Secrets, Lies, and Account Recovery:
Lessons from the Use of Personal Knowledge Questions at Google

Joseph Bonneau*
Stanford University & EFF
jbonneau@cs.stanford.edu

Elie Bursztein
Google
elieb@google.com

Ilan Caron
Google
ilanc@google.com

Rob Jackson
Google
roj@google.com

Mike Williamson
Google
miwilliamson@google.com

ABSTRACT

We examine the first large real-world data set on personal knowledge question's security and memorability from their deployment at Google. Our analysis confirms that secret questions generally offer a security level that is far lower than user-chosen passwords. It turns out to be even lower than proxies such as the real distribution of surnames in the population would indicate. Surprisingly, we found that a significant cause of this insecurity is that users often don't answer truthfully. A user survey we conducted revealed that a significant fraction of users (37%) who admitted to providing fake answers did so in an attempt to make them "harder to guess" although on aggregate this behavior had the opposite effect as people "harden" their answers in a predictable way.

On the usability side, we show that secret answers have surprisingly poor memorability despite the assumption that reliability motivates their continued deployment. From millions of account recovery attempts we observed a significant fraction of users (e.g 40% of our English-speaking US users) were unable to recall their answers when needed. This is lower than the success rate of alternative recovery mechanisms such as SMS reset codes (over 80%).

Comparing question strength and memorability reveals that the questions that are potentially the most secure (e.g what is your first phone number) are also the ones with the worst memorability. We conclude that it appears next to impossible to find secret questions that are both secure and memorable. Secret questions continue have some use when combined with other signals, but they should not be used alone and best practice should favor more reliable alternatives.

Categories and Subject Descriptors

D.4.6 [**Software**]: Security and Protection—*authentication*

General Terms

Security, Privacy, Authentication

*Part of this research was conducted while Joseph Bonneau was at Google.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741691.

Keywords

personal knowledge questions; account recovery

1. INTRODUCTION

Personal knowledge questions (also called "secret questions" or "challenge questions" among other names) have long been used as backup mechanism to reclaim lost accounts [20]. Many academic studies have argued in their favor on the ground that they should be more memorable than passwords [33, 10, 26] for two reasons: the presence of a question makes remembering the answer a *cued recall* task instead of a *free recall* task and the information being asked for is something users inherently remember rather than a secret stored explicitly for authentication. On the security side, previous work has highlighted the potential weakness of secret questions [28, 14] based on laboratory experiments and analysis.

In this work we provide the first large-scale empirical data analysis of secret questions based on their deployment at Google. We studied the distribution of hundreds of millions of secret answers and millions of account recovery claims, demonstrating that in practice secret questions have poor security and memorability. This poor level of security, their unreliability for successful account recovery, and the existence of alternative recovery options with significantly higher success rate motivated Google's decision to favor alternative options (SMS, Email) as a recovery mechanism. Secret questions are now only used as a last resort in conjunction with other signals.

Some of our key empirical findings are:

- Statistical attacks against secret questions are a real risk because there are commons answers shared among many users. For example using a single guess an attacker would have a 19.7% success rate at guessing English-speaking users' answers for the question "*Favorite food?*". Similarly, with a single guess the attacker would have a 3.8% success rate at guessing Spanish-speaking users' answers for the question "*Father's middle name?*". With 10 guesses an attacker would be able to guess 39% of Korean-speaking users' answers to "*City of birth?*" (Section 3.1).

- Questions that are supposedly more secure due to the expectation that each user will have a different answer (e.g phone number) in practice don't exhibit a flat distribution because people provide untruthful answers. As a result the security of these questions is significantly lower than hoped. For example with a single guess an ideal attacker would have a success rate of 4.2% at guessing English-speaking users' answers to the question "*Frequent flyer number?*". Similarly, they would be

able to guess 2.4% of Russian-speaking users' phone number answers with a single try (Section 3.1).

- It is easy and cheap to create answer distributions that closely approximate real answer distributions using crowdsourcing services such as MTurk and CrowdFlower. Our experiment reveals that with as few as 1000 answers from crowdsourced users we are able to build approximate distributions which enable a guess rate that is between 75% and 80% as effective as the true distribution when making up to 100 guesses (Section 3.4).
- Users are no more likely to recover their accounts early or late in the lifetime of their account with the exception of the first 72 hours which see a surge with ≈10% of the recovery claims (Section 4.1).
- Questions that are potentially more secure have worse recall than unsafe questions: For the US English-speaking population the the question "*Father's middle name?*" had a success rate of 76% overall whereas the potentially safer question "*First phone number?*" had a 55% recall. The potentially safest questions have abysmal recall: "*Library card number?*" has a 22% recall and "*Frequent flyer number?*" only has a 9% recall rate (Section 4.2).
- Question memorability decreases significantly over time. For example the success rate for the question "*Favorite Food?*" is 74% after a month, 53% after 3 month and barely 47% after a year (Section 4.2).
- The decay of memorability over time is greater for questions about numbers assigned to people vs personal questions: for the question "*Frequent flyer card number?*" the recall rate decreased by 18% after a month whereas it only decreased by 6% for the question "*Father's middle name?*" (Section 4.3).
- Memorability is greatly impacted by people supplying untruthful answers. For the question "*First phone number?*" US users who supplied an answer with a plausible length of 7 digits have a 55% chance to recall their question while people who supplied an answer with a length of 6 characters only have an 18% chance to answer correctly (Section 4.4).
- The memorability of secret questions is influenced by cultural factors even when the language is the same. For instance the recall for English users from Great Britain is significantly lower than English users from the USA: 52% vs 61% (Section 4.5).
- Surveying the US population using Google Consumer Surveys reveals that people provide untruthful answers to secret questions because they try to make it harder to guess (37% of the 1500 respondents) or easier to remember (15%). Ironically of course, this behavior achieves exactly the opposite effect (Section 5).
- SMS and email-based account recovery have a significantly higher chance of success: 81% for SMS vs 75% for Email vs 61% (US/English) down to 44% (France/French) for secret questions (Section 6).

2. THREAT MODEL & GUESSING METRICS

In this section we discuss the threat model that is addressed by account recovery in general and secret questions in particular. Past research has highlighted many security weaknesses for personal knowledge questions:

- **Questions with common answers.** Many personal knowledge questions have common answers shared by many in the user population which an adversary might successfully guess. Schechter et al. were able to guess approximately 10% of user's answers by using a list of other answers provided by users in the same study [28]. Bonneau et al. [5] considered the difficulty of guessing common pieces of information used in personal knowledge questions using public statistics, for example using census records to estimate the difficulty of estimating of guessing a surname.
- **Questions with few plausible answers.** A number of potential questions, such as "*who is your favorite superhero?*" have very few possible answers. An empirical study by Rabkin of questions used in practice found that 40% had trivially small answer spaces [27]. User-chosen questions appear even worse: Just and Aspinall found that the majority of users choose questions with trivially few plausible answers [21].
- **Publicly available answers.** Rabkin found that 16% of questions had answers routinely listed publicly in online social-networking profiles [27]. Even if users keep data private on social networks, inference attacks enable approximating sensitive information from a user's friends [24]. Other questions can be found in publicly available records. For example, at least 30% of Texas residents' mothers' maiden names can be deduced from birth and marriage records [15].
- **Social engineering.** Users may not appreciate that the information in their security questions is of critical security importance. Karlof et al. were able to extract answers to personal knowledge questions from 92% of users via email phishing in a 2009 study [22].
- **Social guessing attacks.** Users' answers may be easily available to partners, friends, or even acquaintances. Schechter et al. found in a laboratory study that acquaintances could guess 17% of answers correctly in five tries or fewer [28], confirming similar results from earlier user studies [16, 26].

While all of these threats are important, in this work we focus solely on the problem of questions with common answers and the risk of online guessing. We consider the main threat against an account recovery system for an online service to be an adversary that aims to compromise accounts en masse by trying to guess common answers for a large number of accounts. This threat model is similar to that for passwords which are used to defend against account hijacking via mass guessing attacks [11].

Therefore the main security criteria for secret questions is their resistance to *statistical* guessing attacks by an attacker with no knowledge of the individual user. This attack model has not yet been rigorously evaluated. For example, Schechter et al. [28] and Bonneau et al. [5] attempted to estimate guessing difficulty but did not have access to actual answers chosen by real users. It remains unknown exactly how users answer these questions in practice and how different users' answers are from a theoretical distribution such as the population-distribution of surnames. Similarly there has been no published study of how effective distributions built via crowd-sourcing will be at approximating real distributions.

2.1 Evaluation metrics

Throughout this paper we use *statistical guessing metrics* to estimate the difficulty for our online adversary to guess secret answers. We start by using metrics which model an *ideal* attacker who knows the precise distribution of answers across the entire population of

users, but has no per-user data. This ideal attacker provides a lower bound on security for any real attacker, hence we focus primarily on these metrics. We also use distribution comparison metrics to compare the difficulty of guessing secret answers with passwords and PIN codes. Finally we study how effective crowd-sourced distributions are at approximating the true distribution of answers.

Online guessing attacks: The simplest metric is "what percentage of users share the β most common answers?" An attacker able to make β guesses per account can expect to guess roughly this proportion of user's answers. Following the notation of Bonneau [3], we denote this proportion as λ_β for a given number of guesses β. This is the easiest metric to define and interpret when evaluating security against an online attacker who is limited to interacting with the genuine server and can only make a small fixed number of guesses before being locked out. In practice, Google limits users by default to three incorrect answers to prevent brute force attacks while accounting for typos and normal memory failures [8]. Other websites sometimes choose different guessing cutoffs, or require CAPTCHAs or other additional hurdles to increase the cost to attackers of making large numbers of guesses [1].

While Google rate-limits account recovery attempts and performs risk-analysis, to keep our analysis generic, we avoid making any particular assumption about an appropriate rate-limiting policy for personal knowledge questions as used in account recovery and instead present our data for various policies. In Table 1 we show λ_β for $\beta = 1$, 3, 10, 100, and 1000 as representative values.

Extended guessing attacks: We may also attempt to measure the difficulty of guessing a user's secret answer if the attacker is able to guess exhaustively up to their computational limits. Note that most websites have rate-limiting defenses to prevent such attacks. This scenario would only arise if no rate-limiting were in place, for example if a list of hashed answers was leaked from a compromised website, or if personal knowledge questions were used in offline application such as for encryption keys.

The simplest metric is the expected number of guesses before the correct answer is found for a randomly-chosen answer from the distribution. This is typically called *guessing entropy*. Unfortunately, this quantity can be deceivingly large in practice due to the presence of a small number of extremely difficult to guess answers, often due to users entering long random strings. For example, it was estimated that it actually takes over 2^{100} guesses to compromise an average password due to the presence of less than one in a million users choosing 128-bit random strings as passwords [3].

Instead, the best approach is to measure the expected number of guesses required for an adversary have a probability α of compromising a random user's account. This is denoted as G_α and called the α-guesswork. This quantity can also be converted into units of bits by comparing to a uniform distribution which would provide equivalent security (denoted $\tilde{G}_\alpha$) [3], which can be easier to interpret and compare.

A reasonable value for modeling the difficulty for an attacker of breaking a median user's account is $\tilde{G}_{0.5}$. Thus, we provide values of $\tilde{G}_\alpha$ for $\alpha = 0.1$,0.25, and 0.5 in Table 1. We also list the min-entropy H_∞, which represents the limit of $\tilde{G}_\alpha$ as $\alpha \to 0$ and is thus a lower bound on any guessing attack. It is also more simply defined as simply $H_\infty = -\lg_2(p_0)$ where p_0 is the probability of the most commonly chosen answer in the distribution.

Sample size & significance: We use Bonneau's method to determine when estimates for both λ_β and $\tilde{G}_\alpha$ are accurate using bootstrap re-sampling [4]. This technique determines, given a desired confidence level and accuracy, a cutoff point α_* at which estimates for these metrics can be accepted. Generally, we are able to estimate metrics λ_β and $\tilde{G}_\alpha$ at a given sample size for values of $\alpha < \alpha_*$ and β such that $\lambda_\beta < \alpha_*$, as these metrics depend only on higher-probability answers which are most accurately estimated in the sample. Estimating the difficulty of more extended guessing attacks becomes inaccurate because rarer answers may only be observed a small number of times in the sample and hence their frequencies are poorly estimated.

We evaluate using a confidence level of $p = 0.98$ and an error of 0.1, meaning we are 98% confident the relative error of our estimate compared to the true value is less than 0.1. For most of the metrics we estimate the expected relative error is far lower than this. While we can't publish the exact size of our sample size, the data considered contains hundreds of millions of data points and each question analyzed had over 1 million answers. For most of the metrics we wish to compute the sample size we obtained was large enough to compute them to within our desired accuracy. For those which the error was higher than our cutoff, mostly estimates of $\tilde{G}_{0.5}$ for distributions of phone numbers, we simply show '–' in Table 1.

Bonneau also introduced techniques to estimate metrics beyond this bound for password distributions by fitting a parametric model to the observed data [4]. For example, distributions of human-chosen answers are often assumed to be approximately a Zipfian or other simple power-law distribution, although these have proven to be a poor fit to real password distributions. A more complicated distribution (specifically a zero-truncated generalized inverse-Gaussian/Poisson distribution) was used by Bonneau to successfully estimate metrics for passwords [4]. However, this model has not been validated for distributions of answers to personal knowledge questions, hence in this work for simplicity we only use non-parametric estimations in the well-approximated region of the distribution.

3. STRENGTH AGAINST GUESSING

In this section we evaluate the guessing difficulty of distributions of answers provided by Google users over the last 5 years. We use the guessing metrics discussed in Section 2.1.

3.1 Statistical evaluation

Table 1 lists statistics for distributions of answers to several commonly used questions in different languages. Several general trends emerge. First, nearly all questions have very low min-entropy (and hence high λ_β for low values of β) meaning there exist answers which are very common and therefore useful for an attacker to guess. Given the ability to make 10 guesses, for example, an attacker would have at least a 2% chance of answering correctly for any of the questions studied and often over 10%. This suggests that (without further abuse detection systems in place) nearly all questions are vulnerable to *trawling attacks* [5] where an attacker makes a few guesses of common answers for a large number of accounts in hopes of compromising a significant number of (random) accounts. Even a single guess can yield between 0.8% to 19.7% success rate.

As visible in the Table 1, questions about taste such as "*Favorite food?*" are the least secure questions. Questions about places and people are also very weak with a success rate well above 1% in most cases for a single guess. Cultural differences emerge as guessing difficulty varies significantly between countries and languages. For example the "*Place of birth?*" distribution is 10 times less secure given one guess for answers chosen by Korean-speaking users compared to English-speaking users due to the concentration of Korean-speaking users in a few major cities.

Most questions are further highly vulnerable to extended guessing attacks, with $\tilde{G}_{0.5} < 20$ bits for most questions studied. This confirms that personal knowledge questions are completely insecure

question	lang.	*online guessing (success %)* λ_1	λ_3	λ_{10}	λ_{100}	λ_{1000}	*offline guessing (bits)* H_∞	$\tilde{G}_{0.1}$	$\tilde{G}_{0.25}$	$\tilde{G}_{0.5}$
names										
best friend's name	Spanish	1.3%	3.5%	7.8%	27.8%	61.1%	6.3	7.2	8.3	9.6
	French	0.7%	1.7%	4.5%	23.6%	62.4%	7.2	8.2	8.7	9.7
childhood best friend's name	English	0.4%	1.0%	2.7%	13.3%	40.5%	8.1	9.3	10.2	11.7
	Portuguese	1.0%	2.7%	6.4%	27.6%	58.8%	6.7	7.5	8.3	9.7
	Russian	1.9%	4.2%	9.4%	35.8%	65.4%	5.7	6.8	7.5	8.8
	Spanish	1.0%	2.8%	7.2%	28.9%	63.0%	6.6	7.4	8.2	9.4
father's middle name	Chinese	2.2%	6.0%	15.0%	49.9%	85.7%	5.5	5.8	6.6	7.5
	English	2.7%	6.6%	14.6%	40.3%	64.9%	5.2	5.8	6.6	8.6
	Portuguese	2.7%	6.7%	15.4%	44.6%	73.8%	5.2	5.8	6.5	7.9
	Spanish	3.8%	8.9%	21.3%	58.1%	83.8%	4.7	5.1	5.7	6.8
first teacher's name	Arabic	7.7%	14.4%	23.7%	37.4%	61.4%	3.7	4.1	5.7	9.1
	English	0.4%	1.1%	2.8%	9.7%	26.7%	8.0	10.0	11.6	14.1
	Russian	1.5%	4.3%	11.3%	39.4%	61.4%	6.1	6.4	7.0	8.9
	Portuguese	6.0%	8.5%	13.0%	34.7%	65.2%	4.1	5.6	7.5	8.9
	Spanish	2.9%	5.3%	11.3%	37.6%	69.5%	5.1	6.4	7.4	8.5
first manager's name	English	0.9%	2.7%	5.9%	21.6%	46.8%	6.7	7.9	9.0	11.1
favorites										
favorite food	English	19.7%	26.0%	36.5%	59.4%	76.8%	2.3	2.3	3.4	5.9
	Korean	11.8%	30.5%	43.2%	70.0%	85.7%	3.1	3.1	3.3	5.0
	Spanish	7.3%	15.4%	28.1%	59.2%	80.1%	3.8	4.1	4.9	6.4
places										
place of birth	English	1.3%	3.0%	6.9%	24.6%	58.8%	6.2	7.5	8.6	9.9
	Korean	12.0%	25.0%	39.0%	70.1%	87.8%	3.1	3.1	3.7	5.4
high school	English	0.5%	0.9%	1.9%	7.6%	22.6%	7.7	10.7	12.2	13.6
numbers										
first telephone number	Arabic	2.9%	6.3%	13.0%	28.6%	38.5%	5.1	5.9	7.7	15.5
	Chinese	1.2%	2.4%	4.5%	7.9%	10.2%	6.3	12.9	–	–
	Korean	1.2%	2.8%	6.4%	13.0%	18.3%	6.3	8.4	13.8	–
	English	0.4%	1.0%	2.5%	5.5%	8.4%	7.9	14.9	21.5	–
	Portuguese	0.9%	2.2%	4.3%	10.8%	16.7%	6.8	9.5	15.6	–
	Russian	2.4%	4.2%	7.3%	14.6%	21.7%	5.4	7.8	13.5	–
	Spanish	0.6%	1.5%	4.4%	9.7%	14.1%	7.4	10.1	17.9	–
frequent flyer number	English	4.2%	7.8%	13.6%	26.8%	38.6%	4.6	5.5	8.1	13.5
	Portuguese	5.8%	11.8%	21.6%	43.2%	63.2%	4.1	4.6	5.8	8.4
vehicle registration number	English	0.8%	1.5%	2.6%	5.6%	11.2%	7.0	12.7	14.8	–
library card number	English	2.3%	6.4%	12.2%	22.5%	33.0%	5.4	5.9	9.3	15.5
user-chosen secrets (baseline)										
password (RockYou)	–	0.9%	1.4%	2.1%	4.6%	11.3%	6.8	12.8	15.9	19.8
password (Yahoo!) [3]	–	1.1%	1.4%	1.9%	3.6%	8.3%	6.5	14.0	17.6	–
4-digit PIN (iPhone) [6]	–	4.3%	9.2%	14.4%	29.3%	56.4%	4.5	5.2	7.7	10.1

Table 1: Guessing difficulty estimates for distributions of answers to various challenge questions. λ_β represents the success rate of an attacker limited to β guesses, while $\tilde{G}_\alpha$ represents the average amount of work (in bits) to compromise a proportion α of users in an offline attack. The min-entropy H_∞ is also included as a lower-bound on the difficulty of any offline attack. Values which could not be estimated accurately due to an insufficient sample size are represented by '–' (see Section 2.1).

against an attacker able to perform offline brute-force as 20 bits worth of brute force can typically be performed in less than one second and will be effective against over half of the population for most questions.

The impact of untruthful answers: The most surprising observation is that even on questions where answers are presumably unique (e.g. phone number, frequent flyer number) the distribution still contains answers that are shared by a significant fraction of users. For example 4.2% of English-speaking users have the "same" frequent flyer number and 0.4% have the same phone number. These untruthful answers significantly weaken the potentially most secure questions. The security of these questions against extended guessing attacks is considerably higher, as they are very difficult to guess for users who answer honestly. However for certain combinations of language and question this degradation makes those questions even less secure that the one based on name and places. For example, 2.9% of the Arabic-speaking users used the same phone number as their answer and 2.4% of the Russian-speaking users did. As we will discuss in Section 5 users often provide untruthful answers in an attempt to make the answer more secure. Perhaps they achieve this against targeted attackers (for example, a friend who knows their phone number) but this harms security against untargeted guessing.

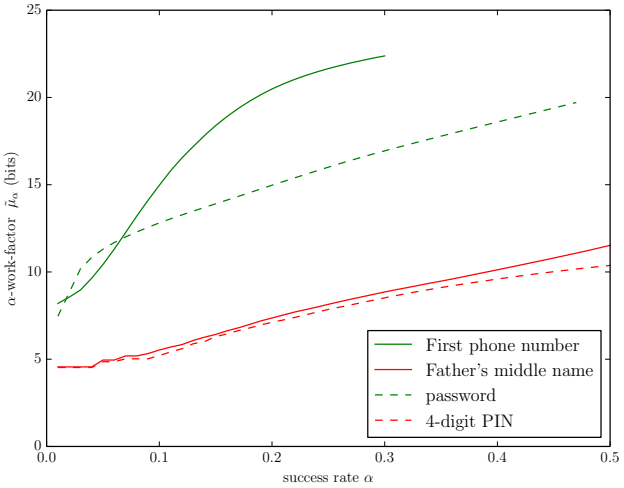

Figure 1: Guessing curve comparison for two representative personal knowledge questions and two user-chosen secrets (passwords from the leaked RockYou dataset and PINs from the leaked iPhone dataset [6].

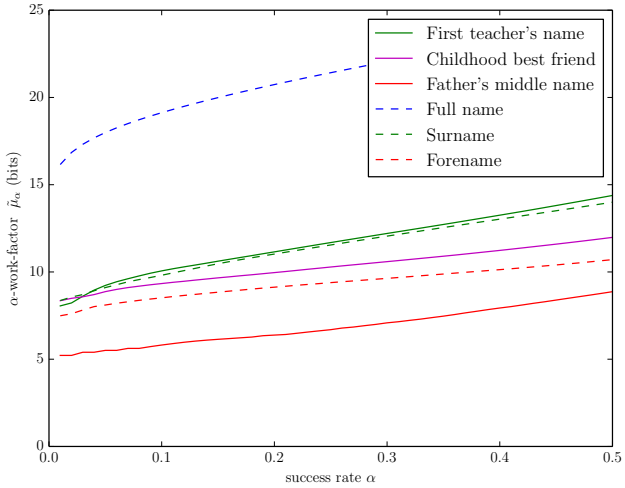

Figure 2: Guessing curve comparison for three personal knowledge questions asking for human names with population statistics on full human names, surnames and forenames collected from a 2009 crawl of Facebook's public directory [5]

3.2 Comparison to user-chosen secrets

We can compare the collected distributions of answers to personal knowledge questions with previously collected statistics on user choices of passwords and PINs. The bottom of Table 1 lists statistics from three datasets as baselines. First is the password distribution leaked from RockYou, a social gaming website, in 2009. This data set consisted of 32 million plaintext passwords and has been frequently used in password research. Second is the password distribution collected by Bonneau from Yahoo! in 2012 [3].[1] Finally there is a distribution of user-chosen 4-digit PINs leaked by an iPhone application developer in 2012 [6].

As a rule-of-thumb, most of the name-based personal knowledge questions produce a distribution with equivalent security to a user-chosen PIN and considerably less security than that of passwords. Numerical questions produce distributions with roughly equivalent security against guessing as user-chosen passwords. This arguably makes them effective for the purposes of account recovery as their use would be no more of a security risk than passwords which are the primary authentication mechanism.

This is shown graphically in Figure 1, showing the complete *guessing curve* for a representative high and low-security personal knowledge question alongside passwords and PINs. Note that "What is your father's middle name?" produces a statistically very similar distribution of answers to user-chosen PINs throughout. However, "What was your first phone number?" produces a very differently curve from a distribution of passwords. The weakest answers to the phone-number question are roughly similar to the weakest passwords, after which the phone numbers quickly become *more difficult* to guess than passwords.

3.3 Comparison to public distributions

We can also compare our figures to expected distributions based on published statistics, particularly for questions asking for human names for which population-wide statistics are readily available.[2] Bonneau et al. [5] evaluated many published distributions of human names; for this work we will take the largest distribution collected in that research, a crawl of over 100 million names in a public directory of Facebook users with separate distributions for complete names, surnames (last or family names) and forenames (first or given names). In Figure 2 we compare these name distributions to distributions of answers to three name-based questions: "*What is your father's middle name?*", "*What was your first teacher's name?*" and "*What was your childhood best friend's name?*"

Interestingly, the answer distribution for "*What was your first teacher's name?*" is statistically very similar to the population distribution of surnames, suggesting most users identify their first teacher by surname. By comparison, "*What was your childhood best friend's name?*" is somewhat between surnames and forenames (perhaps because there are multiple ways of naming this individual). Finally, "*What is your father's middle name?*" is significantly below the natural distribution of forenames, likely due to a significant number of users answering inaccurately.

None of the questions are close to the distribution of full names (first and last name combined), which is considerably stronger than any of the distributions we observed for personal knowledge questions. This suggests that if users could be asked to enter the full name of their childhood best friend, security might be significantly higher, though it appears users rarely do so.

3.4 Effectiveness of crowdsourcing attacks

So far all of the analysis in this section models an ideal attacker who knows the precise distribution of answers in the population; hence it can be considered a lower-bound on security since a real attacker may guess using an inaccurate approximation. To measure how easy it would be for an attacker to approximate the real distributions using crowd-sourcing services we asked 1000 users on CrowdFlower to answer the following two questions: "*Favorite food?*" and "*Father's middle name*". We then compared the effi-

[1]This password distribution was collected anonymously without observing user-chosen passwords; only the password frequencies were observed.

[2]We could attempt a similar exercise for telephone numbers or frequent flier numbers but this would not be interesting; the genuine distribution is nearly flat (with few users sharing an answer).

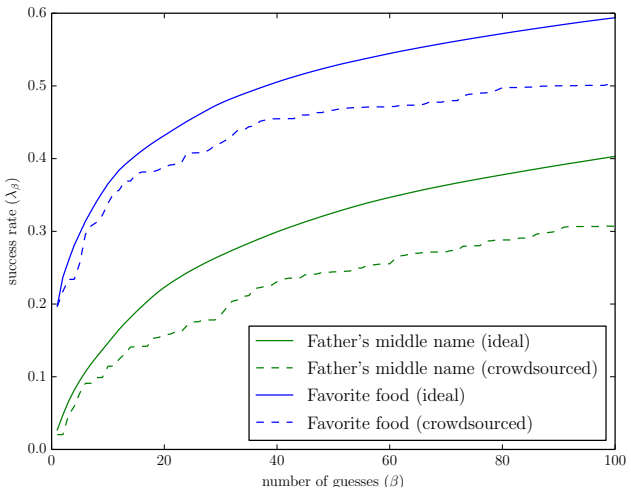

Figure 3: Effectiveness of a crowdsourcing attack. The solid lines demonstrate an attacker's success λ_β for up to $\beta = 100$ guesses. The dashed lines represent an attacker using an approximate distribution obtained by crowdsourcing.

ciency of those approximate distributions to an ideal attacker. As visible in Figure 3, in both cases, the attacker using a crowdsourced distribution does very well: with up to 100 guesses efficiency is at least 75% as high for the father's middle name distribution and at least 80% as high for the favorite food distribution. Thus, we conclude that it is not difficult for attackers to learn a reasonable approximation of the true answer distribution as our simple crowdsourcing attack cost only $100 and took less than a day.

4. QUESTION MEMORABILITY

In this section, we analyze the ability of users to remember their secret answers based on a random sample of 11 million account recovery claims that occurred in 2013, the year that Google started preferring alternative recovery methods and stopped collecting personal knowledge questions during web account signup. We chose this period to have meaningful temporal data: that is, claims that occurs shortly after setting the secret question and its answer. For every slice of data discussed in this section the number of claims by bucket is at least 500, and well in the ten of thousands in most cases. Note that the data presented in this section was filtered to remove blatantly fraudulent recovery attempts and restricted to one claim per user to prevent users that routinely use the account recovery process from skewing the data.

4.1 Time between enrollment and use

We start by looking at the time between users enrolling a secret question and using it in an account recovery claim. Figure 4 shows the distribution of time since enrolling a secret question for all of the claims in 2013 divided into ten deciles; the distribution is almost exactly linear. This demonstrates that people are no more likely to recover their accounts early or late in the lifetime of their secret question, for example because they are not used to their account password for old accounts and are more likely to have forgotten it.

There is one caveat to this linear relation: the first few hours after enrollment are much more likely to see the user attempt to use it for account recovery. This is visible in Figure 5 which shows an hour-by-hour breakdown of account claims in the first week after enrolling a new secret question This surge is related to questions being set during initial account creation and people not remembering the password they just set when logging in for the first time.

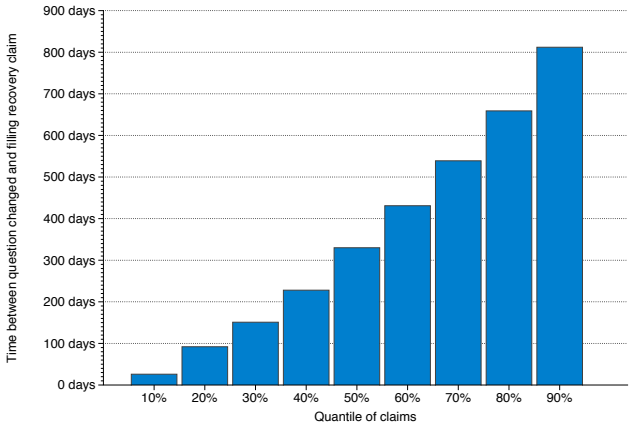

Figure 4: Quantile distribution of the time between the secret answer is set and the time of the claim in day.

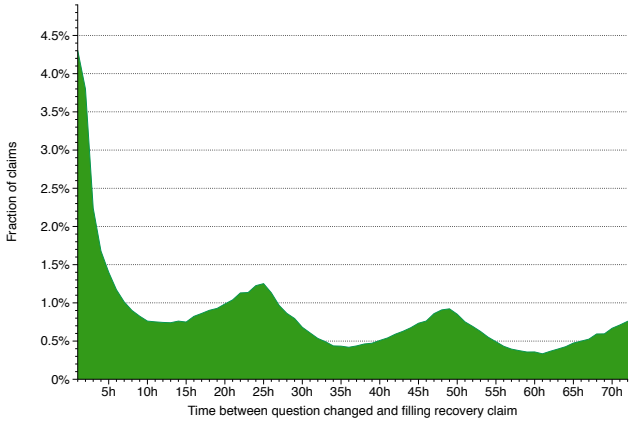

Figure 5: Overall frequency of account recovery claims for the first 72 hours after enrollment

question	overall success	success within n months 1	3	6	12
City of birth?	80.1%	83.9%	79.9%	79.2%	79.5%
Father's middle name?	75.6%	85.9%	75.7%	74.4%	74.3%
Childhood best friend?	68.5%	82.9%	65.0%	64.6%	63.7%
High school name?	67.3%	78.8%	62.8%	62.6%	61.4%
First phone number?	55.2%	70.0%	55.4%	53.3%	50.1%
Favorite food?	48.0%	73.6%	52.8%	50.1%	46.6%
First teacher's name?	47.1%	71.7%	45.9%	43.2%	39.8%
Library card number?	22.5%	49.6%	24.3%	19.9%	17.7%
Frequent flyer number?	9.0%	32.1%	8.5%	6.4%	6.4%

Table 2: Success rate for English questions broken-down by number of months since enrollment

4.2 Effects of question type on memorability

We chart the success rate of users attempting to answer their personal knowledge questions in Table 2 for various English-language questions. Success rates are broken down by the number of months since the question was enrolled; for all questions we observed that

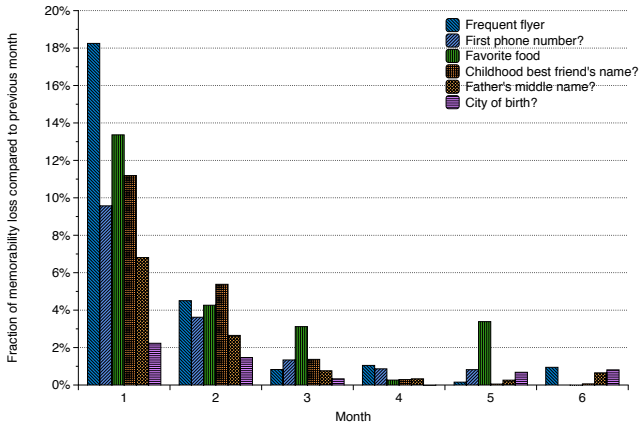

Figure 6: Loss of memorability compared to previous months for various English questions

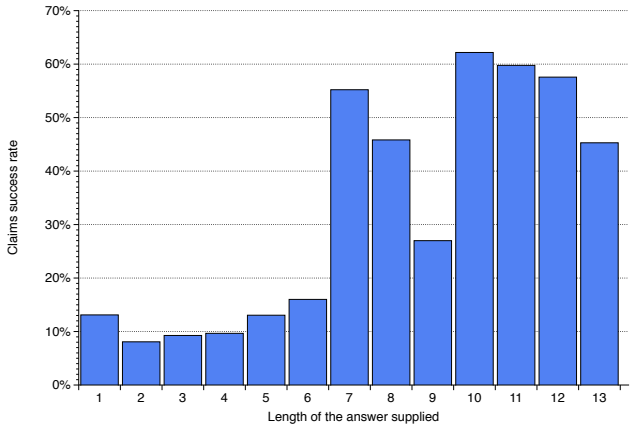

Figure 7: Success rate for the question "First phone number?" for US users broken down by secret answer length

memorability monotonically declined with time since the question was enrolled.

Overall, we can note immediately that the success rate is below 80% for all questions, going against the common wisdom that personal knowledge questions provide a highly reliable means of authenticating users. Generally, questions involving names and places fared fairly well with success greater than 50%, with library card and frequent flyer number faring very poorly and being unreliable questions and suffering the steepest drop-off in memorability over time. Sadly the memorability order is inversely proportional to the security of these questions, as seen in Table 1. The mismatch between security and memorability highlights why personal knowledge questions are inherently difficult to use in practice.

This gap between security and memorability gets even wider when considering that what people remember the most are, unsurprisingly, the things closest to their heart: city of birth and father's middle name. As discussed in Section 2, such distribution are also the easiest to find from public records or online social networks.

Telephone numbers appear to offer the best balance, with memorability over 50% and security comparable to passwords (see Section 3.2). Of course, phone numbers are also liable to exist in public records and are likely known by many of a user's social contacts.

4.3 Impact of time

Table 2 showed a clear trend of decline in memorability over time. In Figure 4.3 we drill down further into this trend by plotting the loss of accuracy users suffer for 5 popular English language questions over the first 6 months of use. This graph suggests the decline in memorability is not gradual or linear but instead users suffer a very sharp decline within 1 month of enrollment for all questions. This is particularly true for "favorite food" and "childhood best friend," questions which are not necessarily factual and to which users may change their minds or have to choose from among several possibilities at the time of enrollment.

Curiously, this is also true for "frequent flyer number" even though this should be an unambiguous fact that is unlikely to change within one month. We speculate that this is due to the large number of inaccurate answers or "don't know" answers, either of which are susceptible to be forgotten quickly after being registered.

		months since registration			
language	country	1	3	6	12
English	US	70.0%	55.4%	53.8%	49.8%
English	UK	68.4%	52.1%	49.7%	44.2%
German	Germany	69.2%	44.6%	42.3%	37.7%
Spanish	US	70.0%	59.2%	59.1%	57.6%
Spanish	Spain	68.6%	47.5%	41.5%	37.9%
French	France	75.6%	59.2%	58.5%	57.0%

Table 3: Recall for the questions "*First phone number?*" for various languages and countries

4.4 Impact of inaccurate answers

As discussed in Section 3, a key reason behind the insecurity of secret question is that people supply inaccurate answers. We hypothesize that this is in fact also a driver of users failing to remember their answers correctly in many cases, as inaccurate answers no longer represent a true memory the user holds.

To measure how this behavior also impairs users' ability to remember their answer, we looked at the success rate of users attempting to answer the question "*what was your first phone number?*" broken down by the length of the answer they initially enrolled. As shown in Figure 7 the set of answers that have plausible length for a phone number, namely 7 or 8 digits (a North American number without an area code, possibly with a space) or 10–13 characters (a North American number with an area code and possibly spaces, dashes or parentheses), exhibit a significantly higher memorability. Note that when verifying users' answers, spaces and punctuation characters are stripped out so the exact format used will not matter if the digits are all correct. Answers with a plausible length of 10 have an accuracy of 62% vs an accuracy of only 28% for answer which have a length of 9. Answers of fewer than 6 characters all had less than a 20% memorability.

This suggests that users who answer accurately are more successful at recovering their account. One possible explanation is that these users care more about security or have a better memory to begin with, but it seems far more likely to indicate that many users can't remember which inaccurate answer they may have provided. Thus, in addition to greatly harming security, inaccurate answers are also a significant problem for memorability.

language	country	months since registration 1	3	6	12
English	US	85.9%	75.7%	75.1%	74.4%
English	UK	81.2%	68.0%	64.6%	64.1%
German	Germany	81.9%	68.0%	64.4%	64.4%
Spanish	US	88.3%	81.3%	82.2%	80.8%
Spanish	Spain	85.3%	71.7%	70.2%	62.8%
French	France	56.8%	39.6%	37.6%	36.9%

Table 4: Recall for the questions "*Father's middle name?*" for various languages and countries

4.5 Impact of cultural differences

To understand the impact of cultural difference on secret questions we compared the success rate for various languages and countries for the same questions. Usually the question is semantically the same, but translated, with one exception: "*Father's middle name*" was replaced by "*Primer apellido del padre?*" which translate as "*Father's first surname?*" in Spanish as it is customary in many Spanish-speaking countries for individuals to use two surnames.

Overall as visible in Table 3 and Table 4, the same question in the same language can have a different recall for various countries. For example US English-speaking users have an easier time remember their first phone number than UK English speakers: 49.8% vs 44.2% after 12 months. The difference between the two is even wider for the question "*Father's middle name?*" where the recall gap reach 10% after 12 months: 74.43% vs 64.12%. The gap between various language/country groups can be very drastic. In particular after a few months the gap between the best performing country/language and the worst is as high as 44%: 36.9% for France/French users vs. 80.8% for US/English users for the question "*Father's middle name?*". Finally it is worth noting that depending on the question the top performing country/language groups are drastically different. France has the best recall for phone number and the worst for father's middle name. These shifts demonstrate that many cultural differences exist which are hard to predict or account for when designing a recovery system that needs to internationalize.

5. UNDERSTANDING USER PERCEPTIONS

In this section we discuss the result of a survey we ran on the US population to better understand users' perceptions of secret questions. The survey was using Google Consumer Surveys [30], a micro-survey platform for asking web users a small number of short questions as a replacement for advertisements.

5.1 Motivation for untruthful answers

As discussed in previous sections, one of the driving factors behind secret questions' poor security (Section 3.1) and usability (Section 4.4) is that a significant fraction of users provide untruthful answers. To better understand this behavior we ran a survey that asked as a filtering question "*When creating your primary email account, how did you answer the secret questions that are used to recover the password?*". For those that admitted to providing fake answers we asked why in a follow-up multi-choice question. As seen in Figure 8 two main reasons reported for providing a fake answer were to improve security (37% of respondents) and to make the answer easier to remember (15%), though the effect is the exact opposite. We also observed that a good fraction of the respondent (31.9%) didn't provide the real answers for privacy reasons. Addressing privacy concerns for various account recovery options is an open question.

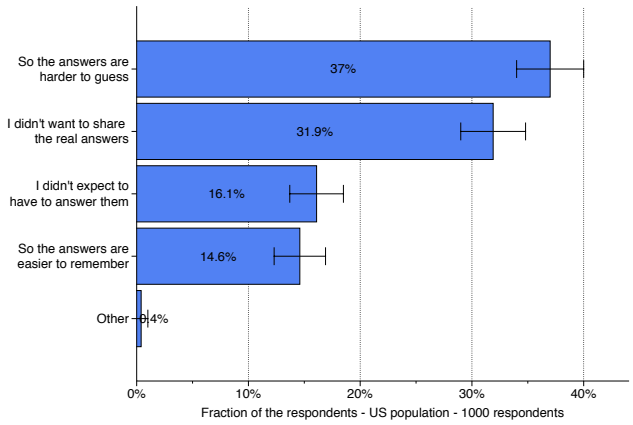

Figure 8: Survey answers for the question "*Why did you provide fake answers to your password recovery question?*"

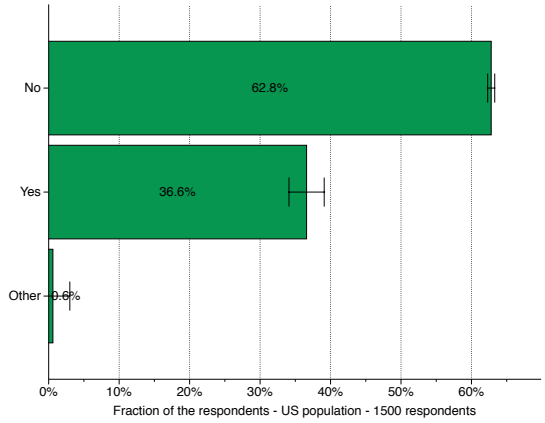

Figure 9: Survey responses for the questions "*Have you ever thought that someone might try to break into your primary personal email account by trying to reset your password?*"

.

5.2 Perceptions of security

Another hypothesis we had about why people provide very easy to guess answers was a false sense of security. To confirm this hypothesis we ran two additional independent surveys. In the first one we asked: *"Have you ever thought that someone might try to break into your primary personal email account by trying to reset your password?"*. As seen in Figure 9, the vast majority of the respondents (62.8%) never considered the possibility that their security question could be used against them. This lack of awareness potentially contribute to a false sense of security that leads users to not pay attention to secret questions.

In a second survey we asked respondents to answer Likert questions to compare how much trust they had in SMS recovery vs. personal questions. As seen in Figure 10, people have more trust in secret questions security than in SMS security which is opposed to reality. Overall this distorted perception of secret questions is yet one more reason to move away from security questions as people are not paying attention to them and are over-confident. Shifting their perception and habits would be very challenging and the energy seems better invested in convincing users to adopt two-factor authentication and switch to better recovery mechanisms.

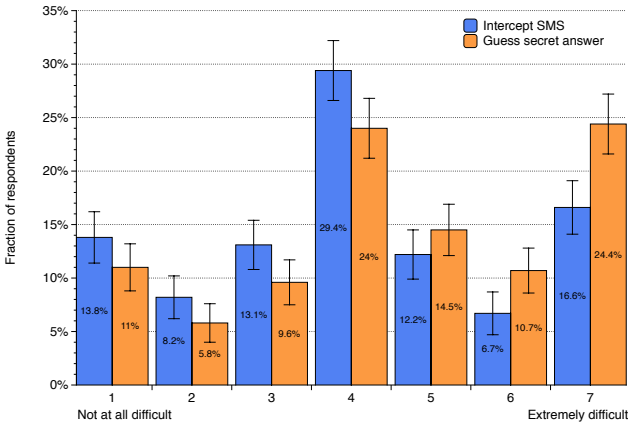

Figure 10: Survey responses for the question "*How difficult would it be for a hacker to _____ to break into your primary personal email account?*"

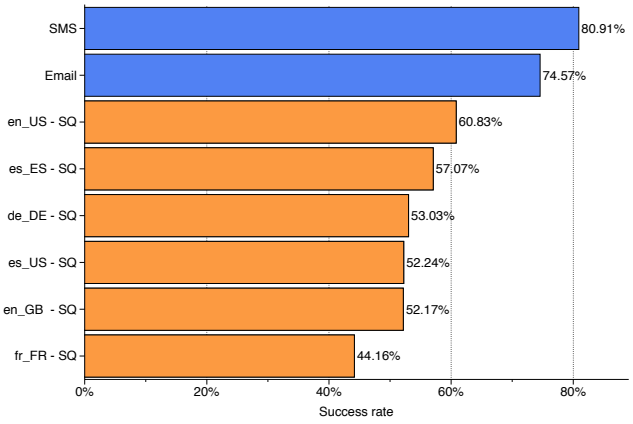

Figure 11: Success rate of various recovery options

6. CONCLUDING DISCUSSION

We analyzed the first large-scale empirical data on secret questions, based on their deployment at Google. Hundreds of millions of secret answers and millions of account recovery claims clearly demonstrate that secret questions have poor security and reliability. For reliability, we can directly compare to SMS and email-based recovery and find personal questions are inferior. The fact that secret questions are relatively less secure led Google to prefer those alternative options and only use secret questions in conjunction with other signals to compensate for their weaknesses.

6.1 Alternatives deployed at Google

Figure 6.1, summarizes the success rate of SMS and email-based recovery compared to secret questions. These statistics were computed based on a full month of account recovery claims. As visible in this figure, SMS recovery has a success rate which is 20% better than even the most successful secret answer language/population bucket (80.9% vs 60.8%). Similarly, email-based recovery increases the odd of a successful recovery by 14.5%.

Beside being more reliable, SMS recovery is also be preferred over email based recovery due to several additional security benefits. First, people might use the same password for their recovery email address as their Google account which they have forgotten. Second, some email providers, including Microsoft [31], expire email addresses after some period of inactivity and allows anyone to register them again, making email recovery sometimes unreliable. As of 2014, we estimate that 7% of the secondary emails our users provided for recovery have since been recycled.

6.2 Other potential alternatives

Beside SMS and email recovery, which are currently deployed, several potential alternatives have been proposed elsewhere. To counter statistical guessing, Jakobsson et al. developed "preference-based authentication" in 2008 [18, 19, 17]. In this scheme, users choose a number of items (16 is suggested) which they strongly like or dislike from a large set of items such as "rap music" or "vegetarian food." These preferences are claimed to not exist in online databases or public records, and a user study suggests a negligible false negative rate can be achieved while limiting statistical guessing to a 0.5% chance of success [19]. However, despite improvements [19] preference-based schemes require considerably more time to enroll users and authenticate them than individual questions making them less attractive to deploy in practice.

Graphical password schemes have also frequently been proposed as a potential alternative for backup authentication [2], particularly recognition-based schemes in which a user is asked to identify previously-seen images from a set of candidates [12]. Such schemes can be designed with firm security guarantees as the user's set of images is randomly chosen. However, while there is evidence that such schemes are more memorable than text passwords [9], they still require additional user training compare to personal knowledge questions and have not seen significant deployment.

Several other proposals have been made to automatically generate questions and answers based on data stored about users. Nosseir et al. suggested querying users' browsing history or location history to generate harder-to-guess questions [25], though such an approach appears to inevitably leak private data. A related proposal specifically for social networks is to require users to identify friends in tagged photographs [32]. While this has been effective in practice in the context of social networks, the risk of face recognition software and publicly-available photos may mean the security of this scheme is too weak for higher security applications [23].

Another proposal is requiring users to select a set of trusted friends to ask as delegates. In the event of a forgotten password, the user must contact a designated threshold of these delegates to receive one-time tokens from the server. Brainard et al. first proposed this idea in 2006 as "vouching-based" authentication [7]. A follow-up usability study by Schechter et al. found that only about 71% of participants could execute this scheme correctly and social-engineering attacks worked against about 10% of users [29]. It has been suggested that in place of an explicit user-conducted protocol, authentication could be performed automatically by communicating with the mobile devices of nearby users to establish a user's social context [13].

6.3 Secret questions' continued role

The ability to quickly confirm a user's identity when we suspect an account hijacking attempt has become an essential part of our login risk analysis system [11]. While Google prefers SMS and email recovery, no mechanism is perfect. For example, SMS will fail if the user doesn't have access to their phone while traveling abroad. In the context of a risk analysis system taking multiple signals into account, we have found personal knowledge questions can still be a useful lightweight signal when the risk level is considered low. Finding more identity confirmation questions that are both secure and easy to answer is an open question.

7. REFERENCES

[1] Mansour Alsaleh, Mohammad Mannan, and P.C. van Oorschot. Revisiting defenses against large-scale online password guessing attacks. *IEEE Transactions on Dependable and Secure Computing*, 2012.

[2] Robert Biddle, Sonia Chiasson, and P.C. van Oorschot. Graphical Passwords: Learning from the First Twelve Years. Technical Report TR-11-01, Carleton University, 2011.

[3] Joseph Bonneau. *The science of guessing: analyzing an anonymized corpus of 70 million passwords*, May 2012.

[4] Joseph Bonneau. *Guessing human-chosen secrets*. PhD thesis, University of Cambridge, May 2012.

[5] Joseph Bonneau, Mike Just, and Greg Matthews. What's in a name? Evaluating statistical attacks against personal knowledge questions. *Financial Cryptography*, 2010.

[6] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A birthday present every eleven wallets? The security of customer-chosen banking PINs. *Financial Cryptography*, 2012.

[7] John Brainard, Ari Juels, Ronald L. Rivest, Michael Szydlo, and Moti Yung. Fourth-Factor Authentication: Somebody You Know. *CCS '06: The 13th ACM Conference on Computer and Communications Security*, 2006.

[8] Sacha Brostoff and Angela Sasse. "Ten strikes and you're out": Increasing the number of login attempts can improve password usability. *CHI Workshop on HCI and Security Systems*, 2003.

[9] Sacha Brostoff and M. Angela Sasse. Are Passfaces More Usable Than Passwords? A Field Trial Investigation. *People and Computers XIV: Usability or Else!: HCI 2000*, 2000.

[10] Julie Bunnell, John Podd, Ron Henderson, Renee Napier, and James Kennedy-Moffat. Cognitive, associative and conventional passwords: Recall and guessing rates. *Computers & Security*, 1997.

[11] Elie Bursztein, Borbala Benko, Daniel Margolis, Tadek Pietraszek, Andy Archer, Allan Aquino, Andreas Pitsillidis, and Stefan Savage. Handcrafted Fraud and Extortion: Manual Account Hijacking in the Wild. *Internet Measurement Conference*, 2014.

[12] Rachna Dhamija and Adrian Perrig. Déjà vu: A user study using images for authentication. *USENIX Security Symposium*, 2000.

[13] A.D. Frankel and M. Maheswaran. Feasibility of a Socially Aware Authentication Scheme. *CCNC '09: IEEE Consumer Communications and Networking Conference*, 2009.

[14] Simson L. Garfinkel. Email-Based Identification and Authentication: An Alternative to PKI? *IEEE Security & Privacy Magazine*, 1(6), 2003.

[15] Virgil Griffith and Markus Jakobsson. Messin' with Texas: Deriving Mother's Maiden Names Using Public Records. *Applied Cryptography and Network Security*, 2005.

[16] William J. Haga and Moshe Zviran. Question-and-Answer Passwords: An Empirical Evaluation. *Information Systems*, 16(3):335–343, 1991.

[17] Markus Jakobsson and Hossein Siadati. Improved visual preference authentication. *Workshop on Socio-Technical Aspects in Security and Trust (STAST)*, 2012.

[18] Markus Jakobsson, Erik Stolterman, Susanne Wetzel, and Liu Yang. Love and Authentication. *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[19] Markus Jakobsson, Liu Yang, and Susanne Wetzel. Quantifying the Security of Preference-Based Authentication. *ACM Workshop on Digital Identity Management (DIM)*, 2008.

[20] Mike Just. Designing and Evaluating Challenge-Question Systems. *IEEE Security & Privacy Magazine*, 2(5), 2004.

[21] Mike Just and David Aspinall. Personal Choice and Challenge Questions: A Security and Usability Assessment. *SOUPS '09: The 5th Symposium on Usable Privacy and Security*, 2009.

[22] Chris Karlof, J. D. Tygar, and David Wagner. Conditioned-Safe Ceremonies and a User Study of an Application to Web Authentication. *SOUPS '09: The 5th Symposium on Usable Privacy and Security*, 2009.

[23] Hyoungshick Kim, John Tang, and Ross Anderson. Social Authentication: Harder than it Looks. *Financial Cryptography*, 2012.

[24] Jack Lindamood and Murat Kantarcioglu. Inferring Private Information Using Social Network Data. Technical Report UTDCS-21-08, University of Texas at Dallas Computer Science Department, 2008.

[25] A. Nosseir, R. Connor, and M.D. Dunlop. Internet Authentication Based on Personal History—A Feasibility Test. *ACM Customer Focused Mobile Services Workshop*, 2005.

[26] Rachael Pond, John Podd, Julie Bunnell, and Ron Henderson. Word Association Computer Passwords: The Effect of Formulation Techniques on Recall and Guessing Rates. *Computers & Security*, 2000.

[27] Ariel Rabkin. Personal knowledge questions for fallback authentication: Security questions in the era of Facebook. *SOUPS '08: The 4th Symposium on Usable Privacy and Security*, 2008.

[28] Stuart Schechter, A. J. Bernheim Brush, and Serge Egelman. It's No Secret: Measuring the security and reliability of authentication via 'secret' questions. *2009 IEEE Symposium on Security and Privacy*, 2009.

[29] Stuart Schechter, Serge Egelman, and Robert W. Reeder. It's Not What You Know, But Who You Know: A social approach to last-resort authentication. *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2009.

[30] Victoria Schwanda-Sosik, Elie Bursztein, Sunny Consolvo, David A Huffaker, Gueorgi Kossinets, Kerwell Liao, Paul McDonald, and Aaron Sedley. Online microsurveys for user experience research. *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, 2014.

[31] The Next Web. Microsoft can recycle your outlook.com email address if your account becomes inactive. `http://tnw.co/1sWsNAU`, 2013.

[32] Sarita Yardi, Nick Feamster, and Amy Bruckman. Photo-Based Authentication Using Social Networks. *WOSN '08: The 1st Workshop on Online Social Networks*, 2008.

[33] Moshe Zviran and William J. Haga. A Comparison of Password Techniques for Multilevel Authentication Mechanisms. *Computer Journal*, 36(3):227–237, 1993.