader data from 40 different individuals. The data set collected on Windows XP machines is the first publicly available data set for masquerade attack detection since the Schonlau dataset [5].

B. Paper Outline

In section II of this paper, we briefly present the results of prior research work on masquerade detection. Section III expands on the objective and the approach taken in this work. In section IV, we present our home-gathered dataset which we call the RUU dataset. Section V describes the user study that we conducted to show the impact of the masquerader's intent on search behavior, and how the malicious intent of a masquerader whose objective is to steal information has a significant effect on their search behavior. In section VI, we discuss experiments conducted by modeling search behavior using the RUU dataset. In Section VII, we discuss potential limitations of our approach and how such limitations could be overcome. Finally Section VIII concludes the paper by summarizing our results and contributions, and presenting our ongoing work to improve and better evaluate the proposed modeling approach.

II. Related Work

In the general case of computer user profiling, the entire audit source can include information from a variety of sources, such as command line calls issued by users, system calls monitoring for unusual application use/events, database/file accesses, and the organization policy management rules and compliance logs. The type of analysis used is primarily the modeling of statistical features, such as the frequency of events, the duration of events, the co-occurrence of multiple events combined through logical operators, and the sequence or transition of events. However, most of this work failed to **reveal** or clarify **the user's intent** when issuing commands or running processes. The focus is primarily on accurately detecting change or unusual command sequences. In this section, we focus on the approaches reported in the literature that profile users by the commands they issue.

Schonlau et al. [1] applied six masquerade detection methods to a data set of 'truncated' UNIX commands for 70 users collected over a several month period. Each user had 15,000 commands collected over a period of time ranging between a few days and several months [5]. 50 users were randomly chosen to serve as intrusion targets. The other 20 users were used as masqueraders. The first 5000 commands for each of the 50 users were left intact or 'clean', the

next 10,000 commands were randomly injected with 100-command blocks issued by the 20 masquerade users. The commands have been inserted at the beginning of a block, so that if a block is contaminated, all of its 100 commands are inserted from another user's list of executed commands. The complete data set and more information about it can be found at http://www.schonlau.net [5]. The objective was to accurately detect the 'dirty' blocks and classify them as masquerader blocks. It is important to note that this dataset does not constitute ground truth masquerade data, but rather simulates impersonation.

The first detection method applied by Schonlau et al. for this task, called 'uniqueness', relies on the fact that half of the commands in the training data are unique and many more are unpopular amongst the users. Another method investigated was the Bayes one-step Markov approach. It is based on one step transitions from one command to the next. The approach, due to DuMouchel (1999), uses a Bayes factor statistic to test the null hypothesis that the observed one-step command transition probabilities are consistent with the historical transition matrix.

A hybrid multi-step Markov method has also been applied to this dataset. When the test data contain many commands unobserved in the training data, a Markov model is not usable. Here, a simple independence model with probabilities estimated from a contingency table of users versus commands may be more appropriate. The method used automatically toggles between a Markov model and an independence model generated from a multinomial random distribution as needed, depending on whether the test data are 'usual', i.e. the commands have been previously seen, or 'unusual', i.e. Never-Before-Seen Commands (NBSCs). We note with interest that our taxonomy of commands reduces, if not entirely eliminates, the problem of modeling 'Never-Before-Seen-Commands' since any command is likely to be categorized in one of the known classes specified in the taxonomy. Hence, although a specific command may never have been observed, members of its class probably were.

IPAM (Incremental Probabilistic Action Modeling), another method applied on the same dataset, and used by Davidson and Hirsch to build an adaptive command line interface, is also based on one-step command transition probabilities estimated from the training data [6], [7]. A compression method has been also applied to the Schonlau data set based on the premise that test data appended to historical training data compress more readily when the test data stems indeed from the same user rather than from a masquerader, and was implemented through the UNIX tool compress which implements a modified Lempel-Ziv algorithm. A sequence-match approach has been presented by Lane and Brodley [8]. For each new command, a similarity measure between the 10 most recent commands and a user's profile is computed.

A different approach, inspired by the Smith-Waterman local alignment algorithm, and known as semi-global alignment, was presented by Coull et al. [9]. The authors enhanced it and presented a sequence alignment method using a binary scoring and a signature updating scheme to cope with concept drift [10]. Oka et al. [11], [12] noticed that the dynamic behavior of a user appearing in a sequence can be captured by correlating not only connected events, but also events that are not adjacent to each other while appearing within a certain distance (non-connected events). To that extent, they have developed the layered networks approach based on the Eigen Co-occurrence Matrix (ECM).

TABLE I
SUMMARY OF ACCURACY PERFORMANCE OF ANOMALY DETECTORS USING THE SCHONLAU DATA SET

Method	True Pos. (%)	False Pos. (%)
Uniqueness [1]	39.4	1.4
Bayes one-step Markov [1]	69.3	6.7
Hybrid multi-step Markov [1]	49.3	3.2
Compression [1]	34.2	5.0
Sequence Match [8], [1]	26.8	3.7
IPAM [6], [7], [1]	41.1	2.7
Naïve Bayes (w. Updating) [2]	61.5	1.3
Naïve Bayes (No Upd.) [2]	66.2	4.6
Semi-Global Alignment [9]	75.8	7.7
Sequence Alignment (with Updating) [10]	68.6	1.9
Eigen Co-occurrence Matrix [12]	72.3	2.5

Maxion and Townsend [2] applied a naïve Bayes classifier, which has been widely used in text classification tasks, and they provided a thorough and detailed investigation of classification errors [13] highlighting why some masquerade victims are more vulnerable than others, and why some masqueraders are more successful than others. Maxion and Townsend also designed a new experiment, which they called the "1v49" experiment, in order to conduct this error analysis. Another approach called a self-consistent naïve Bayes classifier was proposed by Yung [14] and applied on the same data set. Wang and Stolfo used a naïve Bayes classifier and a Support Vector Machine (SVM) to detect masqueraders [3]. Their experiments confirmed, that for masquerade detection, one-class training is as effective as two class training.

These specific algorithms and the results achieved for the Schonlau dataset appear in Table I (with True Positive rates displayed rather than True Negatives). Performance is shown to range from 1.3% - 7.7% False Positive rates, with a False Negative rate ranging from 24.2% to 73.2% (alternatively, True Positive rates from 26.8% to 75.8%). Clearly, these results are far from ideal. The problem of effective and practical masquerade detection remains quite challenging.

Finally, Maloof and Stephens proposed a general system for detecting malicious insider activities by specifically focusing on violations of "Need-to-Know" policy [15]. Although the work is not aimed directly at masquerade detection, such a system may reveal actions of a masquerader. They defined certain scenarios of bad behavior and combined evidence from 76 sensors to identify whether a user is malicious or not. Our approach is more generalizable and does not specify what bad behavior looks like. Instead, we only model normal behavior and detect deviations from that behavior.

III. Objective and Approach

When dealing with the masquerader attack detection problem, it is important to remember that the attacker has already obtained credentials to access a system. When presenting the stolen credentials, the attacker is then a legitimate user with the same access rights as the victim user. Ideally, monitoring a user's actions after being granted access is required in order to detect such attacks. Furthermore, if we can model the user's intent, we may better determine if the actions of a user are malicious or not. We have postulated that certain classes of user commands reveal user intent. For instance, search should be an interesting behavior to monitor since it indicates the user lacks information they are seeking. Although user search behavior has been studied in the context of web usage mining [16], [17], [18], it has not been used in the context of intrusion detection.

We define a taxonomy of Windows applications, DLLs and user commands to readily identify and model search behavior which appear using a variety of system-level and application-specific search functions. Another behavior that is interesting to monitor is remote access to other systems and the communication or egress of large amounts of data to remote systems, which may be an indication of illegal copying or distribution of sensitive information. Once again, the taxonomy defined allows a system to automatically audit and model a whole class of commands and application functions that represent the movement or copying of data. User behavior naturally varies for each user. We believe there is no one model or one easily specified policy that can capture the inherent vagaries of human behavior. Instead, we aim to automatically learn a distinct user's behavior, much like a credit card customer's distinct buying patterns.

Our objective is to model the normal pattern of submitted commands and called processes and applications assuming that the masquerader will exhibit different behavior from the legitimate user and this deviation will be easily noticed. Hence, this approach essentially tracks a user's behavior and measures any changes in that behavior. Any significant change will raise an alarm. In the following subsections, we present the application and command taxonomy that we have developed for the Windows environment and the machine learning algorithm that we use for learning user behavior and building individual user models.

A. Application Taxonomy

We abstract the set of user commands, Windows applications, and DLLs into a taxonomy of command categories as presented in Figure 1. In particular, we are interested in identifying the specific set of commands that reveal the user's intent to search, to change access control privileges, and to copy or print information. Once these commands are identified, we can extract features representing such behavior while auditing the user's behavior.

The taxonomy has 22 different categories: Browsing, Communications, Database Access, Development and Coding, Editing, File Conversion or Compression, File System Management, Games, Application Installation, Media, Networking, Security, Software Management and Licensing, Web Hosting, Information Gathering, Data Mining and Modeling, Desktop Configuration, Learning Applications, I/O Peripherals, System Management, Organization and Project Management, and Other Applications, such as mathematical and scientific applications. Most categories were further classified into subcategories, however some did not require more granularity, such as the *Desktop Configuration* or *Software Licensing* category. The *Information Gathering* or *Search* category includes commands such as **find** and applications such as **Google Desktop Search**.

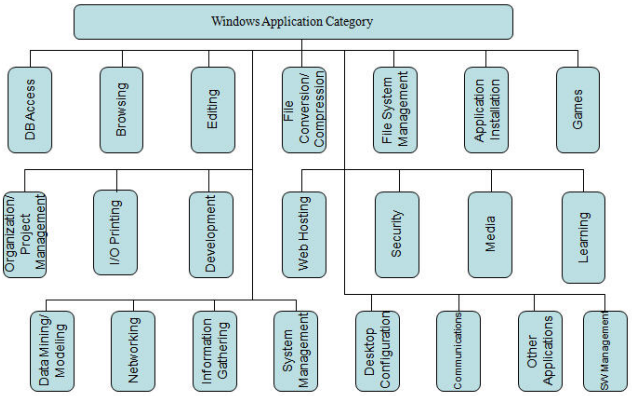


Fig. 1. Taxonomy of Windows applications

B. One-Class Support Vector Machines

We use one-class support vector machines to develop user behavior models. SVMs are linear classifiers used for classification and regression. They are known as maximal margin classifiers rather than probabilistic classifiers. Schölkopf et al. [19] proposed a way to adapt SVMs to the one-class classification task. The one-class SVM algorithm uses examples from one class only for training. Just like in multi-class classification tasks, it maps input data into a high-dimensional feature space using a kernel function, such as the linear, polynomial, or Radial Basis Function (RBF) kernels. The origin is treated as the only example from other classes. The algorithm then finds the hyper-plane that provides the maximum margin separating the training data from the origin in an iterative manner. The kernel function is defined as: $k(x, y) = (\Phi(x).\Phi(y))$,where $x, y \in X$, X is the training data set, and Φ is the feature mapping to a high-dimensional space $X \rightarrow F$. We note that SVMs are suitable for block-by-block incremental learning. As user behavior changes and new data is acquired, updating SVM models is straightforward and efficient. Prior data may be expunged and the support vectors computed from that data are retained and used to compute a new update model using the new data [20], [21].

IV. Data Gathering and "Capture The Flag" Exercise

As we have noted, most prior masquerade attack detection techniques were tested using the Schonlau data set, where 'intrusions' are not really intrusions, but rather random excerpts from other users' shell histories. Such simulation of intrusions does not allow us to test our conjecture that malicious attacker intent will be manifested in the attacker's search behavior. For this reason, we have collected our own dataset, which we will use for testing. However, for completeness, we test our detection approach as a baseline against the Schonalu dataset. The results will be reported in Section VI-C2.

In the following subsections, we describe our home-gathered dataset, consisting of both normal user data and simulated masquerader data, and the host sensor used to collect it.

A. Host Sensor

We have developed a host sensor for Windows platforms. The sensor monitors all registry-based activity, process creation and destruction, window GUI and file accesses, as well as DLL libraries' activity. The data gathered consisted of the process name and ID, the process path, the parent of the process, the type of process action (e.g., type of registry access, process creation, process destruction, window title change, etc.), the process command arguments, action flags (success or failure), and registry activity results. A time stamp was also recorded for each audit record. The Windows sensor uses a low-level system driver, DLL registration mechanisms, and a system table hook to monitor process activity.

B. RUU Dataset

In order to address one of the most significant shortcomings of the Schonlau dataset, namely the lack of 'real' intrusions, we gathered user data and *simulated* masquerader data by conducting a user study under IRB approval. [1] We will refer to this data as the RUU (Are You You?) dataset.

1) Collecting Normal User Data: Eighteen computer science students installed the Windows host sensor on their personal computers. The host sensor collected the data described in the previous subsection and uploaded it to a server, after the students had the chance to filter any data that they were not willing to share. The students signed an agreement for sharing their data with the research community. This unique dataset with a size of more than 10 GBytes is available for download after signing a usage license agreement. The data collected for each student spanned 4 days on average. An average of more than 500,000 records per user were collected over this time. The data collected for each user is approximately the same per day per user. What varies significantly is the amount of "search" behavior exhibited by each user. The impact on the user model performance is described in section VI-C.

[1]Human user studies of insider threat and masquerade detection are approved at Columbia University under Human Subjects protocol IRB-AAAC4240.

2) Collecting Simulated Masquerader Data: To obtain masquerader data, we developed a "capture the flag" exercise in our lab with user volunteers acting as masqueraders. In order to ensure the validity of the experiment and the resulting data, we have decided to control the 'user's intent'. We hypothesize that user search behavior is a behavioral feature that is impacted by the user's intent. If a masquerader is looking to steal information, their intent will be manifested in their search behavior through the volume of the search activities performed by the masquerader. To that extent, we have designed a user study experiment, where the intent to leak information is the independent variable that we control. A total of sixty computer science students were randomly selected to participate in the exercise from a list of volunteers. The participants were randomly assigned to one of three groups. Each group was given a different scenario to execute: a 'malicious attacker' scenario, a 'benign masquerader' scenario, or a 'neutral' scenario.

Malicious Scenario: The participants in this first scenario were asked to perform a specific task, which consisted of finding any information that could be used for financial gain from a coworker's desktop computer. The participants had access to the coworker's computer for 15 minutes while he or she left out for lunch. The scenario clearly described the financial difficulties that the user was going through and the personal problems they were having with the coworker. In particular, they believed that their coworker had been undeservingly promoted instead of them, making any PII information theft or the cause of financial damage to their colleague, that they might engage into, justifiable to them. The task was **goal-specific** in order to **capture the intent** of the users. The participants in this scenario knew that they were being monitored. While they were encouraged to find relevant information for their task, they were also urged to hide their masquerade activity. It is also important to mention that they were not specifically asked to search the file system in the scenario, i.e. some users could decide to look into the computer owner's e-mail, which constitutes another approach for information gathering, instead of searching their files. After completing the task, the participants filled a questionnaire about their experience and behavior during the task, and how different it was from their own normal search behavior. Most participants indicated that they tried to leave the desktop in the same state that they accessed it, including the same list of recently opened documents. Most also avoided using desktop search tools and opted for manually exploring the contents of the file systems by traversing directories of interest. Some students also cleared the browser's cache and history.

Benign Scenario: Similar to the previous scenario, the participants in this scenario acted as employees in a certain organization. While their coworker left out for lunch and stayed logged in to their computer, their own computer experienced a hard drive failure. Because they had to finish working on slides, which they had a copy of on a USB drive, for a meeting with a customer occurring later in the afternoon, they did not hesitate to use the coworker's computer to finish working on

their slides.

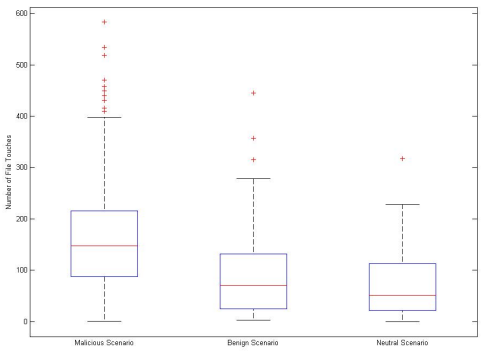
Neutral Scenario: The participants in this scenario too has a co-worker leave for lunch while staying logged in to their desktop computer. However, the participants in this scenario had no compelling reason to access the co-worker's computer. They were left to freely choose whether they wanted to access their coworker's desktop. We observed the behaviors of the participants, and whether they decided to access the coworker's desktop. In particular, we observed what they did if they decided to access it. The participants had also to describe what they did and explain their decision after completing the experiment.

The participants in the user study in all three groups had unlimited access to the same file system for 15 minutes each. The exercise was conducted on different week days and at varying times during the day. None of the users had access to this file system before, which was designed to look very realistic and to include potentially interesting patent applications, personally identifiable information, as well as account credentials. The file system had more than 100,000 files indexed by desktop search tools installed on the system, and totaling a size of more than 70 GBytes. The files were copied from a modern Windows XP machine routinely used in a university research environment. They were created over a period of 5 years. The file system included at least 80 files containing personal information that could be used for identity theft or financial gain. The contents of certain personal files were sanitized, so that no personal information was leaked. We also installed applications that typical computer science students would have on their own machines, such as programming APIs, media players, etc together with code projects, games, music, and video files. The goal was to make this machine look similar to the ones that the normal users in our experiment were using. Special care was taken to make sure that the desktop appeared in the same state to all participants in the experiment. While simulating masquerader attacks in the lab is not ideal, it was the best available option. None of the students who shared their normal usage data were willing to lend their computers in order to conduct masquerade attack experiments on them.

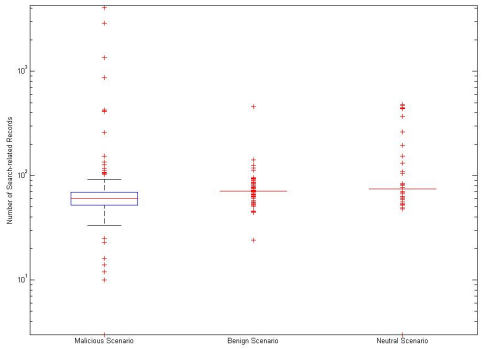
V. User Study Experiment

The objective of this experiment is to provide evidence for our conjecture that the masquerader's intent has a significant effect on their search behavior. We extracted 3 search-related features from the data collected in the user study: The number of files accessed during an epoch of two minutes, the number of search-related actions initiated by the masquerader, and the percentage of file system navigation user actions during the same epoch. We display the distribution of these features for the malicious, benign and neutral scenario participants of the user study in Figure 2. Our experiments show that the manipulation of the user intent has demonstrated a significant effect on user search behavior.

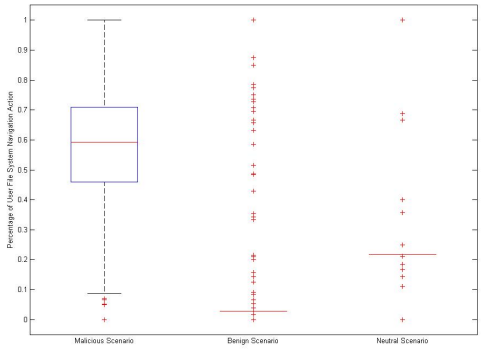
Figures 2(a), 2(b), and 2(c) are box-and-whisker plots of the three user search-related features. The horizontal red line in the middle of each of the boxes in these plots corresponds to the median value of the displayed feature. Fifty per cent of the data falls within this box, while the top and bottom quartiles (25% of the data) of the data are represented by the whisker lines above and below this box. Data points whose value is above 1.5 times the upper quartile or lower than 1.5 times the lower



(a) Distribution of File Touches across the three User Study Groups



(b) Distribution of Search-related Actions across the three User Study Groups



(c) Distribution of the Percentage of File System Navigation User Actions across the three User Study Groups

Fig. 2. Distribution of Search-related Features across the three User Study Groups

quartiles are considered outliers and are represented as small crosses. The short horizontal lines above and below the box represent the maximum and minimum data values exclusing outliers.

Figure 2(a) depicts the distribution of the number of accesses to all files residing on the file system per a 2-minute epoch. It is clear that the average number of file touches amongst the malicious scenario group is much higher than in the other two groups.

Figure 2(b) shows the distribution of search-related queries to the registries and accesses to search-related DLLs and applications for the different groups of the user study. The average number of search-related records in the first group is comparable to that in the other groups, which might be unexpected. However, this could be explained by the fact that the participants in the first group were penalized if their illegal masquerade activities were discovered, which is why most of them decided to refrain from using desktop search tools to find relevant information. This choice is reflected in Figure 2(b) and in Figure 2(c).

In the latter figure, one can clearly notice that the percentage of user actions related to navigating through the file system in order to explore its contents is much more significant in the first group. The participants in the second group did not have the intention to steal any information, which is why they did not engage in file system exploration activities. While most of the participants in the third scenario decided not to access the coworker's computer, some of them did urged by their curiosity, as indicated in the participants' answers to our post-experiment questionnaire. Figure 2(c) shows that for this group, the users explored the directories in the file system in order to satisfy their curiosity.

Finally, Figure 3 shows how the number of personal files accessed by masqueraders varies by user study scenario. The results of this user study provide evidence that search behavior is significantly affected by a masquerader's intent. The question that we attempt to answer next is: Can we model normal user search behavior and use it to detect malicious masqueraders?

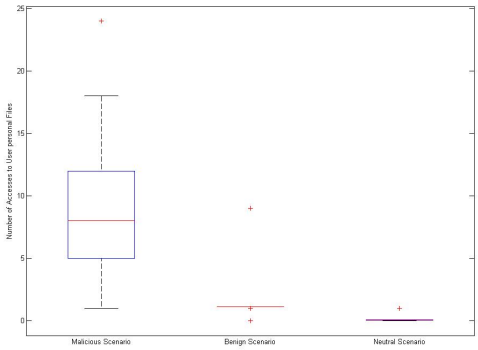


Fig. 3. The personal files accessed by masqueraders

VI. RUU Experiment

In order to evaluate our conjecture that search behavior modeling can provide a means for detecting malicious masqueraders, we use the normal user data to build user search behavior models. We then use the simulated masquerader data gathered for the participants in the 'malicious' scenario of our user study to test these user models. Here we describe our modeling approach, the experimental methodology, and the results achieved in this experiment.

A. Modeling

We have used the taxonomy displayed in Figure 1 in order to capture search and information gathering applications. The data was grouped into 2 minute quanta of user activity, and we counted all events corresponding to each type of activity within each of the 2 minute epochs. Eventually a total of three features were selected for each of those epochs. Each of the features is related to some aspect of the user's search behavior:

1) Number of search-related records: Specific sections of the Windows registry, specific DLL's, specific index files, and specific programs on the system, particularly desktop search tools, are correlated with system searching. For the 2 minute epoch, we model all search-related activity.
2) Number of file touches: Any file fetch, read, write, or copy action results into loading the file into memory. We count the number of times files are touched and loaded into memory by any process within each 2-minute epoch.
3) Percentage of file system navigation user actions: Not all search is performed using a desktop search tool. Navigating through the file system to explore its contents is also a form of user search. We model all manual search and file system navigation related user activity occurring during the 2-minute epoch.

We chose simple search features that characterize search volume and velocity to test our hypothesis. While none of the features could be used to achieve high detection rates alone, the combination of the three features could be very effective. More complex search features that describe user search patterns could be extracted. Such features include, but are not limited to search terms and specific directory traversals. Evaluation of these features is the subject of our future work.

B. Experimental Methodology

For each of the 18 normal users, the first 80% of their data were used for training a one-class SVM model. The user's test data and the masquerader data were kept separate. We have noticed that some of the processes running on users' computers belonged to malware and spyware programs. We have classified those processes (a total of 21 different processes) as *Malware* or *Spyware* respectively and have not included the records that featured such processes for modeling and testing. The computer that we have used to run the 'capture the flag' exercise did not have any such programs running.

Therefore, we did not want such records to skew the accuracy results.

After the baseline models were computed, the same features used in the model were extracted for the test data after dividing them into 2-minute quanta of user activity. The models were tested against these features, and an empirically identified threshold was used to determine whether the user activity during the 2 minute-period was normal or abnormal. If the user activity was performed by the normal user, but was classified as abnormal by the ocSVM model, a false positive was recorded.

C. Detection Accuracy Evaluation

For evaluation purposes we have conducted two experiments. In the first one, we have used one-class SVM models using the three features listed in Section VI-A. In the second experiment, we have used the frequency of applications and processes within the 2 minute epoch as features for the ocSVM models. This is the modeling approach that achieved results comparable to those achieved by the naïve Bayes approach when applied to the Schonlau dataset [3].

1) Accuracy Results: Using the search-behavior modeling approach, 100% of the 2-minute quanta that included masquerader activity were detected as abnormal, while 1.1% of the ones with legitimate user activity were flagged as not confirming to the user's normal behavior. The results achieved are displayed in Table II. The FP rate is significantly reduced compared to the application frequency-based modeling approach, while a perfect detection rate is achieved. These results substantially outperform the results reported in the literature so far.

TABLE II
EXPERIMENTAL RESULTS OF OCSVM MODELING APPROACHES USING SEARCH-BEHAVIOR RELATED FEATURES AND APPLICATION FREQUENCY FEATURES

Method	True Pos. (%)	False Pos. (%)
Search-behavior ocSVM	100	1.1
App.-freq. ocSVM	90.2	42.1

Monitoring file access and fetching patterns proved to be the most effective feature in these models. Note that if the attacker follows a different strategy to steal information, and decides to copy whole directories in the file system to a USB drive for later investigation, instead of identifying files of interest during one user session, then the 'file touches' feature will reflect that behavior.

Since each user has its own model with its own detection threshold, we cannot build a single ROC curve for each modeling approach. However we can compare the ROC curves for individual user models using the two modeling approaches investigated. One way to compare the ROC curves is to compare the Area Under Curve (AUC) scores. The higher the AUC score, the better the accuracy of the model.

Figure 4 displays the ROC scores for all user models. The search-behavior modeling approach outperforms the application frequency based modeling approach for each user model. The average AUC score achieved for all ROC curves when modeling search behavior is 0.98, whereas the average AUC score for the application frequency-based models is 0.63.

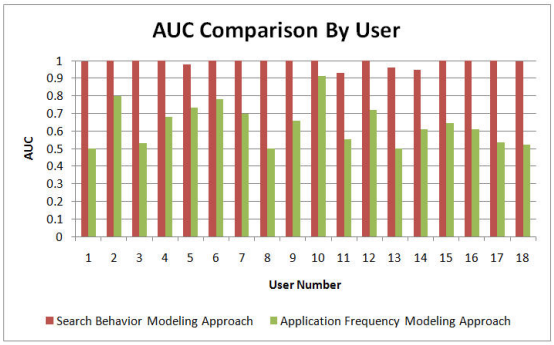


Fig. 4. AUC Scores By User for the Search Behavior and Application Frequency-Based Modeling Approaches using one-Class Support Vector Machines

Figure 5 depicts the number of ROC curves having AUC scores higher than a certain value for both modeling approaches approach. Note that for 38 user search behavior models, the AUC score is equal to 1 indicating the absence of any false positives.

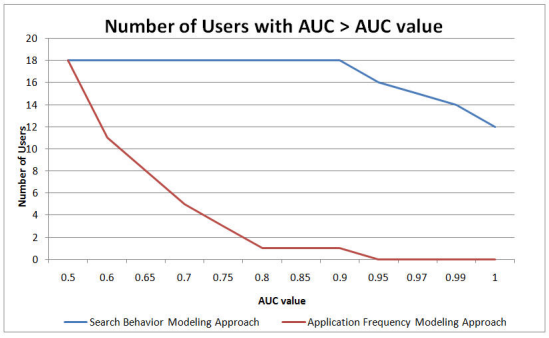


Fig. 5. The number of user models with AUC values greater than the value displayed on the x-axis for the search behavior and the application frequency modeling approaches using one-class SVMs. (The upper-left point shows 18 user models with AUC scores greater than 0.5)

Recall that the RUU data set consists of user data with varying amounts of data for different users. The amount of search behavior information varied from user to user. False positives were higher for users who contributed less data in general and less search-related data in particular than for those for whom we collected a large amounts of such data, such as users 11 and 14. For a 100% detection rate, the FP rate scored by these user models ranged between 11% and 15%, which proves the need for more training data for such users in order to improve the performance of the user models.

In summary, the significant accuracy improvement achieved can be explained by the fact that masqueraders were focused on a clear objective, namely finding information that could

be used for financial gain, and by the tight link between the masquerader's intent and their search behavior, as demonstrated through our user study described in section V. Despite the simplicity of the search features used, which only characterize search volume and velocity, we were able to reliably detect malicious masqueraders. We note that most masqueraders indicated in the post-experiment questionnaires that their strategy for finding relevant information started by quickly scanning the most recently opened documents, or the list of bookmarks. However, they still engaged in a wider search activity eventually when these sources proved fruitless.

2) Accuracy Results Discussion: It is difficult to compare our results with the results achieved using the techniques described in section II. These techniques were evaluated using the Schonlau dataset. Recall that the Schonlau dataset is not a 'true' masquerader dataset, since its 'intrusions' or 'masquerade' command blocks are just sequences of commands generated by random normal users. So search activities may not be significant in this dataset. Furthermore, this dataset does not include any timestamps. However, for completeness we evaluate the impact of using the taxonomy of user commands and applications (as opposed simple commands when building user models) on detection accuracy. To that extent, we developed a taxonomy of Linux commands similar to the one we created for Windows applications and DLLs. We conducted an experiment where we followed the methodology described in [1], [3]. In this experiment, we measured the performance of one-class SVM models using frequencies of simple commands per command block as features, and we compared the performance of ocSVM models sing frequencies of command categories as features. Table III shows the results achieved by the one-class SVM classifiers. The results confirm that the information that is lost by compressing the different user shell commands into a few categories does not affect the masquerader detection ability significantly.

TABLE III
OCSVM SCHONLAU EXPERIMENTAL RESULTS

Method	True Pos. (%)	False Pos. (%)
ocSVM w/ simple cmds	98.7	66.47
ocSVM w/ taxonomy	94.8	60.68

In an operational monitoring system, one would be concerned with the error rate of a detector. The downside of a false positive is essentially annoyance by a legitimate user who may be alerted too frequently. An interesting problem to study is how to calibrate the modeling and detection frequency to balance the detector's false positive rate while ensuring its false negative rate is minimized. False negatives in this context, i.e., an undetected masquerader, are far more dangerous than an annoying false positive. A thorough evaluation of the right model checking and alerting frequency in light of average search times on a file system inter alia is the subject of ongoing research. Another focus of ongoing research is the correlation of search behavior anomaly detection with trap-based decoy files such as [22]. This should provide stronger evidence of malfeasance, and therefore improve the detector's accuracy. Not only would a masquerader not know the file system, they would also not know the detailed contents of that file system especially if there are well placed traps that they cannot avoid. We conjecture that detecting abnormal search operations performed prior to an unsuspecting user opening a decoy file will corroborate our suspicion that the user is indeed impersonating another victim user. Furthermore, an accidental opening of a decoy file by a legitimate user might be recognized as an accident if the search behavior is not deemed abnormal. In other words, detecting abnormal search and decoy traps together may make a very effective masquerade detection system. Ongoing work should establish evidence to corroborate this conjecture.

D. Performance Evaluation

1) Computational Complexity: Our experiment can be divided into four main steps: identifying the features to be used for modeling, extracting the features to build the training and testing files, building a ocSVM model for each normal user, and finally testing each user model against the test data. We discuss the computational complexity of each of these steps for one user model.

Let o be the total number of raw observations in the input data. We use this data to compute and output the training vectors $x_i \in R^n, i = 1, ..., l$ and testing vectors $x_j \in R^n, j = 1, ..., m$ for each user u, where n is the number of features used for modeling.

When using the application frequency features, this step requires reading all training data (about 0.8 of all observations o) in order to get the list of unique applications in the dataset. This step can be merged with the feature extraction step, but it would require more resources, as the feature vectors would have to remain in memory for updates and additions of more features. We chose to run this step in advance in order to simplify our program. This step is not required for the search behavior profiling approach, as all features are known in advance.

In the feature extraction step, we go through all input data once, grouping the observations that fall within the same epoch, and calculate and output n features for that epoch. This operation has a time complexity of $O(o + n \times (l + m))$.

Chang and Lin [23] show that the computational complexity of the training step for one user model is $O(n \times l) \times$#Iterations if most columns of Q are cached during the iterations required ; Q is an $l \times l$ semi-definite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$; $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel; each kernel evaluation is $O(n)$; and the iterations referred to here are the iterations needed by the ocSVM algorithm to determine the optimal supporting vectors.

The computational complexity of the testing step is $O(n \times m)$ as the kernel evaluation for each testing vector y_j is $O(n)$. We experimentally validate the complexity analysis in the next section to determine whether we have improved performance both in terms of accuracy and speed of detection.

2) Performance Results: We ran our experiments on a regular desktop with a 2.66GHz Intel Xeon Dual Core processor and 24GB of memory in a Windows 7 environment. We measure the average running time of each step of the experiment over ten runs. The results are recorded in table IV. As we pointed out in the previous subsection, the very first step is not executed in the our proposed search behavior modeling approach, but it takes more than 8 minutes when using the application frequency modeling approach. The running time of the feature extraction step shows that the number of raw observations in the raw data dominates the time complexity for this step. We point out that the RUU data set contains more than 20 million records of data.

The training and testing vectors are sparse, since only a limited number of the 1169 different applications could conceivably run simultaneously within a 2-minute epoch. This explains why the 389.7 ratio of features does not apply to the running time of the training and testing steps, even though these running times depend on the number of features n. While one might argue that, in an operational system, testing time is more important than training time, we remind the reader that a model update has the same computational complexity as model training. For the latter, the use of a very small number of features as in our proposed approach clearly provides significant advantages.

All of these differences in running times culminate in a total performance gain of 74% when using the search behavior model versus the application frequency model typical of prior work. This computational performance gain coupled with improved accuracy could prove to be a critical advantage when deploying the sensors in an operation environment if a system design includes automated responses to limit damage caused by an insider attack.

TABLE IV
PERFORMANCE COMPARISON OF OCSVM MODELING APPROACHES USING SEARCH BEHAVIOR-RELATED FEATURES AND APPLICATION FREQUENCY FEATURES

Step	**ocSVM app. freq.**	**ocSVM search-beh.**
Identifying Features (min)	8.5	0
Extracting Features (min)	48.2	17.2
Training (min)	9.5	0.5
Testing (min)	3.1	0.5
Total (min) (Rounded to Nearest Min.)	**69**	**18**

VII. LIMITATIONS AND DISCUSSION

An attacker could try to evade the monitoring system by renaming DLLs and applications so that they are assigned to a different category. Although we have not implemented a monitoring strategy to counter this evasive tactic, it is clear that a simple extension to the monitoring infrastructure can account for this case.

We assume that the attacker does not have knowledge about the victim's behavior. However, if the attacker does have such prior knowledge, we propose combining user behavior profiling with monitoring access to well-placed decoy files in the file system (as explained in Section VI-C) in order to limit the success of evasion.

A masquerader could choose to send data over the network to a different host, or copy it to a USB drive for later examination. Our taxonomy could be easily used to monitor this behavior in case the attacker resorts to such strategies. As noted in section VI-C1, the 'file touches' feature already captures some aspect of this behavior. The applications taxonomy could be used to extract 'Networking'-, 'Communications'- and I/O-related features to be included in the user model, so that such masquerader behavior could be detected more reliably.

VIII. CONCLUDING REMARKS

Masquerade attacks (such as identity theft and fraud) are a serious computer security problem. We conjecture that individual users have unique computer search behavior which can be profiled and used to detect masquerade attacks. The behavior captures the types of activities that a user performs on a computer and when they perform them.

The use of search behavior profiling for masquerade attack detection permits limiting the range and scope of the profiles we compute about a user, thus limiting potentially large sources of error in predicting user behavior that would be likely in a far more general setting. Prior work modeling user commands shows very high false positive rates with moderate true positive rates. User search behavior modeling produces better accuracy.

We presented a modeling approach that aims to capture the intent of a user more accurately based on the insight that a masquerader is likely to perform untargeted and widespread search. Recall that we conjecture that user search behavior is a strong indicator of a user's true identity. We modeled search behavior of the legitimate user using three simple features, and detected anomalies that deviate from that normal search behavior. With the use of the RUU dataset, a more suitable dataset for the masquerade detection problem, we achieved the best results reported in literature to date: 100% masquerade detection rate with only 1.1% of false positives. Other researchers are encouraged to use the data set that we have made available for download after signing a data usage agreement [4].

In an operational monitoring system, the use of a small set of features limits the system resources needed by the detector, and allows for real-time masquerade attack detection. We note that the average model size is about 8 KB when the search-behavior modeling approach is used. That model size grows to more than 3 MB if an application and command frequency modeling approach is used. Furthermore, it can be easily deployed as profiling in a low-dimensional space reduces the amount of sampling required: An average of 7 days of training data was enough to train the models and build effective detectors.

In our ongoing work, we are exploring other features for modeling that could improve our results and extend them to

other masquerade attack scenarios. The models can be refined by adding more features related to search, including search query contents, parameters used, and directory traversals etc.. Other potential features to model include the use of bookmarks and most recently opened documents which could also be used by masquerade attackers as a starting point for their search. The models reported here are primarily volumetric statistics characterizing search volume and velocity. We can also update the models in order to compensate for any user behavior changes. We will explore ways of improving the models so that they reflect a user's unique behavior that should be distinguishable from other legitimate users' behaviors, and not just from the behavior of masqueraders. The ultimate objective is to build more secure and dependable systems that authenticate legitimate users by their behavior, rather than exclusively by their possibly stolen credentials.

ACKNOWLEDGMENT

This material is based on work supported by the US Department of Commerce, National Institute of Standards and Technology under Grant Award Number 60NANB1D0127, the US Department of Homeland Security under grant award number 2006-CS-001-000001-02 and the Army Research Office under grant ARO DA W911NF-06-10151. 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 US Department of Commerce, National Institute of Standards and Technology, the U.S. Department of Homeland, or the Army Research Office.

REFERENCES

[1] M. Schonlau, W. Dumouchel, W. Ju, A. F. Karr, M. Theus, and Y. Vardi, "Computer intrusion: Detecting masquerades," *Statistical Science*, vol. 16, pp. 58–74, 2001.

[2] R. A. Maxion and T. N. Townsend, "Masquerade detection using truncated command lines," in *DSN '02: Proceedings of the 2002 International Conference on Dependable Systems and Networks*. IEEE Computer Society, 2002, pp. 219–228.

[3] K. Wang and S. J. Stolfo, "One-class training for masquerade detection," in *Proceedings of the 3rd IEEE Workshop on Data Mining for Computer Security*, 2003.

[4] M. Ben-Salem, "RUU dataset: http://www1.cs.columbia.edu/ids/ruu/data/." [Online]. Available: http://www1.cs.columbia.edu/ids/RUU/data/

[5] M. Schonlau, "Schonlau dataset: http://www.schonlau.net." [Online]. Available: http://www.schonlau.net

[6] B. D. Davison and H. Hirsh, "Predicting sequences of user actions," in *Working Notes of the Joint Workshop on Predicting the Future: AI Approaches to Time Series Analysis, 15th National Conference on Artificial Intelligence/15th International Conference on Machine Learning*. AAAI Press, 1998, pp. 5–12.

[7] ——, "Toward an adaptive command line interface," in *Proceedings of the Seventh International Conference on Human-Computer Interaction (HCI97)*. Elsevier Science Publishers, 1997.

[8] T. Lane and C. E. Brodley, "Sequence matching and learning in anomaly detection for computer security," in *In AAAI Workshop: AI Approaches to Fraud Detection and Risk Management*. AAAI Press, 1997, pp. 43–49.

[9] S. E. Coull, J. Branch, B. Szymanski, and E. Breimer, "Intrusion detection: A bioinformatics approach," in *Proceedings of the 19th Annual Computer Security Applications Conference*, 2001, pp. 24–33.

[10] S. E. Coull and B. K. Szymanski, "Sequence alignment for masquerade detection," *Computational Statistics and Data Analysis*, vol. 52, no. 8, pp. 4116–4131, 2008.

[11] M. Oka, Y. Oyama, and K. Kato, "Eigen co-occurrence matrix method for masquerade detection," in *Publications of the Japan Society for Software Science and Technology*, 2004.

[12] M. Oka, Y. Oyama, H. Abe, and K. Kato, "Anomaly detection using layered networks based on eigen co-occurrence matrix," in *Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection*, 2004.

[13] R. A. Maxion and T. N. Townsend, "Masquerade detection augmented with error analysis," *IEEE Transactions on Reliability*, vol. 53, no. 1, pp. 124–147, 2004.

[14] K. H. Yung, "Using self-consistent naïve bayes to detect masqueraders," in *PAKDD'08: Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004, pp. 329–340.

[15] M. A. Maloof and G. D. Stephens, "elicit: A system for detecting insiders who violate need-to-know," in *RAID*, 2007, pp. 146–166.

[16] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret, "Modeling user search behavior," in *LA-WEB '05: Proceedings of the Third Latin American Web Congress*. IEEE Computer Society, 2005, pp. 242–251.

[17] J. Attenberg, S. Pandey, and T. Suel, "Modeling and predicting user behavior in sponsored search," in *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009, pp. 1067–1076.

[18] M. O'Brien and M. T. Keane, "Modeling user behavior using a search-engine," in *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2007, pp. 357–360.

[19] B. Schölkopf, J. C. Platt, J. Shawe-taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, 2001.

[20] V. N. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, 1999.

[21] N. A. Syed, H. Liu, S. Huan, L. Kah, and K. Sung, "Handling concept drifts in incremental learning with support vector machines," in *In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99*. ACM Press, 1999, pp. 317–321.

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

[23] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf," 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf