Life-Experience Passwords (LEPs)

Simon Woo
University of Southern California
simonwoo@usc.edu

Elsi Kaiser
University of Southern California
emkaiser@usc.edu

Ron Artstein
USC Institute for Creative Technologies
artstein@ict.usc.edu

Jelena Mirkovic
USC Information Sciences Institute
sunshine@isi.edu

ABSTRACT

Passwords are widely used for user authentication, but they are often difficult for a user to recall, easily cracked by automated programs and heavily reused. Security questions are also used for secondary authentication. They are more memorable than passwords, but are very easily guessed. We propose a new authentication mechanism, called "life-experience passwords (LEPs)," which outperforms passwords and security questions, both at recall and at security. Each LEP consists of several facts about a user-chosen past experience, such as a trip, a graduation, a wedding, etc. At LEP creation, the system extracts these facts from the user's input and transforms them into questions and answers. At authentication, the system prompts the user with questions and matches her answers with the stored ones.

In this paper we propose two LEP designs, and evaluate them via user studies. We further compare LEPs to passwords, and find that: (1) LEPs are 30–47 bits stronger than an ideal, randomized, 8-character password, (2) LEPs are up to 3× more memorable, and (3) LEPs are reused half as often as passwords. While both LEPs and security questions use personal experiences for authentication, LEPs use several questions, which are closely tailored to each user. This increases LEP security against guessing attacks. In our evaluation, only 0.7% of LEPs were guessed by friends, while prior research found that friends could guess 17–25% of security questions. LEPs also contained a very small amount of sensitive or fake information. All these qualities make LEPs a promising, new authentication approach.

1. INTRODUCTION

Textual passwords are widely used for user authentication. An ideal password should be easy for a user to remember, but difficult for others to guess. These two requirements are at odds. People remember by relating their passwords to some personally salient facts – but this leads to common and predictable patterns in passwords that make them insecure against automated guessing [95, 66]. Users also tend to reuse their passwords, to achieve memorability. But this lowers security, because passwords stolen from one server can be used to gain access to another one.

Many alternatives to textual passwords have been proposed, such as graphical passwords, cognitive authentication, one-time passwords, hardware tokens, phone-aided authentication and biometric passwords. However, research has shown that none of these can offer convenience, simplicity, and user familiarity [68, 69] comparable to that of textual passwords. We thus focus on trying to improve text-based authentication.

Our insight is that it is highly unnatural to humans to create *new, complex memories* (passwords) and recall them in minute detail (e.g., recall capitalization and placement of special characters) after a long time period, and without any hints. Humans remember by association, relating new facts to existing memories [88]. This makes it difficult to create and recall many new, strong passwords. Humans also recall by reconstructing facts, sometimes imprecisely, from relevant data stored in the brain [88]. Thus a user may recall that they used a family member's name plus their birth year for a password, but forget which family member they chose, if they used their first or last name, how it was capitalized, etc. This makes it difficult to precisely recall passwords.

We propose a new authentication method — *life-experience passwords (LEPs)* — which extracts authentication secrets out of a *user's existing memories*, and uses prompts and imprecise matching at the authentication stage to further improve recall. Our contribution lies in design and implementation of this authentication method, and its evaluation via two human user studies.

A LEP consists of several facts about a user-chosen life experience, such as a trip, a graduation, a wedding, a place, etc. At password creation, the system prompts the user for the experience's title, and for facts relating to this experience, such as names of people and locations, special objects and activities, dates, etc. These facts are transformed by the system into questions – stored in clear – and answers – stored hashed and salted. At authentication, the system prompts the user with corresponding questions, and matches her answers with those stored by the system, allowing for imprecise matches due to extraneous words, capitalization, punctuation and reordering. LEPs could be used for primary or secondary authentication, in cases when high recall and high security are desired.

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

ISBN 978-1-4503-4771-6/16/12...$15.00

DOI: http://dx.doi.org/10.1145/2991079.2991107

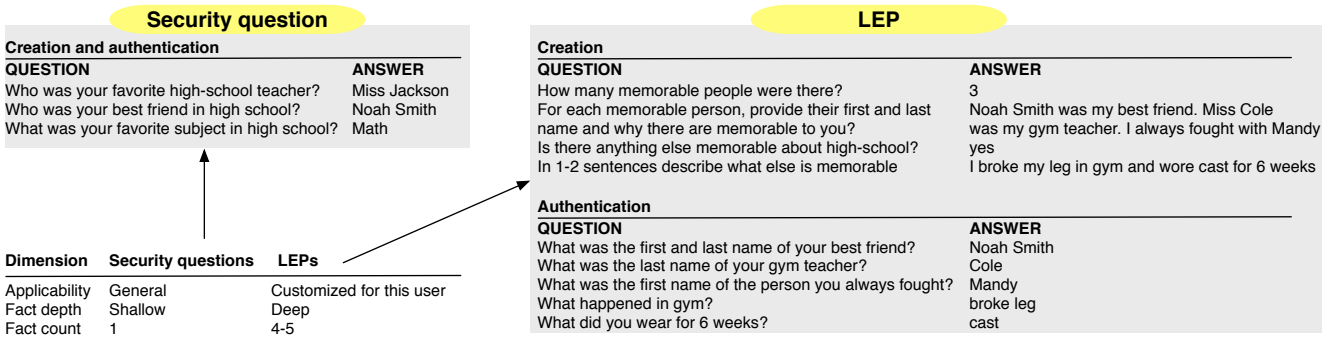


Figure 1: Security questions versus LEP example for "high school" topic.

We designed and evaluated two LEP designs in human user studies, approved by our institutional IRB, to evaluate their recall and security. We found that: (1) LEPs are 30–47 bits stronger than an ideal, randomized, 8-character password, averaging 83–100 bits of strength against statistical attacks (2) LEPs are 2–3× more memorable than passwords, having 73% recall after a week and 54% recall after 3–6 months and (3) LEPs are reused half as often as passwords.

LEPs resemble security questions, in that both use personal experiences for authentication. But LEPs use more diverse, unique, and memorable personal facts than security questions, and these are deeper, more specific facts. This makes LEPs hard to guess by friends or attackers, who use social networks or public sources. When compared with security questions, (1) LEPs are 24–35× harder to guess by friends (only 0.7% of LEPs were guessed in our study), (2) LEPs contain 2.4–3.2× fewer fake answers (11.5% and 15.7% LEPs are potentially fake, versus 37% of security questions).

There are two downsides to LEPs: (1) user burden for creating and authenticating with a LEP is 3–6× higher than when using passwords – this burden may be prohibitive for some applications and devices, (2) LEPs may contain sensitive information, which poses a privacy risk. While these are serious drawbacks, our human user study found that 93.7% of users would use LEPs for high-security content (e.g., banking applications). We further found that only 3% of LEPs contained generally sensitive information, which could be further reduced with better user prompts. We thus believe that LEPs are a promising authentication method for some users and applications. Interested readers can try LEPs out by visiting our project page: http://leps.isi.edu/demo.

We discuss related work to LEPs in Section 2. We present LEP design in Section 3 and we detail the setup of our human studies in Section 4. Section 5 presents the results of our evaluation of LEPs, and Section 6 offers our conclusions.

2. BACKGROUND AND RELATED WORK

There is much research on non-textual alternatives to passwords, such as graphical passwords [83, 41, 75, 42, 43, 82, 9, 65], video passwords [97, 76] and biometric passwords [60]. Due to space limitations, we only survey research that is directly related to textual passwords.

Improving Password Strength. A few recent publications [72, 95, 85, 81] proposed improvements in password design using Markov models, semantic patterns of passwords, and user feedback. For example, Telepathwords [85] help users create strong passwords by comparing user input with popular substrings. When such a substring is detected, Telepathwords provides actionable, real-time feedback to steer a user towards a different choice. In [81] the system randomly generates strong passwords and then allows a user to replace a few characters to make the password more memorable. All these techniques increase password strength, but do not improve memorability nor diversity, and all require users to create new memories of complex strings. Our work on LEPs differs fundamentally from these approaches because we seek to exploit existing memories and thus increase strength, memorability and diversity of textual passwords.

Security Questions. Security questions [80] are often used for secondary authentication, e.g., when a user loses her password, or to supplement password-based authentication for high-security servers (e.g., bank). A user is offered a choice of a small number of fixed questions, such as mother's maiden name, pet names, favorite teacher names, best friend names, etc. While both security questions and LEPs use personal knowledge for authentication, there are significant differences. We discuss them here and summarize them in Figure 1, which also illustrates sample security questions and a LEP on the high school topic.

Applicability. Organizations usually offer a very limited choice of security questions. There may be users to whom no question applies. For example, a user may not have a favorite high school teacher or a best friend, or may have multiple teachers/friends that she likes. When faced with such questions users select answers that they do not recall at authentication time. Schechter et al. [91] and Bonneau et al. [67] measured that 20–40% of security questions cannot be recalled by users. Conversely, during LEP creation users can choose, with very little constraint, the topic they want to talk about and the facts about that topic, which are memorable to them. This leads to more personalized facts and thus higher recall.

Depth of facts. Security questions ask for *shallow* facts (e.g., pet's name, best friend's name), which are generally applicable to many users. Such facts can be mined from public sources [79, 91], or guessed using statistical attacks [67]. Easy guessability leads users to provide fake answers to security questions, which leads to low recall when they cannot remember the fake answers. On another hand, LEPs ask for *deep* facts – memorable people, places, activities or objects in connection with a user-chosen event. Answers to such questions are not easily found on social networks, or

guessed by family and friends, which removes the need for users to lie. In our studies, only 0.7% of LEPs were guessed by friends (compared to 17–25% of security questions [91]) and only 11.5–15.7% of answers were fake (compared to 37% for security questions [67]).

Number of facts. Security questions contain only one fact, which may be easily guessed or obtained from public sources. LEPs contain a larger number of facts, and a user must recall most or all of them for authentication. Thus the barrier for a successful guessing attack is higher.

Another approach to security questions is to let users choose the questions themselves. This allows users to freely choose facts that are relevant to them, but decreases security [84, 91]. While LEPs also allow users to choose which facts to provide, our fact elicitation guides them toward secure, memorable and stable facts. This allows LEPs to outperform user-chosen security questions.

Cognitive Passwords. Similar to LEPs, cognitive passwords are based on personal facts, interests, and opinions that are thought to be easily recalled by a user. Article [12] provides an overview, definitions, and some examples of cognitive passwords. Das et al. [74] and Nosseir et al. [90] explore autobiographical authentication that uses facts about past events, which are captured by smartphones or calendars, without any user input. While such information may be memorable in short intervals after it is collected, humans do not remember ordinary daily events for long periods nor with great consistency [71]. On the other hand, LEPs require more user effort during creation, but elicit more salient facts [89, 71], which is essential for good recall.

Narrative Authentication. Somayaji et al. [94] propose use of narratives for user authentication, but do not evaluate them. Their narratives require users to associate imaginary objects with past memories (e.g., contents of a drawer from a childhood bedroom), and may also be fully fictional. Because these narratives lack personal significance to user, we expect they would be less memorable than LEPs.

3. LIFE-EXPERIENCE PASSWORDS

This section describes the design and the implementation of life-experience passwords (LEPs). We discuss our choices for LEP topics and facts in Section 3.1, attacker models and strength calculation in Section 3.2, the LEP creation process in Section 3.3, LEP-based authentication in Section 3.4 and LEP uses in Section 3.5.

Figure 2 shows the LEP creation process. A user identifies a life-experience she wants to use for a LEP. She then inputs a title for this experience and recounts interesting facts, with some guidance from the system. The system mines the facts from the user's input, and transforms them into question-and-answer pairs. Questions and the title must be stored as clear text, because they are shown to users during authentication. To store the answers, we concatenate either all of them, or several subsets (see Section 3.4), add the salt, and hash the resulting string(s). During authentication, the system displays the title and the questions, and user answers are hashed and compared to those stored by the system using imprecise matching.

3.1 Topics and Facts

In this Section we discuss how much guidance should be provided to users during LEP creation. In general, users need some guidance to remember interesting facts to recount. Further, elicitation must be carefully developed to produce facts, which can be accurately recalled by users, and cannot be easily guessed by others.

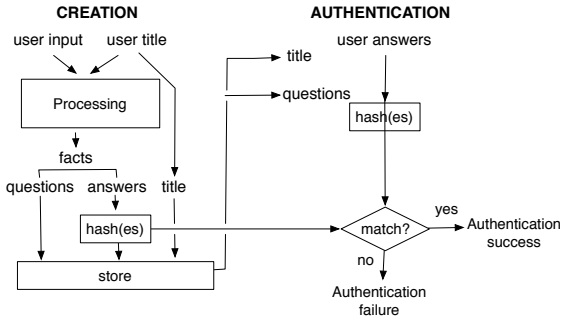


Figure 2: LEP creation and authentication

Which experiences can be used for LEPs? Letting a user freely select an experience to talk about, without any guidance, may not produce secure and memorable input, as shown by research on self-built security questions [84, 91]. This motivated us to build a list of diverse and general topics, to guide password creation (see Table 2).

Category	Subcategory
Event	Engagement, wedding, birth, death, accident, graduation, party, trip
Learning	Driving, skiing, snowboarding, swimming, biking, skill/art, language
About	Person, place

Table 2: LEP topics

How to elicit useful facts? A *useful fact* is a fact that is strong, stable and immutable.

A *strong* fact has many possible answers to the question, which gives it strength against brute-force attacks (see Section 3.2).

A *stable* fact is consistently recalled by a user. We have learned by exploring several LEP designs that stability is influenced the most by the fact's type and our elicitation method. Subjective facts about feelings and opinions are inconsistently recalled by users. We thus ask about objective facts, such as names, locations, times, objects and activities. Further, elicitation specificity makes a large difference. The more specific questions we pose during elicitation, the more stable facts we get. A user may use multiple terms for the same person (e.g., "my mom", "mom", "mother", "Jennifer"). Transforming "who" questions into "first and last name" questions reduces the ambiguity and increases stability of answers. Another source of instability comes from asking for a singular answer to a question that may have a plural answer (e.g., a user has two best friends). Asking too specific questions (e.g., "what is the first and last name ..." but the user only recalls the first name) or questions that do not apply to a given user (e.g., a pet's name when the user does not own a pet) also leads to unstable facts. We believe that stability issues, more than lying, may be responsible for many authentication failures found in the past studies of security questions [91, 67]. We have refined our elicitation process to contain very specific prompts, which depend on the user's prior input. This leads to stable facts.

An *immutable* fact does not change over time. For this reason we do not ask about preferences and opinions (e.g., "What is your favorite band"), which are mutable.

Category	Description	Statistical strength				Brute-force strength
		Lists	Min. size	Max. size	Unique items	
FN	first name (e.g., John)	384	3	38,717	150,695	285,537
LN	last name (e.g., Doe)	80	100	151,671	223,096	6,209,229
FL	first and last name (e.g., John Doe)	combinations of FN and LN				563,335,972,290
PL	place (e.g., UCLA)	48	10	18,467	36,864	1,398,314
CI	city (e.g., Seattle)	8	85	870	2,230	754,450
OBJ	object (e.g., watch)	30	6	19,681	22,210	139,049
ACT	activity (e.g., kayaking)	4	14	276	385	11,539
DT	date (e.g., 2/28/1972)	n/a				18,250
YR	year (e.g., 2001)	n/a				50
RL	relationship (e.g., mom)	n/a				49
HU	approx. 100 choices (e.g., Toyota)	n/a				100
TN	approx. 10 choices (e.g., yellow)	n/a				10

Table 1: Fact categories and their statistical and brute-force strengths (see Appendix for sources)

During LEP creation, we mine facts about **people**, **locations**, **time**, **objects** and **activities**. These facts are objective, and thus immutable. People and location facts have a high strength (see the Section 3.2), while time, objects and activities have a lower but still substantial strength. Further, we have designed our elicitation process to produce very specific questions, and thus stable facts.

Privacy risk. LEP questions and answers contain information about some past event, which may pose privacy risk to a user if questions are observed by others, or if answers are guessed or cracked. We advise users to avoid sensitive or incriminating facts during LEP creation. Our evaluation finds that only 3% of LEPs in our study had sensitive information, which we plan to address with better LEP creation prompts.

3.2 Strength

This section discusses our attacker models and how we calculate strength of LEPs against these attackers.

Strength of a password can be measured as the number of trials a guessing attacker has to make until success – this is known as *guesswork* [66] or a *heuristic measure of password strength* [69]. As Bonneau points out in [66], this definition of strength depends on the size of the attacker's dictionary and is difficult to reproduce across different studies. He proposed several more stable strength measures, such as α-guesswork. However, they require a much larger sample size than we had in our studies. We thus use guesswork as our measure of strength.

A LEP consists of multiple facts. We calculate its strength as a half of the product of individual fact strengths: $S_{LEP} = 1/2 \cdot \prod_{i=1}^{k} S_{f_i}$, where S denotes strength, k is the number of facts in a given LEP, and f_i is the i-th fact. On the average the attacker will go through a half of the possible fact combinations before guessing correctly.

We classify LEP facts into categories, shown in Table 1. We assume an intelligent attacker, who can infer the fact's category from the user's authentication prompt, and guesses answers only within that category.

Attacker Models. We consider the following attacker models [67, 69]. A **brute-force attacker** compiles a dictionary of all possible answers within a fact category (e.g., first name), and tries them in any order. A **statistical attacker** compiles ranked dictionaries of popular answers within a fact category, and tries them in the order of popularity. A **friend attacker** forms guesses using her personal knowledge of the user, and may mine some guesses from the user's social network pages or use search engines. A **password-reuse attacker** has stolen a user's password from one server and attempts to reuse it, in the exact or a slightly modified version, to gain access to another server.

We assume that brute-force, statistical and password-reuse attackers are *offline* attackers [67]. They will use automated programs to crack LEPs, just like they do for passwords today. They can make as many guesses as their dictionaries allow. A friend attacker, is an *online* attacker [67], and will attempt to guess passwords manually. We thus assume that a friend attacker is allowed to make a small number of guesses, before being locked out by the server for excessive failed logins.

We denote the strength of a fact against brute-force attacks as its *brute-force strength*, and measure it as the number of all possible inputs in the fact category. It is challenging to count all possible inputs, since some answers may be drawn from sets that are not fully enumerable. For example, an answer to a "who" question can be a relationship, a first name, a last name, a first and last name pair, a title like Mr. and a last name, a nickname, etc. Even within these subsets there are variations. For example, one could combine relationship and an adjective, e.g., "my favorite aunt" or "my oldest uncle". Further, some subsets may not be fully enumerable. For example, there are publicly available censuses of US names but not of Chinese or Indian names.

We denote the strength of a fact against statistical guessing as its *statistical strength*, and measure it as: (1) the rank of the fact on a ranked list of popular facts in its category, or (2) the brute-force strength if the fact is not found on the popular list. A challenge lies in creating sufficiently large and comprehensive lists of popular facts, and ranking them based on their popularity.

We examined many different data sources, seeking to identify: (1) the total number of possible facts, and (2) the ranked list of popular facts, within each fact category. We provide a brief explanation of our data sources here and refer the readers to the Appendix for more details. Our estimates and popular list sizes are shown in Table 1.

Brute-force strength calculation. To calculate brute-force strength, we needed the total number of possible facts for each category. For the "first name" (FN) and "last name" (LN) categories, we used the estimates from the U.S census [19, 16], U.S. Social Security Administration [21] and popular names available in 67 countries from Wikipedia. The total number of FN and LN is 285,537 and 6,209,229 respectively. For the total number of "full name" (FL) facts, we calculate a product of FN and LN counts for each country, and sum them up arriving at 563 B possible inputs. This overestimates the number of possible full names, since some FN-LN combinations may never occur in practice. But, we

could not find a good public source of full names, and were forced to approximate.

For the "city" (CI) category, we obtained the list of 754,450 locations with population greater than 5,000 people from DBpedia [13]. For the "place" (PL) category, we calculated the sum of the number of restaurants in the US [59] (1,232,016), the number of universities/colleges in US [45] (7,234) and in the world (8,766) [61], the number of elementary, middle, and high schools in US [44] (129,189), and the number of secondary schools in UK [70] (21,109). Note that this estimate does not include other popular attractions, such as amusement parks, hotels and monuments, and is thus an underestimate of the total number of inputs for the PL category. For the "relationship" (RL) category, we built a small list of relationships (49 entries), compiled from a dictionary. For the "object" (OBJ) and "activity" (ACT) categories, we used the size of the Wordnet [78] dictionary for nouns and verbs in English language. For the "year" (YR) category, we assumed that the user will recount experiences, which are at most 50 years in the past. The total number of inputs in the "date" (DT) category is calculated as $365*50$. Finally, the categories "hundred" (HU) and "ten" (TN) encompass facts, which have a limited number of possible answers (e.g., color of a bike, model of a car).

Statistical strength calculation. To calculate statistical strength we needed lists of popular facts in each category, ranked by popularity. We used online domain-specific sources to form these lists. We gathered around 434,000 unique popular list items from more than 530 different online sources. These sources include (1) Wikipedia/DBpedia [63], (2) Freebase [14], (3) U.S. Government sources: U.S. Census, U.S Social Security Administration, Dept. of Education, Dept. of Labor Stat., National Center for Education Statistics, (4) Other domain specific online sources: TripAdviser for popular travel destinations, Forbes and US News for educational institutions, IMDB for movie names, etc. (5) Popular English word lists from Google 20K [17], and 5K nouns, words, and lemma from the Corpus of Contemporary American English (COCA) [51]. We further incorporated popular lists for different categories from the Bonneau et al. dataset [70]. More details are provided in the Appendix.

Some lists had items ranked by popularity, while others did not (e.g. the list of the 100 most popular Chinese names from Wikipedia). For unranked lists, we used the Bing search engine to calculate the number of Web pages containing each list item. We automatically built structured queries as a "fact category + the item" (e.g., "first name Hao") and mined the number of pages from the search engine's reply. We then assumed that the popularity of an item is proportional to the number of Web pages containing it, and used this to rank the items in the list. While this may not be an accurate reflection of the popularity of each item, we believe it is a good approximation for relative rankings.

We have multiple lists of popular facts per category. We calculate the strength of a LEP fact as its *lowest* rank on any list. This approach assumes a strong statistical guessing attacker, which has the best popular list for each input.

Finally, if our popular lists were too small, we would overestimate the statistical strength of LEPs, as we would often use the brute-force strength for off-the-list facts. We show the count of popular lists per category, and their minimum and maximum sizes, as well as the total number of unique items in Table 1. For example, in the FN (first name) category, we have 384 popular lists, ranging from 3 to 38,717 inputs, and containing the total number of 150,695 unique names. We further evaluated how many facts collected in our user studies were covered by our popular lists. We were able to find 75% of FN, 99% of LN, 81% of FL, 63% of CI, 54% of OBJ, 46% of ACT and 34% of PL inputs on our popular lists. Thus our popular lists seem comprehensive enough for statistical strength calculation.

3.3 Creation

LEP creation requires users to actively provide input, from which the system extracts useful facts. In our work we have investigated guided and semi-guided methods for LEP input. These methods are triggered after a user has chosen the topic they want to talk about and provided its title. Figure 3 illustrates these input methods with one specific title "Trip to France".

In the *guided method* a user is prompted with a series of questions, chosen from a fixed set. The questions are displayed one at a time and the choice of the subsequent questions may depend on the user's answers to the preceding ones. This is illustrated in Figure 3(a). Some questions may be open-ended, e.g., "What else do you remember about ...".

In the *semi-guided method* the user is prompted to input a certain number of facts in the given category, and to provide a "hint" for each fact that will be used to form the authentication prompt. This is illustrated in Figure 3(b).

We also investigated a *freeform method*, where a user inputs a paragraph of free narrative, out of which we automatically extracted useful facts [96]. However, we abandoned this approach early since it had a large overhead for the users.

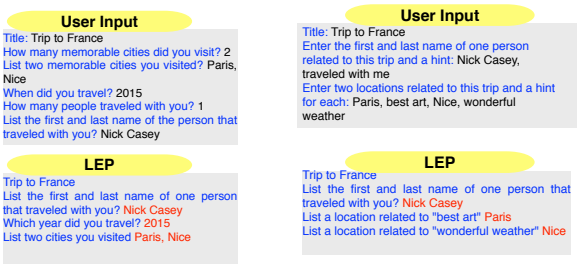
Our input methods guide the user toward useful facts, such as names, locations, objects, etc. and away from facts, which are not useful, such as preferences, opinions and feelings. The semi-guided method allows more freedom to the user to choose facts, which are relevant to her, but this freedom may lead to unstable facts. We evaluate these aspects in Section 5.

In extracting and processing useful facts from user responses, we normalize inputs for capitalization and punctuation. We also use POS tagging [87], dependency parsing [73], noun stemming and semantic role labeling [49] to extract the specific parts of user responses, such as verbs, nouns, subject, object, location, time, action, and person information. This helps us transform facts into question/answer pairs, and to identify, for multi-word inputs, those parts that carry the most meaning for the user (nouns, proper names of people and locations, verbs or adjectives).

3.4 Authentication

During authentication the system shows all the questions to the user, obtains the answers and compares them against one or several stored hashes.

Let a LEP contain N facts. We require that a user recalls M facts for authentication success, where $M \leq N$, and that the strength of the recalled facts be greater than some target value (we use 3class8 strength of 95^8 or 52.55 bits in our evaluation). The smaller the difference between N and M, the stronger the authentication criterion. Further, if $M < N$, the system must store $\binom{N}{M}$ hashes for one LEP. During authentication, the system produces all possible combinations of M user answers, and hashes them. Any match between these and stored hashes leads to authentication success. We



(a) Guided (b) Semi-guided

Figure 3: LEP input methods

have explored different values for N and M, and provide more information about their performance in Section 5.

Authentication may fail not only because a user forgot her answers, but also because she recalled them imprecisely. Imprecise recall of LEPs occurs due to a high redundancy of natural language, as explained below. We address some sources of mismatch through *imprecise matching*, which includes *normalization*, *keyword extraction* and *reorder matching*. While imprecise matching will reduce strength of LEPs, our evaluation shows that resulting LEPs still have high security (Section 5), and that imprecise matching significantly improves recall.

Capitalization, reordering and punctuation. A user might respond to the prompt using different capitalization or punctuation than she did during password creation. We overcome this by normalizing user answers before storage and authentication, by removing all capitalization and punctuation. A user may also list several parts of the answer in a different order, e.g., she may reply "Nice, Paris" where the original answer was "Paris, Nice". We resolve this through reorder matching. We detect when an answer may consist of multiple parts, and try all permutations of these parts in the matching process.

Misspelling. A user may misspell a reply during password creation or authentication. We leave handling of misspelling for future work.

Synonyms. A user may reply to a question with a near-synonym to the extracted term, such as responding with "lake" instead of "pond"; or with a term that is more specific (hyponym) or more general (hypernym) than the expected term, such as "dog" instead of "poodle" . We leave handling of synonyms, hyponyms and hypernyms for future work.

Extraneous words. A user may provide extraneous words in an answer. For example a question "what was red" may lead the user to input "my apple was red" even though we expect just "apple" as an answer. We address this via keyword extraction. We apply keyword extraction both during password creation and during authentication, in the same manner. The extraction method depends on the answer category (see Table 1). From answers in OBJ category we extract nouns only, from PL answers we extract nouns and out-of-dictionary words (likely place names), and from ACT answers we extract verbs only. Other categories do not need keyword extraction.

3.5 Uses of LEPs

LEPs could be used instead of passwords, but they may not be best suited for all authentication tasks, because their creation and authentication are more time-consuming. Extended authentication time may be especially burdensome to users on mobile devices, where keyboard input is slower than on desktops/laptops.

One possibility would be to use LEPs instead of passwords for first-time authentication, when cookies have expired, when the user is accessing an online service from a new machine, or when the user is logging onto his local machine after a logout. In these rare situations the added overhead of LEPs may be acceptable to users, at the benefit of higher security. LEPs could further be activated only for high-value accounts, such as bank or e-mail, where secure access is crucial.

Another possibility would be to use LEPs for secondary or added authentication, instead of security questions. We show in Section 5 that LEPs surpass security questions in security, recall and strength against friend guessing. Many high-value services currently use text messages sent to user phone with a code, to reduce risk of password cracking. LEPs could be used in lieu of the code, when a users does not have access to her phone (e.g., during international travel).

LEPs could also be used for continuous authentication on high-security servers, such as government or bank servers. A logged in user may be prompted for one or several facts after a period of inactivity to verify that someone did not gain physical access to her computer.

4. USER STUDIES

We evaluated LEPs through a series of user studies. We implemented LEP creation and authentication as a Web application, so that it can be used remotely. We then ran multiple user studies over the period of two years, with Amazon MTurk participants [7] and with students at our institution, and used their results to refine and improve both our user interface, and our elicitation process. All studies were approved by our Institutional Review Board (IRB). All communication with participants was in English, and their inputs were also required to be in English. Participants were required to be 18 years of age to participate.

Each participant was first shown the Information Sheet explaining the purpose of the study. Following this she registered in one of three ways, depending on the specific study's design: by entering her MTurk ID, or her E-mail address or the system assigned her a random identifier. A participant then input her demographic information (age, gender, and native language) and proceeded with the study. In this paper we report on the two culminating studies, which we used to evaluate LEP performance. We describe our study design in this section and provide detailed results and their interpretation in the next section.

Performance Study. This study was run in Fall 2015 and Spring 2016. It was designed to evaluate strength against brute-force and statistical guessing attacks, memorability and reuse of LEPs, and compare them to the same qualities of ordinary, 3class8 [85] passwords. We recruited participants from Amazon MTurk, and asked each to create ten LEPs and ten ordinary passwords. This scenario is unnatural, since no user would create so many passwords in such a short time span in real life. However, asking for ten passwords enabled us to study password reuse, because we could measure how many among the ten (passwords or

LEPs) are similar or same.[1] We asked participants to create passwords (LEPs or 3class8 passwords) for the following ten online sites and displayed their logos during creation and authentication: Facebook, Google+, Gmail, Outlook, Bank of America, Chase Bank, Target, WalMart, Wall Street Journal and CNN.

Asking users to create passwords for fictional servers and recall them later will necessarily underestimate recall in real life, because user motivation to remember these passwords is low. We believe that this confounding factor will have a similar effect on LEP recall as on password recall.

We required the ordinary passwords to follow 3class8 policy – being at least 8 characters long, and containing characters from 3 out of 4 character classes: uppercase and lowercase letters, digits and special characters. Each LEP was required to specify at least 5 facts. Users were asked to return for authentication after one week. We allowed three authentication attempts per LEP or password. We paid $1 for the creation task, and $2 for the authentication task.

To minimize password reuse, we did not let participants select a LEP topic from a list. Instead, we offered a topic for each LEP, which was randomly selected from our topic list. A participant could reject the offered topics, until she finds the one that she wants to talk about.

At registration, each participant was randomly assigned to either guided or to semi-guided input category. She then created and authenticated with all 10 LEPs using the same input method. The guided group was asked between 5 and 15 questions per LEP. The semi-guided group was asked to specify two people, one location and two objects for each LEP, and to specify a hint for each fact. After creating all 10 LEPs, each participant was asked to create 10 passwords, for the same servers. Participants were reminded via E-mail to return for authentication after one week. We also invited them to return for authentication after 3–6 months to measure long-term recall.

Friend Guessing Study. This study was run concurrently with our performance study, using the same system. We recruited participants from our institution (University of Southern California) to conduct an in-lab study, with the goal to measure the strength of LEPs against a friend attacker. In addition to personal knowledge, we encouraged the guessers to fully utilize information available from various social network sites and search engines. We observed that many participants in the study indeed made use of these online information sources.

Participants were required to enroll into the study with at least one other friend. We advertised the study via class announcements, wall posters, and flyers. Each participant was paid $10, and took 45-60 minutes to complete the study.

Unlike our previous studies, this study used deception in the Information Sheet (approved by our IRB), by not informing the participants that they will be guessing each others' LEPs. This was necessary to prevent participants from intentionally creating LEPs, which would be either too easy or too hard for a given friend to guess. We designed our study to mimic the real-life password use, where one does not know who may try to guess one's password.

After reviewing the Information Sheet, each participant was asked how close they were with their friend on the scale from 1 to 5 (closest), and how long they knew each other. Next, they were asked to create three LEPs for three different online accounts: Gmail, Facebook and Bank of America. We displayed corresponding logos during creation and authentication. Each participant was randomly assigned to either guided or semi-guided input method, and created all LEPs using this method.

Next, participants were asked to authenticate with each LEP, and were allowed unlimited number of trials, but required to make at least three. We incorporated user authentication in the study to ensure that participants did not make up answers they could not recall themselves.

After authentication, we debriefed each participant about the deception and explained our reasons for this. We informed them that their friend would be guessing their LEP and vice versa. They were offered a chance to quit the study at this time, and still receive the full payment. No participants quit.

Next, each participant attempted to guess her friend's LEP's, and was allowed unlimited number of trials, and required to make at least three. Afterwards, we reviewed successfully guessed answers with participants, asking them about their strategy. We ended the study with a short survey about usability of LEPs. Lastly, we asked the participants not to disclose details about deception to other students on campus, so we could continue recruitment.

Limitations and Ecological Validity. Our study had the following limitations, many of which are common for online password studies. First, it is possible but very unlikely that a participant may enroll into our Performance study more than once. While the same Mechanical Turk user (as identified by her MTurkID) could not enter the study twice, it is possible for someone to create multiple Mechanical Turk accounts. There is currently no way to identify such participants.

Second, we cannot be sure that our Performance study participants did not write down or photograph their LEPs or passwords. We can only claim that they had very low incentives to do so (since we promised payment regardless of authentication success). Our study mechanisms further detected copy/paste actions and we have excluded any participant that used these (for whatever reason) from the study. We also reminded the participants multiple times to rely on their memory only. If any cheating occurred it was likely to affect all the results equally, without bias to LEPs or passwords.

Third, Mechanical Turkers may not be very motivated or focused – this makes it likely that actual recall both of real-world LEPs and of passwords would be higher. While it would have been preferable to conduct our Performance study in the lab, the cost would be too high (for us) to afford as large a participation as we had through the use of Mechanical Turkers. We believe that any effect from participant motivation on recall applies equally to LEPs and to passwords.

Fourth, participants in our Friend Guessing study were all recruited from our university and thus were not very close. Best friends or family members may have higher success at guessing LEPs. We plan to investigate this in our future research.

[1]We initially attempted a phased study design, where a participant created one password (LEP or 3class8) and returned after one week to authenticate. After authentication the participant created another password for the next cycle. However, we had to abandon this study idea due to high attrition rates.

5. STUDY RESULTS

In this Section, we report on the results of our two user studies. We found that: (1) LEPs are 30–47 bits stronger than an ideal, randomized, 8-character password, (2) LEPs are 2–3× more memorable than passwords, (3) LEPs are reused half as often as passwords, (4) LEPs are 24–35× harder for friends to guess than security questions, (5) LEPs contain 2.4–3.2× fewer fake answers than security questions.

5.1 Participant Statistics

Table 3 shows the breakdown of our participants in the first two rows. We show count of participants, who completed both password creation and authentication tasks, and the total number of passwords. We also collected demographics, age and language of participants but found that these factors did not have significant impact on security, memorability or password reuse. With regard to topics chosen by participants, 55% of LEPs talk either about learning (26%), people (18%) or trips (11%), while other topics were less popular.

5.2 LEPs Are Memorable and Secure

In this section we report findings from our performance study, described in Section 4.

Privacy risk. LEPs have more sensitive information than passwords and, since their title and questions are displayed in clear, that may increase privacy risk to a user. It is difficult to accurately measure sensitivity of LEPs, as it depends on a user's subjective assessment how specific information relates to her sense of privacy. Instead, we calculate how many LEPs contain information that *most people would find sensitive*, such as information about illness, incarceration, love affairs (excluding those involving boyfriend/girlfriend and spouse) and indecent or illegal activities. Our result represents a lower bound on sensitive information contained in our dataset. 29 out of 930 LEPs, or 3%, contained sensitive information. Majority of these were LEPs about the death topic, and the sensitive information was divulged in the guided LEPs, because we asked about the cause of death. Better question design can further reduce this privacy risk.

Security. Table 3 shows the total number of LEPs, and their average brute-force and statistical strength in the 3rd and the 4th row. We also calculated the percentage of LEPs, which have the higher strength than a random, 3class8 password. We denote this strength as S_{3c8}. Almost all LEPs (94–95%) exceed S_{3c8}, with the average strength being at least 30 bits higher. Statistical strength of LEPs is also quite high – it is 30–47 bits higher than S_{3c8}.

Short-term recall. We report the percentage of successful authentications, after one week, in Table 3, in the rows 5–12 for LEPs and for passwords. For LEPs, in addition to *all-fact* recall, we investigated three alternative authentication schemes – *five-fact*, *four-fact* and *three-fact* recall. In these, the user must successfully recall 3, 4 or 5 facts, respectively, and the statistical strength of the recalled facts must exceed S_{3c8}.

Authentication success after one week is shown in rows 5–8 of Table 3. Password recall in our study was 26% (others have found 45–70% recall [85, 92, 93], but they asked users to recall only one password after 2 days). Our findings are consistent with psychological literature [64, 77], where Ebbinghaus found that people retain only 25% of new information they learned after six to seven days [77].

Imprecise matching greatly helps to increase LEP recall. With exact matching, all-fact authentication would be 19% for guided and 9.6% for semi-guided LEPs. With imprecise matching, it is 31.6% and 45.7%.

LEP authentication success with all facts is 30–75% higher than that for passwords. When we require fewer facts for authentication the success rate increases significantly. At four facts, LEP success is 2.7× higher than password success rate, and at three facts it is 3.2× higher. Allowing users to authenticate with fewer than all facts lowers security of LEPs. Since we also require that the statistical strength of recalled facts exceeds S_{3c8}, the remaining strength of these "shortened LEPs" is sufficiently large to thwart attacks. In Section 5.3 we will investigate how requiring $M < N$ facts for authentication affects strength against friend attacks.

Security questions have a wide range of recall rates after one month – from 32.1% for frequent flyer number to 83.9% for city of birth [67]. LEP recall with four-fact and three-fact authentication is 70–89.2% and thus resembles recall for memorable security questions. If a LEP were equivalent to a set containing several security questions, its best recall rate would be $83.9^4 = 50\%$ for four-fact and $83.9^3 = 59\%$ for three-fact authentication. The fact that LEP recall exceeds these values shows the power of user-customized questions and imprecise matching, over general questions and exact matching.

To understand reasons for failures to recall a LEP we examine recall rate per fact category (as given in Table 1). Table 4 shows the percentage of correctly recalled facts (in at least one attempt) per fact category in column 2. Relationships and cities are most accurately recalled, followed by items in the HU category, places, and first and last names. Overall, all categories except ACT have recall of more than 70% after one week. While this is quite high, it may be puzzling why a user would fail to recall *all* facts correctly, or at least at higher rates. We note some frequent reasons for failed recall per category in column 3. Many of these could be handled by better NLP techniques, e.g., using stemming for verbs, trying synonyms during matching, building a database of common abbreviations (e.g., gf for girlfriend), etc. This would further improve recall, at some security cost. We leave this direction for future work.

Cat.	Recall	Failure reasons	Guess
FN	77%	Misspelling, nickname, FL	20%
FL	81%	FN, misspelling	5%
PL	82%	Misspelling, abbreviations, synonyms	13%
CI	92%	Misspelling, more/less specific ans	53%
OBJ	79%	More/less specific ans	10%
ACT	51%	Tense mismatch, miss verbs	-
DT	77%	Total miss	16%
YR	73%	Total miss	17%
RL	95%	Abbreviation	48%
HU	82%	Synonym, more/less specific	21%
TN	80%	Synonym	36%

Table 4: Fact recall and guess success per category

Overall there were 18.2% of facts, which a user failed to recall in any authentication attempt. Users provide fake answers to security questions. They may also provide fake answers to LEPs, that they could not later recall. While we cannot accurately establish which facts are fake, and which are not, we estimate incidence of lying by looking for facts, which a user failed to recall, and where failure cannot be attributed to the NLP reasons we listed in Table 4, i.e., it

Row	Measure		Sec. 5.2 Perf. LEPs Mechanical Turk		Sec. 5.3 Friend Guess. USC		Sec. 5.2 Perf. Passwords Mechanical Turk	Literature Security Questions
			Guided	Semi-guid.	Guided	Semi-guid.		
1	Participants		44	49	47	44	93	literature
2	Passwords		440	490	-	-	930	literature
3	Brute-force (avg)		161 bit	132 bit			53 bit	-
4	Statistical (avg)		99 bit	82 bit	-	-		
5	Recall (1 week)	all-fact	31.6%	45.7%	-	-		
6		five-fact	47.7%	45.7%	-	-		
7		**four-fact**	**70.0%**	**73.0%**	-	-	26%	32.1%–83.9% [67]
8		three-fact	82.1%	89.2%	-	-		
9	Recall (3-6 mo)	all-fact	16.5%	32.3%	-	-		
10		five-fact	33.9%	32.3%	-	-		
11		**four-fact**	**53.0%**	**54.0%**	-	-	9%	6.4%–79.2% [67]
12		three-fact	66.5%	73.6%	-	-		
13	Friend guessing	all-fact	-	-	0.7%	0%		
14		five-fact	-	-	0.7%	0%		
15		**four-fact**	-	-	**0.7%**	**0%**	-	17–25% [91]
16		three-fact	-	-	1.3%	4.5%		
17	Fake info.		15.7%	11.5%	-	-	-	37% [67]
18	Identical (avg)		3.1%	2.7%	-	-	5.7%	-
19	Similar (avg)		15.4%	4.6%	-	-	31.6%	
20	Time to create (med)		112.7 s	112.0 s			16.8 s	-
21	Time to succ. auth. (med)		51.9 s	37.3 s			11.3 s	-

Table 3: Participant Statistics and Results of Our Studies

is a total miss. We find that 11.5–15.7% of facts were not recalled by users, and thus may be fake. This rate is less than a half of the fake answer rate for security questions [67]. A finer investigation of these "fake facts" shows that about half of them fail authentication because a user used an initial instead of a last name in LEP creation, but reverted to full name in authentication. We could handle this case with better authentication prompts.

Long term recall. We invited all participants in our performance study to authenticate with their LEPs and passwords once again, in May 2015. A total of 54 participants returned. The time between creation and authentication for these return participants ranged from 104 to 231 days, with a median of 120 days. Table 3 shows the long-term authentication success in rows 9–12. While both LEP and password recall has declined, time lapse affected recall of passwords much more than recall of LEPs. Password recall declined by 66%, while LEP recall declined by 17–47%. We thus conclude that LEPs are more robust with regard to long-term recall, than passwords.

Security questions have a wide range of recall rates after 3–6 months – from 6.4% for frequent flyer number to 79.2% for city of birth [67]. LEP recall with four-fact and three-fact authentication is 53–73.6%, within the range of more memorable security questions.

Reuse. We also explored strength of LEPs and passwords against a password-reuse attacker. The results are shown in Table 3 in rows 18–19. We first investigated how many out of 10 passwords were identical, for each given user. A LEP fact is said to be identical to a fact in another LEP, by the same user, if their answers would match during authentication (accounting for capitalization, reordering, punctuation and extraneous words). A LEP l_1 is identical to the LEP l_2 if all of l_1's facts match the facts l_2. There were 2.7–3.1% identical LEPs, compared to 5.7% identical passwords.

We next investigated how many out of a user's ten passwords were sufficiently similar to each other, so that a password-reuse attacker could easily guess one if he knew the other. Because LEPs are authenticated based on fact matches, and passwords based on the exact string match, it is hard to devise a similarity measure that applies equally well to both concepts. To define similarity of two authentication tokens (LEPs or passwords), we borrow from the Linux Pluggable Authentication Modules (PAM) [46] design. We say that two tokens $t1$ and $t2$ are similar, if more than 1/2 of items in $t2$ also appear in $t1$. For passwords, items are characters and for LEPs, items are facts. This definition allows us to directly apply `pam_cracklib` to detect similar passwords. We say that a password $op1$ is similar to the password $op2$ if one of the following conditions is met: (1) more than half of $op1$'s characters appear in $op2$ (this includes cases when (2) $op1$ is a palindrome of $op2$ or a rotated version of $op2$), or (2) $op1$ differs from $op2$ only in case. There were 4.6–15.4% similar LEPs, in semi-guided and guided categories, respectively, compared to 31.6% similar passwords. Thus passwords were reused more than twice as often as LEPs, and guided LEPs were reused 3 × more often than semi-guided LEPs.

One possible reason for such low LEP reuse is the way we prompted users for LEPs (described in Section 4). Whenever a LEP was to be created, we offered to the participant one topic, randomly selected from our list. The participant could reject the offered topics, until she found the one that she wanted to talk about. These prompts seem to have stimulated recall of diverse memories of past events, and lowered LEP reuse. We expect that servers that adopt LEPs would also use such user prompts to lower reuse.

Time to create and authenticate. We show the median time to create and authenticate a LEP or a password in Table 3, in rows 20–21. LEPs require 6.7× longer to create and 3.3–4.6× longer to authenticate, than passwords. This is expected as they require a user to both read the questions and to provide input that is approximately five times longer than a password.

Storage. LEP answers should be concatenated and stored as one or several hashes. In case of all-fact authentication, we store only one hash per LEP. If we allow for M-fact authentication (M=3, 4 or 5), we would have to create, hash and store $\binom{N}{M}$ combinations of facts for each LEP, where N is the number of all facts. 77% participants would need up to

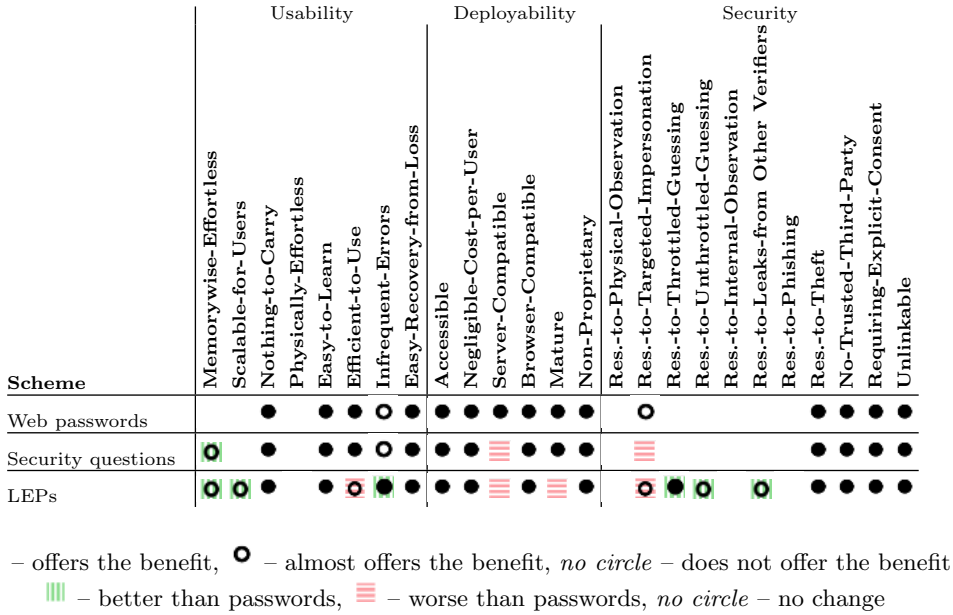


Table 5: Comparing security questions and LEPs to passwords using the UDS framework [68]. LEPs outperform passwords in six categories related to usability and security, and underperform in two deployability categories. LEPs are also more usable and more secure than security questions.

35 hashes, 88% would need up to 70, 92% would need up to 126 hashes, and the worst case scenario would require 1,287 hashes. Even in the worst case, the storage cost would only amount to several Kilobytes per user, which is negligible for today's server storage. Because one-way hashing is fast, the processing cost should also be acceptable. We could further limit the storage cost of LEPs by discarding all but N strongest facts, at creation time. For $M = [3, 4]$ and $N = 8$ each LEP would require at most 70 hashes.

5.3 LEPs Are Strong Against Guessing

This section discusses results of our friend guessing study, described in Section 4. We recruited a total of 91 participants and a few participants came in groups of 3 or more people, forming 100 different pairs. All of the participants completed the study in the same sitting. The participants were students from our institution from freshmen to graduate students, with majors in engineering, social science, theater, biology, math, international relations, music, business, economics, psychology, and linguistics.

Participants knew each other between 3 months to 6 years. All of them were already friends on at least one social network (e.g., Facebook, Instagram, etc) and were encouraged to use social networks and public sources during guessing. For participant pairs that were not from the US, both participants were from the same country and shared the same cultural background, which helped them make more educated guesses. Average closeness rating on the scale 1–5, with 1 being the lowest, was 2.6 for the guided group, and 3 for the semi-guided group.

Most users were able to fully authenticate with their LEPs in the second stage of our study. A few authentication failures we observed were due to misspelling, synonyms and more/less specific answers provided during authentication than during creation.

We show friend guessing success in the Table 3, rows 13–16 using the same authentication schemes as in Section 3.4. Guess success rate is very low (0–0.7%) for all-fact authentication, and climbs up to 1.3–4.5%, for three-fact authentication. Taken together with recall results, this data suggests that the four-fact authentication seems to strike the right balance between achieving high authentication success (70-73%), and reasonable strength, while keeping the friend guessing success low (0–0.7%).

Compared to security questions, where friends could guess 17–25% [91], LEP's are 3.7–35× stronger, assuming four-fact or three-fact authentication.

Guess success rate per fact category, for facts described in Table 1, is shown in column 4 of Table 4. Friends could successfully guess more than half of the cities, and more than 20% of relationships, places, first names, and facts in HU and TN categories. Other categories, such as FL, PL, OBJ, DT and YR had lower guess success. This shows another strength of LEPs over security questions. Bonnie et al. found that security questions could not strike the right balance between security and memorability, because facts that were memorable for users were also easily guessed [67, 91]. Conversely, LEPs have multiple fact categories with high user recall and low friend guess rate (e.g., FN, FL, PL, OBJ).

Participant Feedback. We carefully observed participant behavior during guessing and interviewed them about their strategy after the study. Except two participants, who attempted random guesses, the rest invested significant effort in looking for possible guess options online. They reported using the following information sources for guessing: personal knowledge, Facebook, Google search engine, Instagram, RenRen, WeChat, QQ, Line, LinkedIn, and Spokeo. Overall, 78–83% of participants reported using personal knowledge, 76–79% used social networks and 50–56% used search engines. At social networks, participants checked the personal profile and friend lists first, and then scanned the re-

cent wall posts. They complemented this with online searches for popular items which appeared in LEP questions.

We further collected participant feedback on why guessing LEPs was hard. About half of responses (48%) stated that LEPs involved too much *detail* about the topic, which was hard to mine from online sources or from personal knowledge. Unless the friend participated in the same event, it was difficult to mine correct answers online. In addition, 26% of participants acknowledged that many facts were too *personal* and thus not shared among friends, nor used in social network posts, such as events from early childhood and elementary school. These *private* and *unique* facts make guessing difficult, unlike facts used for security questions, which can be easily found in online sources.

When asked about what they liked about LEPs, 58.2% of participants stated that LEPs were much harder to guess than the current security questions, and more memorable than ordinary passwords since they were built from personal experience. In addition, 11.4% participants said that the variety of question sets and their detail were another advantage of LEPs over security questions. The downsides reported by participants were the time it took to come up with answers (6.3% of participants) and the concern about memorability (19% of participants).

More than 90% of participants reported that they would consider using and adapting LEPs for different online accounts, while 6.3% said that they did not plan to use LEPs due to time overhead and concerns about memorabilty. 44.3% were willing to use LEPs for an online banking account, which needs high security, and has less frequent logins, 31.6% said they would use LEPs for secure and professional email accounts, and 17.7% would use them for government accounts and health records.

5.4 LEPs Are Usable and Secure

In [68], Bonneau et al. present the usability-deployability-security (UDS) framework – they define 25 properties of Web authentication schemes and use them to rate 35 password-replacement schemes. We reproduce the rating of passwords and personal knowledge (security) questions in Table 5, and add our rating for LEPs.

LEPs outperform passwords in six categories. Regarding usability, LEPs are *Quasi-Memorywise-Effortless*, as users still forget some LEP facts. They are *Quasi-Scalable-for-Users*, since users have abundance of memories that servers can elicit by randomizing topic offerings, as we did in our user studies. They are further *Infrequent-Errors*, as our imprecise matching takes care of many causes of authentication failures. Regarding security, LEPs are *Resilient-to-Throttled-Guessing* and *Quasi-Resilient-to-Unthrottled-Guessing*, as our evaluation shows. Thanks to low LEP reuse, they are also *Quasi-Resilient-to-Leaks-from Other Verifiers*. LEPs do worse than passwords in one usability, two deployability and one security category. They are less *Efficient-to-Use*, as they take longer to create and authenticate. They are not *Server-Compatible* nor *Mature*, because they are a research technology and are not widely deployed. Security-wise, LEPs are *Quasi-Resilient-to-Targeted-Impersonation* – friends can guess a very small portion of LEPs.

LEPs outperform security questions in six categories. They are *Quasi-Scalable-for-Users*, due to their wide applicability and broad range of topics, while security questions are not. They are further *Infrequent-Errors*, because we apply imprecise matching. LEPs are further *Quasi-Resilient-to-Targeted-Impersonation*, *Resilient-to-Throttled-Guessing*, *Quasi-Resilient-to-Unthrottled-Guessing*, and *Quasi-Resilient-to-Leaks-from-Other-Verifiers*, while security questions have none of these qualities. There are only two categories where LEPs lag after security questions – they are less *Efficient-to-Use* and not *Mature*.

Comparing LEPs to other 33 authentication technologies in publication [68] (Table I in [68]), only federated authentication schemes offer both better usability and security, but at the cost of lower deployability. Thus LEPs are competitive against other authentication schemes, and can offer a unique combination of features.

6. CONCLUSIONS

Textual passwords are a widely-used form of authentication and suffer from many deficiencies because users trade security for memorability. Users create weak passwords because they are memorable, and reuse the same password across many sites. Forcing users to create strong passwords does not help, as these are easily forgotten.

We have proposed life-experience passwords (LEPs) as a new authentication mechanism, which strikes a good balance between security and memorability. We investigated several LEP designs, and evaluated them in two user studies. Our results show that LEPs are much more memorable and secure than passwords, they are less often reused, and they are strong against friend guessing. While they take more time to create and input during authentication, we believe their benefits may make them a viable primary authentication mechanism for high-security servers, or a much better secondary authentication mechanism than security questions.

7. ACKNOWLEDGEMENTS

We would like to thank Jason Hong (CMU) for valuable feedback on previous versions of this manuscript. We are also grateful to anonymous reviewers for their helpful comments.

8. REFERENCES

[1] Occupational Employment Statistics . http://www.bls.gov/oes/current/oes_nat.htm.
[2] Party. https://en.wikipedia.org/wiki/Party.
[3] 10 Tried-And-True Wedding Flowers. https://www.theknot.com/content/top-10-wedding-flowers.
[4] 100 Greatest Actors of All Time. http://www.imdb.com/list/ls000034841/.
[5] 50 College Graduation Gift Ideas. https://www.universityparent.com/topics/parent-posts/50-college-graduation-gift-ideas-for-parents-2/.
[6] Academic Ranking of World Universities . https://en.wikipedia.org/wiki/Academic_Ranking_of_World_Universities.
[7] Amazon mechanical turk. https://www.mturk.com/.
[8] Best Places to Propose. http://www.brides.com/honeymoons/2014/04/best-honeymoon-destinations#slide=2.
[9] Brain Authentication. *http://brainauth.com/testdrive/*.
[10] Chinese girl names. http://www.top-100-baby-names-search.com/chinese-girl-names.html.

[11] Chinese male names. http://www.top-100-baby-names-search.com/chinese-male-names.html.
[12] Cognitive password. *http://en.wikipedia.org/wiki/Cognitive_password/*.
[13] DBpedia. *http://wiki.dbpedia.org/*.
[14] Freebase. *http://www.freebase.com/*.
[15] Frequently occurring surnames from census 1990 - names files. http://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html. Accessed: 2015-10-14.
[16] Frequently occurring surnames from the census 2000. http://www.census.gov/topics/population/genealogy/data/2000_surnames.html. Accessed: 2015-10-14.
[17] google-10000-english. https://github.com/first20hours/google-10000-english/. Accessed: 2015-10-14.
[18] Highest Rated TV Series With At Least 5,000 Votes . http://www.imdb.com/chart/top.
[19] How Common is Your Last Name? http://www.pbs.org/pov/thesweetestsound/popindex.php.
[20] Indian name. https://en.wikipedia.org/wiki/Indian_name. Accessed: 2015-10-14.
[21] John the ripper password cracker. https://www.ssa.gov/OACT/babynames/limits.html. Accessed: 2015-10-14.
[22] List of best-selling books. https://en.wikipedia.org/wiki/List_of_best-selling_books.
[23] List of best-selling fiction authors. https://en.wikipedia.org/wiki/List_of_best-selling_fiction_authors/.
[24] List of best-selling music artists. https://en.wikipedia.org/wiki/List_of_best-selling_music_artists.
[25] List of cities in China. https://en.wikipedia.org/wiki/List_of_cities_in_China.
[26] List of cities proper by population. https://en.wikipedia.org/wiki/List_of_cities_proper_by_population.
[27] List of common chinese surnames. https://en.wikipedia.org/wiki/List_of_common_Chinese_surnames. Accessed: 2015-10-14.
[28] List of hobbies . https://en.wikipedia.org/wiki/List_of_hobbies.
[29] List of largest hotels in the world. https://en.wikipedia.org/wiki/List_of_largest_hotels_in_the_world.
[30] List of most common surnames in asia. https://en.wikipedia.org/wiki/List_of_most_common_surnames_in_Asia. Accessed: 2015-10-14.
[31] List of most common surnames in north america. https://en.wikipedia.org/wiki/List_of_most_common_surnames_in_North_America. Accessed: 2015-10-14.
[32] List of most common surnames in oceania. https://en.wikipedia.org/wiki/List_of_most_common_surnames_in_Oceania. Accessed: 2015-10-14.
[33] List of most common surnames in south america. https://en.wikipedia.org/wiki/List_of_most_common_surnames_in_South_America. Accessed: 2015-10-14.
[34] List of most commonly learned foreign languages in the United States. https://en.wikipedia.org/wiki/List_of_most_commonly_learned_foreign_languages_in_the_United_States.
[35] List of most popular given names. https://en.wikipedia.org/wiki/List_of_most_popular_given_names.
[36] List of national parks of the United States. https://en.wikipedia.org/wiki/List_of_national_parks_of_the_United_States.
[37] List of sports attendance figures. https://en.wikipedia.org/wiki/List_of_sports_attendance_figures/.
[38] List of the most common surnames in europe. https://en.wikipedia.org/wiki/List_of_the_most_common_surnames_in_Europe. Accessed: 2015-10-14.
[39] List of towns in India by population. https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_India_by_population.
[40] List of United States cities by population. https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population.
[41] Microsoft corporation, sketch-based password authentication. US Patent number 8,024,775.
[42] Mnemonic Guard. *http://www.mneme.co.jp/english/index.html*.
[43] Mnemonic Guard Blog. *http://mnemonicguard.blogspot.com/*.
[44] National Center for Education Statistics, How many educational institutions exist in the United States? . https://nces.ed.gov/fastfacts/display.asp?id=84.
[45] Number of U.S. Colleges and Universities and Degrees Awarded, 2005. http://www.infoplease.com/ipa/A0908742.html.
[46] Pluggable authentication modules for linux (pam). http://www.linux-pam.org/. Accessed: 2015-10-14.
[47] Popular Indian Boy Names. http://babynames.extraprepare.com/boy-popular.php.
[48] Popular Indian Girl Names. http://babynames.extraprepare.com/girl-popular.php.
[49] Semantic role labeling. https://en.wikipedia.org/wiki/Semantic_role_labeling. Accessed: 2015-10-14.
[50] Sports in the United States. https://en.wikipedia.org/wiki/Sports_in_the_United_States.
[51] The corpus of contemporary american english (coca). http://corpus.byu.edu/coca/. Accessed: 2015-10-14.
[52] The World's Top 20 Honeymoon Destinations. http://www.brides.com/honeymoons/2014/04/best-honeymoon-destinations#slide=2.
[53] Top 10 Colors for Bridesmaid Dresses. http://www.tulleandchantilly.com/blog/top-10-colors-for-bridesmaid-dresses/.
[54] Top 100 Chains: U.S. Sales. http://nrn.com/us-top-100/top-100-chains-us-sales.
[55] Top 25 Most Popular Sports/Recreational Activities in the U.S. . https://www.sfia.org/.
[56] Top Rated Movies . http://www.imdb.com/search/title?num_votes=5000,&sort=user_rating,desc&title_type=tv_series.
[57] Top Ten Most Requested Cake Combos. http://nymag.com/shopping/guides/weddings/planner/features/topten_cakes.htm.
[58] Tripadviser. http://www.tripadvisor.com/.
[59] U.S. Total Restaurant Count Increases by 4,442 Units

over Last Year, Reports NPD. http://www.nytimes.com/interactive/2015/01/11/travel/52-places-to-go-in-2015.html?_r=0.

[60] Why haven't biometrics replaced passwords yet? http://www.digitaltrends.com/android/can-biometrics-secure-our-digital-lives/.

[61] World List of Universities, 25th Edition: And Other Institutions of Higher Education (World List of Universities & Other Institutions of Higher Education). http://www.amazon.com/World-List-Universities-25th-Edition/dp/1403992525.

[62] Your pick: World's 50 best foods . http://travel.cnn.com/explorations/eat/readers-choice-worlds-50-most-delicious-foods-012321.

[63] Wikipedia, the free encyclopedia, 2004. [Online; accessed 12-Feb-2016].

[64] L. Averell and A. Heathcote. The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, 55(1):25–35, 2011.

[65] H. Bojinov, D. Sanchez, P. Reber, D. Boneh, and P. Lincoln. Neuroscience meets cryptography: designing crypto primitives secure against rubber hose attacks. In *Proceedings of the 21st USENIX conference on Security symposium*, pages 33–33. USENIX Association, 2012.

[66] J. Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*, May 2012.

[67] J. Bonneau, E. Bursztein, I. Caron, R. Jackson, and M. Williamson. Secrets, lies, and account recovery: Lessons from the use of personal knowledge questions at google. In *Proceedings of the 24th International Conference on World Wide Web*, pages 141–150. International World Wide Web Conferences Steering Committee, 2015.

[68] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *2012 IEEE Symposium on Security and Privacy*, May 2012.

[69] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. Passwords and the Evolution of Imperfect Authentication. *Communications of the ACM*, July 2015.

[70] J. Bonneau, M. Just, and G. Matthews. What's in a name? In *Financial Cryptography and Data Security*, pages 98–113. Springer, 2010.

[71] N. M. Bradburn, L. J. Rips, and S. K. Shevell. Answering autobiographical questions: the impact of memory and inference on surveys. *Science*, 236(4798), 1987.

[72] C. Castelluccia, M. Dürmuth, and D. Perito. Adaptive password-strength meters from markov models. In *NDSS*, 2012.

[73] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, pages 740–750, 2014.

[74] S. Das, E. Hayashi, and J. I. Hong. Exploring capturable everyday memory for autobiographical authentication. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 211–220. ACM, 2013.

[75] D. Davis, F. Monrose, and M. K. Reiter. On user choice in graphical password schemes. In *USENIX Security Symposium*, volume 13, pages 11–11, 2004.

[76] T. Denning, K. Bowers, M. van Dijk, and A. Juels. Exploring implicit memory for painless password recovery. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 2615–2618. ACM, 2011.

[77] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Number 3. University Microfilms, 1913.

[78] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, 1998.

[79] V. Griffith and M. Jakobsson. Messin' with texas: Deriving mother's maiden names using public records. In *Applied Cryptography and Network Security*, pages 91–103. Springer, 2005.

[80] N. E. A. Guideline. Nist special publication 800-63 version 1.0. 2, 2006.

[81] J. H. Huh, S. Oh, H. Kim, K. Beznosov, A. Mohan, and S. R. Rajagopalan. Surpass: System-initiated user-replaceable passwords. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 170–181. ACM, 2015.

[82] G. Inc. Facial recognition. US Patent number 8,457,367.

[83] I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A. D. Rubin. The design and analysis of graphical passwords. In *Proceedings of the 8th USENIX Security Symposium*, pages 1–14. Washington DC, 1999.

[84] M. Just and D. Aspinall. Personal choice and challenge questions: a security and usability assessment. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 8. ACM, 2009.

[85] S. Komanduri, R. Shay, L. F. Cranor, C. Herley, and S. Schechter. Telepathwords: Preventing weak passwords by reading users' minds.

[86] F. Magazine. The World's Most Valuable Sports Teams.

[87] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[88] L. Mastin. The human memory. http://www.human-memory.net/.

[89] A. Niedzwienska. Distortion of autobiographical memories. *Applied Cognitive Psychology*, 17(1):81–91, 2003.

[90] A. Nosseir, R. Connor, and M. Dunlop. Internet authentication based on personal history – a feasibility test. In *Proceedings of Customer Focused Mobile Services Workshop*, 2005.

[91] S. Schechter, A. B. Brush, and S. Egelman. It's no secret. measuring the security and reliability of authentication via 'secret' questions. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 375–390. IEEE, 2009.

[92] R. Shay, P. G. Kelley, S. Komanduri, M. L. Mazurek, B. Ur, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proceedings of the eighth symposium on usable privacy and security*, page 7. ACM, 2012.

[93] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. Can long passwords be secure and usable? In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2927–2936. ACM, 2014.

[94] A. Somayaji, D. Mould, and C. Brown. Towards narrative authentication: or, against boring authentication. In *Proceedings of the 2013 workshop on New security paradigms workshop*, pages 57–64. ACM, 2013.

[95] R. Veras, C. Collins, and J. Thorpe. On the semantic patterns of passwords and their security impact. In *Network and Distributed System Security Symposium (NDSS'14)*, 2014.

[96] S. Woo, Z. Li, and J. Mirkovic. Good automatic authentication question generation. In *Proceedings of 9th International Natural Language Generation (INLG)*, 2016.

[97] T. Yamamoto, A. Harada, T. Isarida, and M. Nishigaki. Improvement of user authentication using schema of visual memory: Exploitation of "schema of story". In *Advanced Information Networking and Applications, 2008. AINA 2008. 22nd International Conference on*, pages 40–47. IEEE, 2008.

APPENDIX

A. APPENDIX

Table 6 shows data sources, which we used to build our popular item lists.

Subcategory	Possible ans.	Sources
Last Name (LN)		
US	$\geq$ 88,879	[16, 15]
China	$\geq$ 100	[27]
India	$\geq$ 580	[20]
43 European Countries	$\geq$ 77,000	[38, 70]
23 Other Countries	$\geq$ 116,000	[30], [33], [31], [32]
First Name (FN)		
US	$\geq$ 5,494	[15, 21]
China	$\geq$ 259	[11, 10]
India	$\geq$ 1833	[47, 48]
43 European Countries	$\geq$ 72,408	[35, 70]
23 Other Countries	$\geq$ 1,240	[35]
Cities (CI)		
US	$\geq$ 298	[40]
China	$\geq$ 642	[25]
India	$\geq$ 870	[39]
U.K Cities	$\geq$ 100	[70]
Top 85 largest Cities in the world by Population	85	[26]
Popular Places (PL)		
US and World Tourist Attraction Places (Best Places to visit in US and World)	$\geq$ 568	[58, 36]
Other Places (Schools, Hotels, Hospitals, Restaurants, etc)	$\geq$ 452	[54, 6, 29, 52, 8]
Activities/Actions (ACT)		
Top 25 activities in US	25	[55]
List of Hobbies	276	[28]
Objects (OBJ)		
Top Hobby, Popular jobs in US, Popular Grad. Gift, Top Activities, Best Movies, Best Singers, Best TV Series, Popular Sports in US, Popular Sports in World, Best Writer, Best Books, Popular Wedding flower, Popular Wedding cake, Popular color of bride maid dress, Top 50 food in the world	$\geq$ 1,162	[18, 4, 56, 24, 50, 37, 86, 23, 22, 3, 57, 53, 2, 62, 5, 34, 1]
Google 20,000 words (*after removing stopwords)	19,000	[17]
COCA	5,000	[51]

Table 6: Popular lists, their sizes, and sources