Augmenting Authentication with Context-Specific Behavioral Biometrics

Haoruo Zhang
Information Security Institute
Johns Hopkins University
harol@jhu.edu

Digvijay Singh
Information Security Institute
Johns Hopkins University
dsingh34@jhu.edu

Xiangyang Li
Information Security Institute
Johns Hopkins University
xyli@jhu.edu

Abstract

Behavioral biometrics, being non-intrusive and cost-efficient, have the potential to assist user identification and authentication. However, user behaviors can vary significantly for different hardware, software, and applications. Research of behavioral biometrics is needed in the context of a specific application. Moreover, it is hard to collect user data in real world settings to assess how well behavioral biometrics can discriminate users. This work aims to improving authentication by behavioral biometrics obtained for user groups. User data of a webmail application are collected in a large-scale user experiment conducted on Amazon Mechanical Turk. Used in a continuous authentication scheme based on user groups, off-line identity attribution and online authentication analytic schemes are proposed to study the applicability of application-specific behavioral biometrics. Our results suggest that the useful user group identity can be effectively inferred from users' operational interaction with the email application.

1. Introduction

Companies are exposed to great risk of compromised user accounts and insider attacks. In a recent global study featuring 208 companies, 69% of enterprise security professionals admitted having experienced theft or corruption of company information at the hands of trusted insiders [1]. Sharing credentials among employees makes it more difficult for an IT system to identify such attempts through authentication [2]. Conventional authentication scheme often does not verify users' authenticity continuously during active sessions, leaving unattended computer systems vulnerable to unauthorized use [3]. It has been widely realized by the security community that traditional identity and access management controls, which are static and rigid, can no longer effectively protect valuable information assets once user account credentials are compromised [4]. There is a higher risk of abuse for mobile devices when they fall in wrong hands.

Role-based access control (RBAC) regulates access to enterprise system resources based on the roles of individual users within an enterprise [5]. RBAC enables users to carry out a wide range of authorized tasks by dynamically managing their actions according to functions, relationships, and constraints that can be flexibly defined for user account groups [6]. This contrasts with other methods of access control, which grant or revoke user access on an object-by-object basis. However, misconfiguration of RBAC systems, such as orphaned accounts and shared accounts, exposes enterprise assets to an increased level of risk of insider attack who can steal data beyond a user's access privilege [7].

Biometrics-assisted authentication schemes help with identity recognition by matching physiological or behavioral traits to users who are be authenticated. However, several issues exist with current biometrics-based authentication systems: (1) required special hardware or software can be expensive and intrusive to existing authentication processes, which causes privacy and reliability issues; (2) a user's identity stays unverified beyond initial authentication, posing a risk as described previously; (3) behavioral biometrics analysis is isolated from specific application contexts, information of which can likely contribute to the effectiveness of biometrics-assisted authentication solutions. It is of necessity to study behavioral biometrics in a specific application context, i.e., taking in consideration knowledge of its operations.

In this work, we propose a *Context-Specific Behavioral Biometrics Augmented Authentication Scheme* by assuming that inherent differences exist for a special type of task among dissimilar groups (roles) of users that can augment authentication and access control. Application-specific behavioral biometrics recognize patterns in high-level human computer interaction under a specific application context [8], e.g., UI designs and system architecture. A behavioral user

URI: https://hdl.handle.net/10125/60165
ISBN: 978-0-9981331-2-6

profile can be developed based on a user's behavior that is associated with the typical usage of one type of application.

We explored the feasibility and applicability of using application-specific behavioral biometrics by the example of a common webmail application, instead of application-independent behavioral biometrics, e.g., voice, eye movement, keystrokes and scrolling or clicking activities. We customized logging functionality on top of existing web event-listening mechanisms of the webmail application and saved behavioral logs of specific user operations. We extracted interpretable context-specific features from such behavioral log data, which were used for group-based authentication. The labelling of user groups was derived from clustering these behavioral biometric features. We investigated a set of classifiers for offline identify attribution to infer the user group of each user from behavioral log data and applied a sliding window approach to support continuous online authentication. Evaluation results suggest that the introduced behavioral biometrics can be applied to effectively classifying users' group identities in authentication during an active web session.

We aimed to make three-fold contributions in this research effort. First, we extended the concept of behavioral biometrics to incorporate the consideration of a specific application or system and designed a scheme of augmenting authentication utilizing such behavioral biometrics. Second, we proposed and examined offline identity attribution by mining user log data and online authentication to recognize user behavior patterns continuously. Third, we evaluated our methods through k-fold cross-validation for accuracy to test their feasibility.

2. Background

The challenge of identity recognition and verification for authentication in information systems is essentially the tradeoff between security of such a scheme and its overall cost and usability. While physical biometrics are mostly stored as credentials on the server or locally at the risk of being stolen, user behavioral profiles used for authentication usually possess no value to attackers [9]. Moreover, physical biometrics leaves more room for spoofing attempts. Therefore, behavioral biometric systems recently have been explored extensively in addition to physical biometrics.

2.1. Continuous authentication

Furnell et al. [10] found that users have a reasonable expectation of continuous or periodical authentication throughout their daily use of an IT system that tried to maintain confidence in identity management. Both physical biometrics and behavior indicators were investigated and received positive feedback from users. Location-based access control [11] and behavioral biometrics, notably keystroke dynamics and mouse movement [12-14] can provide common forms of implicit authentication. More recently, accelerometers and other sensors in mobile devices have been used to profile and identify users. Chang et al. [15] used accelerometers in television remote controls to identify individuals. Kale et al. [16] and Gafurov et al. [17] used gait recognition to detect whether a device is being used by its actual owner. These biometrics and location-based approaches are relevant to our work to demonstrate the potential applicability of behavioral biometrics to authentication.

The convergence of multiple biometric indicators of identity is another current development that combines multiple biometric factors to support an authentication decision [18, 19]. For example, Greenstadt and Beale [20] formulated a concept of "cognitive security" for personal devices. Specifically, they proposed a multi-modal approach "in which many different low-fidelity streams of biometric information are combined to produce an ongoing positive recognition of a user." These works have enriched the knowledge body of behavioral biometrics. Though they did not specifically touch on how behavioral biometrics can be further formulated under certain software or hardware context.

2.2. Keystroke dynamics

Users interact with a computer through I/O devices in specific ways. Patterns associated to individual users can be recognizable in scenarios where there is a repetition of interactions, such as typing one's credentials on a regular basis. Keystroke dynamics refers to a mechanism of recording one's behavioral biometrics in during typing, which provides an accessible manner for individual user authentication and identification [21]. Investigated features for keystroke dynamics vary from simple metrics of key press interval and dwelling times, e.g., the up-up time, up-down time, and down-down time, to multi-key features, e.g., bi-graph and trigraph [22].

Classification methods have been researched extensively to use these features in making authentication decisions, including initial and continuous authentication. Less researched keystroke features include overall typing speed, frequency of errors (i.e., use of backspace), use of the numpad, the order in which a user presses the shift key for capital letters, and possibly the force with which keys are hit using special keyboards [23]. Keystroke dynamics aims to model user typing patterns independent of the

application context to suit computer systems that involve the use of keyboard. Studies mostly ignore how the contextual information of different applications influences such behavioral biometrics.

2.3. Mouse movement

Mouse dynamics, as a behavioral biometric for analyzing behavior data from pointing devices, e.g., mouse or touchpad, can aid authentication in an accessible and convenient manner. Hashia et al. [24] and Bours et al. [25] presented preliminary results on using mouse dynamics for user authentication. They both asked participants to perform fixed sequences of mouse operations and analyzed behavioral characteristics of mouse movement to identify a user during the login stage. Distance-based classifiers were established to compare the validation data with the enrollment data. Hashia et al. collected data from 15 participants using one computer, while Bours et al. collected data from 28 participants using different computers; they achieved equal-error rates (EERs) of 15% and 28% respectively. Aksari et al. [26] presented an authentication framework for verifying users based on a fixed sequence of mouse movement. Features were extracted from nine movements among seven squares displayed consecutively on a computer screen. They built a classifier based on a scaled Euclidean distance using data from both legitimate users and impostors. The researchers reported to achieve an EER of 5.9% over 10 users collected from on the same computer. GUI design and different computer platforms can have a significant impact on user mouse movement. One-size-fit-all behavioral biometrics should be examined in different meaningful application contexts.

3. Research design

Our work is based on an empirical user study that collected user operations of fine granularity using a modified web application [27, 28]. This user study instructed participants recruited online to sort 40 emails using a webmail system. Each email was classified as either legitimate or phishing. The main goal was to enable these participants to interact unbiasedly with a web-based application and to capture realistic user behaviors in doing so. Email sorting embodies one of the most typical computer applications. It is viable if not ideal to establishing the validity of the proposed authentication scheme. Such context-specific behavioral biometrics are based on one applicable platform among many other possibilities. The nature of user operations is expected to vary for different applications, which makes research findings specific to intended applications.

3.1. Behavior data collection

Existing literatures often give out just a few details about the instrumentation used to collect user behavioral data and most studies hosted human subject experiments in a highly controlled lab environment [20, 22, 24-26]. There have been successful examples of employing participants remotely, e.g., using the Amazon Mechanical Turk human subject pool, to perform research-focused tasks [29-31]. Specifically, Bartneck et al. [29] showed that recruiting such participants is efficient and affordable for certain types of user tasks. In our case, we hosted the user study experiment on Amazon Mechanical Turk and recruited participants of differing backgrounds from over twenty US states. We instructed them to classify 40 emails into two categories, i.e., legitimate and phishing, within a given amount of time on Roundcube, a webmail system (http://www.roundcube.net].

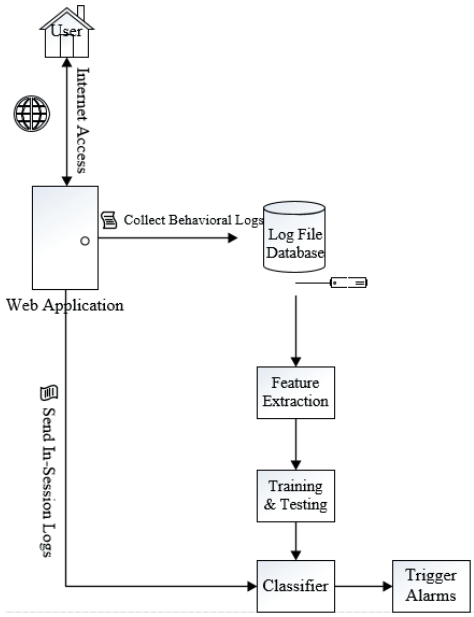

Figure 1. Data collection and continuous authentication framework overview

The Roundcube webmail application is implemented via HTML, JavaScript and PHP. The primary way to interact with the application is through mouse hovering over and clicking on buttons, links, email address, etc. on a web browser. We embedded special JavaScript code in HTML files dynamically generated by Roundcube to capture users' interaction with the system, including mouse hovering and clicking as two primary event types. We utilized AJAX to send captured

user behavior information remotely to a logging service running on the Roundcube server. Each participant had a separate log file generated on the server, recording all the operations throughout his/her active session using the webmail application.

As shown in Figure 1, extracted behavioral features from the log data can be used to develop a classifier of the user group identify for authentication purposes. A user is initially authenticated using stored credentials, e.g., a password, and associated with a user group, e.g., system administrators or regular users. During online authentication, sub-samples of in-session log data are sent continuously to the classifier in real time that updates the prediction of the group identity for a current user. Alarms will be triggered or challenges issued to this user if a mismatched group identity is suspected.

3.2. Overview of the data

We conducted the user experiment with 205 human subjects recruited from the Amazon Mechanical Turk, ending up with 177 users completing all the experiment phases and only 146 participants completely sorting all the 40 emails. There was data loss due to issues of the server, network transmission, and the client end. Furthermore, these participants were assigned to four conditions determined by the presence of a secondary question-answering tasks and/or a monetary incentive. After a careful examination of log files, we selected the data for 35 participants that were in coherent condition.

Event/Operation Name	Captured Data
1.1.1 Clicking on Buttons on MenuBar	UserID;Time Stamp;Event Type;Button Name
1.2.1 Clicking on Inbox Folder	UserID;Time Stamp;Event Type;Folder Name(Inbox)
1.2.2 Clicking on Keep Folder	UserID;Time Stamp;Event Type;Folder Name(Keep)
1.2.3 Clicking on Suspicious Folder	UserID;Time Stamp;Event Type;Folder Name(Suspicious)
1.3.1 Clicking on Each Email Item	UserID;Time Stamp;Event Type;Email Title;Sender's address
1.4.1 Hovering in Sender's Address	UserID;Time Stamp;Event Type;Sender's address
1.4.2 Hovering out Sender's address	UserID;Time Stamp;Event Type;Sender's address
1.4.3 Clicking on Sender's address	UserID;Time Stamp;Event Type;Sender's address
1.4.4 Hovering in Attachment	UserID;Time Stamp;Event Type;Attachment Information
1.4.5 Hovering out Attachment	UserID;Time Stamp;Event Type;Attachment Information
1.4.6 Clicking on Attachment	UserID;Time Stamp;Event Type;Attachment Information
1.4.7 Hovering in URLs	UserID;Time Stamp;Event Type;URL Information
1.4.8 Hovering out URLs	UserID;Time Stamp;Event Type;URL Information
1.4.9 Clicking on URLs	UserID;Time Stamp;Event Type;URL Information
1.4.10 Rating Trusting Confidence	UserID;Time Stamp;Email Title;Sender's Address;Ratings
1.4.11 Completing Email Classification	UserID;Time Stamp;Event Type;Classification Folder Name

Figure 2. Data collection description

Figure 2 provides a brief description of the data collected of user interactions with the emails. We defined the behavioral log format based on the HTTP common log format, which includes five fields of server timestamp, client timestamp, user identifier, action event, and action object. Clicking and hovering, two essential ways of interacting with the webmail application, can be further classified as hovering in and hovering out and clicking on application buttons and email related information, such as sender address, links and attachment.

Demographic information was collected for each of the participants, including age, gender, education level, language background, education of cyber awareness, etc. An ANOVA test was conducted to determine if there was any significant difference among different demographic groups in terms of the biometric features listed later in Section 3.3. The tests result suggested demographic groups showed no correlation with these features.

3.3. Exacting features

The features extraction stage characterizes a user's behavioral biometric information. Unlike traditional behavioral biometric systems, the proposed application-specific scheme extracted user behavioral features that directly interpretable within the context of the Roundcube webmail application. Four features were developed to represent distinctive behavioral characteristics of users based on available web interface APIs that capture email processing operations.

1) *Processing Time* is defined as the time taken by each participant from the moment they open an email to the moment they assign a rating of confidence level of classifying the email.
2) *Reaction Time* is defined as the time taken by each participant from the moment they assign a rating of confidence level to the moment they classify the email into one category.
3) *Phishing Tells Checking Bit* is a binary value to indicate whether the participant has checked an email is coming from a legitimate source, i.e., hovering over the sender's address.
4) *Rating* is defined as the confidence level assigned by participants of how strongly they believe the email falls into the chosen category of legitimate or phishing. A participant was mandated to give a confidence rating to each email before the email was moved into the classified category. This value ranges from 1 to 10, with 10 to denote the highest confidence.

These four independent features were extracted and each feature was represented by a numerical variable. All these values can be derived from the log file for each user and each instance of email processing. In this manner, unstructured textual log files were converted into vectorized features.

3.4. Offline identity attribution and online authentication

Our approach combines classical unsupervised and supervised machine learning algorithms to instrument offline identity attribution and online authentication.

Attribution is defined as the assignment of a user's behavior to a group identity. We differentiate offline attribution and online classification based on not merely different feature sets but also their purposes of identification and authentication respectively. Identification involves user profiling to generate distinct user classes or groups based on their distinguishable characteristics, while authentication is the verification of claimed identification.

In offline identity attribution, the entire log file generated by each user is used to evaluate the group that this user belongs to. It combines all the features in email processing, i.e., 4 features by 40 emails. Therefore, group identity attribution is based on complete user data to profile user behavioral patterns. More specifically, the group identity of each user was derived through unsupervised clustering.

The online authentication scheme works in a way that an alert is triggered when a user's real-time behavioral biometric features deviate drastically from the supposed group identity, e.g., being associated to the user through initial authentication using a password. This scheme presents a viable approach of continuous authentication with minimal computational cost and response time. A sliding window approach was used to feed a classifier features in the most recent segment of processing emails during user interactions.

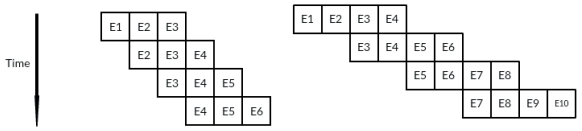

Figure 3. A sliding email window where training and testing instances are created by shifting a window frame over processed emails (*l*=3, *r*=1; *l*=4, *r*=2)

As shown in Figure 3, a window frame contains a certain number of emails, defined as the window length ***l***. This data processing window moves over the emails being processed in time order to create multiple instances of vectorized behavioral biometric features in the active session of a user. Each vector instance consists of the 4 features averaged over the number of emails in the current window frame. The other parameter of this sliding window approach, the shift ***r***, is the number of emails being replaced for each frame. In our scheme, the sliding window cannot be designed using a fixed time interval. Moreover, the window length ***l*** is an indicator of how soon this continuous authentication scheme can detect irregularities in real implementation.

3.5. Summary of the analysis methodology

The multi-stage analysis is illustrated in Figure 4. The pre-processing stage vectorizes log data in pre-defined format to extract the features as defined. We first tried out three clustering algorithms to determine a plausible number of user groups based on their distinctive behavioral patterns. We then assigned the group labels derived from clustering results to each user. After that we employed a set of classification algorithms to go through training and testing in both offline and online modes.

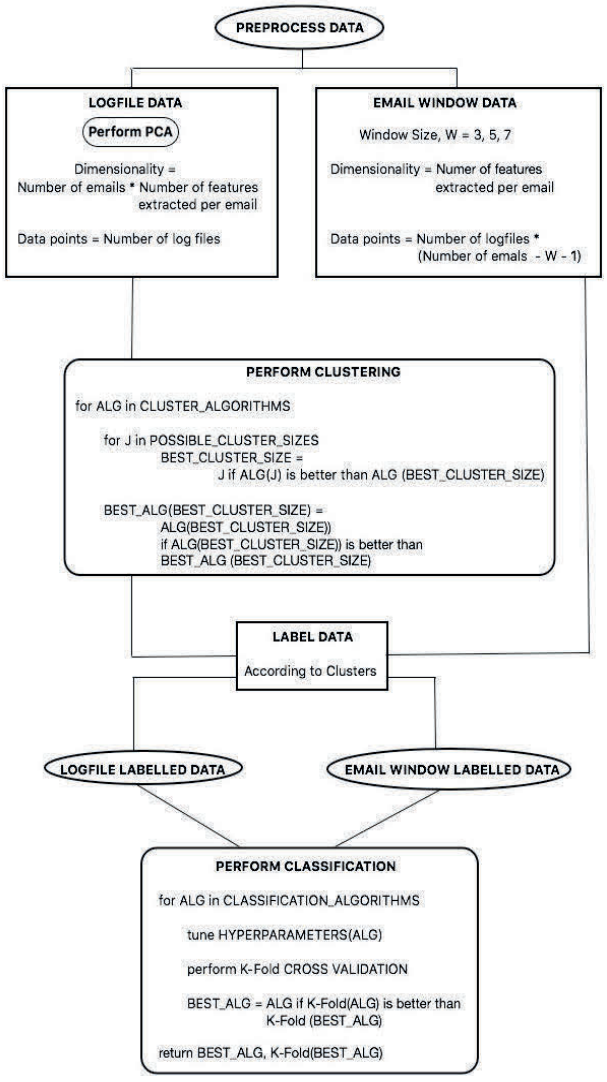

Figure 4. Training and testing stages for identity attribution and online authentication

In online authentication, if consecutive email window frames of emails processed by a user generate

a classification different from the originally associated user group, it would be compelling evidence that the current user might not be the claimed user.

4. Clustering, training, and testing

4.1. Dimensionality reduction

We conducted statistical analysis to determine the effect that each feature had on the unsupervised clustering results. Specifically, we performed an exploratory principal component analysis (PCA) of all the features of the users, to remove redundant features as the first step before clustering and classification to extract most useful components. The PCA algorithm automatically compares the number of data points and the number of features in determining the resulting components.

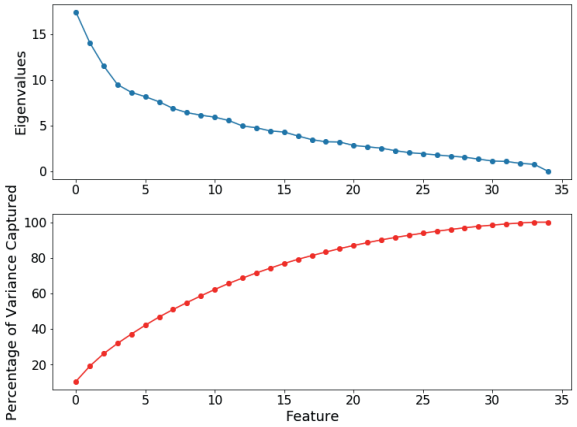

Figure 5. Dimensionality reduction by PCA

As in Figure 5, we arrived at this number by plotting the percentage of variance captured by the *k*th component. By examining the plots, we determined that 85% of the resulting components capture nearly 100% of the variance. For the group consisting of 35 data points, 29 components are kept. So, the data was whitened to remove correlated variables.

4.2. Clustering and initial labeling

Users of similar behavioral patterns are allocated into clusters using three clustering algorithms, i.e., k-means, hierarchical (agglomerative), and mean shift, on the feature vectors.

Mean Shift clustering grouped most points into a single cluster, with some points as their own clusters. Both k-means and hierarchical clustering gave out 2 almost identical clusters. Hierarchical clustering did not always give consistent number of clusters, i.e., it was more susceptible to finding different local optima each time the algorithm was run. Therefore, it looks that the best clustering algorithm for the dataset is k-Means. Specifically, as shown in Figure 6, 21 out of 35 participants were found belonging to one cluster, and 14 out of 35 belonging to the other.

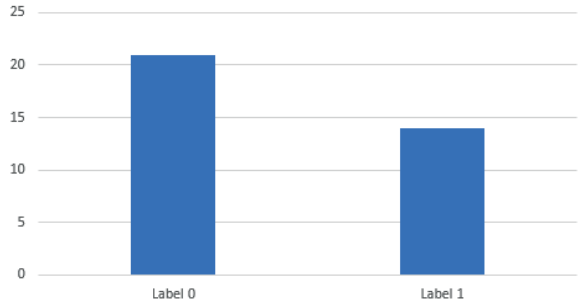

Figure 6. Identity attribution clustering result

In further evaluation, we considered the two cluster labels potentially reflect two group identities in the original dataset and conducted in-sample validation after PCA. However, it is hard to interpret exactly what would these group identities translate into. A probable explanation would be some users are more security conscious as they perform the email sorting tasks with relatively more time invested, and better email sorting performance achieved. However, it is hard to find proof due to the lack of knowledge of the participant users. The optimum number of clusters was evaluated only using the Silhouette Coefficient metric. Details about evaluation metrics will be addressed in Section 5.

4.3. Data generation for online authentication

Differences between legitimate and phishing emails could affect user behavior of email processing when implementing our email window approach for online authentication. On the one hand, training using binary email classification as features does not encode enough information to do good classification predictions. On the other hand, it is inherently unrealistic to know the nature of each email in real world and to control the order in which the emails are coming. As the result, we took a sliding window over emails listed in the randomized order that these 40 emails came for these participants. And we assumed that if the window size is large enough, we should be able to capture adequate information to smooth out effect of mixed normal and phishing emails. In one window frame, the email processing features are averaged over the contained emails, which results in a data point. Another advantage of this method is that this approach generates significantly more data points for

training purposes compared to using the entire log file for a user as one data point. This can thus help to build better classifiers. The chosen sliding email window influences the key parameters in classifier algorithms, which will be reported in Section 5.

After creating the sliding email window dataset, we assigned cluster labels to the resulting dataset. Each data point in the email window dataset will inherit the label of the user it belongs to.

4.4. Classification and evaluation

We now have 2 datasets with corresponding group labels for supervised learning or classification. We did classification using 7 different classifiers. The method for testing accuracy is k-fold cross validation, where k = 3 for offline identity attribution and k = 5 for online authentication using the sliding email window dataset. The final accuracy for each classifier is calculated by averaging the k-fold accuracy results. The 7 classifiers we have used are:

1) *K-Nearest Neighbors*
2) *Logistic Regression*
3) *Support Vector Machine*
4) *Linear Discriminant Analyses*
5) *Gaussian Naïve Bayes*
6) *Decision Tree*
7) *Random Forest*

5. Result analysis

Extensive analysis was conducted to examine the reliability and efficacy of the proposed approach. The evaluation of group identity is essentially two-class classification based on the clustering results. For offline identity attribution, we compared the classification performance before and after PCA. For continuous online authentication, we examined the classification performance of these classifiers with different parameters.

We considered performance metrics of accuracy (ACC), Area under Curve (AUC), as well as the time delay. ACC measures how well a binary classification test correctly identifies an observation and is defined as the ratio of the number of correct classifications to the testing sample size. AUC is defined as the area under a Receiver Operating Characteristic (ROC) curve, which is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [32]. For convenience, the definition of true positive and false positive here is based on the two clusters, not in the original sense of legitimate users and intruders. Specifically, in plotting and calculation, cluster label 0 represents the negative class while label 1 the positive.

We like to make a couple notes here. A misclassification is classifying one user wrongly to a different group. So, there is no need to further define false positive and false negative. Therefore, essentially every misclassification could result in a legitimate user being unauthenticated, comparable to the effect of a false alarm. The delay required by an online authentication scheme is the amount of time to collect and process sufficient behavioral data for a decision. In our scenario, this overhead is impacted by the size of the used email sliding window, which corresponds to time relative to the number of email processing, instead of an absolute time duration. The time spent in filling out the data needed for an email window frame and the time needed for executing the classifier together decide the time overhead for online authentication.

5.1. Offline identity attribution

The training phase for identity attribution takes all the data of all the users as input. For ACC comparison. we applied seven classifiers both before and after a PCA analysis and observed an improvement in accuracy as shown in Figure 7. The PCA is used to filter out non-discriminating features with less variance to contribute to classification applications. Such variables therefore can confuse the classifiers by having information that is not relevant. PCA removes these variables and only keeps valuable information, thus reducing noise and improving classification models.

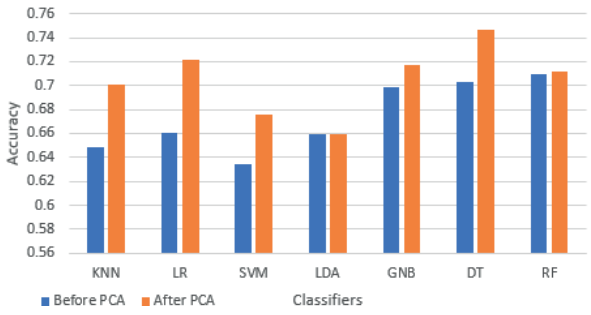

Figure 7. Before/after PCA performance

In evaluating the performance of offline identity attribution, we used 3-fold cross validation for ACC. The data set was divided into 3 subsets and the holdout method is repeated 3 times. Each time, one of the three subsets was used as the test set and the rest two subsets are put together to form a training set. Then the average ACC across all the three trials was calculated. Lastly, we applied traditional validation by setting training and testing data ratio to be 75:25 to calculate the AUC of

classifiers only after PCA. As shown in Table 1, decision tree and random forest present ACC of above 70% before PCA is applied. After PCA decision tree achieving ACC and AUC of 74.7% and 87.5%, respectively, proves to be the best classifier for identity attribution. That means the model can correctly classify the users of both classes with a probability of 74.7% while it erroneously classifies the users with a probability of approximately one fourth.

Table 1. Identity attribution performance

Classifier	Before PCA	After PCA	
	ACC	ACC	AUC
K-Nearest Neighbors	0.648	0.701	0.679
Logistic Regression	0.66	0.721	0.775
Support Vector Machine	0.634	0.676	0.667
Linear Discriminant Analysis	0.659	0.659	0.417
Gaussian Naïve Bayes	0.699	0.717	0.75
Decision Tree	0.703	0.747	0.875
Random Forest	0.7	0.712	0.75

5.2. Online authentication

We were more interested in the performance of this approach for online authentication. For this, evaluation was conducted over of the same 7 classifiers with different sliding email window parameters, as shown in Tables 2, 3, and 4. Overall, decision tree and random forest are two classifiers with the best performance.

Table 2. Online authentication performance with *l*=3, *r*=1

Classifier	ACC	AUC
K-Nearest Neighbors	0.731	0.608
Logistic Regression	0.67	0.568
Support Vector Machine	0.6	0.5
Linear Discriminant Analysis	0.719	0.611
Gaussian Naïve Bayes	0.794	0.784
Decision Tree	0.82	0.865
Random Forest	0.776	0.814

Table 3. Online authentication performance with *l*=5, *r*=1

Classifier	ACC	AUC
K-Nearest Neighbors	0.673	0.611
Logistic Regression	0.682	0.604
Support Vector Machine	0.62	0.655
Linear Discriminant Analysis	0.684	0.605
Gaussian Naïve Bayes	0.761	0.781
Decision Tree	0.813	0.786
Random Forest	0.848	0.799

Table 4. Online authentication performance with *l*=7, *r*=1

Classifier	ACC	AUC
K-Nearest Neighbors	0.692	0.612
Logistic Regression	0.698	0.657
Support Vector Machine	0.6	0.5
Linear Discriminant Analysis	0.689	0.662
Gaussian Naïve Bayes	0.67	0.607
Decision Tree	0.767	0.762
Random Forest	0.815	0.796

Since we had much more data points for online authentication, we employed 5-fold cross validation for ACC estimation. With the window size set at 5 and the shift at 1, decision tree and random forest achieved an ACC of over 81%. And the ACC of the random forest classifier was up to 84.8%, being the best.

The random forest classifier slightly outperforms the decision tree classifier in terms of ACC. This may be due to that a weighted random forest classifier puts more weight on the minority class, inflicting a heavier penalty on misclassifying the minority class. Additionally, its classification and "randomness" rules employ bootstrapping of data and random feature selection [33]. This enables the random forest classifier to find informative information in small subsets of the data.

Similarly, we set training and testing data ratio to 75:25 for AUC estimation. The AUC values range from 76.2% to 86.5% for the two best classifiers, i.e., decision tree and random forest, in the three settings of sliding window, again a promising performance.

These rates of misclassification are comparatively higher than many biometrics systems reported. A probable explanation is that our user study was conducted in a highly uncontrolled manner and the participants coming from varying backgrounds might not present strong group identities in such behaviors.

On the other hand, comparing the results derived from using different sliding email window parameters ***l*** and ***r***, there does not exist a clear pattern of how these parameters impact the performance of classifiers. Empirically we can determine the best parameters on a chosen specific application platform, although we have not fully examined them in a systematic manner. An additional complicating factor is how the nature of emails, phishing or legitimate, could impact the use of sliding email window.

Different window sizes can exert an impact on the best parameters for KNN, decision tree, and random forest. Specially, for Random Forest, the number of classification trees is the key parameter affecting its performance, which performs the best among a series of classifiers. For random forest, in order to get a better performance in actual application scenario, the number of trees should be more than the number of classes,

which is the number of user groups, which is 2 in this case. Different window sizes influence Random Forest the number of trees needed. Table 5 shows the number of trees needed decreases when the size of email window grows. Although the number of trees declines by one when the window size changes from three to five, the average accuracy increases by 7.2% with the AUC only declined by 1.5%.

Table 5. Classifier parameter and sliding email window size, where KNN-K is the number of neighbors; DT-D is the depth of the decision tree; RF-N is the number of trees in the forest

Window Size	KNN-K	DT-D	RF-N
$l = 3$	9	8	10
$l = 5$	5	9	9
$l = 7$	7	10	8

The number of trees also affects the training speed and complexity: processing speed would slow down, and complexity would escalate when the number of trees is growing. The time spent in filling out the data needed for the email window and the time needed for executing classifier algorithms together decide the delay for online authentication.

6. Conclusion and discussion

Behavioral biometrics deals with hardware platforms, software environment, and applications. Current authentication schemes based on physical or behavioral biometrics tend to be independent of different application contexts. They can be further assisted by behavioral biometrics that take the advantage of rich contextual information to a specific application.

This work is among the very few that studied user behaviors on web applications. We hosted our user study experiment on Amazon Mechanical Turk, remotely collected user behavioral data, extracted features from real user interactions with emails, and investigated a set of classifiers for offline group identify attribution and online authentication. The study demonstrated effectiveness of our methods measured by ACC and AUC.

The preliminary result has shown that the proposed authentication scheme is promising, although further research is warranted for a real-world implementation. Application-dependent user behavioral biometrics can encode distinctive identity information that can be used to assist and augment traditional authentication schemes. The results should also incentivize future studies aiming to detect insider attacks based on application-specific behaviors.

In an active authentication setting, a few challenges and further considerations arise. First, our experiment presents a relatively high misclassification rate, which risks forcing legitimate users to respond to challenges unnecessarily. Moreover, for the sliding email window approach, the tradeoff between the amount of data collect by an email window frame and the authentication efficiency is delicate to balance. In addition, hosting our user study experiment on a remote web platform may introduce more variables that need to be put under control. A limited scale of dataset and a choice of just an application can raise concerns to its validity too. Findings or challenges presented in this paper deserve a continuing effort in this novel direction of studying application context-specific behavioral biometrics.

7. Acknowledgement

The authors would like to thank Anton Dahbura from the Johns Hopkins University Information Security Institute for organizational support. This work has received partial support from NSF through Award No. 1544493.

8. References

[1] Accenture, “The State of Cybersecurity and Digital Trust 2016”, https://www.accenture.com/us-en/_acnmedia/PDF-23/Accenture-State-Cybersecurity-and-Digital-Trust-2016-Report-June.pdf, accessed 01.05.2018.

[2] Tripwire Guest Authors, “Insider Threats as the Main Security Threat in 2017”, https://www.tripwire.com/state-of-security/security-data-protection/insider-threats-main-security-threat-2017/, accessed 01.06.2018.

[3] Niinuma, K. and Jain, A. K., “Continuous user authentication using temporal information”, Proc. SPIE 7667, Biometric Technology for Human Identification VII, Orlando, Florida, United States, 2010.

[4] CA Technologies, “Enterprise Data Security: The Basics of User Behavior Analytics”, https://www.ca.com/us/collateral/white-papers/enterprise-data-security-the-basics-of-user-behavior-analytics.html, accessed 01.08.2018.

[5] Sandhu, R., Ferraiolo, D., and Kuhn, R., “The NIST Model for Role-Based Access Control”, RBAC '00 Proceedings of the fifth ACM workshop on Role-based access control, Berlin, German, 2000, pp. 47-63.

[6] Sandhu, R., Coyne, E., Feinstein, H., and Youman, C., “Role-based access control models”, IEEE Computer, 29(2), IEEE Computer Society Press, Los Alamitos, CA, USA ,1996, pp. 38-47.

[7] Fuchs, L. and Pernul, G., “Minimizing insider misuse through secure Identity Management.”, Security and Communication Networks, 5(8), John Wiley and Sons, Inc., New York, NY, USA, 2011, pp. 847-862.

[8] Schmit, A., "The Encyclopedia of Human-Computer Interaction (2nd Ed.)", https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction, accessed 01.08.2018.

[9] Shulkind, R., "Behavioral Biometrics 101: Behavioral Biometrics vs. Behavioral Analytics", https://blog.securedtouch.com/behavioral-biometrics-101-an-in-depth-look-at-behavioral-biometrics-vs-behavioral analytics, accessed 01.08.2018.

[10] Clarke, N., Karatzouni, S., and Furnell, S., "Flexible and Transparent User Authentication for Mobile Devices.", Emerging Challenges for Security, Privacy and Trust IFIP Advances in Information and Communication Technology, vol 297, Springer, Berlin, Heidelberg, 2009.

[11] Sastry, N., Shankar, U., and Wagner, D., "Secure verification of location claims", WiSe '03 Proceedings of the 2nd ACM workshop on Wireless security, San Diego, CA, USA, 2003, pp. 1-10.

[12] Yampolskiy, R. V. and Govindaraju, V., "Behavioral biometrics: A survey and classification.", International Journal of Biometrics, 1(1), Inderscience Publishers, Geneva, Switzerland, 2008, pp. 81-113.

[13] Buschek, D., Luca, A. D., and Alt, F., "Improving Accuracy, Applicability and Usability of Keystroke Biometrics on Mobile Touchscreen Devices", Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15, Seoul, Republic of Korea, 2015, pp. 1393-1402.

[14] Yu, E. and Cho, S., "Keystroke dynamics identity verification—its problems and practical solutions", Computers and Security, 23(5), Elsevier Advanced Technology Publications, Oxford, UK, 2004, pp. 428-440.

[15] Chang, K., Hightower, J., and Kveton, B., "Inferring Identity Using Accelerometers in Television Remote Controls.", Lecture Notes in Computer Science Pervasive Computing, 2009, pp. 151-167.

[16] Kale, A., Rajagopalan, A., Cuntoor, N., and Kruger, V., "Gait-based recognition of humans using continuous HMMs", Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, IEEE Computer Society Washington, DC, USA, 2002, pp. 336.

[17] Gafurov, D., Helkala, K., and Søndrol, T., "Biometric Gait Authentication Using Accelerometer Sensor", Journal of Computers,1(7), Academy Publisher, 2006, pp. 51-59.

[18] Brunelli, R. and Falavigna, D., "Person identification using multiple cues", IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(10), 1995, pp. 955-966.

[19] Bigun, J., Fierrez-Aguilar, J., Ortega-Garcia, J., and Gonzalez-Rodriguez, J., "Combining Biometric Evidence for Person Authentication", ASB'03 Proceedings of the 1st international conference on Advanced Studies in Biometrics, Springer-Verlag Berlin, Heidelberg, pp. 1-18.

[20] Greenstadt, R. and Beal, J., "Cognitive security for personal devices", AISec '08 Proceedings of the 1st ACM workshop on Workshop on AISec, ACM, New York, NY, USA, 2008, pp. 27-30.

[21] Patil, R.A. and Renke, A. L., "Keystroke Dynamics for User Authentication and Identification by using Typing Rhythm", International Journal of Computer Applications, 144(9), 2016, pp. 27-33.

[22] Dowland, P. S. and Furnell, S. M., "A Long-Term Trial of Keystroke Profiling Using Digraph, Trigraph and Keyword Latencies", Security and Protection in Information Processing Systems, vol 147, Springer, Boston, MA, 2004, pp. 275-289.

[23] Ilonen, J., "Keystroke dynamics", http://www.it.lut.fi/kurssit/0304/010970000/seminars/Ilonen.pdf, accessed 01.12.2018.

[24] Hashia, S., Pollett, C., and Stamp, M., "On using mouse movements as a biometric," in Proc. Int. Conf. Computer Science and Its Applications, Singapore, 2005, pp. 143–147.

[25] Bours, P. and Fullu, C. J., "A Login System Using Mouse Dynamics", 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto, 2009, pp. 1072-1077.

[26] Aksari, Y. and Artuner, H., "Active authentication by mouse movements", 2009 24th International Symposium on Computer and Information Sciences, Guzelyurt, 2009, pp. 571-574.

[27] Muthal, S., Li, S., Huang, Y., Li, X., Bos, N., Dahbura, A., and Molinaro, K., "A phishing study of user behavior with incentive and informed intervention", National Cyber Summit'17, 2017.

[28] Zhang, H., Singh, S., Li, X., Dahbura, A., and Xie, M., "Multitasking and Monetary Incentive in a Realistic Phishing Study", British Human Computer Interaction Conference, 2018.

[29] Bartneck, C., Duenser, A., Moltchanova, E., and Zawieska, K, "Comparing the Similarity of Responses Received from Studies in Amazon's Mechanical Turk to Studies Conducted Online and with Direct Recruitment", Plos One, 10(4): e0121595, 2015.

[30] Kittur, A., Chi, E. H., and Suh, B., "Crowdsourcing user studies with Mechanical Turk", CHI '08, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2008, pp. 453-456.

[31] Layman, L. and Sigurdsson, G, "Using Amazons Mechanical Turk for User Studies: Eight Things You Need to Know", ACM / IEEE International Symposium on Empirical Software Engineering and Measurement, Baltimore, MD, 2013, pp. 275-278.

[32] Fawcett, T., "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers", http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf, accessed 04.12.2018.

[33] Osanaiye, O., Cai, H., Choo, K.K.R., Dehghantanha, A., Xu, Z., and Dlodlo, M., "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing", EURASIP Journal on Wireless Communications and Networking, 2016, pp. 130-139.