Faces in the Distorting Mirror: Revisiting Photo-based Social Authentication

Iasonas Polakis
Dept. of Computer Science
Columbia University
New York, NY, USA
polakis@cs.columbia.edu

Panagiotis Ilia
Institute of Computer Science
FORTH
Heraklion, Greece
pilia@ics.forth.gr

Federico Maggi
DEIB
Politecnico di Milano
Milan, Italy
federico.maggi@polimi.it

Marco Lancini
CEFRIEL
Milan, Italy
marco.lancini@cefriel.com

Georgios Kontaxis
Dept. of Computer Science
Columbia University
New York, NY, USA
kontaxis@cs.columbia.edu

Stefano Zanero
DEIB
Politecnico di Milano
Milan, Italy
zanero@elet.polimi.it

Sotiris Ioannidis
Institute of Computer Science
FORTH
Heraklion, Greece
sotiris@ics.forth.gr

Angelos D. Keromytis
Dept. of Computer Science
Columbia University
New York, NY, USA
angelos@cs.columbia.edu

Abstract

In an effort to hinder attackers from compromising user accounts, Facebook launched a form of two-factor authentication called *social authentication* (SA), where users are required to identify photos of their friends to complete a log-in attempt. Recent research, however, demonstrated that attackers can bypass the mechanism by employing face recognition software. Here we demonstrate an alternative attack that employs image comparison techniques to identify the SA photos within an offline collection of the users' photos.

In this paper, we revisit the concept of SA and design a system with a novel photo selection and transformation process, which generates challenges that are robust against these attacks. The intuition behind our photo selection is to use photos that fail software-based face recognition, while remaining recognizable to humans who are familiar with the depicted people. The photo transformation process creates challenges in the form of photo collages, where faces are transformed so as to render image matching techniques ineffective. We experimentally confirm the robustness of our approach against three template matching algorithms that solve 0.4% of the challenges, while requiring four orders of magnitude more processing effort. Furthermore, when the transformations are applied, face detection software fails to detect even a single face. Our user studies confirm that users are able to identify their friends in over 99% of the photos with faces unrecognizable by software, and can solve over 94% of the challenges with transformed photos.

CCS'14, November 3–7, 2014, Scottsdale, Arizona, USA.

ACM 978-1-4503-2957-6/14/11 ...$15.00.
http://dx.doi.org/10.1145/2660267.2660317.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection—*Authentication*

General Terms

Security, Human Factors

Keywords

Social Authentication; Image Analysis; Face Recognition; CAPTCHAs

1. INTRODUCTION

The abundance of personal information uploaded to online social networks (OSN), coupled with the inherent trust that users place into communications received from their contacts, has rendered compromised profiles a lucrative resource for criminals [23]. Moreover, the widespread adoption of single-sign on services offered by popular OSNs, makes user profiles even more valuable. Consequently, researchers designed various systems for detecting compromised accounts [10, 13]. However, preventing unauthorized access in a user-friendly manner remains an open problem. To safeguard profiles against attackers that have stolen user credentials Facebook deployed a countermeasure called social authentication (SA) [21]. This is basically a variant of the traditional two-factor authentication scheme (e.g., [3, 5]), which requires users to identify their contacts in a series of photos.

Social authentication is a promising approach, as it offers a user-friendly mechanism to strengthen the login process. Researchers, however, have analyzed [16] its weaknesses, and demonstrated [20] that the existing system is vulnerable to attacks that employ face recognition software. We further

demonstrate that SA is vulnerable to an attack that previous work has overlooked; the adversary first builds a collection of the photos uploaded by the victim and his online friends. The adversary can then solve the challenges by identifying the photos within the collection via image comparison, and using the tag information to select the correct answer. Compared to the previous attack, this attack has an important advantage: the identification of photos within a collection based on image comparison techniques is far more accurate than face recognition, and effective even when no faces are present.

In this paper, we revisit the concept of SA and build a system that retains the usability of the existing mechanism, while being robust to attacks employing image analysis software. We conduct a user study that provides us with valuable information regarding a critical aspect of our approach to SA; the ability of users to identify their friends in photos taken under realistic, non-ideal conditions. The participants' photos are processed by state of the art face recognition software and categorized as "simple" , "medium" or "difficult", based on the quality of the faces found (if any). While SA picks *simple* photos, we focus on the *medium* and *difficult* categories. Users solve over 99% of the *medium* and 82% of the *difficult* challenges, indicating their ability to identify their friends even when their faces are not clearly visible, based on secondary features (e.g., posture, hair), associative information (e.g., pets, objects) or memory retrieval (users remember having seen the photos). On the other hand, face recognition software fails to identify the users in such photos.

Analysis of the results of the user study and the characteristics of the two attacks provides significant insight that allows us to design a secure, yet usable, system that renders the attacks ineffective. First, we redesign the photo selection procedure, which processes photos with face recognition software, and selects those that contain faces but are not recognizable by face recognition software (i.e., *medium* photos). Next, we apply a novel transformation process for creating the SA challenges, which hinders image comparison techniques from mapping them to the original photos.

Our prototype implementation creates SA challenges by superimposing the selected *medium* faces (tags) over the faces of a random "background" photograph. The overlayed faces are made transparent so as to blend in with the underlying faces. Then, a perspective transformation is performed on the photo, which prohibits even highly resilient pattern matching approaches, like template matching, from mapping them to the original photos.

Subsequently, we conduct an extensive set of experiments using real data, to measure how various levels and combinations of the transformations impact the attacks. The results demonstrate the robustness of our challenges, as our system completely hinders the face recognition attack that fails to detect any faces. The challenges are even robust against the image comparison attack that employs three template matching algorithms; *all three pass less than 2% of the challenges with two tags, and 0.4% of those with three tags, while requiring four orders of magnitude more processing effort than against the non-processed photos.*

To verify that people depicted in the photos remain identifiable after the transformations, we conduct a preliminary user study where users are asked to identify famous people in a series of challenges. Results verify the usability of our system, with users solving *94.38%* of the challenges.

Finally, we discuss the applicability of our approach as a security service offered by an OSN to other websites. In our vision, this mechanism can be adopted by web services as a user-specific CAPTCHA service, or even by high value services (e.g., banking websites) as a security measure additional to two-factor authentication. We discuss its robustness against attacks that break traditional CAPTCHAs, like outsourcing attacks, and argue that it is a user-friendly and secure alternative to existing schemes.

In summary, the key contributions of this work are:

- We demonstrate a novel attack technique against social authentication, that is more effective and efficient than the one previously presented.
- We conduct the first user study that explores the ability of humans to identify their acquaintances in photos taken under realistic, non-ideal conditions. Our results demonstrate that humans solve this task effectively, even when no faces are in the photos.
- Based on the insights derived from our experiments, we design a secure, yet usable SA mechanism, that relies on a novel tag selection and transformation process. We experimentally evaluate our proof-of-concept implementation which completely hinders the face recognition, and reduces the success of the image comparison attack to 0.4%, while requiring four orders of magnitude more processing effort from the attacker.
- To verify the usability of our system, we conduct a preliminary user study where users solve 94.38% of the challenges with transformed photos.

2. ATTACKING SOCIAL AUTHENTICATION

Photo-based authentication in OSNs was first presented in 2008 by Yardi et al. [27]. In 2010 Facebook deployed its SA application in an effort to prevent adversaries from using stolen credentials. In a nutshell, when a login attempt is considered suspicious, the system presents the user with a series of 7 pages, each containing 3 photos of a friend and 6 potential answers. The user is required to correctly identify the friends depicted in at least 5 of the pages.

Face Recognition Attack. In previous work we demonstrated that practically anybody can solve SA challenges, by collecting publicly available data and employing off-the-shelf face recognition software [20]. The photos and their tags are used to train face recognition classifiers, which can identify the friends depicted in SA challenges. Our estimations showed that 84% of Facebook users are susceptible to this attack.

Image Comparison Attack. Attacking SA needn't rely on face recognition, as more effective photo matching techniques can be used instead. The attacker first creates a collection with all the victim's friends' photos he can access, along with the tag information. When SA is triggered, the attacker identifies the presented photos within the collection, and uses the tag information to answer the challenge.

The advantage of this attack is its effectiveness even when the challenges contain faces that cannot be identified via face recognition. Regardless of the content the adversary can pass the challenge if some of the photos are in the collection (at least 1 of the 3 photos, in 5 of the 7 pages). The attack success is proportional to the coverage of photos. This can

Collection size	5K	10K	20K	30K	40K
Identified photos	98.8%	98.4%	98.4%	98.4%	98.4%

Table 1: Identified photos in image comparison attack, for different collection sizes.

(a) Simple (b) Medium (c) Difficult

Figure 1: Sample photo from each category.

be increased by employing fake accounts to befriend the victim's friends. However, publicly available photos are a good resource for the attack. According to [18], the authors found that 63% of the photos have privacy settings different from what the users intended. Alarmingly, the privacy setting for 51% of those photos was set to public, allowing anyone to access them. Thus, overall, about 1 out of every 3 photos will be publicly viewable by accident. Nonetheless, previous work (e.g., [6, 7, 14, 25]) has extensively demonstrated the effectiveness of employing fake accounts to befriend users of OSNs, and have reported success rates of up to 90%.

Various image comparison techniques can be used for this attack. Here we demonstrate its effectiveness even when employing a simplistic pixel comparison. We assume the attacker has access to all of the user's photos (we retain the same assumption when defending against this attack in Section 4.3). Experiments were conducted on a 4-core Intel® Core™ i7-4770 CPU @ 3.40GHz equipped with an SSD.

We build collections of varying sizes, and create 100 SA challenges from each collection. The collections are up to 40,000 photos, which is higher than the average from our user study (Section 3.2). We employ a simple and fast image comparison approach to identify the presented photos within our collection: we crop the top left corners of the presented photos, and match them to those of the photos in the collection. In certain cases this might result in false positives (e.g., the top left corner of the photo has only black pixels), however it did not affect our success rate. Table 1 presents the results of our experiments, indicating the average ratio of photos identified correctly in each SA challenge (21 photos). We passed all the challenges with at least 18 of the photos identified, and had identified at least 98.4% of the photos in all scenarios.

Our approach is very efficient as we are able to identify the 21 photos within 40K photos in 1.24 seconds ($\sim$ 0.06 per photo). One could possibly improve performance by intersecting the suggested names and the tags after each photo identified within a page.This could decrease times, as the tags from one photo might be enough to infer the answer.

3. MEASURING USER ABILITIES

To design a secure SA system that exploits noisy and unidentifiable faces, we need to verify our intuition that humans are capable of recognizing their friends in photos taken under natural conditions. Although previous work [12, 22] has explored the ability of people to discern *human faces* or their features, we are the first to focus on the ability of *recognizing friends* (as opposed to unknown faces), even under conditions where the faces may not be clear or even present at all.

Measurement Application. We created a Facebook app that replicates the SA mechanism, which require users to identify their friends in SA challenges, and complete a questionnaire for each photo. We opted for a Facebook app for two reasons: first, they inspire trust in users as they are deployed within a sandbox and are governed by a series of permissions that clearly state the user data accessed. Second, a Facebook app enables direct access to some user profile data (e.g., their social circle). This enables us to respect user privacy and minimize collection of potentially sensitive information, since we use data stored on Facebook rather than having users upload it to our own infrastructure.

IRB Approval. We issued an IRB protocol request to the review board of our institution, that clearly described the parameters of our study, and the data we would be gathering. After it was approved we invited users to participate.

Recruiting Users. We explored and experimented with the possibility of reaching human subjects through the Amazon Mechanical Turk (AMT) service [1]. However, asking Turks to install an app, or directing them to a URL outside Amazon to fill out a survey, explicitly violates the AMT terms of services. Our tasks were rejected by Amazon because of this purely technical incompatibility. The nature of our system, where challenges are crafted specifically for each user, prohibited us from taking advantage of such crowdsourcing platforms. Therefore we resorted to recruiting users directly by sharing our app with the online contacts of our personal Facebook accounts, and posting flyers around the university campus. We also offered prizes as an incentive for user participation. This allowed us to collect and analyze a significant amount of user data, regarding over 4 million photos, and over 1,000 solved SA challenges.

3.1 Measurement Workflow

Once a user installs our app, it collects the social graph and related metadata (i.e., friends, URLs of photos, tags). It processes the collected photos with state of the art face recognition software, and categorizes them as *simple* , *medium* or *difficult* , based on the quality of the faces found. Photos of each category are selected to create challenges of increasing difficulty and measure the user's ability to solve them.

Step 1: Face Extraction. We use the `face.com` online service, which has since been acquired by Facebook [4]. Its effectiveness when using photos taken under realistic conditions has been demonstrated [20], and it performs better than other state of the art solutions [24]. We focus on two specific metrics assigned to the detected faces.

Confidence: when detecting faces, `face.com`'s classifier returns its level of confidence that the tagged area contains a face. Tags assigned a low confidence level have a high probability of not containing a face.

Recognizable: not all faces are suitable candidates for training (or being recognized by) a classifier: `face.com` returns a boolean value to indicate this; "true" when faces can be recognized or are suitable to be used as part of a training set, and "false" otherwise. Even if a face is present, due to various rea-

Type	Total	Passed	Success	Per User
Simple	362	358	98.89%	3.98
Medium	347	344	99.14%	3.81
Difficult	335	275	82.09%	3.68
Total	1044	977	93.58%	11.47

Table 2: Number of challenges taken from each category, and percentage of successfully passed ones.

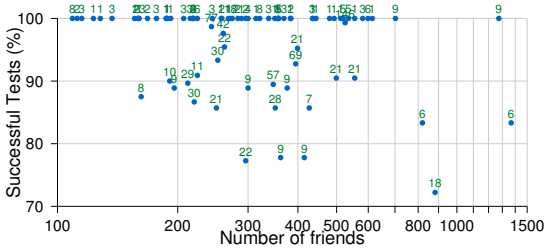


Figure 2: Correlation between number of friends and challenges solved. Each point's label indicates how many challenges the user has taken.

sons (e.g., angle, obstacles) proper face-classification features (e.g., haar, eigenfaces, fisherfaces) cannot be extracted.

Step 2: Photo Categorization. Based on these metrics, our app assigns photos to the following categories:

Simple - Figure 1(a): photos containing tags that most likely frame a human face. This is our baseline category, as it replicates the existing SA mechanism, and provides a reference for comparison. According to [20], 80% of the photos presented in SA challenges by Facebook had a face in the tagged area that was detectable by software. Therefore, we select photos with high confidence ($\geq$80%) which have been classified as recognizable (recognizable=`T`).

Medium - Figure 1(b): photos with a high probability of containing a face (confidence$\geq$80%), which are classified as bad candidates for training/recognition (recognizable=`F`).

Difficult - Figure 1(c): photos classified with a confidence below 40%. This category is to measure how effective people are at recognizing their friends even if their face is not visible.

Step 3: Photo Description. After a user selects the name of each depicted friend, our app informs them if they were right or wrong, and requires them to answer 4 questions per photo describing: the content, the position and visibility of the user's face and other faces within the tagged area, and reasons why the photo was useful or not (see Appendix A).

3.2 User Study Results

Our goal is to measure the users' ability to recognize their friends, and demonstrate that humans can solve this task in conditions where the automated attacks would fail, as we demonstrate in Section 4.3.

Collected dataset and demographics. 141 users installed our app which led us to a total of 4,457,829 photos and 5,087,034 tags. However, 90 of the users actually completed challenges, out of which 79 were listed as male and 11 as female, from 6 different countries. Of the 82 that reported their age, 63 were between 20 and 30 years old and 15 were between 30 and 40. On average, users had 347 friends each.

Recognizing Friends. Table 2 presents the number of challenges (each has 7 pages with 3 photos of the same user and 6 suggested names) per category, and the percentage that users were able to solve (recognize at least 5 out of 7). Results are surprisingly high and consistent for *medium* challenges, as users solve over 99% of the challenges.Thus, *even for photos with faces that cannot be identified by state of the art face recognition software, users' success rates are not impacted.* Users also score surprisingly well for the *difficult* challenges, with an 82% success rate.

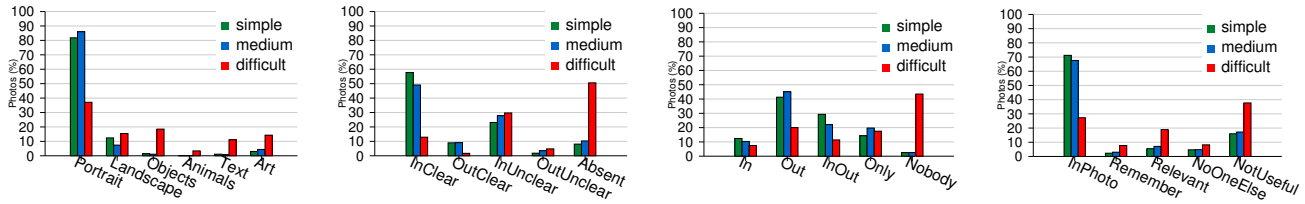
Influence of the Social Circle Size. Figure 2 shows the number of friends that a user has and the success rate for solving SA challenges. Each point refers to the overall success rate of a user for all 3 categories, and the label indicates the total number of challenges that the user completed. As the number of friends increase, we expect users to score significantly lower. However, the results do not demonstrate such an effect, and no strong correlation is evident. The suggestion of names is important, as users can decide based on content that can be associated to certain friends. This result is very encouraging, as it disproves concerns regarding the applicability of SA for users with a large number of friends. Here we only visualize users that have completed at least 3 challenges for reasons of visual clarity, without the overall distribution being affected in any way.

Photo Content. Figure 3(a) shows the distribution of answers regarding the content of the photos, which also offers an evaluation of the quality of the photo-detection process. As expected, for the *simple* and *medium* categories, the vast majority of photos (over 80%) are labeled as portraits, meaning that the focus of the photos are human faces. In contrast, in the *difficult* one they account for 37%. These numbers verify how accurate the face detection process of `face.com` is, as the confidence levels we have set in our photo categorization process are verified by users. The *difficult* category is more evenly distributed.

Face Position. Figure 3(b) plots the distribution of answers about the placement of the friend's face with respect to the tagged area. 49% of the *medium* photos contain a clearly visible face inside the tagged area (InClear) and an unclear face in 27.8% (InUnclear), cumulatively approaching the 80% confidence criteria. The *difficult* photos do not contain the friend's face (Absent) in about half the photos.

Presence of Other Faces. Figure 3(c) shows the distribution of other faces being visible in the photo, and their placement in regards to the tagged area. The *simple* and *medium* categories contain some other face in 83% and 77.5% of the photos with faces being outside the tag in 41% and 45% of the cases respectively. For the *difficult* category, 43.5% of the photos contain no human faces (Nobody).

Usefulness of the Photo. Figure 3(d) plots the distribution of photos regarding their usefulness. The selected friend was present in about 70% of the *simple* and *medium* photos, which is less than the photos containing the friend's face according to Figure 3(b). This is due to users selecting other options such as "remembering the photo" or "relevant to this friend", even though the friend's face is in the photo. An interesting aspect of the *difficult* category, where photos have a low probability of containing the face, is users relying on other information to correctly select the friend. This category has a higher percentage of answers that rely on other



(a) Content of photo. (b) Position and clarity of friend's face. (c) Position and clarity of other faces. (d) Usefulness of photo.

Figure 3: Distribution of answers given by the users in our study.

Type	Portrait	Landscape	Objects
Simple	97.4% (1133)	94.9% (59)	0% (1)
Medium	97.6% (1225)	90% (30)	0% (0)
Difficult	92.1% (267)	76.9% (26)	64.5% (31)

Table 3: Success rates (and total number) for pages where all 3 photos were labelled as the same type.

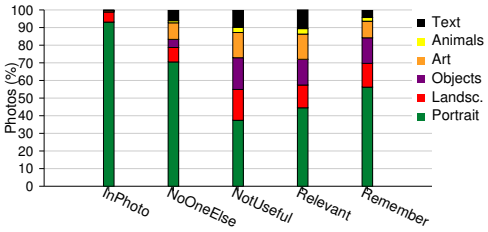


Figure 4: Correlation between content of the photos and usefulness as reported by users.

types of visual clues and context for excluding (NoOneElse) or inferring (Relevant) suggested names.

Absolute Success Rate per Category. Table 3 shows the statistics for pages (each SA challenge contains 7 pages with 3 photos each) in which all 3 photos are assigned to the same category. We present the percentage of pages in which the depicted friend was correctly identified, and the total number of pages in parentheses. For *difficult* portraits, users were able to identify their friends in 92.1% of the pages. *Thus, people can identify their friends in photos where software cannot even detect a face.* In the *medium* category, users were successful in more than 97% of the pages, validating our initial intuition and the applicability of our approach. Even though the number of the remaining *difficult* challenges is too small to draw concrete conclusions, it is surprising that success rates are over 64% in all cases, and users even identified their friends in 77.7% of the pages that contained photos of animals. Thus, associative correlations assist users when presented with a photo of objects or pets.

Absence of Friend's Face. To verify the ability of users to infer the answer even when the user is not present, Figure 4 breaks these numbers down. It becomes evident, as the cumulative ratios for the Landscape, Objects, Text and Art photos account for 44% and 55.5% of Relevant and Remember respectively. Thus, almost half of the photos for which users relied on inference or memory, are not of faces.

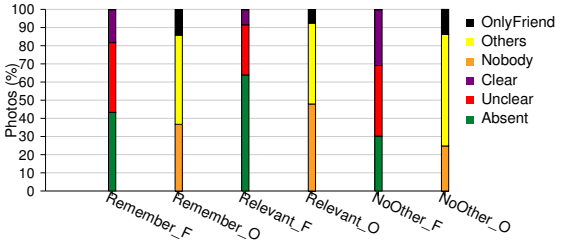


Figure 5: Breakdown of the photos that were useful for inferring or excluding answers, in regards to the friend's face (_F) and other people's faces (_O).

We then focus on photos where the depicted friend's face was absent. When the users remember the photo, they are almost always able to correctly identify the friend (only one case with a wrong answer). Surprisingly, users achieve high success rates when selecting the depicted friend based on the relevance of the content, being 97.4% for the difficult photos. When users try to exclude suggested friends based on the content until they can select the correct answer (NoOneElse), they still achieve very high accuracy (albeit lower than the Relevant ones) with a combined 89.6% across the 3 difficulty categories. As expected, when the photo is not considered useful the success rates are lower. However, in the *simple* and *medium* categories, these photos belong to pages that were correctly answered in more that 82% of the cases, due to the high probability of the other photos containing a face.

Total absence of faces. We wanted to explore whether the existence of other people in the photo might influence the results, i.e., even if the requested friend is not in the photo but other common friends are, the user might still be able to infer the correct answer. When we focus on photos where the friend was absent, and no other faces were contained either, the results remain almost identical. However, the *simple* and *medium* photos that were flagged as NotUseful present a measurable decrease (6.5%, 16.3%). As can be seen in Figure 3(c), these categories have a much higher ratio of photos that contain other faces compared to the *difficult* category. Thus, while photos might not contain any faces, the content can assist the user in inferring the correct answer or excluding suggestions until left with a plausible one.

Exclusion and Inference. To explore the effect of other people's faces on users excluding suggestions or inferring the answer, we plot Figure 5, which provides interesting insight on the strategy users follow for identifying the depicted friend based on the presence of the friend's (columns with _F) and other users' (columns with _O) faces. Users tend to remember

the photo when the depicted friend's face was either Unclear or Absent (Remember_F), as is the case for users inferring the correct answer (Relevant_F). Users also tend to remember photos where other people (common friends) or no people are present at all (landscapes, pets and important objects). Furthermore, in Relevant_O we can see that users infer the correct friend almost equally from photos where Nobody (47.9%) is present (due to relevant objects and pets) or where Other (44.5%) people are present (people that the user knows are friends with the requested friend).

When excluding suggestions (NoOther_F), the absence of the friend's face (Absent) or its poor quality (Unclear) have a similar effect. However, the presence of other people's faces has a major impact, as it accounts for 61.4% of the photos in NoOther_O. This is a strong indication that when users are presented with photos of unknown individuals they tend to follow an approach of excluding suggested friends until left with a plausible answer. If the users know the depicted people are acquaintances of the requested friend, they select Relevant. In the case of unknown individuals, they exclude suggestions of close friends and select less known contacts that have a higher probability of being friends with people unknown to the user. Combined with the information that can be extracted from the other photos contained in a challenge page, this approach can be very effective, as users correctly answered 88.5% of all the pages that contained a photo for which NoOther friend matched.

4. SECURE SOCIAL AUTHENTICATION

Based on the results of our user study we proceed with establishing guidelines for creating a secure SA mechanism that is robust against the attacks we presented.

4.1 Tag Selection

The goal is to filter out faces that can be of use to adversaries that employ face recognition software, and select a subset of the remaining tags that have a high probability of containing a face. We use two criteria for selecting tags after being analyzed by face recognition software: a high *confidence* level for containing a human face, and a false *recognizability* flag, i.e., *medium* tags. While our user study demonstrated that users are highly effective even if the friend's face is not in the photo (i.e., *difficult* photos), we do not use such photos in our SA challenges.

Even though we build our selection process using `face.com`, our tag selection can be completed with any accurate face recognition software. Once all the tags have been analyzed, the system selects the tags that could not be recognised.

The OSN can also use several types of information to narrow down the set from which friends are selected for the SA challenges. This set needn't be small, but in the order of 200-300 friends, that have a minimum level of interaction with the user. All friends from this set must have the same chance of being selected, and all suggestions must be from this set, so as not to aid the attacker by limiting the number of suggestions that seem plausible as answers.

4.2 Tag and Photo Transformations

To defend against the image comparison attack, tagged areas should not be identical to the areas in the original photos, to prevent the attacker from mapping them to the photos in the collection. Our approach blends the faces the user has to identify with the faces on a "background" photo.

Figure 6: An example output photo, with rotation, opacity, and perspective transformations performed. The other faces have been blurred for privacy.

If we simply overlay the tagged areas containing the faces onto a new photo, an adversary could still resort to the image comparison attack. To prevent this, we perform a sequence of transformations on the extracted areas.

Tag transformation: First, we rotate the face, a transformation that can impact face detection. Second, we edit the tag's `alpha level` (a) to make it translucent and blend it with the underlying faces ($0 \leq a \leq 1$, where 0 is fully transparent). Thus, the tag will not contain any parts from the photos in their initial form.

Photo transformation: Each challenge contains one photo of N friends, with M tagged faces for each friend. We select a background photo that contains at least $N * M$ faces, and overlay the existing faces with the processed tags we created in the previous step. We then apply a perspective transformation, which is challenging even for complex feature or template matching techniques. According to Gauglitz et al. [11], "perspective distortion is the hardest challenge". The perspective transformation we perform is variable by P, with P denoting the ratio by which the bottom of the photo is "compressed" from each side; e.g., for $P = 3$, the bottom is compressed from both the left and right by 1/3 of the photo. The user is presented with N menus, each one containing the name of one of the depicted friends, along with the names of $S - 1$ other friends. The user is required to correctly identify the N depicted friends. To demonstrate that familiar faces remain recognizable after our transformations, Figure 6 shows an example output, with $a = 0.6$ and $P = 3.2$, for two well-know politicians[1].

Prototype Implementation. We implemented a prototype, which comprises of a Facebook app for the data collection process, and a back end for the photo processing. We implemented the back-end in Python, using SimpleCV and OpenCV for the face detection and image processing.

To create a SA challenge, the back-end first selects N distinct friends of the target user. For each friend, it finds M random tags of that friend, and fetches the corresponding photos. The tags are extracted, transformed and overlayed on a random background photo, which is then also transformed. The *tag processing* part randomly rotates and applies an alpha opacity filter that ensures that none of the original pixels of the tag are preserved. This is implemented

[1]Challenge solution: Barack Obama, Vladimir Putin.

rotations	CCOEFF	CCORR	SQDIFF	time_photo
None	12.8%	11.0%	10.6%	6.61
7 (30°)	67.8%	36.4%	43.8%	46.23
13 (15°)	91.0%	60.0%	68.8%	87.85
19 (10°)	95.2%	67.4%	77.8%	130.65
37 (5°)	97.8%	75.8%	90.8%	244.68
time_rotation	2.24	2.19	2.18	-

Table 4: Attack success rate of each algorithm for various rotation approaches, time required (sec) for each algorithm to process one rotated version, and the total time for all rotated versions of a photo.

with SimpleCV's `blit()` function, which takes the background photo, tag image, position of the overlay and opacity percentage as input, and returns a collage image. The rotation is implemented with `rotate()`, the perspective transformation is based on the `getPerspectiveTransform()` and `warpaffine()` functions of OpenCV. We set a time-window of one minute for users to solve the challenge.

4.3 Experimental Evaluation

Threat model. We assume the attacker has knowledge of our system, and has created a collection containing all the photos of the victim and his friends. We also assume he can apply the same categorization to photos as we do, and identify *medium* faces. Furthermore, as each tag in the challenge has a suggestion of 6 users, the attacker will only compare the photo to the tags of those users. In our user study, we found that each user's friend has ∼12 *medium* tags on average. Thus, in our experiments, for each tag, the attacker simply has to identify which tag out of a set of 72 (12 for each suggested user) is contained in the photo, to pass the challenge.

Image comparison attack. We employ 3 different *template matching* methods: the normalized versions of the correlation coefficient (`CCOEFF`), cross correlation (`CCORR`) and squared difference (`SQDIFF`) algorithms. To verify their accuracy, we first run the attack against a set of 500 photos, where we have overlayed a tag but have not performed any transformations. All three algorithms correctly identify the 500 tags, requiring ∼6.89 seconds per photo.Compared to the simplistic pixel-comparison attack from Section 2, the template matching algorithms identify every tag but have two orders of magnitude processing overhead. However, the simplistic attack cannot handle the transformations.

First, we measure the impact of rotating the tag. We create a set of 500 photos each containing a *medium* tag that has been randomly rotated within [-90°, 90°]. In the first attack scenario, the attacker does not perform any rotations on the photo and simply selects the tag out of the 72 with the highest similarity score with the photo. In the other scenarios the attacker performs a series of 30°, 15°, 10° and 5° rotations (in the range [-90°, 90°]). For each photo, the attacker selects the best match among all the tags and all the rotations. While the template matching algorithms can handle certain rotations, the results of our experiment shown in Table 4, demonstrate that effectiveness is greatly increased when multiple rotated versions of the photo are processed before selecting the best-matching tag. `CCOEFF` yields the best

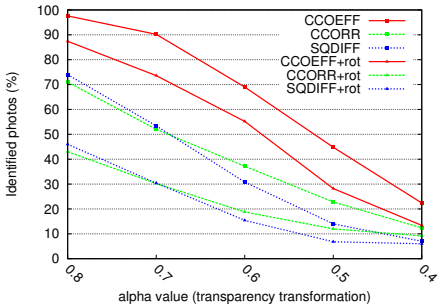


Figure 7: Attack success, against transparency and transparency+rotation transformations.

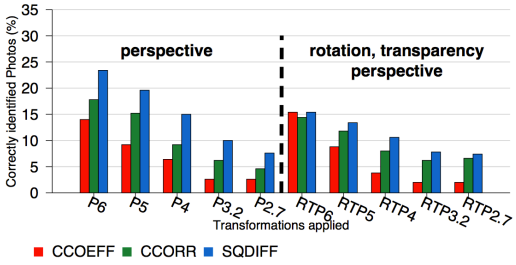


Figure 8: Attack success against perspective (P), and all transformations combined (RTP), for $a = 0.8$.

results, with a success rate of up to 97.8% when the attack performs 5° rotations. However, the increased accuracy comes at a cost, as it has a linear computational overhead for the attacker. While processing the 15° rotations requires an order of magnitude more computational effort compared to no rotations, we will use that to test the robustness of our system (unless otherwise stated), as it is sufficiently effective and 3 times faster than the attack with 5° rotations.

Next, we compare the combined impact of the rotation and transparency transformations. We create five sets of 500 challenges, each with a different `alpha level` (transparency) transformation, and run the attack without rotations. We then create another five sets of 500 challenges each, with a randomly rotated tag and different `alpha level` transformations. As can be seen in Figure 7, the `CCOEFF` algorithm is the most effective in all scenarios, proving to be the most effective against both transparency and rotation transformations. Nonetheless, we can see that transparency has a significant impact, dropping the success rate of all three algorithms to $6 - 10\%$ when $a = 0.4$. Since such a transparency level may prove to be difficult for users to identify the tags, we employ more conservative transparency transformations in the following experiments. Even with an 0.8 `alpha level`, which is a mild transformation, the success rate of two of the algorithms drops to less than 50%.

To compare the combined impact of all our transformations, we create two sets of 500 challenges each, that contain photos with one tag; one set with tag transformations and one with tag and photo transformations. We also place the tags randomly in the photos, to explore the significance of the "background" photo. We maintain a constant 0.8 `alpha level` (which is a very conservative transformation and remains easily identifiable to humans), and experiment by

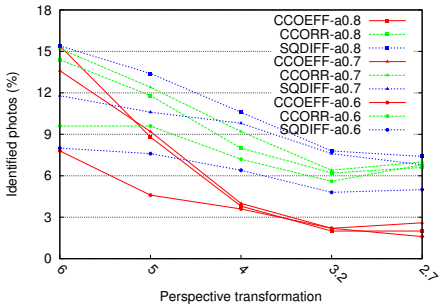


Figure 9: Attack success rate against different perspective (P) and alpha level transformations.

varying the perspective transformation ($P = 2.7, .., 6$). Figure 8 shows that the perspective transformation (P) has the most significant impact, but the tag transformations (RTP) also contribute to the overall robustness, especially when the perspective transformation is not as extreme, i.e., for larger values of P. `SQDIFF` proves to be the most effective method against the perspective transformation, while `CCOEFF` is the least effective despite being the most robust against transparency and rotation transformations. We achieve our best results when $P = 3.2, 2.7$, with the highest scoring method (`SQDIFF`) identifying only ∼7.4% of the photos that have undergone all transformations, compared to the 98.4%-98.8% success rate of our simplistic attacker (see Section 2), or the 100% rate of the template matching attack against non-transformed photos. Depending on the placement of the tag in the photo and the rotation effect, the perspective transformation has a combined impact that cannot be handled by the attacker even when conducting multiple rotations.

To further explore the combined impact of the transformations, in Figure 9 we show the percentage of identified photos for each algorithm with varying `alpha level` and perspective transformations. Our results clearly show a correlation between the identified tags and the levels of transformation. As the *alpha* and P values decrease (i.e., the tag becomes more transparent and the photo's perspective is more distorted) the effectiveness of the attack decreases by up to a factor of two for `SQDIFF` and by up to a factor of five for `CCOEFF`. Apart from the impact on the success rate, the transformations require a significant increase of processing effort from the attacker. Attempting to match the tags to the transformed photo requires ∼87.8 as opposed to ∼0.06 required for the simplistic attack against the non-transformed photos.

Surprisingly, we found that for a perspective transformation of 2.7 we didn't see a significant decrease compared to 3.2 and `CCORR` actually scored higher. As such, we set $P = 3.2$ for our next experiments as it is less distorting and also yields slightly better results. Furthermore, we manually inspected the identified tags and found them to be cases where they had been placed randomly on a position of the background photo with almost "no noise" (e.g., on a clear sky, wall, etc.). Thus, we should strongly enforce a restriction of placing the tags on faces in the background photos, which will further decrease success rates.

We measure the effect of increasing the number of tags per challenge. We create three sets of 1,000 photos that contain two tags ($N = 2$), with $P = 3.2$ and varying `alpha levels`. The attacker now has to compare each photo to 144 tags (72 per depicted friend), and correctly identify both tags to pass the challenge. To calculate the maximum success rate the attacker can achieve, the attack conducts 5° rotations which produce the best results, even though the processing time is unrealistic. Table 5 shows the results. With a 0.6 `alpha level`, `CCOEFF` fails to pass any challenges, `CCORR` passes 15 and `SQDIFF` passes 10 challenges, while processing a photo requires ∼ 401.5 sec. We create three sets of 250 photos with 3 tags, where one algorithm fails and the rest solve a single challenge, requiring over ∼ 663.6 seconds. *Thus, we reduce the attacker's success rate to* 0.4%*, while requiring four orders of magnitude more processing effort on his part.*

Tags	`alpha`	`CCOEFF`	`CCORR`	`SQDIFF`	Time (sec)
2	0.8	0.0%	2.1%	1.8%	397.7
	0.7	0.0%	1.9%	1.6%	400.6
	0.6	0.2%	1.5%	1.0%	401.5
3	0.8	0.0%	0.0%	0.0%	663.6
	0.7	0.0%	0.0%	0.4%	675.0
	0.6	0.0%	0.4%	0.0%	695.9

Table 5: Attack success with 5° rotations against challenges with 2 tags, for $P = 3.2$.

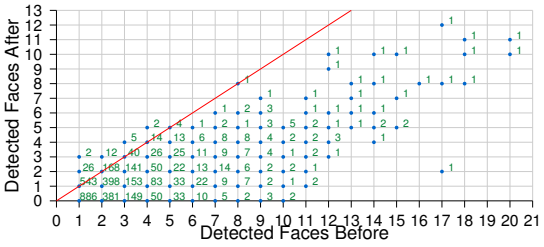


Figure 10: Faces detected before and after rotation+transparency. Point labels correspond to the number of photos. The line shows the $X = Y$ axis.

We also explored the possibility of combining the results of the 3 methods for achieving better success rates, i.e., compare the output of each method and select the photo that receives the highest confidence out of the 3. This, however, is infeasible because in our experiments the "confidence" returned by `CCORR` is always higher than the one by `CCOEFF`, even when `CCORR`'s identification is wrong. Also, `SQDIFF` returns a differential result which is not comparable to the other two.

Face Recognition attack. We also evaluate the robustness of our approach against face detection and, by extension, recognition. To explore how our tag selection and transformation process impacts face detection, we calculate the number of faces detected in 3,487 "background" photos before and after we transform the tags and superimpose them on the background photo (no perspective transformation performed). We first detect the faces in the photo, then superimpose a transformed tag over every face and, finally, execute the face detection process again. We perform a conservative transparency transformation with an `alpha level` of $a = 0.8$ that can easily be solved by users. Figure 10 shows the detected faces before and after, and the label of each point indicates the number of photos with that (before, after) tuple. The

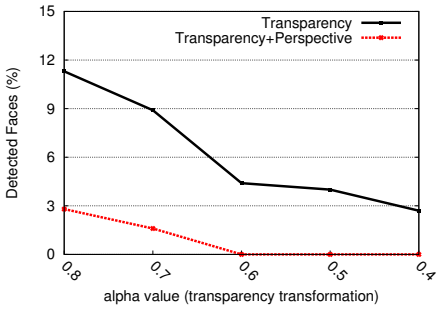


Figure 11: Faces detected for various `alpha levels`, against transparency and transparency+perspective transformations.

red line denotes the $X = Y$ axis, upon which are the cases where the same number of faces are detected. Interestingly, even though the photos now contain double the number of faces, an extra face is detected only in 47 (1.3%) cases (points over the red line). Everything below the red line, indicates improvement as less faces are detected. Due to our tag transformations, no faces are detected in 43.6% of the photos, which significantly impacts face recognition, as faces have to be detected before compared to facial models.

While the rotation transformation increases the processing effort of the attacker, it cannot hinder an attacker as a standalone transformation, as the attacker can perform rotations upon the crafted photo to increase detection rates. Thus, we want to explore the combined impact of our two other transformations. To do so, for each experiment we create two versions of 250 transformed photos with one *medium* tag each. In the first version, we only apply the transparency transformation, and in the second both the transparency and perspective transformation. We then manually remove any photos with tags that do not contain a human face, and use our face detection method to see how many faces are detected in the remaining tags. We test various `alpha levels` of transparency, with a constant value of $P = 3.2$ for the perspective transformation, as it had the best effect in our previous experiments. We present our results in Figure 11. While the transparency transformation has a very significant impact on the attack, by combining it with perspective transformation, *we are able to completely hinder face detection and, thus, recognition in all the photos, for* $a \leq 0.6$. Again, we found that any faces that were detected had been placed on a "clear" background.

Security through distortion. Overall, our experiments demonstrate the robustness of our transformations against pattern matching and face recognition attacks. Even under an extreme threat model where the attacker has collected *every single photo* of the users and has knowledge of our approach, we are able to completely deter one class of attacks and decrease the success rate of the other by over 99%, while incurring a massive overhead to the required processing effort.

4.4 Usability Evaluation

To evaluate the usability of our approach we conduct a preliminary study using challenges with tags of famous people. While users might be more familiar with certain famous people than with some of their contacts, they may not be familiar with others as well (or know their name). The goal

	Departmental				*AMT*
	Initial		**Normalized**		**Success**
Gender	NoHelp	Help	NoHelp	Help	94.38%
Female	75.0%	95.8%	76.4%	97.8%	
Male	73.9%	87.5%	78.5%	92.6%	**Time**
Total	74.2%	89.7%	77.9%	94.0%	7.93s

Table 6: Usability Evaluation. Initial and normalized success rates per gender, before and after suggestions are presented (Departmental). Success rate and seconds per challenge (AMT).

of this study is to measure the impact of the transformations on the ability of users to identify the depicted people. As such, we have selected fairly recognizable individuals.

We selected photos of them wearing sunglasses, with their face at an angle, taken under normal conditions and not edited. We used the values that we found in our experimental evaluation to be secure, but not to extreme so as to impede users' ability for identification. Specifically, we used all the possible combinations of: $N = \{1, 2\}, a = \{0.6, 0.7, 0.8\}, P = \{2.7, 3.2\}$. We conduct two separate studies, one with participants from our department, and one with Amazon Turk workers, each case allowing us to measure different aspects of our system; the impact of suggestions and the time required to solve the challenges.

Departmental. We recruited 30 participants from our department (students and staff), 8 of which were female and 22 were male, with ages ranging from 19 to 37. The same set of 12 challenges (one for every possible combination) was presented to all participants, to identify whether a specific transformation or person could not be identified by multiple users. Users were first shown a photo, and if they were able to provide the correct answer they would move on to the next. If not, they were given 6 suggestions for each tag, which allowed us to measure the impact of the suggestions. After the initial results, we also calculated a normalized rate where we filtered out challenges where the participant did not know the depicted person at all, as our goal is to measure the impact of the transformations alone.

As shown in Table 6, users solved 89.7% of the challenges, which increases to 94% if normalized. Surprisingly, users immediately identified over 75% of the challenges without help. Suggestions do offer a significant boost, with 14.1% and 21.4% for male and female participants respectively. There was no strong correlation between the values of transformations and the success rates. On the contrary, less transformed photos had higher scores in several cases. We found, however, that the face was an important factor as there were a few tags that users could not identify even without transformations, as they were barely familiar with certain individuals. Nonetheless, suggestions are an important aspect of the system as they help users solve the challenges.

AMT study. We recruited AMT workers (at least 95% approved, correct HITs and 1000 HITs completed) and presented them with two batches of 72 distinct challenges generated automatically by our system. The first batch had $N = 1$ celebrity face and the second one had $N = 2$. For each batch we included all the possible combinations of a and P as described above, and 6 suggested names (of which 1 was the correct answer). Overall, 49 workers solved 1,556 challenges,

with at least 20 workers solving each challenge. As Table 6 shows, AMT workers confirmed the previous results (on a larger scale) with a 94.38% success rate (not normalized), taking 7.93 seconds on average (7.19 standard deviation).

Conclusions on Usability. A crucial metric for the applicability of such a mechanism is the ability of users to solve the challenges. The results of the user study demonstrate the usability of our system, as users solved ∼94% of the challenges. Bursztein et al. [8] reported an 84–87% success rate for image CAPTCHAS, which is lower than our results. For the easiest category of text CAPTCHAs results were slightly higher than ours, with 95–98%. Furthermore, they reported an average completion time of ∼ 9.8 seconds for the image CAPTCHAs (∼ 8.9 when solved correctly), which is almost two seconds slower than ours.

We believe that our initial results demonstrate that our system provides a viable solution for securing the login process against adversaries that have stolen user credentials. Nonetheless, we plan on conducting an extensive user study with challenges that depict actual contacts of the users, to fully evaluate the usability of our approach.

5. SOCIAL AUTHENTICATION SERVICE

We discuss how an OSN can deploy our secure SA as a service for other websites. By adding a simple piece of code, services can access an API and fetch challenges specifically crafted for each individual user. This can be adopted as a user-gnostic CAPTCHA, or as an additional security mechanism.

To outline the benefit of employing such a service in addition to a traditional two-factor authentication scheme, we describe the following scenario. An attacker steals a user's smartphone, which contains the credentials to an e-banking service and is also the device that receives the security token (via SMS or a token-generation app). Normally, the attacker will be able to complete any transaction as he possesses both tokens needed to pass the two-factor authentication. Similarly, attacks in the wild, have passed two-factor authentication by tricking the user to install a malicious app (e.g., the Eurograbber malware [2]). However, if the e-banking service employs this service for additional security, attackers that don't belong to the victim's social circle will fail to complete any transaction. Even if the device has an existing session with the OSN, they will not be able to pass the challenges (see outsourcing attacks below).

Dataset. Another important feature of our system, is that it builds upon a dataset of photos and tag information that is readily available to the OSN. Thus, it doesn't face the challenge of creating a correct and extensive dataset as other image-based CAPTCHA schemes do [15, 26].

Someone could argue that SA is too restricted in terms of the number of challenges that can be created, in comparison to text-based CAPTCHA schemes than can create infinite combinations of numbers and letters. However, such an argument is far from true for our approach. In our study we found that users have an average of 347 friends each with approximately 12 *medium* tags, resulting in 4,164 suitable tags per user. These can produce over 8 million different permutations of 2 tags, and 1.2e+10 for 3 tags, which is more than enough for a single user profile. Also, the huge number of photos that can be used as "backgrounds" increases the number of possible challenges even more.

Privacy. A user-gnostic CAPTCHA service may raise privacy concerns, as the OSN acquires information regarding websites visited by users. However, this information is also acquired through the "social plugins" or single sign-on services offered by many popular OSNs. These services have been widely adopted, and [17] reports that over 35% of the top 10,000 Alexa websites include the "Like" button.

Security Properties. We discuss the effectiveness of our approach against typical CAPTCHA-breaking attacks.

Guessing Attacks. Our scheme allows automated bots to pass the challenge with a probability of $1/S^N$, where N is the number of friends depicted and S the number of suggestions per friend. The threshold adopted by previous work [28] is that bots should not have a success rate higher than 0.6%. By selecting 3 friends and providing 6 suggestions per friend, we are able to achieve an even lower probability of 0.46%. Furthermore, our system provides an extra level of security. As each CAPTCHA is created for a specific user, it is trivial to identify automated attacks that try to guess the answer. A legitimate user can request a new challenge when not able to identify the friends, without providing an answer, until presented with one he feels confident about answering. An automated script trying to guess the answer will continuously provide wrong answers until eventually guessing one.

Outsourcing/Laundering Attacks. Existing approaches create challenges that are user-agnostic, i.e., are not created for a specific user. They are built upon the notion that virtually any human should have a high probability of successfully carrying out the tasks required to pass the challenge. However, these approaches are susceptible to "laundering" attacks [19] where adversaries relay the challenge to CAPTCHA-breaking services with human solvers that provide the correct answer, and 'smuggling" attacks [9] that trick users into solving CAPTCHAs by injecting them in interactions with benign sites. Our approach is robust against such attacks, as challenges are *user-gnostic*: they cannot be outsourced to others, as they wouldn't be familiar with the user's friends. Solving them would require the worker to first familiarize with the friends' faces. This, of course, is impractical as it would require too much time to pass a single CAPTCHA challenge (might not even be possible within the allowed time window).

6. LIMITATIONS AND FUTURE WORK

The attacker could create a collection of processed tags and compare those to the presented SA challenge. However, various characteristics of our approach render it robust against such a scenario. First, the completely random background photo, which blends in with the tags, introduces significant noise which can't be predicted by the attacker. Second, the placement of the tag on a photo significantly affects the result of its perspective transformation. Finally, as the transformations' values can be selected from a range, the attacker would have to create a massive collection of processed tags with various combinations of transformations and backgrounds. Even then, identifying the tag might not be feasible.

The attacker could attempt to reverse the perspective transformation. However, that requires knowledge of the exact value of the transformation, which can be randomly selected from a range of values. Furthermore, various types of perspective transformations exist (which we plan to explore), and our system could randomly select one for each challenge. Even in the implausible scenario where an attacker could guess the correct combination of algorithm and exact value,

the transparency and rotation transformations provide a very strong defence against the face detection attack and can reduce the success of the image comparison attack by over 80% for two of the algorithms.

7. RELATED WORK

The first to analyze SA and discuss potential vulnerabilities were Kim et al. [16], who presented a formal quantification of the risks that threaten SA. A key observation is that tightly connected communities exhibit higher risks. In [20] we demonstrated the feasibility of attacks against SA: a casual attacker with access only to publicly available information would pass 22% of the challenges and significantly improve the chances for 56%; a determined attacker, with access to social-circle information, would pass 100% of the challenges, when training classifiers with 120 faces per friend.

Bursztein et al. [8] conducted an extensive study to measure the success rate of users when presented with CAPTCHAs from various services. An important observation the authors make is that the difficulty of CAPTCHAs is often very high, resulting in their solution being a troublesome process for users. A significant aspect of our approach is that it is user-friendly, as users are required to identify their friends.

Previous work [12, 22] has explored the use of human faces for creating CAPTCHAs. Goswami et al. [12] built upon the failure factors of face detection algorithms to propose a CAPTCHA that uses distorted human and cartoon faces as a challenge, and users must identify all human faces without making any mistakes. The pool of faces is compiled from publicly available databases and distortions are applied, such as adding stripes to cover key face characteristics, rotating the image and overlaying the actual face over a background of random noise. While this approach may be robust against face detection, it can be bypassed with image comparison algorithms (like template matching), that map the visible parts of the faces to the photos within the public database.

Previous work (e.g., [26]) has attempted to defend against face recognition attacks, based on semantic properties of the photo's content. This presents a significant drawback typical of various image-based CAPTCHA systems; the creation of the dataset labelled with the semantic description cannot be automated. On the contrary, SA takes advantage of an enormous, ever-expanding, labelled dataset created by users.

8. CONCLUSIONS

In this paper we revisited the concept of social authentication, and proposed a novel approach that retains its usability while being robust against attacks that utilize image analysis techniques. The key concept is to filter out faces that can be identified by face recognition software, and craft the challenge photos in a manner that obfuscates any areas with sensitive information from the initial photos. We conducted a measurement to explore the ability of users to identify their friends in photos taken under realistic conditions. Our results demonstrated that users are very effective at identifying their friends even when their face is not clearly visible or present at all. Based on our study results and a series of observations, we implemented a prototype and evaluated it against the face recognition and image comparison attacks. We also verified the usability of our system through a user study. Finally, we discussed the benefits of employing such a system as a user-gnostic CAPTCHA service, or an additional security mechanism to two-factor authentication.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported in part by the FP7 project SysSec funded by the EU Commission under grant agreement no 257007, the FP7 Marie-Curie ITN funded by the European Commission under grant agreement no 316808, and by the MIUR under the FIRB2013 FACE grant. This work was also supported by the NSF Grant CNS-13-18415. Author Keromytis was also supported by (while working at) the NSF during the conduct of his work. Any opinions, fundings, conclusions, or recommendations expressed herein are those of the authors, and do not necessarily reflect those of the US Government or the NSF.

References

[1] Amazon Mechanical Turk. `https://www.mturk.com/mturk/`.

[2] Eurograbber. `https://www.checkpoint.com/products/downloads/whitepapers/Eurograbber_White_Paper.pdf`.

[3] Facebook Introducing Login Approvals, . `https://www.facebook.com/note.php?note_id=10150172618258920`.

[4] Facebook Acquires Face.com, . `http://mashable.com/2012/06/18/facebook-acquires-face-com/`.

[5] Google 2-step. `http://www.google.com/landing/2step/`.

[6] Leyla Bilge, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 2009.

[7] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proceedings of the Annual Computer Security Applications Conference*. ACM, 2011.

[8] Elie Bursztein, Steven Bethard, Celine Fabry, John C. Mitchell, and Dan Jurafsky. How good are humans at solving CAPTCHAs? A large scale evaluation. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*. IEEE, 2010.

[9] Manuel Egele, Leyla Bilge, Engin Kirda, and Christopher Kruegel. Captcha smuggling: Hijacking web browsing sessions to create captcha farms. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10', pages 1865–1870. ACM, 2010.

[10] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. COMPA: Detecting Compromised Accounts on Social Networks. In *ISOC Network and Distributed System Security Symposium (NDSS)*, 2013.

[11] Steffen Gauglitz, Tobias Höllerer, and Matthew Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *Int. J. Computer Vision*, 94(3):335–360, 2011.

[12] Gaurav Goswami, Brian M. Powell, Mayank Vatsa, Richa Singh, and Afzel Noore. FaceDCAPTCHA: Face detection based color image CAPTCHA. In *Future Generation Computer Systems (September 2012)*.

[13] Junxian Huang, Yinglian Xie, Fang Yu, Qifa Ke, Martin Abadi, Eliot Gillum, and Z. Morley Mao. Socialwatch: detection of online service abuse via large-scale social graphs. In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, ASIA CCS '13', 2013.

[14] Danesh Irani, Marco Balduzzi, Davide Balzarotti, Engin Kirda, and Calton Pu. Reverse social engineering attacks in online social networks. In *Proceedings of the 8th international conference on Detection of intrusions and malware, and vulnerability assessment*, DIMVA'11, 2011.

[15] Elson Jeremy, John R. Douceur, Jon Howell, and Jared Sault. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. In *Proceedings of the 14th ACM conference on Computer and communications security (CCS)*. ACM, 2007.

[16] Hyoungshick Kim, John Tang, and Ross Anderson. Social authentication: harder than it looks. In *Proceedings of the 2012 Financial Cryptography and Data Security conference*. Springer.

[17] Georgios Kontaxis, Michalis Polychronakis, Angelos D. Keromytis, and Evangelos P. Markatos. Privacy-preserving social plugins. In *Proceedings of the 21st USENIX conference on Security symposium*, Security'12. USENIX Association.

[18] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing facebook privacy settings: User expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11'. ACM, 2011.

[19] Marti Motoyama, Kirill Levchenko, Chris Kanich, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. Re: Captchas: understanding captcha-solving services in an economic context. In *Proceedings of the 19th USENIX conference on Security*, USENIX Security'10. USENIX Association, 2010.

[20] Iasonas Polakis, Marco Lancini, Georgios Kontaxis, Federico Maggi, Sotiris Ioannidis, Angelos Keromytis, and Stefano Zanero. All your face are belong to us: Breaking facebook's social authentication. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC '12'. ACM, 2012.

[21] Alex Rice. Facebook - A Continued Commitment to Security, Jan 2011. `http://www.facebook.com/blog.php?post=486790652130`.

[22] Yong Rui and Zicheng Liu. Artifacial: Automated reverse turing test using facial features. In *In Multimedia*, pages 295–298. ACM Press, 2003.

[23] Amichai Shulman. The underground credentials market. *Computer Fraud & Security*, 2010(3):5–8, March 2010.

[24] Yaniv Taigman and Lior Wolf. Leveraging billions of faces to overcome performance barriers in unconstrained face recognition. *CoRR*, abs/1108.1122, 2011.

[25] Blase E. Ur and Vinod Ganapathy. Evaluating attack amplification in online social networks. In *Proceedings of the 2009 Web 2.0 Security and Privacy Workshop*.

[26] Shardul Vikram, Yinan Fan, and Guofei Gu. SEMAGE: A New Image-based Two-Factor CAPTCHA. In *Proceedings of 2011 Annual Computer Security Applications Conference (ACSAC'11)*, December 2011.

[27] Sarita Yardi, Nick Feamster, and Amy Bruckman. Photo-based authentication using social networks. In *Proceedings of the first workshop on Online social networks*, WOSN '08'. ACM, 2008.

[28] Bin B. Zhu, Jeff Yan, Qiujie Li, Chao Yang, Jia Liu, Ning Xu, Meng Yi, and Kaiwei Cai. Attacks and design of image recognition captchas. In *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10'. ACM, 2010.

APPENDIX

Appendix A

Below is the questionnaire that users were requested to fill in for each photo presented to them during the Social Authentication challenges. Assuming the depicted friend's name is X, the user will be presented with the following questions.

1. Type of photo?
   - Portrait
   - Landscape
   - Objects
   - Text
   - Animals
   - Art
2. Where is X's Face?
   - Within the tag and is clearly visible.
   - Outside the tag and is clearly visible.
   - Within the tag, but not clearly visible.
   - Outside the tag, but not clearly visible.
   - Not in the photo at all.
3. Are there other faces in the photo?
   - There are other people's faces both outside and inside the tag.
   - There's someone else's face within the tag.
   - There's someone else's face outside of the tag.
   - There are no other faces in this photo.
   - There are no faces in this photo.
4. Why was this photo useful for identifying X?
   - I remember seeing this photo from X.
   - The content of the photo is relevant to X.
   - None of the other suggested friends matched.
   - This photo was not useful.
   - X is in the photo.