Exploring Implicit Memory for Painless Password Recovery

Tamara Denning,[†][*] Kevin Bowers,[†] Marten van Dijk,[†] and Ari Juels[†]
RSA Labs,[†] University of Washington[*]
tdenning@cs.washington.edu, {kevin.bowers, marten.vandijk, ari.juels}@rsa.com

ABSTRACT

Knowledge-based authentication systems generally rely upon users' explicit recollection of passwords, facts, or personal preferences. These systems impose a cognitive burden that often results in forgotten secrets or secrets with poor entropy. We propose an authentication system that instead draws on *implicit memory*–that is, the *unconscious* encoding and usage of information. In such a system, a user is initially presented with images of common objects in a casual familiarization task. When the user later authenticates, she is asked to perform a task involving a set of degraded images, some of which are based upon the images in the familiarization task. The prior exposure to those images influences the user's responses in the task, thereby eliciting authentication information. We ran a user study to investigate the plausibility of our system design. Our results suggest that implicit memory has potential as a basis for low-cognitive-overhead, high-stability, knowledge-based authentication.

Author Keywords

Authentication, implicit memory, password recovery, priming, security.

ACM Classification Keywords

H.5.m Information Interfaces and Presentation: Miscellaneous

General Terms

Human Factors, Security.

INTRODUCTION AND BACKGROUND

Explicit memory forms the basis of most "something-you-know" authentication systems in use today. Most commonly, password and PIN systems and password-recovery systems based on 'life' questions [7] set users the explicit task of recalling a piece of secret information. Most alternative systems, such as graphical passwords [2, 5, 10, 11] and preference-based authentication [3] rely on the same approach: the user enrolls by committing a secret to memory or presenting some personal secret, and authenticates by recalling or recognizing that secret. All of these systems are explicit in the sense that the user is consciously attempting to recall or recognize the authentication secret.

Authentication schemes based upon explicit memory suffer from a tension between security and usability. User-generated passwords (visual or word-based) can be difficult to remember and are subject to reuse, having low entropy, and being written down for easy reference. Increasing the length or randomness of secrets increases the cognitive burden on the user. Life-history and preference questions avoid some of these pitfalls, but can be subject to low entropy, data-mining, or changing over time [7].

Implicit memory, in contrast to explicit memory, unconsciously influences or controls people's actions even when they are not attempting to retrieve the memory in question. Motor skills are one notable type of implicit memory; habituated physical actions—such as riding a bike—do not require explicit mental effort. Cognitive studies have shown that explicit and implicit memories have different biological mechanisms; patients with brain damage that makes them perform poorly when tested on their explicit memory may still perform well on motor and other implicit memory tasks [6].

One particular kind of implicit memory—priming—involves exposing a user to a particular set of stimuli in order to observe its effects in later testing. In short-term priming, a user unconsciously completes open-ended tasks with a bias towards recently viewed stimuli. Certain priming effects, however, have been shown to persist for weeks, months, and even years [1, 4]. We propose harnessing this priming effect for non-conscious, low-effort authentication.

In this paper, we devise a new implicit-memory approach to authentication with the following benefits: (1) Retention of secrets is potentially very long-lasting [1, 4]; (2) Registration and authentication effort by users does not require memorization; (3) Secrets are truly random and have precisely quantifiable entropy; and (4) Enrollment stimuli are distinctly different from authentication stimuli, thereby avoiding image set intersection attacks.

The protocol we explore here does not immediately yield a workable system: authentication would be intolerably slow for most users. But our experiments do suggest the potential of a novel approach—using implicit memory to authenticate a user—and point the way to more practical variants.[1]

[1] Previous work published in the community by Weinshall makes reference to the potential of implicit memory and priming for authentication; however, the systems and experiments described in the paper utilize explicit memory, not implicit memory [10].

SYSTEM CONCEPT

For our authentication system, we propose using pairs of images—complete images and their degraded counterparts—in a long-term visual priming activity. When the images are presented in their degraded form, it is difficult for a user to identify the content of the image. Exposure to a complete image makes identification of the counterpart degraded image easier—even after a substantial lapse of time [1]. Mitchell observed a priming effect lasting 17 years [4], even when the participants reported no conscious memory of participating in the original experiment.

In our design, a set of complete images is assigned to each user, she is exposed to the stimuli, and she is later authenticated based on her ability to identify the images' degraded counterparts. In the enrollment phase, we select a random subset of images from the corpus for each user. We expose the user to the complete variant of each image: we ask her to label each of these images by providing the word(s) that best describe the object. In keeping with terminology in the psychology literature, we refer to this exposure process as *priming*, and the presented images as *primed*.

Later, in the authentication phase, we present users with degraded versions of a combination of primed and random non-primed images, and again ask for labels. A user's *score* is the number of accurately identified primed images plus the number of incorrectly labeled (or skipped) non-primed images. Her score thus indicates the degree to which her identification of images aligns with the priming she received. To clarify, a user's score is not simply the number of degraded images that she is able to label: a user is scored on how well her labeling success aligns with her primed image set, meaning that she is penalized for correctly labeling non-primed images and rewarded for incorrectly labeling non-primed images. In an authentication system, a user's score would need to exceed a certain, predetermined threshold for the user to be accepted. Therefore, an adversary does not benefit from having complete knowledge of the image corpus and its labels. A primed stimulus is only used once in an authentication prompt. After this different primed stimuli must be used, otherwise the user would potentially be contaminated by previously viewed image fragments.

THEORETICAL ANALYSIS OF THE SYSTEM

Let p_i be the probability that a randomly selected user, who was primed on image i, is able to correctly label image i. Let n_i be the probability that a randomly selected user, who was not primed on image i, is able to correctly label image i. We assume that all images respond positively to priming and thus expect $p_i \geq n_i + \alpha$ where α is a measure of the priming effect and is ≥ 0.

If there are t images in a test corpus, users are primed on an arbitrary subset U of size $t/2$. During password recovery, the user is presented with all t images and attempts to correctly label them. The password recovery system counts the number c of images $i \in U$ that are correctly labeled plus the number of images $i \notin U$ that are not correctly labeled. If c is at least some threshold value τ, password recovery succeeds. An image i is counted in c with probability p_i if $i \in U$ and with probability $1-n_i$ if $i \notin U$. For a user who is primed on U, c has mean $\mu = \sum_{i\in U} p_i + \sum_{i\notin U}(1-n_i)$ and standard deviation $\sigma = \sqrt{\sum_{i\in U} p_i(1-p_i) + \sum_{i\notin U} n_i(1-n_i)} \leq \sqrt{t}/2$.

The expectation of μ over all users, that is, over random choices of U, is equal to $\hat{\mu} = Exp[\mu] = \sum_i (p_i+1-n_i)/2 \geq t(1+\alpha)/2$. We may rewrite $\mu = \sum_i X_i(1-n_i)+(1-X_i)p_i$, where $X_i = 0$ for $i \in U$ and $X_i = 1$ for $i \notin U$. Since half of the images i have $X_i = 1$, we may approximate (for large enough t) μ as a sum of weighted binomial distributions by taking X_i to be a binary uniformly distributed random variable. Since $X_i(1-n_i) + (1-X_i)p_i$ has standard deviation $(1-n_i-p_i)/2$, μ has standard deviation $\hat{\sigma} = \sqrt{Exp[\mu^2] - Exp[\mu]^2} \approx \sqrt{\sum_i (1-n_i-p_i)^2}/2 \leq \sqrt{t}/2$.

Taken over all users, value c has mean μ and standard deviation $\sigma \leq \sqrt{t}/2$, where μ is distributed with mean $\hat{\mu} \geq t(1+\alpha)/2$ and standard deviation $\hat{\sigma} \leq \sqrt{t}/2$. By combining both distributions, we obtain that taken over all users, c has mean $\hat{\mu} \geq t(1+\alpha)/2$ and standard deviation $\sqrt{\hat{\sigma}^2+\sigma^2} \leq \sqrt{t}/\sqrt{2}$.

Adversarial Strategy: We now look at the optimal adversarial strategy given the strong assumption that an adversary knows the value of p_i and n_i for each image i, as well as the correct label for that image. The adversary then attempts to maximize his probability of password recovery without knowledge of the set U on which the user was primed. Without knowing U, his best strategy is to label each image correctly with probability $1/2$. The value c computed for the adversary then will have a binomial distribution with mean $t/2$ and standard deviation $\sqrt{t}/2$.

Suppose that an adversary is allowed to recover a user's password with probability at most 0.5% (false positive rate). This implies a threshold $\tau = t/2 + 3\sqrt{t}/2$. For valid password recovery to succeed at least 97.5% of the time (false negative rate of 2.5%), we need $\hat{\mu} - 2\sqrt{\hat{\sigma}^2+\sigma^2} \geq t(1+\alpha)/2 - 2\sqrt{t}/\sqrt{2} \geq \tau = t/2 + 3\sqrt{t}/2$, or, $\alpha \geq 5.8/\sqrt{t}$. If we assume $\alpha = 0.5$, this implies $t \geq 135$. A user would need to attempt to label 135 images to recover their password.

USER STUDY

We performed a user study in order to investigate the viability of our system design. Our implementation draws on an image corpus created by Snodgrass, Vanderwart, and Corwin (SVC) [8, 9] consisting of 150 line drawings of familiar objects (e.g., animals, vehicles, and tools). Each line drawing has two variants: a complete image and a degraded version (see Figure 1).

Experiment: Each subject was randomly assigned 75 of the 150 images from the SVC corpus on which to be primed,[2] with image assignments randomly and evenly distributed across the set. The priming task was distributed as a spreadsheet

[2] Not all of the primed images were used in the authentication tasks.

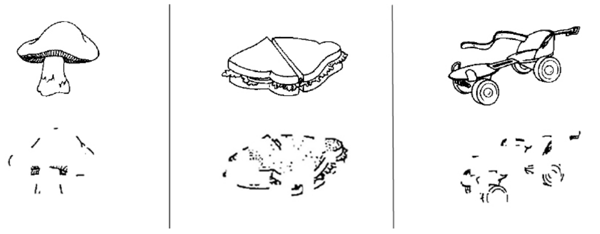

Figure 1. Complete and fragmented line drawing pairs from the SVC corpus. From left to right: a mushroom, a sandwich, and a roller skate.

via email. Two rounds of testing were administered via a web application that participants completed in situ on their computers. Participants were invited via email to complete test 1 approximately 14 days after priming and to complete test 2 approximately 21 days after completing test 1. Due to our sample size, the fact that participants varied in their completion times after receiving invites, and the exploratory nature of this study, we combined the results from the two tests in our analysis (min = 9 days, max = 49 days, average = 26.8 days, median = 28 days). Participants' image labels were coded as either correct or incorrect based upon both the original labels provided by Snodgrass and Corwin and the priming labels provided by participants.[3]

Participants: Participants were recruited from within RSA via email and an internal message board. 103 participants completed the priming task. Of those 103, 77 participants completed the first authentication test and 69 completed the second authentication test. 38 participants returned demographic surveys. The surveyed population is 21% female (79% male) and has the following age breakdown: 21–30 (24%), 31–40 (34%), 41–50 (21%), 51–60 (16%), 61+ (5%).[4] 89% of the population are employed in a technical profession (e.g. computer engineer). 68% of the survey respondents reported English as their primary language, with the remainder reporting predominantly dialects of India and China.

RESULTS

Looking at the experimental data from all images (treating them as a single image), we find that users correctly labeled images on which they were primed 984 out of 2149 times (45.8%). Users correctly labeled 831 out of 2143 images on which they were not primed (38.8%). This means primed users label images correctly 7% more often than non-primed users ($\alpha = 0.07 \pm 0.042$ at a 95% confidence interval, approximated using a normal distribution). This suggests that a priming effect does exist, even in the presence of a large number of images with negligible priming.

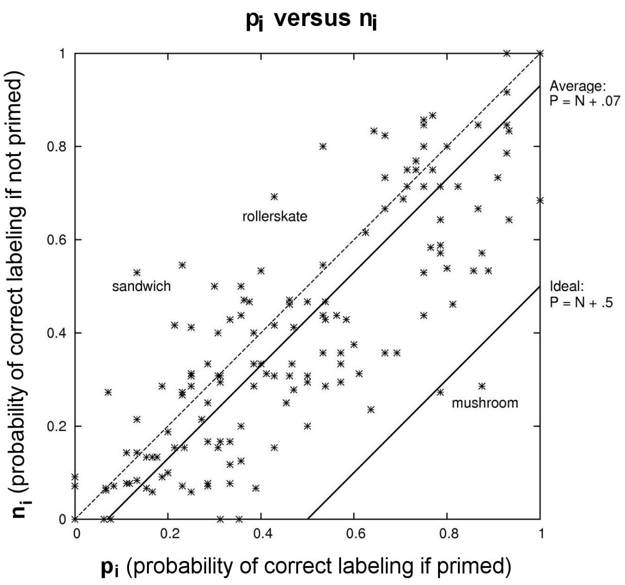

Figure 2. Priming effect as seen across all images.

Figure 2 plots points (p_i, n_i) for each image as measured by our experimental set up. Each p_i and n_i is estimated by using the labels of approximately 14 users leading to wide confidence intervals. This potentially explains why for $\approx 1/3$ of all images the measured $p_i < n_i$. For example, if in truth $p_i \approx n_i$—meaning that the image demonstrates a negligible priming effect—then estimates for p_i and n_i have a standard deviation of $\sqrt{p_i(1-p_i)}/\sqrt{14}$. For $0.2 \leq p_i \approx n_i \leq 0.8$, this standard deviation is 0.1. Thus it is not surprising that 27 of the 150 images graphed above have $p_i \leq n_i \leq p_i + 0.1$.

As an example, Figure 1 depicts both a roller skate and a sandwich, which have a measured value $p_i < n_i$. The first was commonly mislabeled as a car, and the second as a fish. The later mislabeling can in part be attributed to cross-contamination (users who were primed on the fish).

Most of the images do show a positive priming effect, although only two images are measured to have $\alpha \geq 0.5$. Figure 1 shows one of these images (the mushroom). If 135 such images were created, then we would have a viable authentication system with a false positive rate of 0.5% and a false negative rate of 2.5%. Participants showed great variability in the amount of time that it took them to label an image; however, the median image labeling time of 3.9 seconds would result in a password recovery time of 8.8 minutes.

DISCUSSION

The results from the user study suggest that an authentication system based upon implicit memory might be viable; however, the system as implemented in the user study would take an impractical amount of time to authenticate and its images demonstrate an undesirably low priming effect (α). The viability of the system concept is dependent upon being able to systematically identify or create images with a sufficiently strong priming effect.

System Strengths: This system concept provides some usability and security improvements over current password,

[3] While some accommodations were made for mispellings, superlabels (e.g. "peregrine falcon" for "falcon"), and additional adjectives, codings were relatively literal comparisons. An actual implementation could use spell-checking (e.g. edit distance), synonyms, or adapt a database over time based on inputted labels.

[4] Due to an error, the actual survey listed 21–30, 30–40, 40–50, 50–60, and 60+.

life-question, or image-based alternatives. The usability advantage is fairly obvious: by not requiring the user to explicitly remember facts or passwords, the system removes a cognitive burden from the authentication process. Unless a user selects a random password (or creates a random answer to a question), it is difficult to accurately analyze or arbitrarily set the security of current authentication systems. In contrast, the core of our system consists of an arbitrarily set secret: the primed images assigned to the user. The entropy provided by this secret can be calculated and tuned according to the requirements of a particular system deployment.

System Weaknesses: The main problem with the current system design—and a strong obstacle to its being used in practice—is that effectively only one bit of entropy can be extracted per each image provided to the user at authentication. Since the priming effect is not overwhelmingly strong, it is necessary to show users a number of images before being able to authenticate them. As a result, this concept is most likely only suitable in an application such as password recovery, where authentication takes place seldom enough that the time required is not prohibitive. Ideally, we would want to modify the system design so that multiple bits of entropy can be extracted per image task.

CONCLUSIONS

We present the design of an authentication system that is based upon a user's implicit memory of images, which suggests the potential of an arbitrarily secure authentication scheme that does not burden the user with remembering explicit facts. We implemented a user study to perform a preliminary evaluation of such an authentication system and provide a formal evaluation of the properties of such a system. Results from the user study indicate that while such a system is plausible, there are two major obstacles to overcome before such a system could be deployed. First, the specific characteristics of images in the system's corpus affect how strongly they elicit a priming effect; a working authentication system would need a sufficient supply of complete and degraded image pairs that provide strong differentiation between primed and non-primed users. Second, the current system design only extracts one bit of entropy from every image provided to the user, which limits the amount of secret information that can be extracted from the user in a reasonable time frame. In order to produce a viable system with this concept, the design must be altered to collect additional information from the user for every displayed image.

FUTURE DIRECTIONS

There are a number of potential directions for future research relating to this system concept. One obvious goal would be to determine what image features or characteristics contribute to a strong priming effect; this information could then be used to auto-generate or select images for the authentication corpus. Crowdsourcing platforms such as Amazon's Mechanical Turk are potentially a useful mechanism for creating or collecting data on potential images, particularly for an automated degradation process.

Our current system design extracts a maximum of one bit of entropy per primed image: whether the user answers as expected (a correct label when primed or an incorrect label when unprimed) or not. Strong authentication thus requires the use of many images. Since the length of the registration and authentication tasks directly affects the usability of such a system, it would be preferable to change the system design so that more than one bit of entropy is extracted per image. One promising direction is to create multiple complete images that map to the *same* fragmented image. For example, we partition all complete images into groups of 16 images each, and for each group we map its 16 different complete images to one unique fragmented image. Priming each user on one random complete image within each group will give $\log 16 = 4$ bits of entropy per primed image.

There are also many opportunities to explore different registration and authentication tasks within the general system concept. Any task which familiarizes the user with an image set could potentially be used instead of the initial labeling task. Similarly, any task which successfully draws upon the user's implicit memory of the primed images can be used as an authentication task; while asking the user whether she recognizes an image or providing image label options via multiple choice most likely introduces too much noise into the authentication process, one might ask users to select an answer that indirectly demonstrates whether or not they can accurately identify the image (e.g., "Are you most likely to see this object in a bedroom, the kitchen, or the ocean?").

While we leverage users' visual memory in this scheme, priming and implicit memory can also be observed in other modalities (e.g. verbal, motor, and aural). An implicit-memory authentication system could potentially be built on top of one of these modalities if they show a sufficient long-term priming effect without requiring habituation.

REFERENCES

1. C. B. Cave. Very Long-Lasting Priming in Picture Naming. *Psychological Science*, 8:322–5, 1997.
2. E. Hayashi, R. Dhamija, N. Christin, and A. Perrig. Use Your Illusion: secure authentication usable anywhere. In *SOUPS '08*, 2008.
3. M. Jakobsson, E. Stolterman, S. Wetzel, and L. Yang. Love and authentication. In *CHI '08*, 2008.
4. D. B. Mitchell. Nonconscious Priming After 17 Years : Invulnerable Implicit Memory? *Psychological Science*, 17:925–9, November 2006.
5. Passfaces Corporation. Passfaces^{TM}. http://www.realuser.com/.
6. D. L. Schacter, C. Chiu, and K. N. Ochsner. Implicit Memory: A Selective Review. *Annu. Rev. Neurosci.*, 16:159–82, 1993.
7. S. Schechter, A. J. B. Brush, and S. Egelman. It's No Secret. Measuring the Security and Reliability of Authentication via "Secret" Questions. In *S&P '09*, 2009.
8. J. G. Snodgrass and J. Corwin. Perceptual Identification Thresholds for 150 Fragmented Pictures from the Snodgrass and Vanderwart Picture Set. *Perceptual and Motor Skills*, 67:3–36, 1988.
9. J. G. Snodgrass and M. Vanderwart. A Standardized Set of 260 Pictures: Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6:174–215, 1980.
10. D. Weinshall and S. Kirkpatrick. Passwords you'll never forget, but can't recall. In *CHI '04*, 2004.
11. S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. PassPoints: design and longitudinal evaluation of a graphical password system. *Int. J. Hum.-Comput. Stud.*, 63(1-2):102–127, 2005.