Gametrics: Towards Attack-Resilient Behavioral Authentication with Simple Cognitive Games

Manar Mohamed
University of Alabama at Birmingham
Birmingham, AL, USA
manar@uab.edu

Nitesh Saxena
University of Alabama at Birmingham
Birmingham, AL, USA
saxena@uab.edu

ABSTRACT

Authenticating a user based on her unique behavioral biometric traits has been extensively researched over the past few years. The most researched behavioral biometrics techniques are based on keystroke and mouse dynamics. These schemes, however, have been shown to be vulnerable to human-based and robotic attacks that attempt to mimic the user's behavioral pattern to impersonate the user.

In this paper, we aim to verify the user's identity through the use of *active, cognition-based user interaction* in the authentication process. Such interaction boasts to provide two key advantages. *First*, it may *enhance the security* of the authentication process as multiple rounds of active interaction would serve as a mechanism to prevent against several types of attacks, including zero-effort attack, expert trained attackers, and automated attacks. *Second*, it may *enhance the usability* of the authentication process by actively engaging the user in the process.

We explore the cognitive authentication paradigm through very simplistic interactive challenges, called *Dynamic Cognitive Games*, which involve objects floating around within the images, where the user's task is to match the objects with their respective target(s) and drag/drop them to the target location(s). Specifically, we introduce, build and study *Gametrics* ("Game-based biometrics"), an authentication mechanism based on the unique way the user solves such simple challenges captured by multiple features related to her cognitive abilities and mouse dynamics. Based on a comprehensive data set collected in both online and lab settings, we show that *Gametrics* can identify the users with a high accuracy (false negative rates, FNR, as low as 0.02) while rejecting zero-effort attackers (false positive rates, FPR, as low as 0.02). Moreover, *Gametrics* shows promising results in defending against expert attackers that try to learn and later mimic the user's pattern of solving the challenges (FPR for expert human attacker as low as 0.03). Furthermore, we argue that the proposed biometrics is hard to be replayed or spoofed by automated means, such as robots or malware attacks.

ACSAC '16, December 05-09, 2016, Los Angeles, CA, USA

DOI: http://dx.doi.org/10.1145/2991079.2991096

1. INTRODUCTION

Behavioral biometrics is one of the most active research topics in the area of user authentication. The most studied behavioral biometrics technique is keystroke dynamics [2, 20], a method in which the typing patterns are used to build a unique signature of a user, that can be used for point-of-entry authentication (e.g., when combined with the dynamics involved in typing a password). Other widely studied behavioral biometrics techniques include mouse movement patterns [35], and swiping patterns [11] for touchscreen devices. However, these existing behavioral biometrics approaches have two fundamental limitations, which have possibly prevented their transition into real-world application despite significant research:

- The false negative rate (i.e., the possibility of falsely rejecting the legitimate user) is relatively high, which undermines the overall usability and real-world acceptability, as users may fail to login to their accounts on a relatively frequent basis.
- The false positive rate (i.e., the possibility of falsely accepting a "different user") is relatively high, which weakens the overall security. In addition, and perhaps more seriously, a determined attacker (a human or a bot) can deliberately/actively mimic the user's activities (e.g., typing or swiping) and compromise the authentication functionality, for instance, based on the global characteristics of typing patterns as shown in [29], or a user-specific (previously leaked) authentication template as shown in [31]. Existing schemes have also been shown vulnerable to internal attacks where the device from which the users logs in is itself compromised with a malware [19].

These limitations stem from the fact that existing behavioral authentication approaches lack enough randomization and identifying "cues", especially when the duration of input is short (e.g., for point-of-entry authentication), necessary to uniquely identify the user, negatively affecting user experience and facilitating passive and active adversarial spoofing.

In this paper, to overcome the limitations of the current behavioral biometrics systems, we propose *Gametrics*, a novel system of interactive game-based behavioral biometrics. Whenever users wish to authenticate to a device or service, *Gametrics* would simply request them to play a short and simple cognitive game. Once identified, permission to access an account or device can be granted via a back end database as is done with existing behavioral biometric solu-

tions. Games are a good platform for the purpose of authentication since web browsers and touch screen devices fully support them.

In contrast to traditional behavioral biometrics, due to their randomized, dynamic, interactive and cognitive nature, cognitive games offer an attractive platform using which sufficient cues in a short period of time could be extracted. In our proposed *Gametrics* system, the users will be authenticated based on their *multiple multi-modal* gameplay patterns as well as mouse dynamics. Specifically, *Gametrics* utilizes various characteristics of the users' interactions with the games: (1) *active and idle time* [6]; (2) *cognitive abilities* [1], such as visual search, and working memory & information processing speed, and (3) *mouse dynamics*, such as mean click length, average click rate, as well as distance, speed, and angle at which the mouse is moved [24, 26]. This type of data is already being collected and mined by video game companies for marketing [10, 12] and quality control purposes, which supports the plausibility of gaming-based biometrics as a general authentication solution deployable in the near future. As an example, the Valve Corporation collects extensive information on users through its Steam platform and publishes real time statistics for its most popular games online [32].

Gametrics provides significant advantages in terms of security and usability of the user authentication process. *First*, it can significantly help improve the security of authentication in comparison to existing solutions. *Gametrics* can be utilized as a stronger behavioral biometrics [2, 20] since the active, multi-round, interaction with the games is unique per user. Moreover, by using games that contain moving objects or objects that are placed at random locations when combined with passwords, the level of randomization may serve as a defense against key loggers [19] or side-channel attacks (i.e., attacks that try to deduce the entered password on touch screen devices based on the locations the user has pressed on the screen [15, 21]), replay attacks and spoofing attacks [19, 29].

Second, engaging the user in the user authentication process via interactivity may enhance the level of user experience. For example, interactive solutions may be more suitable for small touchscreen devices, where reading/entering text might be challenging. Moreover, the interactivity makes it possible to extract enough information to identify the user within a short period of time.

To summarize, there are many unique advantages of *Gametrics* over known biometrics systems (behavioral or otherwise). *First*, the use of multiple multi-dimensional, explicit-implicit, game play features when "fused" together could significantly reduce the false negative rates and thus improve usability, when compared to existing behavioral biometrics. *Second*, such fusion could significantly reduce the false positive rates, thereby improving the security. Especially, spoofing the user (either automatically or manually) may become very hard given that the attacker would have to *simultaneously* mimic multiple subtle user interaction patterns (corresponding to the different underlying features). Moreover, the dynamic and interactive nature of the game makes it difficult for an attacker to simply "replay" a previously learned template or game session, unlike static passwords or keyboard dynamics mechanisms. *Third*, the interactive element of the underlying games may further enhance the usability and promote user acceptability.

While *Gametrics* can be built using different forms of cognitive games and puzzles, in this paper, as our authentication object, we use the Dynamic Cognitive Game (DCG) notion recently introduced in [18] for the purpose of building CAPTCHA (*not* user authentication) schemes. DCG games involve objects floating around within the images, where the user's task is to match the objects with their respective target(s) and drag/drop them to the target location(s) (examples shown in Figure 1). We investigate the applicability of using such simple constructs to extract a user's unique biometric information based on multi-modal user interactions. We model and analyze the security of our multi-modal game biometrics with respect to spoofing attacks, where the attacker deliberately attempts to mimic the victim user's game play interaction patterns. We argue that attacking the proposed biometrics with automated means would be hard as the bot would require solving a CAPTCHA as well as mimicking the user interaction with the challenge. Moreover, the randomization in the challenges would prevent attacks that involve recording the user interaction with the challenge and then replaying the recorded data later to authenticate the attacker [18]. As authentication applications, *Gametrics* is suitable for point-of-entry login. It could also be a promising and a natural solution for the difficult problem of *fall-back authentication* (e.g., needed to retrieve a forgotten password) [25, 27, 28].

Our Contributions: We believe that this paper makes the following key contributions to the field of user authentication in general and behavioral authentication in particular:

1. *Gametrics Design and Implementation*: We design and implement a *Gametrics* system based on simple DCGs to capture the unique user interactions. Our system is built using machine learning techniques and extracts a total of 64 features from each game challenge solving instance that capture the multiple unique cognitive abilities and the mouse dynamics of the users.

2. *Evaluation of Gametrics under Benign Settings and Zero-Effort Attacks*: We collect a comprehensive data set from a total of 118 users (98 Amazon Mechanical Turk (AMT) online workers and 20 University lab participants), and show that *Gametrics* can identify the legitimate users and the zero-effort attackers ("different users") with a high accuracy (average False Positive Rate = 0.02, and False Negative Rate = 0.02) within a short period of time (average around 15 seconds).

3. *Evaluation of Gametrics under Active Attacks*: We show that *Gametrics* can thwart active attackers that deliberately attempt to mimic a user's interaction with the challenges in an observation-based attack (attack success rate as low as 0.03). Furthermore, we argue that attacking *Gametrics* using automated mechanisms, internal or external, is also a hard task.

Paper Outline: The rest of this paper is organized as follows. In Section 2, we lay out the design goals and threat model for our *Gametrics* system. In Section 3, we describe the authentication game object (DCGs) used in our system. This is followed by Section 4, where we describe our data collection methodology and procedures. Next, in Section 5, we elaborate on our machine learning techniques and feature extraction methods to build the *Gametrics* authenti-

cation model, and provide the classification results in benign setting and against zero-effort passive attackers. In Section 6, we evaluate *Gametrics* against active adversarial attacks that deliberately attempt to mimic a user's game play pattern to defeat the authentication system. In Section 7, we discuss further aspects of our work and provide future research directions. In Section 8, we provide a literature review on different forms of prior behavioral biometric systems. Finally, in Section 9, we conclude our work highlighting the main take away points.

2. DESGIN GOALS & THREAT MODEL

A core objective of *Gametrics* is to improve the usability and the security of user authentication process (especially that of behavioral biometrics authentication). As such, our aim is to design and develop an interactive behavioral biometrics system that possesses the following properties:

1. **Usability:** The user has to be identified within a short time and with high accuracy.

2. **Security against Zero-Effort Attacks**: Any biometrics scheme should be able to distinguish between different users. That is, one user (potentially an attacker) should not be able to log in as another user (a victim).

3. **Security against Shoulder-Surfing Attacks**: An external attacker who monitors the user while she is authenticating herself to the system, should not be able to mimic and impersonate the user at a later point of time.

4. **Security against Automated Attacks:** We aim to provide security against sophisticated attacks where the attacker steals a user's authentication template (e.g., by hacking into the device or server that stores this template) and tries to authenticate itself in an automated manner to the system.

5. **Security against Internal Attacks:** We aim to provide security against internal attacks, such as a malware residing on the authentication terminal itself that records the user's valid authentication token/template and replays it later, or tries to learn the template by recording one or multiple valid authentication sessions and then creates an authentication token to authenticate itself as the user. Other forms of behavioral biometrics schemes have been shown to be vulnerable to such attacks [19].

3. GAME COGNITIVE TASK

In this section, we elaborate on the design and the implementation of the interactive DCG constructs we utilized in our study.

3.1 Cognitive Task Design

We embed the cognitive task in simple web-based games, following the design presented in [18]. In this design, each of the game challenges has three target objects and six moving objects. The user's task is to drag a subset of the moving objects (answer objects) to their corresponding target objects. Solving a challenge require the user to: (1) understand the content of the images, (2) find the semantic relationship between the answer objects and the target objects, and (3) drag the answer objects to their corresponding targets. We impose a time limit of 60 second to complete each challenge.

We aim to identify the user based on her interaction with the challenge. Basically, we aim to identify the user based on her cognitive ability (i.e., the time it takes her to recognize the objects and perform the required task) and mouse interaction (i.e., mouse movement characteristics such as mouse movement speed and acceleration).

3.2 Cognitive Task Implementation

We implemented the challenges using Adobe Flash ActionScript3 and the web server using PHP. The challenge image/frame size is 500×300 pixels, the size of each of the moving object is 75×75 pixels and the size of the target objects is 90×90 pixels. The challenge starts by placing the objects in random locations on the image. Then, each object picks a random direction in which it will move. A total of 8 directions were used, namely, N, S, E, W, NE, NW, SE and SW. If the chosen direction is one of E, W, S, or N, the object will move (across X or Y axis) by 1 pixel per frame in that direction. Otherwise, the object will move $\sqrt{2} = 1.414$ pixels per frame along the hypotenuse, corresponding to 1 pixel across both X and Y axes. This means that on an average the object moves 1.207 $[= (1 \times 4 + 1.414 \times 4)/8]$ pixels per frame. We set the number of frames per seconds to 40 FPS. The object keeps moving in its current direction until it collides with another object or with the challenge border, whereupon it moves in a new random direction.

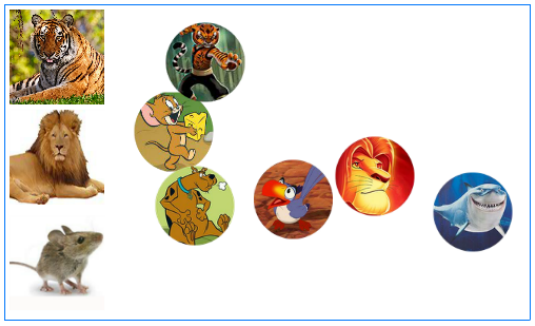
The challenge starts when the user presses a "Start" button on the screen center. The challenge ends when the user drags all the answer objects and drops them onto their corresponding targets, in which case a "Game Complete" message is provided or timeout is reached, in which case a "Time Out" message is provided.

After the user performs an object drag/drop, the challenge code sends to the server the identifier of the object and the drop location. The server checks the correctness of the drag/drop and gives feedback to the challenge code. If the web server confirms that the object was dropped on its corresponding target, the object disappears giving feedback to the user that he performed a correct action. After the user drags and drops all the answer objects to their corresponding targets, the challenge code sends to the server the log of the gameplay. The gameplay log contains the objects locations, the mouse location and status (up/down) at each time interval to the server. The server utilizes this log to authenticate the user. The timestamps were generated from multiple events listeners: MouseEvent.CLICK, MouseEvent.MOUSE_UP, and MouseEvent.MOUSE_MOVE.

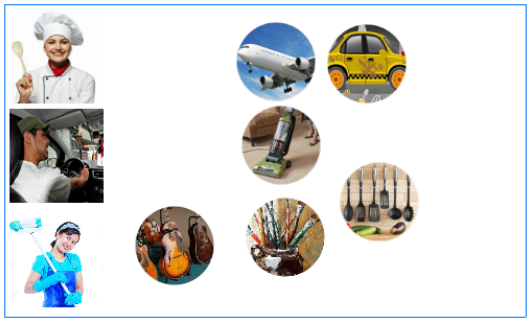
For the purpose of our study, we implemented six instances of the explained challenges that can be categorized into three categories (two instances of each category) described below. A sample of each of the implemented categories is shown in Figure 1.

- **Brands:** The targets are popular worldwide brands and the objects are commercial products (e.g., Nike and Nike shoes).
- **Animals:** The targets are real animals and the moving objects are cartoon animals (e.g., lion and Lion King).

(a) Brands



(b) Animals



(c) Professions

Figure 1: Challenges instances. Targets, on the left, are static; moving objects, on the right, are mobile. The user task is to drag-drop a subset of the moving objects (answer objects) to their corresponding targets

- **Professions:** The targets are professionals and the moving objects are tools (e.g., taxi driver and taxi).

4. DATA COLLECTION

As a pre-requisite to building and testing our *Gametrics* system, we pursued data collection from human users, in both online and lab settings. In this section, we elaborate on our data collection methodology, and the characteristics of the collected data set.

The participation in our two studies was voluntary, and standard ethical procedures were fully followed, e.g., participants being informed, given choice to discontinue, and not deceived. The studies was approved by our university's Institutional Review Board. The data collection experiments were divided into four phases. First, we subjected the participants to a consent form. Then, we asked the participants to go through a tutorial and fill up a demographics form. Next, we asked the participants to solve several instances of the game challenges explained in Section 3.2. At the end of the study, we asked the participants to fill-out a survey form about their experience. The survey contained the 10 System Usable Scale (SUS) [4] standard questions, each with 5 possible responses (5-point Likert scale, where strong disagreement is represented by "1" and strong agreement is represented by "5"). SUS is a standard questionnaire to measure the usability of software, hardware, cell phones and websites, and it has been deployed in many prior security usability studies. Moreover, specific to our study, we added two questions to the survey in order to measure the easiness and playfulness of the challenges. As the participants played the game challenges, all of their gameplay mouse events were recorded in the background.

Table 2: Summary of the Collected Data Sets

		# Users	Solving Time(s) Mean (std)	Completed Challenges
Online Study	**Day 1**	98	7.39 (3.55)	5839
	Day 2	62	7.23 (2.77)	2209
	Day 3	29	7.65 (2.98)	1028
Lab Study		20	7.66 (3.45)	1200

Table 2 summarizes the characteristics of the data collected during the two studies. The total number of participants is 118 (98 in online study and 20 in lab study). The participants successfully completed a total of 10276 challenges (9076 in online study and 1200 in lab study). The average time to complete a game challenge was around 7.5 seconds.

For our online data collection study, we utilized the Amazon Mechanical Turk (AMT) service to recruit the participants. The aim of our online study was to evaluate the applicability of identifying the user based on the way she interacts with the posed game challenges. Moreover, we wanted to determine how our system would perform in a longitudinal setting, over multiple sessions/days. Therefore, we created a total of three Human Intelligence Tasks (HITs) distributed over three days. The first HIT was created with 100 assignments to have 100 unique workers. We gathered 98 valid submissions until the HIT expired. The workers were directed to the website hosting the study. They were required to solve a tutorial, fill a demographics form and play 60 instances of our challenges. The order of presenting the challenges to the participants was random. Finally, the participants filled out the survey. On the next two days, we sent the participants emails asking them to participate in the follow-up study. However, we asked them to solve 36 challenges rather than 60 challenges in this round. 62 participants performed the study on the second day and 29 performed the study on the third day. We paid each participant $1.0 for the first HIT, and $0.5 each for the second and third HIT.

The participants in our online study were from various age groups, education levels and backgrounds. Age group: 1% $<$ 18, 20.4% 18-24, 38.8% 25-34, 32.7% 35-50 and 7.1% $>$ 50. Gender: 58.2% male and 41.8% female. Education: 26.5% high school graduate, 58.2% hold bachelor degree, 14.3% hold master degree and 1% hold a PhD degree. The participants were from various backgrounds such as Computer Sci-

Table 1: The Features Utilized for Classification

	Feature		Description
Cognitive	Time	number	Time taken to complete the challenge
	Time first action	number	The timestamp of the first mouse event after the game start
	Time first drag	number	The timestamp of the first drag
	Time between drags	mean, std, min, max	Times between drops and start of drags
Mouse interaction	Speed drag	mean, std, min, max	Speed while dragging
	Speed move	mean, std, min, max	Speed while moving
	Acceleration drag	mean, std, min, max	Acceleration while dragging
	Acceleration move	mean, std, min, max	Acceleration while moving
	Difference timestamp	mean, std, min, max	The difference between each consecutive recorded timestamps
	Move silence	mean, std, min, max	The times between consecutive timestamps while the mouse is moving
	Drag silence	mean, std, min, max	The times between consecutive timestamps while dragging
	Pause and drag	mean, std, min, max	The times between approaching the object and click on it
	Pause and drop	mean, std, min, max	The times between approaching the target and drop
	Angle	mean, std, min, max	The angles between each three consecutive points
Mixed	Drag distance to real distance	mean, std, min, max	The difference between the distance traveled while dragging and the straight line connecting the start and end points of the drag
	Move distance to distance	mean, std, min, max	The difference between the distance traveled while moving and the straight line connecting the start and end points of the move
	Distance click object center	mean, std, min, max	Distances of the clicks and objects' centers
	Distance drop target center	mean, std, min, max	Distances of the drops and targets' centers
	Total distance	number	Total distance

ence, Engineering, Medicine, Law, Social Science, Finance, Business, Mathematics, Art, etc. (detailed demographics information is populated in Table 6 in the Appendix)

For our lab-based study, we collected data from some volunteers recruited from our University. It followed a similar protocol as the online study, but using a lab computer. We asked the volunteers to perform a similar task as the task performed by the AMT workers on the first day. A total of 20 undergraduate and graduate students as well as some employees participated in the study. The age of the participants ranged between 19 and 50, 13 of them are male and 7 are female, 5 are high school graduate, 8 have bachelor degree and 7 have master degree. The majority of the participants are from Computer Science background (Table 6 in the Appendix). We asked the volunteers to play 60 instances of the challenges using the same computer and same setting. The aim of this study was to validate the results of the AMT study. In particular, we mainly wanted to ensure that the acquired results are not based on the platform and the setting used in performing the experiment rather than the different characteristics of an individual's unique way of interacting/solving the game challenges.

5. SYSTEM DESIGN & RESULTS

In order to evaluate the applicability of the *Gametrics* as an authentication scheme, we utilized the machine learning approach. In this section, we present the features we extracted from the user's gameplay logs collected during our data collection campaign. Then, we discuss the classification models and the classifier employed. Finally, we present the classification results for the benign setting and the zero-effort attack.

5.1 Feature Extraction

From each instance of the gameplay logs we collected during the data collection phase, we extracted a total of 64 features that captures the cognitive abilities as well as the mouse interaction characteristics of the participants while they are interacting with the challenges. (The extracted features are summarized in Table 1.)

As described in Section 3.1, in order to solve a challenge, the user has to match the answer objects to their corresponding targets. In order to do that, the user has to understand the content of the images representing the targets and the moving objects, find the relationship between the moving objects and the target objects, and then select a subset of the moving objects (the answer objects) and finally drag/drop them to their corresponding targets. By monitoring the users while solving the challenges (lab study), we found different users take different approaches to solve the challenges. For example, some users start by trying to comprehend the whole challenge and then start the object matching, while some try to find the answer objects corresponding to the target in certain order (i.e., always try to search for the answer object that corresponds to the top most target, and then the second and so on), while some try to pick the object closest to the mouse cursor and then check if it matches with any of the targets. For visualization purposes, these differences in the cognitive characteristics of different users are illustrated in Figure 2.

These different mechanisms of solving the game challenges are related to the cognitive characteristics of individuals. We capture these characteristics based on the following features:

1. The time between the user pressing on the start button and the first recorded mouse event and the time of the first click/drag. These timing measures capture the time the participates take to comprehend the challenge and start solving it.

2. The times between each of the drops and the start of the next drag (these capture the time the user takes to find the next answer object).

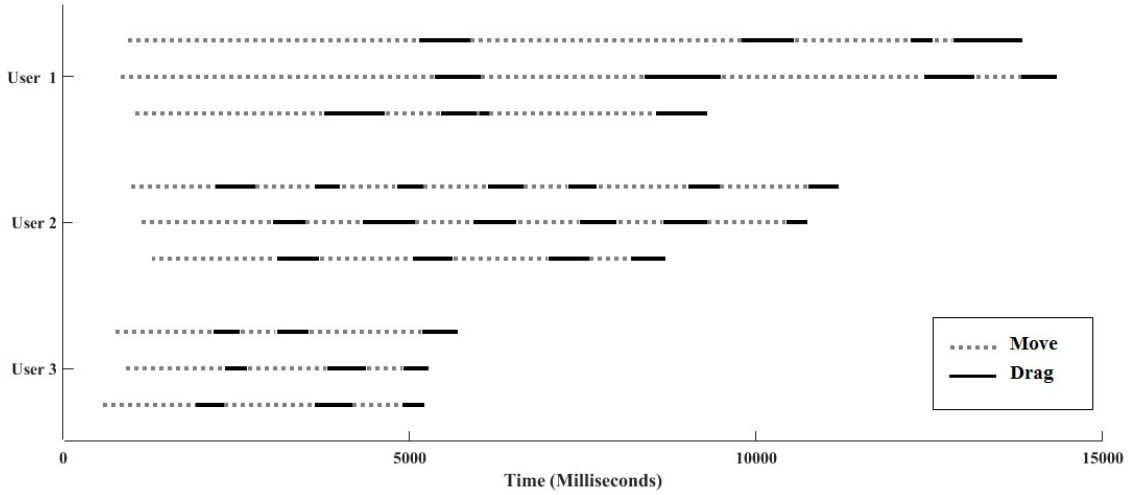


Figure 2: An example for illustration of different cognitive characteristics among different users while playing the game challenges: the time for completing the games and the time spent in drag and time spent in moving the mouse around. We can see that User 1 took a long time to understand the game (long move segment before the start of the first drag), also took on average a long time to locate each of the answer objects and to start dragging. User 2 took shorter time to complete the challenges but committed many mistakes (the user performed exactly 3 drags and drops to complete each challenge, however, User 2 performed on average more than 5 drags), User 3 completed the games in short time with shorter on average times to locate the answer objects.

3. The total time taken by the user to complete the challenge.

The mouse movement characteristics of the users are captured by following features:

1. The speed and acceleration while the user is searching for an answer object and while the user is dragging the object.
2. The duration between each two consecutively generated timestamps and the "silence" during move and during drag.
3. The time duration between reaching an object and clicking on it, and the time duration between approaching a target object and dropping an answer object on it.
4. The angles between the lines that connect each 3 consecutive points in the mouse movement trajectory.

Other mixed features are also extracted that relate to both cognitive and mouse movement characteristics of the participants such as the total distance the mouse moved within a game challenge, the difference between the straight line connecting the start and the end of a move or a drag and the real distance traveled. The distance between a click and the object center, and a drop and the target center.

5.2 Classifier and Metrics

In our analysis, we utilized the Random Forest classifier. Random Forest is an ensemble approach based on the generation of many classification trees, where each tree is constructed using a separate bootstrap sample of the data. In order to classify a new input, the new input is run down all the trees and the result is determined based on majority voting. Random Forest is efficient, can estimate the importance of the features, and is robust against noise [16]. Several other classifiers were tested during the course of study such as SVM, Bayes Network, Neural Networks, but Random Forest outperformed all of them.

In our classification task, the positive class corresponds to the gameplay of the legitimate user and the negative class corresponds to the impersonator (other user / zero-effort attacker). Therefore, true positive (TP) represents the number of times the legitimate user is granted access, true negative (TN) represents the number of times the impersonator is rejected, false positive (FP) represents the number of times the impersonator is granted access and false negative (FN) represents the number of times the correct user is rejected.

As performance measures for our classifier, we used false positive rate (FPR), false negative rate (FNR), precision, recall and F-measure (F1 score), as shown in Equations (1) to (5). FPR and precision measure the security of the proposed system, i.e., the accuracy of the system in rejecting impersonators. FNR and recall measure the usability of the proposed system as high FNR leads to high rejection rate of the legitimate users. F-measure considers both the usability and the security of the system. To make our system both usable and secure, ideally, we would like to have FPR and FNR to be as close as 0, and recall, precision and F-measure to be as close as 1.

$$FPR = \frac{FN}{TN + FN} \tag{1}$$

$$FNR = \frac{FN}{TP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

Table 3: AMT Study Results: Performance for the classifier for three different classification models. The first part shows the performance of the classifier using all the features. The next part shows the results of using the features subset that provides the best average results. The last part shows the result of using the best features subset for each user. For each of the models, we show the results of using a single challenge and merging of two challenges. Highlighted cells emphasize the most interesting results.

			FPR	FNR	Precion	Recall	F-Measure
			Mean (Std)				
All features	**Single**	**Day 1**	0.12 (0.10)	0.12 (0.16)	0.88 (0.09)	0.88 (0.16)	0.87 (0.12)
		Day 2	0.11 (0.09)	0.25 (0.31)	0.81 (0.24)	0.75 (0.31)	0.76 (0.28)
		Day 3	0.10 (0.07)	0.22 (0.27)	0.86 (0.14)	0.78 (0.27)	0.80 (0.24)
	Merge	**Day 1**	0.10 (0.13)	0.11 (0.18)	0.91 (0.11)	0.89 (0.18)	0.88 (0.14)
		Day 2	0.09 (0.11)	0.20 (0.30)	0.85 (0.23)	0.80 (0.30)	0.80 (0.27)
		Day 3	0.08 (0.10)	0.22 (0.30)	0.86 (0.25)	0.78 (0.30)	0.80 (0.27)
Average overall best	**Single**	**Day 1**	0.11 (0.09)	0.11 (0.15)	0.89 (0.08)	0.89 (0.15)	0.89 (0.11)
		Day 2	0.18 (0.13)	0.17 (0.15)	0.83 (0.15)	0.83 (0.15)	0.82 (0.12)
		Day 3	0.10 (0.07)	0.19 (0.26)	0.85 (0.18)	0.81 (0.26)	0.82 (0.23)
	Merge	**Day 1**	0.10 (0.13)	0.09 (0.16)	0.91 (0.11)	0.91 (0.11)	0.90 (0.13)
		Day 2	0.12 (0.12)	0.19 (0.20)	0.88 (0.12)	0.81 (0.20)	0.83 (0.14)
		Day 3	0.12 (0.18)	0.18 (0.18)	0.88 (0.10)	0.82 (0.18)	0.84 (0.13)
User specific	**Single**	**Day 1**	0.06 (0.06)	0.02 (0.04)	0.95 (0.05)	0.98 (0.04)	0.96 (0.04)
		Day 2	0.09 (0.09)	0.07 (0.10)	0.91 (0.09)	0.93 (0.10)	0.92 (0.09)
		Day 3	0.07 (0.06)	0.07 (0.10)	0.93 (0.06)	0.93 (0.10)	0.93 (0.07)
	Merge	**Day 1**	0.02 (0.05)	0.02 (0.05)	0.98 (0.05)	0.98 (0.05)	0.98 (0.04)
		Day 2	0.05 (0.09)	0.04 (0.09)	0.96 (0.08)	0.96 (0.09)	0.96 (0.08)
		Day 3	0.04 (0.06)	0.03 (0.05)	0.96 (0.05)	0.97 (0.05)	0.96 (0.04)

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F\text{-}measure = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

5.3 Classification Models & Feature Selection

We studied various models of classifications. In the first model, we utilized all the features explained in Table 1 for training and later testing the classifier. Second, in order to improve the accuracy of the classification, we ran a program to find the subset of features that produces the best classification results, as using many features can cause over fitting of the classifier and therefore reduce the accuracy of the future prediction, thus removing some features may improve the accuracy. Therefore, we report, in the next subsection, the results obtained by using the subset of features that produces the best average results across all the participants (users being authenticated) in the study. Third, we find the best subset of features that produces the best classification results per user.

For each of the three classification models, we study the identification of the user based on a single game challenge as well as on combining two challenges. As the average time for solving a challenge is around 7.5 seconds, we believe that utilizing two instances of the game challenges to identify the user is not much of an overhead. However, it may improve the accuracy by doubling the amount of captured interactions between the user and the challenges. In a real-life authentication application, posing the user with two consecutive game challenges captures this scenario.

5.4 Classification Results

Inter-Session Analysis: As mentioned in Section 4, we collected data from 98 AMT workers during the first day of our data collection experiment. Each of them completed 60 challenges. We divided the collected data into 98 sets based on the users' identities (ids). In order to build a classifier to authenticate a user based on her gameplay biometrics, we defined two classes. The first class contains the gameplay data from a given user (to be identified), and the other class contains randomly selected gameplay data from other users.

Then, we divided the data into two sets, one for training and the other for testing. The first 40 gameplay instances of each participant and 40 gameplay instances of the randomly selected set were used to train the classifier, while the other 20 are used for testing. We have tested our three classification models in two settings to evaluate our system. In the first setting, we used a single gameplay instance to authenticate the user while in the second setting, we used two instances of the gameplay to authenticate the user. The merging is done by averaging all the features from the two instances.

The results are shown in the first row ("Day 1")of each block in Table 3. We see that utilizing two gameplay instances is consistently better than using a single instance. Also, we find that the user-specific model outperforms both the other models (using all the features and using the features that provide the best average over all results). Thereby, the best results are acquired by using the user-specific model and merging two challenge instances in which both the false positive rate and false negative rate = 2%.

Intra-Session Analysis: Our other main goal was to check the accuracy of the classifier over multiple sessions. As mentioned in Section 4, 62 AMT workers participated in the study in the second day and 36 participated in the study in the third day. For each of these users, we used the data of the gameplay of the previous day(s) to train the classifier and then we tested the classifier with the data collected in the next day(s).

Table 4: Lab-Based Study Results: Performance for the classifier for three different classification models. The first part shows the performance of the classifier using all the features. The next part shows the results of using the features subset that provides the best average results. The last part shows the result of using the best features subset for each user. For each of the models, we show the results of using a single challenge and merging of two challenges. Highlighted cells emphasize the most interesting results.

		FPR	FNR	Precion	Recall	F-Measure
		Mean (Std)				
All features	**Single**	0.20 (0.12)	0.23 (0.14)	0.80 (0.10)	0.77 (0.14)	0.78 (0.10)
	Merge	0.15 (0.18)	0.16 (0.16)	0.87 (0.15)	0.84 (0.16)	0.84 (0.13)
Average overall best	**Single**	0.18 (0.13)	0.22 (0.14)	0.82 (0.11)	0.78 (0.14)	0.80 (0.10)
	Merge	0.14 (0.15)	0.16 (0.14)	0.88 (0.12)	0.84 (0.14)	0.85 (0.10)
User specific	**Single**	0.11 (0.09)	0.08 (0.09)	0.90 (0.08)	0.92 (0.09)	0.91 (0.06)
	Merge	0.04 (0.08)	0.05 (0.08)	0.97 (0.06)	0.95 (0.08)	0.95 (0.05)

The results are shown in the second and third rows ("Day 2" and "Day 3") in each block in Table 3. We find that the performance of the classifier degrades slightly compared to the first day, inter-session analysis. Also, we still found that merging two instances provides better results than using a single instance. The best results are again acquired by using the user-specific model and merging 2 instances. For the second day, False Positive Rate = 0.05 and False Negative Rate = 0.04 and for the third day False Positive Rate = 0.04 and False Negative Rate = 0.03.

Lab-based Study Analysis: Our lab experiment involved 20 participants who were asked to perform the study in controlled settings. All of the participants were asked to solve 60 challenges using the same PC and same setting with minimal distraction. The results of the lab based study are summarized in Table 4. The results indicate that merging two challenges and using the user specific model can identify the user with high accuracy (0.05 False Negative Rate) and reject the zero effort attackers with high accuracy (0.04 False Positive Rate). The results are in line with the results acquired from the AMT study, which show that the performance of the classifier was related to the ability of the classifier to distinguish users' unique way of solving the challenges rather than the platform and the settings they used while solving the challenges.

5.5 Summary of Results

The results obtained from the classification models show that *Gametrics* is a viable form of behavioral biometrics. The results show that the classifier can identify the users and reject a zero effort attacker with a high overall accuracy, especially when user-specific models are employed and two game instances are merged together.

6. IMPERSONATION ATTACKS

In Section 5.4, we demonstrated that *Gametrics* is robust against zero-effort attacks, reflected in the low False Positive Rate. In this section, we analyze the security of *Gametrics* against deliberate impersonation attacks

We first considered shoulder-surfing impersonation attacks. During the lab-based study's data collection, a researcher in our group served the role of an attacker, and monitored, through video recording, the participants while they were solving the challenges. For the impersonation attack analysis, the attacker picked one of the participants who had the most similar features, such as the time duration and mouse movement speed, as that of the attacker, and tried to mimic that participant by solving the challenges in a similar way as the participant did for 60 times. Making a selection in this fashion is representative of a powerful scenario where the attacker targets victims who are easier to attack. If we can show that our *Gametrics* system can be resistant to such a powerful attacker, it may be even more resistant to other weaker, more realistic attackers who may not have the capability to make such selections.

Table 5: Shoulder-Surfing Impersonation Attack Results

		FPR
All features	**Single**	0.15
	Merge	0.07
Average overall best	**Single**	0.20
	Merge	0.10
User specific	**Single**	0.31
	Merge	0.03

The performance of this attack is enumerated in Table 5. For the user-specific model, the attack success rate came out to be 0.31 when single instance of the challenges was used by the classifier, and decreased drastically to 0.03 when merging of two instances is used by the classifier. There are two main reasons for the increase in security when merging two instances. First, the features that were used for the classification in the single instance model (i.e., the features subset that yielded the highest classification accuracy in the benign and zero-effort case) all related to the mouse movement characteristics, namely, the features used were the drag speed, the move and the drag acceleration and the drag silence. However, in the merged instance model, more features were used by the classifier that relate to both of the cognitive as well as the mouse movement characteristics of the user, which made mimicking the victim much harder. Second, the classifier performs better as using two challenges involve more interaction between the user and the challenges, and make the mimicking task much harder for the attacker. In all the other classification models, we found that the security provided by merging two challenges was also much higher than its corespondent in using a single challenge. This suggests that our *Gametrics* system can defeat powerful shoulder-surfing attacks with a high probability when two game instances are merged and when user-specific model is used.

In practice, it is possible that the attacker resorts to an automated strategy, for example, the use of robots, rather than manual shoulder-surfing (which may be a tedious attack anyway). A robotic attack to compromise behavioral authentication schemes, specifically touchscreen dynamics, has been proposed in [29]. Such robots can be built to mimic the user's way of interacting with the authentication construct based on the leaked authentication template. These attacks have been shown to be able to significantly decrease the performance of touch-based authentication systems. In contrast to tradition behavioral biometrics where the authentication construct is static (i.e., PIN or pattern unlock), *Gametrics* involves randomization in the object movements as well as solving a game-based CAPTCHA (DCG) [18]. Thereby, to build a robot that is able to mimic the user's interaction with the games, the robot is required to not only repeat a previously recorded interaction between the user and the authentication construct, but also to understand the underlying challenge as fast as a human user and then try to mimic the user's interaction with the challenge. Although it is shown in [18] that DCG CAPTCHA can be attacked using a *dictionary-based attack*, if the server incorporates a large database of the challenges and display the challenges randomly to the user, this task would become hard for the bot as the dictionary search and the matching between each of the moving objects and the stored answer objects in the database would significantly slow down this process. Furthermore, matching each of the answer objects with the answer objects stored in the dictionary requires some amount of time, for instance in [18] the authors proposed to click on the object to hold it while performing the object matching. This would make it hard to mimic the user's "pause and drag" feature. Based on this analysis, we therefore conclude that even automated shoulder-surfing attacks against *Gametrics* will not be effective.

The authors of [19] showed that most of the currently proposed behavrioral biometrics schemes (including keystroke and touchscreen dynamics) are vulnerable to internal, malware-based attacks. Malware installed on the device (authentication terminal/phone) can record the user's valid authentication template and replay it later to authenticate itself as the user (e.g, replay a "pattern unlock" biometrics [7]), or learn from multiple interactions between the user and the device, and then reproduce the new data that has similar features to the user's valid interactions with the device in order to fool the authentication system (e.g., learn the user's typing pattern and then enter another text mimicking the user's typing style). In contrast to other behavioral biometrics schemes, the multi-round randomization embedded in the *Gametrics* challenges as well as the requirement of solving the underlying game-based CAPTCHA will make *Gametrics* robust against such attacks. That is, even having access to the authentication template or a prior authentication session data will not be sufficient for the attacker to impersonate the user in the *Gametrics* system.

To sum up, *Gametrics* promises to address many of the attacks that are known to be a significant concern for traditional password-based authentication systems as well as existing behavioral biometrics systems, including:

- *User-side attacks*, where the attacker observes the victim as she logs in, through manual or automated mechanisms, to learn the user's input (password in password system) or learn the way the user provides the input (biometrics data from the current session in behavioral biometrics systems). The attacker then attempts to replay the information in an authentication session at a later point of time.

- *Server-side attacks*, where the attacker hacks into the web server databases to learn the stored authentication token (e.g., hash of passwords in password systems and biometrics template in behavioral biometrics systems). The attacker then uses this information to run an offline dictionary attack against passwords, or reproduce the biometric data that matches with the template.

- *Client-side attacks*, where the attacker hacks into the authentication terminal using which the user is logging in and learn the user's input. The attacker then attempts to replay the information in an authentication session at a later point of time.

7. DISCUSSION AND FUTURE WORK

Efficiency: The proposed *Gametrics* system can fit well for many applications noting the short time the user took to solve the challenges (around 7.5 seconds for a single challenge and 15 seconds for two challenges). Moreover, the enrollment phase consisted of 40 challenges (around 5 minutes on average) and provided a reasonably high identification accuracy. In short, building the classifier model, updating the model with the new data over time (e.g., as the user logs in by playing new game instances) and testing a new instance, all take a short amount of time.

User Experience: The *Gametrics* system also seems to offer high usability, as the average SUS score came to be 86.11 (standard deviation = 14.12) in the lab-study and 73.95 (standard deviation = 17.14) for the web study. SUS scores above 70 are indicative of good overall usability. The score for the playfulness of the challenges came to be 3.36 (standard deviation = 1.40) and the easiness of the challenges was 4.58 (standard deviation = 0.77). This suggests that the participants found the game challenges to be very easy (although not necessarily playful). These results overall bode well for the user experience of *Gametrics*.

Application Scenarios: *Gametrics* can be utilized as a point-of-entry mechanisms, such as to authenticate the user to a web server.

Graphical passwords were founded on a psychological principle that the human brain has superior memory for processing visual rather than textual information (see two excellent surveys [3,30]). They can be based on recognition, such as those involving Random Arts images [23], objects (PassObjects) [34] and faces (PassFaces) [22], as well as on recall or cued recall, such as those involving drawings [8, 14] and selection of points on an image (PassPoints) [33]. *Gametrics* can be integrated with graphical passwords as a second factor authentication, which would enhance the security of graphical passwords against shoulder surfing and spoofing attacks. Further work is needed to realize such two-factor designs.

Gametrics can be also used as a fall-back authentication mechanism. In such use case, multiple instances of the challenges can be presented to the user, since fall-back does not happen frequently. However, in order to build an up-to-date classification model for the user, the system may need to ask

the user to solve challenges periodically. Future investigation is necessary to study *Gametrics* in the context of such fall-back authentication applications.

Given the popularity of touchscreen games, *Gametrics* would fit well on touchscreen devices. Here, *Gametrics* can utilize the various sensors, such as accelerometer and gyroscope, available on these devices to measure the users' implicit interactions with these devices, which when combined with other explicit touchscreen interaction features may enhance the overall classification accuracy and resistance to attacks. In our future work, we will study the effectiveness of *Gametrics* for authenticating the users on such devices.

Recently, Google has announced a plan to eliminate passwords by introducing a Trust API that uses a fusion of multiple biometrics indicators to verify the user's identity, such as face recognition and voice patterns, and other behavioral biometrics such as the gait biometrics [13]. In the future, we hope that *Gametrics* can be added to the Trust API by asking the user to play a game challenge on a periodic basis.

Limitations: *Gametrics* is similar to any other behavioral biometrics in that, we believe, it will suffer from a degradation in the accuracy of user identification when the user's behavior is undergoing a significant variation, such as changing emotions [9], falling sick or getting injured. The effect of behavioral changes on the performance of *Gametrics* should be subject to future work.

Future work may also need to be conducted to test the accuracy of the *Gametrics* classification models when switching devices or hardware (e.g., different kinds of mouse or screens).

The results obtained from our study are promising, however, we believe that further work is needed in order to improve the overall accuracy of user identification. One avenue in this direction is to add a little more complexity to the game challenges in order to increase the level of interaction between the challenge and the user, and thereby improving the overall usability (False Negative Rate) and security (False Positive Rate) of the interactive authentication.

8. RELATED WORK

The main aim of behavioral biometrics is to solve the problems associated with the traditional authentication systems, such as password leakage or sharing, requirements for extra hardware in case of traditional biometrics (for example, fingerprint readers or iris scanners). However, most of the proposed behavioral biometrics suffers from various problems, such as low-level of uniqueness among the users, which yields to high acceptance rate of illegitimate users (high False Positive Rate). Moreover, some of the behavioral biometrics require long interaction time to identify/recognize the users (up to 20 minutes [5]), which would allow attackers to interact with the system and may cause harm to the system during that period of time without getting detected.

The most studied approaches for behavioral biometrics are keystroke analysis and mouse dynamics. Keystroke biometrics identifies the user based on her typing characteristics. The verification is performed either based on static text (i.e., password) or random text (i.e., free text to continuously authenticate the users [17]). The features that are mostly used are the timing information of key down/hold/up events, time between the release of a key and the pressing on the next key, overall typing speed, and frequency of errors [36]. Mouse dynamics [35] is another most studied behavioral biometrics. It is mostly used for continuous authentication by recording the user interaction with the device transparently. The user is authenticated based on the general movement, drag and drop, stillness, point and click.

Other recent research studied user authentication based on user's cognitive abilities [1]. In this work, the authors studied the ability of authenticating the users based on their cognitive process captured by visual search, working memory and priming effect on automatic processing. The game they utilized to capture the users' cognitive abilities provides a challenge-response task. In each instance of the challenge-response, the user is given a challenge, which is an object. The user's task is to drag the challenge object onto the matching object inside the search set. After a valid drop, the user then receives a gold coin as a reward and deposits it in a bank. On a correct deposit, the user is challenged with a new object and the game continues as before. From the interaction with the challenges, the authors extracted several features that captures the cognitive abilities of the users, however, they did not look into the mouse dynamics biometrics of the users. Although the proposed method can authenticate the user with high accuracy 0-7.8% FNR and 0-2.3% FPR, the verification time and the enrollment time is much longer than our *Gametrics* system. They require 76.7 seconds – 2.5 minutes for verification and 9.8 min – 24.3 min for enrollment.

The authors in [5] proposed a method to solve account hijacking and share problems in an online gaming environment. They propose identifying the user based on her gameplay activities. They show that the idle time distribution is a representative feature of game players. They propose the relative entropy test RET scheme, which is based on the Kullback-Leibler divergence between idle time (i.e., the idle periods between successive moves of a player controlled character) distributions, for user identification. Their evaluations shows that the RET scheme achieves higher than 90% accuracy with a 20-minute detection time given a 200-minute history size. Their detection time is much higher than that in our *Gametrics* system.

9. CONCLUSION

In this paper, we introduced *Gametrics*, an interactive biometrics system based on the gameplay pattern of the users embedded in very simple game constructs. We showed that incorporating the mouse dynamics with the cognitive mechanisms can identify the users with high accuracy within a short period of time. Moreover, *Gametrics* provides security against multiple forms of vulnerabilities ranging from zero-effort attacks to expert attacks who try to mimic the user, even those who hack authentication templates and employ automated mechanisms such as robots and malware. The time taken for enrollment and authentication are both reasonably short. The system seems to provide good user experience as reflected in the participants' responses to the survey.

Acknowledgments

The work has been partially supported by an award from COMCAST.

10. REFERENCES

[1] A. Al Galib and R. Safavi-Naini. User authentication using human cognitive abilities. In *Financial Cryptography and Data Security*, pages 254–271. Springer, 2015.

[2] F. Bergadano, D. Gunetti, and C. Picardi. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):367–397, 2002.

[3] R. Biddle, S. Chiasson, and P. V. Oorschot. Graphical passwords: Learning from the first generation. In *Technical Report TR-09-09, School of Computer Science, Carleton University*, 2009.

[4] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[5] K.-T. Chen and L.-W. Hong. User identification based on game-play activity patterns. In *Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games*, pages 7–12. ACM, 2007.

[6] Chen, Kuan-Ta and Hong, Li-Wen. User Identification based on Game-Play Activity Patterns. In *Workshop on Network and Systems Support for Games*, 2007.

[7] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. Touch me once and i know it's you!: Implicit authentication based on touch screen patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 987–996, 2012.

[8] P. Dunphy and J. Yan. Do background images improve "draw a secret" graphical passwords? In *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*, pages 36–47. ACM, 2007.

[9] C. Epp, M. Lippold, and R. L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 715–724. ACM, 2011.

[10] Erick Schonfeld. Turiya Media: Data Mining Social Games To Find The Most Valuable Players. In *Tech Crunch*, Available at: http://techcrunch.com/2010/04/06/turiya-media-games, 2010.

[11] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *Information Forensics and Security, IEEE Transactions on*, 8(1):136–148, 2013.

[12] Gabriel Goldwasser. Collecting Data (and Strangers) Online. In *The Faster Times*, Available at: http://thefastertimes.com/videogames/2010/02/21/collecting-data-and-strangers-online, 2010.

[13] A. Hern. Google aims to kill passwords by the end of this year. https://www.theguardian.com/technology/2016/may/24/google-passwords-android.

[14] I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A. D. Rubin. The design and analysis of graphical passwords. In *SSYM'99: Proceedings of the 8th conference on USENIX Security Symposium*, 1999.

[15] C.-C. Lin, H. Li, X.-y. Zhou, and X. Wang. Screenmilker: How to milk your android screen for secrets. In *NDSS*, 2014.

[16] R. A. Maxion and K. S. Killourhy. Keystroke biometrics with number-pad input. In *Dependable Systems and Networks (DSN), 2010 IEEE/IFIP International Conference on*, pages 201–210. IEEE, 2010.

[17] A. Messerman, T. Mustafic, S. A. Camtepe, and S. Albayrak. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8. IEEE, 2011.

[18] M. Mohamed, N. Sachdeva, M. Georgescu, S. Gao, N. Saxena, C. Zhang, P. Kumaraguru, P. C. van Oorschot, and W.-B. Chen. A three-way investigation of a game-captcha: automated attacks, relay attacks and usability. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, pages 195–206. ACM, 2014.

[19] M. Mohamed, B. Shrestha, and N. Saxena. Smashed: Sniffing and manipulating android sensor data. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 152–159. ACM, 2016.

[20] F. Monrose and A. Rubin. Authentication via keystroke dynamics. In *Proceedings of the 4th ACM conference on Computer and communications security*, pages 48–56. ACM, 1997.

[21] E. Owusu, J. Han, S. Das, A. Perrig, and J. Zhang. Accessory: password inference using accelerometers on smartphones. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, page 9. ACM, 2012.

[22] T. S. B. Passfaces. http://www.realuser.com/. Last access, December 2008.

[23] A. Perrig and D. Song. Hash visualization: a new technique to improve real-world security. In *CrypTEC*, 1999.

[24] Pusara, Maja and Brodley, Carla E. User Re-Authentication via Mouse Movements. In *Workshop on Visualization and Data Mining for Computer Security*, 2004.

[25] A. Rabkin. Personal knowledge questions for fallback authentication: security questions in the era of facebook. In *SOUPS '08: Proceedings of the 4th symposium on Usable privacy and security*, 2008.

[26] Ryan Kaminsky, Miro Enev, and Erik Andersen. Identifying Game Players with Mouse Biometrics. Available at: http://abstract.cs.washington.edu/~miro/docs/mouse_ID.pdf, 2008.

[27] S. E. Schechter, A. J. B. Brush, and S. Egelman. It's no secret. measuring the security and reliability of authentication via "secret" questions. In *IEEE Symposium on Security and Privacy*, pages 375–390, 2009.

[28] S. E. Schechter and R. W. Reeder. 1 + 1 = you: measuring the comprehensibility of metaphors for configuring backup authentication. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS)*, 2009).

[29] A. Serwadda and V. V. Phoha. When kids' toys breach mobile phone security. In *Proceedings of the 2013 ACM SIGSAC conference on Computer &*

communications security, pages 599–610. ACM, 2013.

[30] X. Suo, Y. Zhu, and G. S. Owen. Graphical passwords: A survey. In *ACSAC*, 2005.

[31] C. M. Tey, P. Gupta, and D. Gao. I can be you: Questioning the use of keystroke dynamics as biometrics. The 20th Annual Network & Distributed System Security Symposium (NDSS 2013), 2013.

[32] Valve Corporation. Steam: Game and Player Statistics. Available at: http://store.steampowered.com/stats, 2010.

[33] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. D. Memon. PassPoints: Design and Longitudinal Evaluation of a Graphical Password System. In *International Journal of Human Computer Studies*, 2005.

[34] S. Wiedenbeck, J. Waters, L. Sobrado, and J.-C. Birget. Design and Evaluation of a Shoulder-surfing Resistant Graphical Password Scheme. In *Proceedings of the working conference on Advanced visual interfaces (AVI)*, 2006.

[35] N. Zheng, A. Paloski, and H. Wang. An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 139–150. ACM, 2011.

[36] Y. Zhong, Y. Deng, and A. K. Jain. Keystroke dynamics for user authentication. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 117–123. IEEE, 2012.

APPENDIX

Table 6: Participants demographics

	AMT	**Lab**
# Participants	**98**	**20**
Age (%)		
<18	1.0	0
18-24	20.4	40
25-34	38.8	45
35-50	32.7	15
>50	7.1	0
Gender (%)		
Female	41.8	35
Male	58.2	65
Education (%)		
High School	26.5	25
Bachelor	58.2	40
Masters	14.3	30
PhD	1.0	5
Background (%)		
Computer Science	18.4	80
Engineering	7.1	15
Medicine	6.1	0
Law	6.1	0
Social Sciences	3.1	0
Journalism	3.1	0
Finance	4.1	0
Business	20.4	0
Other	31.6	5