What's in a Name?
Evaluating Statistical Attacks on Personal Knowledge Questions

Joseph Bonneau[1], Mike Just[2], Greg Matthews[2]

[1] University of Cambridge
[2] University of Edinburgh

Abstract. We study the efficiency of statistical attacks on human authentication systems relying on personal knowledge questions. We adapt techniques from guessing theory to measure security against a trawling attacker attempting to compromise a large number of strangers' accounts. We then examine a diverse corpus of real-world statistical distributions for likely answer categories such as the names of people, pets, and places and find that personal knowledge questions are significantly less secure than graphical or textual passwords. We also demonstrate that statistics can be used to increase security by proactively shaping the answer distribution to lower the prevalence of common responses.

1 Introduction

Secret knowledge stored in human memory remains the most widely deployed means of human-computer authentication. It is often referred to as *something you know* in contrast to biometrics (something you *are*) or hardware tokens (something you *have*). While human memory is limited, the high deployment costs of alternatives mean we will continue to rely on it for the foreseeable future.

The most common human-memory systems require recalling data specifically remembered for authentication. Passwords and PINs are the most well-known, but there exist a variety of graphical and textual schemes to aid in recalling secret data [28,26,19,6]. Among other problems, passwords are forgotten frequently enough [28] that many deployed systems also use personal knowledge for backup authentication. In contrast to passwords, personal knowledge questions such as "who was my first-grade teacher?" query facts known independently of the system so they are hoped to be recalled successfully when passwords fail.

In the majority of online banking, e-commerce, webmail and social networking websites, users register a question-answer pair on enrolment which can later be used to authorise a password reset. These systems can be no more secure than the difficulty of guessing the answers to these questions. This risk was highlighted in the past year as hackers exploited personal knowledge questions to compromise accounts of politician Sarah Palin and top executives at Twitter.

Despite their ubiquity, personal knowledge questions have received relatively little attention from the security community until recently. User studies have demonstrated the ability of friends, family, and acquaintances to guess answers correctly [23,12,21,27], while other research has found some questions used in practice have a tiny set of possible answers [14,22]. Many common questions have also been shown to have answers available in public databases or online social networks [16]. For example, at least 30% of Texas residents' mothers' maiden names can be deduced from birth and marriage records [11].

Designers may be able to avoid easily looked-up questions, but it remains an open question as to how secure typical questions are against a *statistical attacker* that attempts to break into a small fraction of anonymous accounts by guessing the most likely answers. While this threat has been briefly touched on in previous research [14,23], we contribute a formal security model based on the information-theoretic model of guessing developed over the past decade. We then examine a range of public statistics that we collected to bound the efficiency of statistical attacks. Our results show most questions to be highly insecure, calling into serious question the continued use of personal knowledge questions.

2 Security Model

2.1 Authentication Protocol

Most deployed systems use personal knowledge questions in a simple challenge-response protocol. The party seeking access, called the *prover* or *claimant*, first sends its identity i to the *verifier*. The verifier then responds with a challenge question q, to which the prover sends back an answer x. Unlike most challenge-response protocols, the prover's secret knowledge x is usually revealed to the verifier. Replay attacks can be partially addressed by having the verifier include a nonce r along with q, and having the prover respond with $\mathsf{H}(x, q, r)$ for some one-way function H. However, an eavesdropper still gains the ability to perform offline search of likely values of x using H as an oracle (and as we shall see, few personal-knowledge questions are resistant to offline search).

Additionally, while the challenge from a verifier is typically a fresh random nonce for cryptographic challenge-response, the set Q of personal knowledge questions registered with the verifier is often very small or even a single question. Some non-traditional question types may increase $|Q|$, such as "preference-based authentication" [13], but the upper limit appears low due to the fundamental requirement of human effort to select and answer questions on enrolment.

Finally, unlike many challenge-response protocols, the verifier must maintain a counter t_i of failed authentication attempts from each prover i to limit the number of guesses an attacker can make. Such a protocol is said to be *online*, in contrast to stateless protocols in which the attacker can make as many guesses as bandwidth allows. Offline protocols rarely use personal knowledge questions due to the difficulty of preventing brute-force attacks, though systems have been proposed for personal password-backup which require simultaneously answering many questions [7,10].

2.2 Threat Model

Our attacker's goal is to impersonate some legitimate prover i and successfully complete the protocol. The attacker may only desire to gain access on behalf of one specific user in a *targeted* attack, or may be content to gain access on behalf of any user in a *trawling* attack. In the former case, the attacker knows the account i represents some real-world person Peggy, enabling *research* attacks using search engines, online social networks, or public records. An active attacker could conduct more advanced research by dumpster diving, burgling Peggy's home, or social engineering to trick Peggy into revealing her answer. Targeted attacks may also be performed by somebody who knows Peggy personally. Schechter et al. explored this attack in a laboratory setting and found a high rate of success by acquaintances at guessing personal knowledge questions [23].

Targeted attacks are powerful but do not scale. Trawling attacks, in contrast, require little per-user work and can be used to simultaneously attack many accounts. We assume that a trawling attacker, although they must provide a value for i when initiating the protocol, has no information about the real-world person behind i and must guess answers based on population-wide statistics.

A *blind attacker* guesses without even understanding the question q [14]. This scenario arises if the question is either not transmitted in the clear [18], is transmitted in a CAPTCHA-ised form, or is user-generated and difficult to automatically process.[3] We argue that a more successful attack strategy is to use a weighted combination of answers to likely questions.

An attacker who is able to correctly understand q but not i is a *statistical* attacker (called a *focused* attacker in [14]), whose strategy is to guess the most likely answers to q. Our main goal is to evaluate the security of common questions against statistical attack. While some questions (e.g., "What is my favourite colour?") obviously have too few plausible answers to be secure, the most common classes of answer found repeatedly in practice are the "proper names" of people, pets, and places, whose security against guessing is not obvious.

3 Quantifying Resistance to Guessing

3.1 Mathematical Formulation of Guessing

We now turn to the mathematical problem of quantifying how secure a personal knowledge question q is against guessing. This problem has been previously considered abstractly [4,20,3,17] and in the case of PINs [2], graphical passwords [6,26,19], and biometrics [1]; we synthesise previous analysis and define new metrics most applicable to trawling attackers.

Because a statistical attacker will respond equally to "what is my boss' last name?" or "who was my kindergarten teacher?" by guessing common surnames, we seek to measure security of the underlying answer space. We consider the correct answer to be a random variable X drawn from a finite distribution $\mathcal{X}$ which is known to the attacker, with $|\mathcal{X}| = N$ and probability $p_i = P(X = x_i)$

[3] Some users may even purposefully obfuscate their questions, such as "What do I want to do?" [14].

for each possible answer x_i, for $i \in [1, N]$. We assume that $\mathcal{X}$ is arranged as a monotonically decreasing distribution with $p_1 \geq p_2 \geq \cdots \geq p_N$. Our attacker's goal is to guess X using as few queries of the form "is $X = x_i$?" as possible.

Intuitively, we may first think of the *Shannon entropy*

$$H_1(\mathcal{X}) = -\sum_{i=1}^{N} p_i \lg p_i \tag{1}$$

as a measure of the "uncertainty" of X. Introduced by Claude Shannon in 1948, entropy has entered common cryptographic parlance as a measure of security, with "high-entropy" secrets being considered advantageous [7,10]. As has been argued previously [4,20,3,17,2,6,1], H_1 is a poor estimator of guessing difficulty for security purposes, as it quantifies the expected number of subset membership queries of the form "Is $X \in \mathcal{S}$?" for arbitrary subsets $\mathcal{S} \subseteq \mathcal{X}$.[4]

Because cryptographic protocols are specifically designed to require sequential guessing, a better metric is the expected number of attempts required to correctly guess X if the attacker takes up the obvious strategy of guessing each possible event in order of its likeliness, known as the *guessing entropy*:

$$G(\mathcal{X}) = E\left[\#_{\text{guesses}}(X \overset{R}{\leftarrow} \mathcal{X})\right] = \sum_{i=1}^{N} p_i \cdot i \tag{2}$$

This measure was introduced by Massey [17] and later named by Cachin [4].

3.2 Marginal Guessing

Guessing entropy models an attacker who will never give up in her search, and thus it can be skewed by exceedingly unlikely events. A simple thought experiment demonstrates why this is inadequate for our purposes. Suppose Eve must sequentially guess k challenge questions with answers drawn from $\mathcal{X}$. Some questions will have uncommon answers, and Eve must make $\sim k \cdot G(\mathcal{X})$ guesses.

Now consider a second adversary Mallory whose goal is to guess the answers to k questions from a set of $m > k$ total questions. Her optimal strategy is to first guess the most likely value for each question in sequence, then the second-most likely value for each question, and so on. Mallory's efficiency will greatly increase as m increases, as she may never need to guess uncommon answers. Guessing entropy is inadequate as it doesn't account for Mallory's willingness to give up on the questions which have less probable answers.

To bound an attacker who only requires some probability α of guessing correctly, we define the *marginal guesswork* μ_α:

$$\mu_\alpha(\mathcal{X}) = \min\left\{ j \in [1, N] \middle| \sum_{i=1}^{j} p_i \geq \alpha \right\} \tag{3}$$

[4] The proof of this is a straightforward consequence of Shannon's source coding theorem. Symbols $X \overset{R}{\leftarrow} \mathcal{X}$ can be encoded using a Huffman code with average bit length $\leq H_1(\mathcal{X}) + 1$, and the adversary can learn one bit at a time with set queries.

This function, introduced by Pliam [20], is also referred to as the *α-work-factor*. We define a similar metric λ_β, the *marginal success rate*, slightly adapted from Boztaş [3], as the probability of success after β guesses have been made:

$$\lambda_\beta(\mathcal{X}) = \sum_{i=1}^{\beta} p_i \tag{4}$$

3.3 Effective Key-Length Metrics

While it is important to remember that μ_α and λ_β are not measures of entropy, we nonetheless find it convenient to convert them into units of bits. This makes all the metrics H_1, G, μ_α and λ_β directly comparable and has an intuitive interpretation as (logarithmically-scaled) attacker workload. We convert each metric by calculating the logarithmic size of a discrete uniform distribution $\mathcal{U}_N$ of size $|\mathcal{U}_N| = N$ with $p_i = \frac{1}{N}$ for all $1 \leq i \leq N$, which has the same value of the guessing metric. This can be thought of as the "effective key-length" as it represents the size of a randomly-chosen cryptographic key which would give equivalent security. The guessing entropy of $\mathcal{U}_N$ is:

$$G(\mathcal{U}_N) = \sum_{i=1}^{N} p_i \cdot i = \frac{1}{N}\sum_{i=1}^{N} i = \frac{1}{N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2}$$

The entropy of this distribution is $\lg N$, so given the guessing entropy of an arbitrary distribution $G(\mathcal{X})$ we can find the logarithmic size of a uniform distribution with equivalent guessing entropy as:

$$\tilde{G}(\mathcal{X}) = \lg[2 \cdot G(\mathcal{X}) - 1] \tag{5}$$

The quantity $\tilde{G}(\mathcal{X})$ can then be interpreted as the effective key-length of $\mathcal{X}$ with respect to guessing entropy. We can similarly derive formulas for effective key-length with respect to marginal guesswork and marginal success rate:

$$\tilde{\mu}_\alpha(\mathcal{X}) = \lg\left(\frac{\mu_\alpha(\mathcal{X})}{\alpha}\right) \qquad \tilde{\lambda}_\beta(\mathcal{X}) = \lg\left(\frac{\beta}{\lambda_\beta(\mathcal{X})}\right) \tag{6}$$

Example Calculation Consider a distribution $\mathcal{Z}$ with $P_\mathcal{Z} = \{\frac{1}{3}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \ldots\}$. Regardless of the tail probabilities, an attacker will have a 50% chance of successfully guessing a random variable drawn from $\mathcal{Z}$ after 4 attempts, so $\lambda_4(\mathcal{Z}) = \frac{1}{2}$. The distribution $\mathcal{U}_8$ with eight equally likely events would also have $\lambda_4(\mathcal{U}_8) = \frac{1}{2}$, so these two distributions are equivalent with respect to λ_4. Since $\lg|\mathcal{U}_8| = \lg 8 = 3$, we expect $\tilde{\lambda}_4(\mathcal{Z}) = 3$, and we can verify that by our formula:

$$\tilde{\lambda}_4(\mathcal{Z}) = \lg\left(\frac{4}{\lambda_4(\mathcal{Z})}\right) = \lg\left(\frac{4}{\frac{1}{2}}\right) = \lg 8 = 3$$

3.4 Relationship Between Metrics

A natural question is whether $\tilde{\mu}_\alpha$ and $\tilde{\lambda}_\beta$ are bounded by H_1 or $\tilde{G}$; unfortunately this is not the case. Pliam [20] proved in a strong way the incomparability of marginal guesswork and entropy in a result we re-state here:

Theorem 1 *(Pliam) Given any $m > 0$, $\beta > 0$ and $0 < \alpha < 1$, there exists a distribution $\mathcal{X}$ such that $\tilde{\mu}_\alpha(\mathcal{X}) < H_1(\mathcal{X}) - m$ and $\tilde{\lambda}_\beta(\mathcal{X}) < H_1(\mathcal{X}) - m$.*

The intuition is that H_1 (and similarly $\tilde{G}$) can be inflated by very unlikely events which don't affect $\tilde{\mu}_\alpha$ or $\tilde{\lambda}_\beta$. A distribution $\mathcal{X}$, whose effective key size with respect to marginal guessing is far smaller than its entropy, is simple to construct: assign $p_1 = \frac{1}{2}$ and $p_i = \frac{1}{2^{2m+2}}$ for the remaining symbols ($|\mathcal{X}| = 2^{2m+1} + 1$). We have $H_1(\mathcal{X}) > m + 1$, but $\tilde{\mu}_{\frac{1}{2}}(\mathcal{X}) = \tilde{\lambda}_1(\mathcal{X}) = 1$, since an adversary need guess only 1 value to have a 50% chance of success. This has clear security implications: a distribution with several very likely events may be completely insecure against guessing even though its Shannon entropy is high.

It is worth noting that $\tilde{G}(\mathcal{X})$ is also high in this example, and in general can be arbitrarily higher than $\tilde{\mu}_\alpha$ or $\tilde{\lambda}_\beta$. This follows from Massey's proof that $\tilde{G}$ is bounded from below by $(H_1 - 2)$ [17]. Boztaş proved a similar result to Pliam's Theorem showing that $\tilde{G}$ can be higher than $\tilde{\mu}_{\frac{1}{2}}$ by any fixed m [3].

3.5 Applicability to Personal Knowledge Questions

Assuming that a targeted attacker is likely to use victim-specific research, we are most concerned with a trawling attacker who will never guess uncommon answers, simply trying a new target if common answers fail. The most useful metric we have is the marginal success rate λ_β. Assuming the system imposes a limit of $t_{\max}$ incorrect guesses for each account, the critical value is the fraction of accounts the attacker can expect to compromise, which is $\lambda_{t_{\max}}$. In the limit of an attacker trying only the single most likely answer for multiple accounts, our security is $\tilde{\lambda}_1(\mathcal{X}) = -\lg(p_1)$, which is also called the min-entropy $H_\infty(\mathcal{X})$.

For offline attacks, λ_β is less meaningful because an attacker won't limit their guessing nearly as much. In this case, $\tilde{\mu}_{\frac{1}{2}}$ is a reasonable metric in that it avoids $\tilde{G}$'s dependence on very unlikely events, while still measuring the cost for an attacker to compromise a majority of available accounts.

3.6 Estimation from Statistics

A final subtlety is estimating our metrics from publicly available statistics based on random sampling from $\mathcal{X}$ and not on complete knowledge of the distribution. This, too, strongly favours the use of μ_α and λ_β because they only reflect the most likely events and not affected by large uncertainty on the tail probabilities of $\mathcal{X}$. It is possible to compute a p-confidence interval for μ_α or λ_β by computing p-confidence intervals for each individual event probability, and using all of the minimum (eq. maximum) estimates to compute minimum estimates μ_α^- and λ_β^- (eq. μ_α^+ and λ_β^+). This technique strictly overestimates uncertainty, but in practice we've found most of the statistics which influence μ_α or λ_β have a strong enough set of statistical support that the confidence interval is quite tight.[5]

In contrast, since H_1 and $\tilde{G}$ depend on the entire distribution, they are much more difficult to reliably estimate from statistics. If we don't a priori know $|\mathcal{X}|$, it is impossible to provide any upper bound because we cannot know the number of events which haven't been observed by sampling. As a lower bound for security purposes, we simply assume no unobserved events exist.

A second problem is that unlikely events are often suppressed for privacy or brevity in published census data. Again in the name of a lower bound, we simply take the least-likely observed event and insert copies of it until the probability space is filled. In the case of surname data, for instance, which is given exactly for names shared by at least k people but suppressed for less common names, we repeatedly insert fictitious names shared by k people until the data set contains as many people as the target population. This crude approximation lowers our estimates of H_1 and $\tilde{G}$, but doesn't influence $\tilde{\mu}_\alpha$ or $\tilde{\lambda}_\beta$.

4 Information Sources

4.1 Question Types and their Use

Based upon recent research into deployed personal knowledge authentication systems, we focus our analysis on questions that ask for proper names, as summarised in Table 1. Rabkin collected 216 questions used by 11 financial institutions [22], and Schecter et al. collected 29 questions used for webmail services provided by AOL, Google, Yahoo!, and Microsoft [23]. These provide some hints at the type of questions used—Rabkin found approximately $\frac{1}{3}$ soliciting a person's name and $\frac{1}{5}$ asking for place names, while Schechter et al. found $\frac{1}{4}$ soliciting a person's name and $\frac{1}{6}$ asked for a place name. Unfortunately, this research provides no insight into which questions users actually select. For example, relatively few questions asked for pet names, though this may be because there is only one way to phrase this question and not because it is unpopular.

Just and Aspinall collected approximately 500 user-generated challenge questions and categorised these questions into a small number of types [14], which we consider to be a more insightful data set. Most notably, they found that 34% of user questions asked for a human name, 15% asked for a pet name and 20% asked for a place name. Of the remainder, 22% asked for a user's favourite item among amongst films, singers, car brands, etc., 5% asked for a time, date, or number, and the remainder were ambiguous. Thus, we estimate that a few simple categories of proper names cover roughly 70% of real-world questions, and the remainder appear trivially vulnerable to guessing attacks.

One subtlety with name data is that it is not always clear if users will respond with a forename (also called a 'first name' or 'given name'), surname (also 'last

[5] Indeed, for $\alpha \leq \frac{1}{2}$ and $\beta \leq \frac{N}{2}$ we were always able to calculate $\tilde{\mu}_\alpha$ and $\tilde{\lambda}_\beta$ to within 0.1 bit with $p > 99\%$. We expect errors from divergence between the population distribution and answers which humans actually choose to use to be so much greater than sampling error that we ignore it in the remainder of this paper.

Category	Example Questions
Forename	What is your grandfather's first name?
	What is your father's middle name?
Surname	What is your mother's maiden name?
	What was the last name of your favourite school teacher?
General Name	Who was your childhood best friend?
Pet Name	What was your first pet's name?
Place	In what city were you born?
	Where did you go for your honeymoon?
	What is the name of your high school?
Other	What was your grandfather's occupation?
	What is your favourite movie?

Table 1: Common answer categories

name'), or both. In such cases, a statistical attacker can simply estimate what probability of users will respond with which, and then combine the two probability distributions, scaling each by its sampling frequency. This should slow down attacks by no more than a factor of two. We also assume that middle names (though less commonly asked for) are reasonably approximated by forenames. In reality, middle names probably have slightly higher diversity, but the most common names are likely the same and an attacker can use a forename table in an attack without much slowdown.

4.2 Data Collection

To our knowledge, this is the first time a breadth of data has been collected for analysing personal knowledge questions. We collected data from government sources where possible, as many developed nations keep near-complete records of citizens' names. In some cases the data is not made publicly available but is acquired and published by media organisations, as in the case of pet registration lists which are compiled by smaller local government bodies. We were also able to gather school and city data from official sources.

Official sources often omit items occurring less than some minimum threshold. As mentioned in Section 3.6, we used estimates of the total population to overcome the missing data. A complete list of our data sources, as well as scripts used for calculations on the data, is made available on our project website.[6] We also provide a summarised list of official sources used in Appendix A.

We found no official sources which provide lists of full names, so we collected names from 65 million randomly-crawled public profiles on the popular online social network Facebook. The demographic for this data is less clearly delineated, but can be used to roughly approximate the global Internet user population.

Source	H_0	H_1	$\tilde{G}$	H_2	$\tilde{\mu}_{\frac{1}{2}}$	$\tilde{\lambda}_3$	H_∞	x_0
Full Names								
Full name	25.1	24.0	24.4	20.8	23.3	14.4	14.4	Maria Gonzalez
Surnames								
South Korea	7.5	4.6	4.5	3.5	3.3	2.7	2.2	Kim
Chile	6.8	6.6	6.3	6.3	6.0	4.9	4.5	González
Spain	9.6	8.9	9.1	7.6	8.8	5.4	5.0	Garcia
Japan	14.5	11.3	12.0	9.0	9.2	6.2	6.0	Satō
Finland	13.8	12.2	12.3	10.5	10.5	7.9	7.8	Virtanen
England	17.4	13.3	14.6	10.2	11.0	6.7	6.4	Smith
Estonia	11.9	11.7	11.7	11.3	11.6	7.9	7.6	Ivanov
Australia	18.6	14.1	15.3	10.9	11.8	7.4	6.8	Smith
Norway	13.7	12.5	13.0	9.9	11.9	6.5	6.4	Hansen
USA	19.1	14.9	16.9	10.9	12.3	7.2	6.9	Smith
Facebook (SF)	19.8	14.9	16.8	11.0	12.4	7.3	7.2	Gonzalez
Surname	21.5	16.2	18.1	12.1	13.7	8.1	7.7	Smith
Forenames, Mixed								
Iceland	8.9	8.5	8.3	7.9	7.7	5.9	5.8	Jón
Spain	9.7	9.0	8.9	8.1	7.9	6.0	5.9	Jose
Facebook (SS)	17.5	11.0	13.4	8.6	8.4	6.0	5.8	Maria
USA	16.7	11.2	14.0	8.7	8.6	6.2	5.9	Michael
Belgium	15.0	10.2	10.3	8.8	8.7	6.1	5.7	Maria
Forename	20.6	12.4	15.7	9.9	9.8	7.4	7.3	David
Forenames, Female (♀)								
Iceland	7.9	7.5	7.3	6.9	6.8	5.1	4.9	Guðrún
Spain	8.3	7.9	7.8	7.3	7.1	5.3	5.1	Maria
Belgium	15.2	10.1	10.9	8.1	8.2	5.5	4.9	Maria
USA	15.1	10.9	12.9	8.7	8.3	6.5	6.3	Jennifer
Forenames, Male (♂)								
Spain	8.6	7.8	7.8	6.9	6.6	4.9	4.8	Jose
Iceland	7.9	7.5	7.3	6.9	6.8	5.0	4.8	Jón
USA	15.2	9.4	12.0	7.2	6.9	5.2	5.0	Michael
Belgium	15.0	9.7	10.4	8.2	7.8	6.1	5.7	Jean
Forenames by Birth Decade								
USA, 1950 (♀)	11.8	8.6	9.1	7.1	6.8	5.2	5.0	Mary
USA, 1950 (♂)	11.7	7.7	8.3	6.2	5.8	4.6	4.6	James
USA, 1960 (♀)	11.9	9.1	9.5	7.6	7.1	5.6	5.2	Lisa
USA, 1960 (♂)	11.9	7.9	8.6	6.4	5.9	4.7	4.6	Michael
USA, 1970 (♀)	12.1	9.7	10.3	7.7	7.6	5.5	4.8	Jennifer
USA, 1970 (♂)	12.1	8.4	9.3	6.7	6.3	5.0	4.6	Michael
USA, 1980 (♀)	12.2	9.7	10.4	7.7	7.6	5.4	5.3	Jessica
USA, 1980 (♂)	12.2	8.6	9.6	6.9	6.4	5.1	4.9	Michael
USA, 1990 (♀)	12.3	10.3	10.8	8.4	8.3	6.1	6.0	Jessica
USA, 1990 (♂)	12.3	9.3	10.0	7.5	7.1	5.7	5.5	Michael
USA, 2000 (♀)	12.4	10.8	11.1	9.1	9.0	6.6	6.5	Emily
USA, 2000 (♂)	12.2	9.9	10.4	8.2	7.8	6.4	6.2	Jacob
Pet Names								
Los Angeles	15.8	11.7	13.1	9.2	9.4	6.5	6.4	Lucky
Des Moines	13.6	11.6	12.4	9.4	9.7	6.5	6.2	Buddy
San Francisco	13.7	11.6	12.0	9.6	9.8	6.7	6.7	Buddy
Place Names								
UK High Schools	8.7	8.5	8.2	8.3	8.0	7.4	7.3	Holyrood
UK Primary Schools	14.0	13.8	13.5	13.6	13.3	12.1	12.1	Essex
School Mascots (US)	11.8	8.1	9.3	6.2	5.7	4.5	4.1	Eagles
UK Cities	9.2	8.5	8.8	5.9	8.7	4.4	3.0	London
Tourist Destinations	13.0	12.0	12.5	9.5	12.4	6.3	5.9	London

Table 2: Summary of statistics on real data

5 Results and Discussion

Our calculations of the metrics defined in Section 3 are displayed in Table 2.[7] For online attacks, the marginal success rate $\tilde{\lambda}_3$ models an attacker limited to 3 guesses on each available account. For almost all of our data (exclusive of full names and primary schools), we have $\tilde{\lambda}_3 \lessapprox 8$, indicating that the majority of deployed challenge questions systems are insecure against trawling attackers. If 3 guesses are allowed, an attacker can compromise roughly 1 in 80 accounts. For offline attacks, we mostly find $\tilde{\mu}_{\frac{1}{2}} \lessapprox 12$, meaning an attacker can compromise the majority of accounts with only a few thousand guesses per account.

Our analysis demonstrates significant *weak subspaces* in the answer space of most personal knowledge questions. This can be directly compared to other authentication systems for which weak answer spaces are known to exist. In Figure 1 we plot the name distributions from the Facebook dataset against weak subsets found for textual passwords [15,25,24], the Pass-Go user-drawn password system [19], the Passfaces graphical PIN system [26], the PassPoints visually-cued clicked password system [6] and a handwriting-recognition biometric system [1]. Aside from the badly-broken Passfaces system, personal knowledge questions compare unfavorably to other methods unless full names are required.

We summarise further interesting trends below:

Diversity effects The difficulty of guessing surnames correlates with ethnic diversity. American surnames were the most difficult to guess in our survey, presumably because the population is a blend of immigrants from many ethnicities. Facebook provides even more diversity as a blend of users from around the world. Surnames from Japan and South Korea, which are ethnically homogeneous and have relatively few immigrants, provide low resistance to guessing.

Naming trends Given names are a matter of fashion and vary in several interesting dimensions. In the countries studied, female names seem to provide slightly higher resistance to guessing than male names.[8] Over the past 6 decades in the USA, diversity of forenames has been increasing slowly but steadily. Curiously, pet names are slightly harder to guess than human names.

Ethnic correlations The Facebook data provides ample evidence that forenames and surnames are not independent variables. They are correlated via an individual's ethnicity and possibly further in that some name combinations are considered more pleasing to the ear. Maria Gonzalez and Jose Rodriguez are the most statistically over-represented names in our data set given the independent frequency of the forename and surname component. Each appears with extremely high statistical significance ($p \ll 0.001$ in a χ^2 test). Similarly, there

[6] `http://groups.inf.ed.ac.uk/security/KBA/`

[7] We also include the Rényi entropy $H_\alpha(\mathcal{X}) = \frac{1}{1-\alpha} \lg \left(\sum_{i=1}^{N} p_i^\alpha \right)$ for $\alpha \in \{0, 2, \infty\}$. As predicted by Boztaş [3], H_2 seems to provide a good estimate for $\tilde{\mu}_{\frac{1}{2}}$.

[8] Security increases, of course, if a question doesn't specify gender.

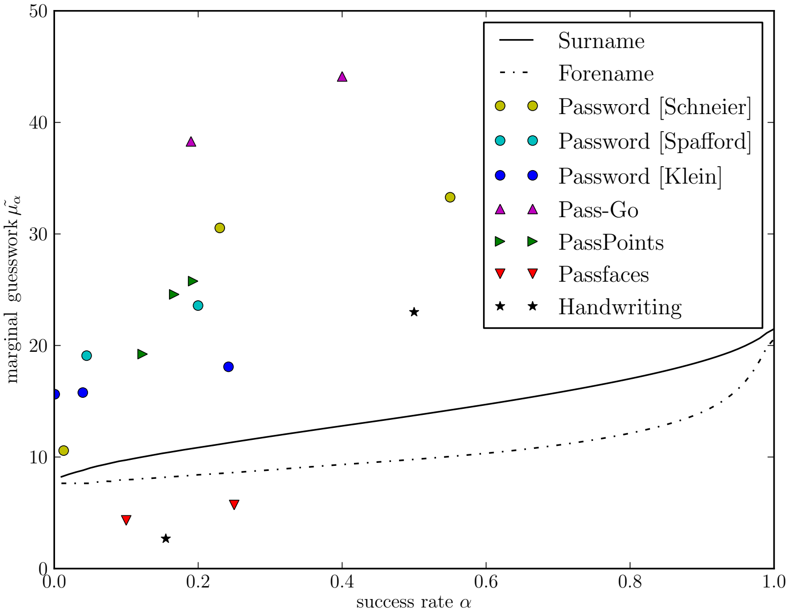

Fig. 1: Comparison of weak subspaces in name distributions (Facebook dataset) to those found in other authentication systems [15,25,24,26,19,6,1].

are a number of highly statistically under-represented name pairs, mostly curious cross-cultural pairings like Francesco Smith or Juan Khan. Frequent names like Maria Gonzalez appear because both components share a common ethnicity (Hispanic). A χ^2 test on the entire forename distribution given a Spanish surname such as Gonzalez confirms with high significance ($p \ll 0.001$) that naming patterns change amongst individuals of this ethnicity.

This dependence between forenames and surnames indicates that guessing difficulty will be lower if an attacker knows the target's ethnicity. To quantify this, we clustered the names and identified a set of 150 common Spanish surnames, which cover 6.2% of all individuals in the dataset. The guessing difficulty for these 4 million individual's forenames is shown in Table 2 under "Facebook†".[9] We similarly took 150 common Spanish forenames, representing nearly 12 million people, over 18% of our data, and computed the guessing difficulty of their surnames. In both cases $\tilde{\mu}_{\frac{1}{2}}$ and $\tilde{\lambda}_3$ drop by about a bit, indicating that identifying an individual's ethnicity may roughly doubles a statistical attacker's efficiency.

Power-law models The frequencies of English surnames have previously been posited to be well-fitted by a discrete Pareto distribution [9], with the probability that a surname X's frequency $f(X)$ is greater than x being proportional to

[9] Note that this entry is not the difficulty of guessing a forename known to be Spanish, it is the difficulty of guessing a forename when the surname is known to be Spanish.

$x^{-(c+1)}$. Fox et al. found this to hold for $c \approx 1.4$. This is thought to occur because surnames are inherited but don't strongly correlate to reproductive fitness, leading to a Pareto-like distribution through random genetic drift.

We found the Pareto distribution with $c \approx 1$ to be a reasonable model for the Facebook surname dataset, though the head of the distribution skewed significantly away from the Pareto model, with the most common names being less popular than expected. Still, support for a power-law model of surname frequency suggests the inappropriateness of this distribution for security purposes.

Interestingly, our forename and pet name distributions were also approximated well by the Pareto distribution, with $c \approx 0.8$ in the Facebook data set. The reasons for this fit are less well-understood, though this is close to the classic Zipf distribution ($c = 1$) which is known to model many natural-occurring phenomena such as word frequency in natural languages. If it is true that humans naturally produce names following the Zipf distribution, this too suggests that human-provided name spaces will not provide adequate guessing resistance.

6 Countermeasures

Up to this point, we have assumed a passive enrolment server which accepts any answers and has no influence on the resulting answer distribution $\mathcal{X}$. If we assume the server knows $\mathcal{X}$, it is possible to actively *shape* the answer space into a more secure distribution $\mathcal{X}'$ by probabilistically rejecting some users' answers. There is a growing literature on proactively encouraging users to select diverse textual [8] or graphical [5] passwords. Bentley et. al previously considered the problem of "grooming" a skewed probability distribution to uniform [2],

The process of a user answering is equivalent to randomly drawing $X \overset{R}{\leftarrow} \mathcal{X}$. The server can examine the result and if $X = x_i$, reject with probability r_i and force the user to answer a differently-worded question with the same answer-space, in practice re-drawing $X \overset{R}{\leftarrow} \mathcal{X}$. We assume the process is recursive: the user's second answer x_j may also be rejected with some probability r_j. This process results in a modified distribution $\mathcal{X}'$ of answers which are accepted.

If we are constrained by a maximum-allowable overall rejection probability r_*, it is simple to find the optimum rejection probabilities $r_0, \ldots r_N$ which will most increase security. This comes from the observation that, given the ability to lower any single p_i by a fixed Δ, lowering p_0 will result in the greatest increase for each of H_α, $\tilde{G}$, $\tilde{\mu}_\alpha$ and $\tilde{\lambda}_\beta$. The optimal $r_0, \ldots r_N$ are thus computed by an iterative algorithm. First r_0 is increased until $p'_0 = p'_1$, namely by setting $r_0 = 1 - \frac{p_1}{p_0}$. Next, we increase r_0 and r_1 together until $p'_0 = p'_1 = p'_2$. We repeatedly increase r_0 through r_{m-1} so that $p'_0 = \cdots = p'_m$, stopping when we have reached our maximum overall rejection probability.[10] The m most likely events are equiprobable in $\mathcal{X}'$. The remaining events are never rejected; their probabilities each increase by $\frac{1}{1-r_*}$.

[10] The algorithm may terminate early if the distribution has reached uniformity, though this is probably impractical. For example, the Facebook surnames corpus requires a rejection rate of 95.5% to be shaped to uniformity.

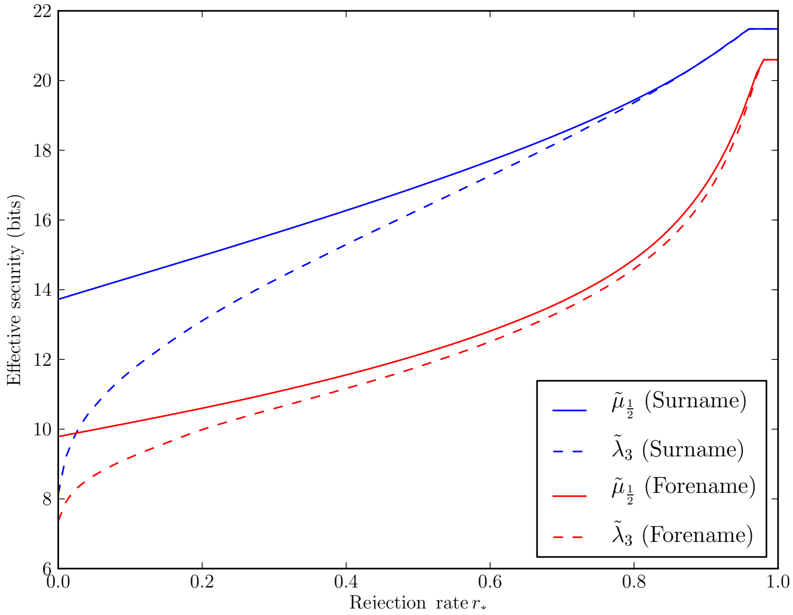

Fig. 2: Effectiveness of shaping a distribution as a function of r_*

Shaping is very effective at increasing $\tilde{\lambda}_\beta$ as the most likely events are greatly reduced in probability. As shown in Figure 2, shaping the name distributions in the Facebook corpus drives $\tilde{\lambda}_3$ close to $\tilde{\mu}_{\frac{1}{2}}$ even for reasonable $r_* < 0.5$. Even relatively mild shaping with $r_* = 0.1$ increases $\tilde{\lambda}_3$ by 3.6 bits for surnames. Although the overall rejection rate is low, though, it is highly unequal: for $r_* = 0.1$ the rejection rate r_0 for the surname "Smith" is 94.3%.

7 Concluding Remarks

We have applied marginal guessing metrics to the security analysis of common personal knowledge questions. We then used a diverse collection of real-world statistical data to estimate the strength of these questions against a trawling attacker with a large number of accounts to test. We believe this is an increasingly important attacker model and our methods provide a useful framework for evaluating human-computer authentication.

We have not assessed a ground-truth answer space; the actual distribution of surnames provided to a deployed authentication server will vary based on the precise question wording and specific user population. Still, we have found strong evidence that across a broad range of cultures and contexts, human-created names simply don't have enough diversity to provide serious resistance to guessing attacks. In combination with recent results demonstrating vulnerability to targeted attacks, our work casts serious doubt on the continued use of personal knowledge questions for backup authentication.

Acknowledgements

We thank our anonymous referees and our shepherd Lucas Ballard for their detailed and helpful comments. We also thank David Aspinall, Claudia Diaz, Andrew Lewis, and Hyoungshick Kim for assistance drafting our report. Just and Matthews were funded by UK EPSRC, Grant No. EP/G020760/1.

References

1. L. Ballard, S. Kamara, and M. K. Reiter. The Practical Subtleties of Biometric Key Generation. In *SS'08: Proceedings of the 17th Conference on Security*, pages 61–74, Berkeley, CA, USA, 2008. USENIX Association.
2. J. Bentley and C. Mallows. How Much Assurance Does a PIN Provide? In *2nd International Workshop on Human Interactive Proofs*, pages 111–126, 2005.
3. S. Boztas. Entropies, Guessing, and Cryptography. Technical Report 6, Department of Mathematics, Royal Melbourne Institute of Technology, 1999.
4. C. Cachin. *Entropy Measures and Unconditional Security in Cryptography.* PhD thesis, ETH Zürich, 1997.
5. S. Chiasson, A. Forget, R. Biddle, and P. C. van Oorschot. Influencing Users Towards Better Passwords: Persuasive Cued Click-Points. In *BCS-HCI '08: Proceedings of the 22nd British HCI Group Annual Conference on HCI 2008*, pages 121–130, Swinton, UK, 2008. British Computer Society.
6. D. Davis, F. Monrose, and M. K. Reiter. On User Choice in Graphical Password Schemes. In *SSYM'04: Proceedings of the 13th conference on USENIX Security Symposium*, Berkeley, CA, USA, 2004. USENIX Association.
7. C. Ellison, C. Hall, R. Milbert, and B. Schneier. Protecting Secret Keys with Personal Entropy. *Future Gener. Comput. Syst.*, 16(4):311–318, 2000.
8. A. Forget, S. Chiasson, P. C. van Oorschot, and R. Biddle. Improving Text Passwords Through Persuasion. In *SOUPS '08: Proceedings of the 4th Symposium on Usable Privacy and Security*, pages 1–12, New York, NY, USA, 2008. ACM.
9. W. R. Fox and G. W. Lasker. The Distribution of Surname Frequencies. In *International Statistical Review*, pages 81–87, 1983.
10. N. Frykholm and A. Juels. Error-Tolerant Password Recovery. In *CCS '01: Proceedings of the 8th ACM conference on Computer and Communications Security*, pages 1–9, New York, NY, USA, 2001. ACM.
11. V. Griffith and M. Jakobsson. Messin' with Texas: Deriving Mother's Maiden Names Using Public Records. *Applied Cryptography and Network Security*, 2005.
12. W. J. Haga and M. Zviran. Question-and-Answer Passwords: an Empirical Evaluation. *Inf. Syst.*, 16(3):335–343, 1991.
13. M. Jakobsson, L. Yang, and S. Wetzel. Quantifying the Security of Preference-based Authentication. In *DIM '08: Proceedings of the 4th ACM workshop on Digital identity management*, pages 61–70, New York, NY, USA, 2008. ACM.
14. M. Just and D. Aspinall. Personal Choice and Challenge Questions: A Security and Usability Assessment. In L. Cranor, editor, *SOUPS*, ACM International Conference Proceeding Series. ACM, 2009.
15. D. Klein. "Foiling the Cracker": A Survey of, and Improvements to, Password Security. In *Proceedings of the 2nd USENIX Security Workshop*, pages 5–14, 1990.
16. J. Lindamood and M. Kantarcioglu. Inferring Private Information Using Social Network Data. Technical Report UTDCS-21-08, University of Texas at Dallas Computer Science Department, July 2008.

17. J. L. Massey. Guessing and Entropy. In *Proceedings of the 1994 IEEE International Symposium on Information Theory*, page 204, 1994.
18. L. O'Gorman, A. Bagga, and J. L. Bentley. Call Center Customer Verification by Query-Directed Passwords. In A. Juels, editor, *Financial Cryptography*, volume 3110 of *Lecture Notes in Computer Science*, pages 54–67. Springer, 2004.
19. P. C. v. Oorschot and J. Thorpe. On Predictive Models and User-Drawn Graphical Passwords. *ACM Trans. Inf. Syst. Secur.*, 10(4):1–33, 2008.
20. J. O. Pliam. On the Incomparability of Entropy and Marginal Guesswork in Brute-Force Attacks. In *Progress in Cryptology-INDOCRYPT 2000*, 2000.
21. R. Pond, J. Podd, J. Bunnell, and R. Henderson. Word Association Computer Passwords: The Effect of Formulation Techniques on Recall and Guessing Rates. *Computers & Security*, 19(7):645–656, 2000.
22. A. Rabkin. Personal knowledge questions for fallback authentication: Security questions in the era of Facebook. In L. F. Cranor, editor, *SOUPS*, ACM International Conference Proceeding Series, pages 13–23. ACM, 2008.
23. S. Schechter, A. J. B. Brush, and S. Egelman. It's No Secret: Measuring the Security and Reliability of Authentication via 'Secret' Questions. In *IEEE Security and Privacy*. IEEE, 2009.
24. B. Schneier. Real-World Passwords. *Schneier on Security*, December 2006.
25. E. Spafford. Observations on Reusable Password Choices. In *Proceedings of the 3rd USENIX Security Workshop*, 1992.
26. J. Thorpe and P. C. van Oorschot. Human-Seeded Attacks and Exploiting Hot-Spots in Graphical Passwords. In *SS'07: Proceedings of 16th USENIX Security Symposium*, Berkeley, CA, USA, 2007. USENIX Association.
27. M. Toomim, X. Zhang, J. Fogarty, and J. A. Landay. Access Control by Testing for Shared Knowledge. In M. Czerwinski, A. M. Lund, and D. S. Tan, editors, *CHI*, pages 193–196. ACM, 2008.
28. J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password Memorability and Security: Empirical Results. *IEEE Security and Privacy Magazine*, 2(5):25, 2004.

A Sources of Statistical Data

Below is a summary of statistical data sources used in compiling this paper. Complete information on the data sets is provided on our project website `http://groups.inf.ed.ac.uk/security/KBA/`.

- Chile Civil Identification and Registration Service
- Des Moines Register
- Eeski Ekspress
- Euromonitor International
- Finland Population Register Center
- Intellectual Property Australia
- Japanese Surname Dictionary
- Los Angeles Department of Animal Licensing
- San Francisco Animal Licensing Department
- Scottish Government School Education Statistics
- Spanish National Institute of Statistics
- Statistics Belgium
- Statistics Iceland
- Statistics Korea
- Statistics Norway
- United Kingdom Department for Children, Schools, and Families
- United Kingdom Office for National Statistics
- United States Census Bureau
- United States Social Security Administration