Exploring Capturable Everyday Memory for Autobiographical Authentication

Sauvik Das
Carnegie Mellon University
sauvik@cmu.edu

Eiji Hayashi
Carnegie Mellon University
ehayashi@cs.cmu.edu

Jason Hong
Carnegie Mellon University
jasonh@cs.cmu.edu

ABSTRACT
We explore how well the intersection between our own everyday memories and those captured by our smartphones can be used for what we call *autobiographical authentication*—a challenge-response authentication system that queries users about day-to-day experiences. Through three studies—two on MTurk and one field study—we found that users are good, but make systematic errors at answering autobiographical questions. Using Bayesian modeling to account for these systematic response errors, we derived a formula for computing a confidence rating that the attempting authenticator is the user from a sequence of question-answer responses. We tested our formula against five simulated adversaries based on plausible real-life counterparts. Our simulations indicate that our model of autobiographical authentication generally performs well in assigning high confidence estimates to the user and low confidence estimates to impersonating adversaries.

Author Keywords
Autobiographical Authentication; Smartphones; Everyday Memory; Capturable Everyday Memory; Android

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
In this paper, we examine a new kind of authentication leveraging *something you know*: one's own everyday autobiographical data. It is motivated by the observation that smartphones know a lot about their users, such as their call logs, location traces, and browser history. In turn, users might accurately remember only some of the details stored in these logs. We call this intersection of what smartphones can capture and what humans can remember *capturable everyday memory,* and explored how well it can be used for autobiographical authentication challenges.

There are several potential advantages to this approach. Unlike most other forms of authentication, autobiographical authentication can be scaled by context. For example, users might have to answer one challenge if authenticating in their homes, but five if in a city they have never visited before. This scaling property also allows for non-binary authentication. Rather than pivot access to all of a smartphone's data and permissions through one or two passwords, access can be granted in tiers. In other words, a user who wants to access banking information may have to answer several questions, while a user who wants to check the weather may have to answer one easy question. Furthermore, authentication challenges are automatically adjusted as a person goes about his daily life, making many attacks harder to successfully execute—for example, shoulder surfing, replay attacks and phishing.

This paper makes two contributions. First, we report on three studies—two on MTurk and one field study—to construct an empirical model of capturable everyday memory. In our two MTurk studies, we narrowed down the broad search space of candidate autobiographical questions. In our field study, we deployed an Android app that asks users to answer questions constructed from ground-truth data. We analyzed 2167 question-answer responses collected from 24 users over 2 weeks, and found that users answered approximately 64% of questions correctly, overall. Furthermore, the type of question mattered: questions about phone usage—about facts such as app usage and website visits—were answered correctly less often than questions about communications and location. Also, performance was stable over time: users performed as well at the end of two weeks as they did at the beginning.

Second, we offer a framework for autobiographical authentication that accounts for users' systematic response error to compute a confidence estimate that an attempting authenticator is the user. We evaluated our framework against five simulated adversaries based on different threat models. We found that while choice of adversary to optimize against mattered, our framework shows promise: generally, users obtained high confidence ratings while impersonators obtained low confidence ratings.

RELATED WORK
Textual Passwords and Alternatives
The core strengths of textual passwords are speed, convenience and challenge [7]. However, many studies have documented their weaknesses, as well. For example, Adams and Sasse [1] discuss the burdens that textual passwords impose on users, and how people tended to circumvent it and undermine security to get their work

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.

ACM 978-1-4503-1770-2/13/09…$15.00.
http://dx.doi.org/10.1145/2493432.2493453

done. Allan [2] states that the increase in computational power predicted by Moore's law continually increases the lower bound of acceptable password complexity. Equally, studies of password policy (e.g. [7,11,18]) have concluded that while users can be encouraged to select safer secrets, new attacks such as phishing and keylogging make high password strength a dubiously effective security measure.

Graphical passwords [19], such as Android's 9-dot password, is one alternative form of authentication that has been gaining popularity. However, while generally easy and quick to use, graphical passwords can be insecure as they are vulnerable to shoulder-surfing attacks and have small search spaces [19]. Challenge questions [10] are the most related alternative authentication technique. Challenge questions authenticate users by matching their answer to a preset query to one previously supplied. Questions typically focus on persistent facts relating to a user's life, such as "what is your mother's maiden name?" or "what is your birthday?" Studying these question systems, Rabkin [13] identified the increasing amount of personal information available online as a weakness: using online sources, attackers can retrieve basic information about a user to answer a wide range of the most commonly used questions. Furthermore, Schechter et al. [17] pointed out that challenge questions are easily guessed by acquaintances, that some answers are relatively predictable, and that many users forget their responses over time.

On the other hand, the increasing online presence of many users is providing new approaches to question systems. For example, in Love and Authentication [9] questions were derived from responses to surveys on online matchmaking and dating services. The authors found that responses to these questions were highly memorable, stable over time, and hard for others to guess. Facebook also implemented social authentication [15], where users authenticate to the site by identifying a number of their friends based on photos that those friends have posted. This approach provides advantages over traditional challenge question systems in that it requires minimal enrolment costs (questions are generated automatically from data stored on Facebook) and has low challenge reoccurrence—any photo from any friend can be used.

Autobiographical authentication can offer improvements over traditional passwords and challenge question schemes. By relying on automatically captured data it hopes to ensure a high diversity of presented challenges, while limiting how easily correct responses may be guessed. It also avoids explicit user enrollment costs; data is generated by users through their day-to-day activities and captured automatically via sensor-equipped smartphones.

Autobiographical Memory

Conway & Pleydell-Pearce [3] provide an informative perspective on autobiographical memory, suggesting that our "self-memory system" is divided by granularity: lifetime periods, general events, and event-specific knowledge. Most relevant to our cause is the latter: *Event-specific knowledge* (ESK) relates to vivid memories about specific event details, for example the act of text messaging a friend. In that regard, Conway and Pleydell-Pearce's [3] notion of ESK is highly salient for autobiographical authentication, as it refers to highly specific memories of events that have a short shelf life—they fade from memory in a matter of days or weeks. These memories are ideal for autobiographical authentication, as they likely correspond to memories that users find easy to recall, but are ephemeral and consequently relatively hard for either strangers or friends to guess, discover or deduce.

However, encoding autobiographical memories is a complex process. Conway and Pleydell-Pearce argue it depends on a range of unobservable factors such as emotion, affect and age [3]; others suggest that gender and vocabulary also exert an effect [20]. In a recent update to this literature, Kristo et al. [12] conducted an Internet-based diary study. They found that different aspects of everyday memories have different retention rates. For example, the content and time of a memory were better remembered than the details; less regularly occurring events were more likely to be remembered; and, pleasant events were better remembered than unpleasant events.

While these findings offer us guidance, there is little presently in the literature about the intersection between human memory and smartphone logs—*capturable everyday memory*. To better conceptualize the bounds of capturable everyday memory, we ran a series of studies: two moderate-scale MTurk mini-studies and one two-week long field study. From this data, we construct an empirical model of capturable everyday memory and derive a Bayesian framework for computing confidence estimates that an attempting authenticator is the user.

USER STUDIES

Mechanical Turk Questionnaires

Study 1: What comprises capturable everyday memory?

We first wanted a qualitative categorization of capturable everyday memory and to formulate candidate questions. We constructed a questionnaire utilizing the Galton-Crovitz cueing technique [4,5], a method frequently employed by cognitive psychologists who study autobiographical memories [12]. Participants were asked to recall the first memory that comes to mind associated with a keyword. Careful selection of keywords allowed us to nudge participants' memories to those capturable by smartphones. Thus, 28 keywords were selected by considering the broad categories of information a smartphone might know about its users. Example keywords include "alarm clock", "SMS", and "phone call." We omit the full list for brevity.

Participants had to answer the questionnaire for five keywords. For each keyword, participants had to describe, in at least 100 characters, a *specific*, recent memory associated with the keyword in relation to digital technology. If they were unable to think of such a memory,

Memory Category	N	Examples
Communication	53	SMS, SNS usage, phone calls
Content Consumption	30	Viewing photos, reading articles
Tech Failures	27	Battery failures, connectivity failures
Scheduling / Events	22	Scheduling & attending events
Travel / Transportation	28	Driving, public transit, GPS navigation
Internet Activity	24	General internet usage (browsing)
Technology Usage	64	Using apps / software, or hardware
Content Search	13	Searching on the net
Weather	10	Memories about weather

Table 1. Concept mapping response categorizations, along with representative examples.

they could enter any other recent memory associated with digital technology.

Results

We obtained 272 valid keyword questionnaires from 70 participants. Thirty-five of the participants were female, and the average age was 36 years (s.d. 12.9).

We constructed a data-driven categorization of user responses, given that the keywords we selected did not necessarily elicit different types of memories (e.g., phone call and SMS). We applied concept mapping [8], a mixed-methods analysis technique, with six coders as a means of generating our categorizations. From the responses, we identified 9 distinct categorizations of everyday memory that were also capturable (see Table 1). The categories are neither mutually exclusive nor exhaustive, but they are the more salient types of capturable everyday memories.

We also formulated a set of 50 questions distributed across these categories based on the questionnaire responses, which we used in our next study.

Study 2: Can people answer autobiographical questions?

Next, we wanted to gauge how well people believed they could answer the autobiographical questions that we formulated from the first study. We asked these questions on MTurk before running a field study for three reasons: (1) to ask a relatively large sample of users a large set of questions to narrow down the list of possible questions to ask in the final application; (2) to ask questions that are not presently feasible to ask on smartphones, but might be asked given more complete data stores (e.g., "what did you eat for lunch yesterday?"); and, (3) to establish hypotheses to guide our analysis of the field study data. Example questions we asked include: *Which wireless network did you connect to yesterday?*, *Name an article you read recently*, and *Who did you last SMS message?* We omit the full list of 50 questions here for brevity.

Participants had to answer five autobiographical questions. For each question, they were also asked how strongly they agreed with a set of prompts on a Likert-scale of 1 (Strongly Disagree) to 7 (Strongly Agree). These prompts are listed in Table 2. The first three prompts are from previous work [12,16]. Rubin et al. [16] found that respondents' answers to a set of Likert-scale questions were indicators of the accuracy of a memory. Consequently, we use a subset of these same questions as a rough surrogate for ground truth. The last prompt was of our own inclusion, presented to control for answer uniqueness on memorability.

Likert-scale prompts in Study 2.
I can mentally relive the event in my answer.
I actually remember the event in my answer, rather than just knowing they happened.
I am confident in my answer.
The event in my answer is a unique event in my life.

Table 2. List of all Likert-scale prompts participants were asked to respond to with each question answered.

Results

We gathered 632 question-answer responses from 145 participants, ranging in age from 18 to 64 (mean: 33.7, s.d.: 10.1). Eighty-three were female and 62 were male.

We modeled the memorability of responses using a generalized linear mixed effects model [14] with the user as a random effect because we collect multiple responses from each user. Specifically, we utilize a random-intercepts model to allow different users to have different base memorability scores. The memorability score we model is the sum of the responses to the Likert-scale questions associated with each response that we borrowed from [12,16]. As there were three supplementary 7-point Likert-scale questions, the range of the response varied from 3 to 21, with a higher score indicating greater memorability.

Table 3 shows the model coefficients. For numeric variables (i.e., age, time elapsed, uniqueness), the coefficient is the effect on memorability from a 1-unit increase in the variable, holding all other numeric variables at their mean and all other categorical variables at their baseline.

The intercept in the model was high at 15.3 out of 21, suggesting that users generally believed they could answer these questions confidently. Furthermore, controlling for uniqueness, age, ethnicity, gender and time-elapsed since the event of the answer, users report relatively high scores for questions about the technology usage and scheduling/attending events memory categories, and low scores for questions about the technology failures, weather information, and content search categories. Note that the significance indicator for categorical variables in Table 5 is relative only to the baseline level. Thus, for categorical variables with more than two levels, like the MemCat (memory category) variable, the significance indicators in Table 5 are not too important. Overall, different memory categories did elicit statistically different scores, confirmed by a repeated measures ANOVA: $F(8,106)=2.5$, $p=0.01$. The coefficients in Table 5 estimate their relative effects.

These findings tell us that users believe they can generally answer recent autobiographical questions confidently, but that the type of question matters. Questions about content search, for example, should be harder to answer. However, it is unclear how well these perceptions reflect reality. To

Feature	Coefficient	Baseline
Intercept	15.31 *	
Age	0.04 *	
Ethnicity: Asian	0.58	White
Ethnicity: Black	0.77	White
Ethnicity: Other	-0.82	White
Ethnicity: Pacific Islander	-2.55	White
Gender: Male	-0.48	Female
Uniqueness	0.47 *	
Time Elapsed (hours)	-0.006 *	
MemCat: Content Consumption	0.13	Communication
MemCat: Content Search	-1.25	Communication
MemCat: Internet Activity	0.06	Communication
MemCat: Scheduling Events	0.52	Communication
MemCat: Technology Failures	-1.52 *	Communication
MemCat: Technology Usage	0.28	Communication
MemCat: Travel / Transportation	0.16	Communication
MemCat: Weather	-0.65	Communication

Table 3. Coefficients for the HLM for Study 2. Significance is at p < 0.05, designated by a * next to the coefficient.

answer that question, we built an Android application to ask users questions for which we had ground-truth data.

Field Study of Capturable Everyday Memory Questions

We built *myAuth,* an application that asked users questions constructed from ground-truth data, on Android 2.3 (see Figure 1). While we wanted to ask as many questions from the second MTurk study as was feasible, technical barriers limited what data we could access. For example, there was no way to access calendar information natively on the phone in Android 2.3. Other questions were not possible to ask given incomplete data stores—for example, what the user ate for lunch yesterday.

We indexed ground truth data about users' phone usage, communications, and location traces. Location data was collected with every location update; communication data was updated twice daily; and, phone usage (e.g., which application was being used) data was collected every 30 minutes, when possible. From these data, we were able to ask 13 questions, listed in Table 4. For quick identification, each question was also given an abbreviated "question type" value. The first eight questions listed were questions with one specific answer about a particular fact—*fact-based questions.* For example, "What application did you use on Thursday, March 14th at 2:53pm?" We also kept track of potential "near miss" answers to these questions—for example, if the user answered the app he or she used at 3:30pm, instead.

The last five questions—*name-a* questions—did not ask about any specific fact; rather, these questions asked the user to recall any fact in the past 24 hours of the sort queried by the question. Thus, these questions could have multiple correct answers. For example, for the question "Name an applicaion you used in the past 24 hours.", if the user used the "Email" and "What's App" apps in the past 24 hours, both answers should be correct.

We also varied the input method of the answer. For non-location *fact-based* question, we presented the user with a

QType	Likert-scale prompts in Study 2.
FBApp	What application did you use on <time>?
FBLoc	Where were you on <time>?
FBOCall	Who did you call on <time>?
FBInCall	Who called you on <time>?
FBOSMS	Who did you SMS message on <time>?
FBInSMS	Who SMS messaged you on <time>?
FBIntSrc	What did you search the internet for on <time>?
FBIntVis	What website did you visit on <time>?
NAOSMS	Name someone you SMS messaged in the last 24 hours.
NAInSMS	Name someone who SMS messaged you in the last 24
NAOCall	Name someone you called in the last 24 hours.
NAInCall	Name someone who called you in the last 24 hours.
NAApp	Name an application you used in the past 24 hours.

Table 4. List of all questions asked by the myAuth app along with their corresponding question type (QType). QTypes starting with "FB" represent *fact-based* questions; those starting with "NA" represent *name-a* questions.

set of 10 options to choose from (recognition) or an empty text-box (recall). We chose 10 options for the recognition question to make it sufficiently hard to guess randomly. The answer options comprised of the correct answer and up to three other "near-miss" answers, depending on how many near-miss answers were available. The remainder were drawn randomly from a list of plausible answers, which varied by question type. For example, for questions where the answer was a person, other "plausible" anwers included anyone in the user's contact list. For recall questions, users had to enter an answer into a textbox. However, even the recall questions had an "auto-complete" option—included primarily because we wanted to avoid misspellings or non-recorded aliases. For location questions, users were presented with a map and asked to pin their best-guess estimate of their location.

When asking users questions, *myAuth* attempted to maximize the entropy of the questions asked and answer methods presented. In other words, *myAuth* would try and ensure users were asked the greatest variety of questions and answer methods in a single session, data permitting.

Users were also provided with an option to skip any question they felt uncomfortable answering. Finally, with every question answered, users were asked to rate their agreement with three Likert-scale prompts on a scale from 1 (Strongly Diasgree) to 5 (Strongly Agree). The prompts were: "I am confident in my answer.", "It was easy for me to remember the answer to this question.", and "No one else would know the answer to this question."

Study Design

We recruited users who owned a phone running Android 2.3 or higher to participate in a two-week long field study. Users were instructed to install *myAuth* and answer at least five questions every day for 14 days. Skipped questions were counted towards their daily totals. The app would remind users to complete this task every day at 8pm. We offered users a reward of $1.00 per day for every day they answered at least five questions. We offered users an additional $0.20 for every question answered correctly, up

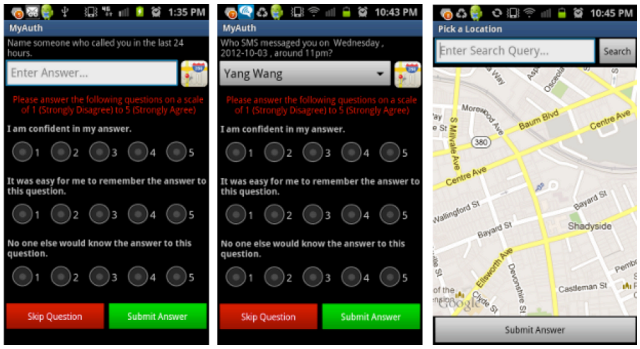

Figure 1. myAuth Android Application Screenshots showing a recall question, recognition question and location entry question. Screenshots are blurred to preserve anonymity.

to a total of an additional $6.00. Users were not made aware of whether their answer was correct till the end of the study, however. Finally, at the end of the 14 days, users were asked to complete an exit survey for an additional $5.00.

Results

We collected 2167 valid question-answer responses from 24 users over 14 days. Users had an average age of 25 (s.d. 6.25, range 18-43). Fourteen of the 24 were male (58%).

Overall, 1381 out of the 2167 (~64%) questions were answered correctly. An additional 168 (~8%) responses were near misses. Users tended to be over-confident in their performance. Both the median and mode values of the responses to the prompts "I am confident in my answer" and "It was easy for me to remember the answer to this question" were the maximum agreement values of 5.

We used a mixed-effects logistic regression [14] to model response correctness with features describing both users and responses. Once again, our model incorporated random intercepts: Each user had his or her own baseline likelihood to correctly answer a question. Table 5 lists the fixed-effect model coefficients.

Coefficients represent a change in "log-odds", or $\ln\frac{P}{1-P}$, where P represents the probability that a question is answered correctly. A positive coefficient implies that the log-odds ratio increases, or that the probability that the answer is answered correctly, P, increases. A negative coefficient implies the opposite, that $1 - P$ increases. As with the model for Study 2, coefficients for numeric features represent the change that would occur with a one unit increase in the feature, holding all other features at their mean. A categorical variable coefficient represents change relative to the baseline level of the variable.

The Intercept suggests that, holding all numeric variables at their mean and categorical variables at their baseline, there is a $\frac{e^{0.71}}{1+e^{0.71}} = 67\%$ chance that a response is correct. The standard deviation of the random effects intercept across all users was also fairly high at 0.47. Thus, there was a lot of variability between users.

Feature	Coeff..	Baseline
Intercept	0.71 *	
Answer Type: Recog	0.68 *	Recall
Age	-0.04 *	
Gender: Male	0.01	Female
Time to Answer (seconds)	-0.004	
Time since Correct Answer	-0.007	
Day of Study	0.01	
Correct Answer Entropy	0.15	
Answer Uniqueness	-0.32 *	
Confidence	0.13	
Ease of Remembering Answer	0.31 *	
Difficulty of Others Guessing	-0.04	
QType: FBApp	-1.70 *	FBOCall
QType: FBLoc	0.66 *	FBOCall
QType: FBInCall	0.58	FBOCall
QType: FBOSMS	0.55 *	FBOCall
QType: FBInSMS	0.52	FBOCall
QType: FBIntSrc	-1.75 *	FBOCall
QType: FBIntVis	-1.32 *	FBOCall
QType: NAOSMS	0.02	FBOCall
QType: NAInSMS	0.35	FBOCall
QType: NAOCall	0.11	FBOCall
QType: NAInCall	-0.17	FBOCall
QType: NAApp	-1.60 *	FBOCall

Table 5. Coefficients for the mixed-effects model for the field study. Significance is at $p < 0.05$, designated by a * next to the coefficient. Features are described as they as discussed.

Unsurprisingly, recognition questions are answered correctly more often than recall questions. Also expectedly, the negative coefficient for age suggests that older users were less likely to answer questions correctly. This finding runs counter to the model from Study 2, however, suggesting that while older users do not perform as well, they are more confident in their responses. Gender had no significant effect in predicting response correctness.

Many response specific attributes had no significant effect on the model's outcome. Neither the amount of time a user took to answer the question, nor the time elapsed since the event of the correct answer appeared to affect response correctness. Part of the reason for the latter may be because we only asked questions of events within the past 24 hours.

Performance was stable over time. Indeed, the insignificant "Day of Study" feature coefficient suggests that questions answered towards the end of the study were answered correctly at the same rate as those answered in the beginning. This finding was confirmed by comparing the relative rates of correctness for the first 20% and last 20% of responses for each user (60.3% vs. 61.7%, chi. sq. = 0.13, df = 1, p = 0.72). In other words, users do not improve at answering questions over time. Similarly, the effect of response entropy—the Shannon entropy of the correct answers for a particular question type for a user—was also insignificant, though it was close (p=.07). The direction of the estimated effect was surprising nonetheless: Questions with more response entropy were answered more correctly.

The strong, significant effect of answer uniqueness is more puzzling. We quantified answer uniqueness as the inverse of the ratio of times a specific question's correct answer was the correct answer to all questions of the same type for the responding user, in general. For example, if a user was asked a question 10 times and the correct answer was "Gmail" twice, the answer uniqueness of the response would be 1/(2/10)=10/2=5. Surprisingly, questions with more unique answers were more likely to be answered incorrectly. One explanation is that users did not *remember* answers rather, they just *knew* or could *guess* the answer. For example, a user may not remember text messaging a friend at 4:53pm, but may deduce that the answer is likely John. For more unique answers, these alternative pathways to an answer may be unavailable or misleading.

Finally, question type does effect how likely a user is to answer a question correctly (chi. sq. = 384.78, df = 12, p = 2.2e-16). Keep in mind that the significance values marked in Table 5 for QType, or "question type", are in relation to the baseline—i.e., they denote if the coefficient for one question type significantly differs from the baseline, not from all other questions. The baseline was the question type with the median rate of correctness: the *fact-based* question about outgoing phonecalls. Recall that questions are listed in Table 4.

Questions about phone usage, such as what apps a user used or which website a user visited, were far less likely to be answered correctly. On the other hand, questions about communication—phonecalls and sms messages—and location were far more likely to be answered correctly. Surprisingly, questions with a single answer at a specific time were answered correctly at the same rate as questions with several answers spanning the previous 24 hours. Indeed, there was no difference between the rates of success for *fact-based* and *name-a* questions (59% vs. 63%, chi. sq.=2.25, df=1, p = 0.13).

In summary, users are decent at answering questions about capturable everyday memories. They are equally good at answering questions about specific events as they are about questions spanning the entire past 24 hours. Questions about social interactions and location are answered correctly more often than questions about smartphone usage, but questions with unique answers are more likely to be answered incorrectly. Also, users' performances are stable over the short term, but older users, are less likely to answer questions correctly.

MODELS FOR AUTOBIOGRAPHICAL AUTHENTICATION

Given that only 64% of questions were answered correctly, it seems that the straightforward model of autobiographical authentication—relying on users to answer all challenges correctly—is insufficient. But there is another viable model: One that accounts for systematic response error.

With the response error model, users need not answer all questions correctly, but consistently. For example, a user who answered application usage questions correctly and communication questions incorrectly in the past would be expected to repeat this pattern in future attempts. In other words, the intuition behind the response error model is to allow users to answer questions naturally, while forcing adversaries to both guess answers *and* replicate their intended victim's error patterns.

In fact, we did find evidence that users answer relatively consistently over time. As we saw from our empirical model, users' overall rate of answering questions correctly was stable. Furthermore, users' performance over time was roughly stable even within question types: For all users, comparing the first half of responses to a question type to the second half, the mean absolute change in response correctness was only around 15%. Consequently, we pursued the response error model for our evaluation.

Autobiographical Response Error Model

Let's assume we have sufficient training data for a user to construct an empirical probability distribution that the user gets a question correct or incorrect given m response features: $P_u(corr(r)|f_1, \dots, f_m)$. Example response features include the question type, the answer type (recog vs. recall) and the amount of time it takes the user to answer the question. Let's also assume that we have a sequence of n question-answer responses from an attempted authentication session, $(r_1,\dots,r_n)$, where each response, r_i, can be represented by the list of response features $(f_{1,i},\dots,f_{m,i})$ along with whether or not the question in the response was answered correctly, $corr(r_i)$. From these data, we are trying to compute a confidence rating that the responses came from the user.

It is simple to calculate $P(r_1,\dots r_n \mid u)$—the probability that we would observe this response sequence from the user. It is the cumulative product of the empirical probabilities that a response is correct or incorrect, given its features:

$$(1)\ P(r_1, \dots, r_n|u) = \prod_{i=1}^{n} P_u\big(corr(r_i)\big|f_{1,i}, \dots, f_{m,i}\big)$$

However, we want the opposite: $P(u \mid r_1,\dots r_n)$. Assuming independence between responses, apart from their common origin from the user, Bayes theorem tells us that:

$$(2)\ P(u|r_1, \dots, r_n) = \frac{P(r_1, \dots, r_n|u)P(u)}{P(r_1, \dots, r_n)}$$

In the above equation above, P(u) represents the prior probability that the authenticator is the user. For personal devices like smartphones, we can treat this as a high constant—close, but not quite equal to 1. The denominator is tricky, representing the probability that we observe a given sequence of responses. We can break this value down into two components:

$$(3)\ P(r_1, \dots, r_n) = P(r_1, \dots, r_n|u) + P(r_1, \dots, r_n|\breve{u})$$

In other words, the probability that we observe a given sequence of responses is the sum of the probability that we observe the responses from the user and the probability that we observe the responses from a non-user—our modeled adversary. For simplicity, we consider this second term to be a specific adversary, though it should theoretically be all possible non-users. However, it is infeasible to enumerate and model all possible non-users. There are many possible adversary models we can adopt, which we will cover in the following section. Substituting (3) into the denominator of (2) and treating P(u) as a constant, *k*, we get:

$$(4)\ P(u|r_1,\dots,r_n,\breve{u}) = \frac{kP(r_1,\dots,r_n|u)}{P(r_1,\dots,r_n|u) + P(r_1,\dots,r_n|\breve{u})}$$

However, we are not yet done. As the equation stands, $P(u|r_1,\dots,r_n,\breve{u})$, the probability that the authenticator is the user given the sequence of responses and an adversary model, is inversely proportional to $P(r_1,\dots,r_n|\breve{u})$, the probability that we observe this sequence of responses from our modeled adversary. This property can be exploited by a cunning impersonator who knows the adversary model being used. In that case, as long as the impersonator answers to minimize $P(r_1,\dots,r_n|\breve{u})$, even if $P(r_1,\dots,r_n|u)$, the probability that we observe a sequence of responses from the user, is low, $P(u|r_1,\dots,r_n,u)$ will be high. In other words, using just equation (4) an impersonator need only act unlike the modeled adversary to achieve a high confidence rating, even if he acts nothing like the user.

To avoid this exploit, we add an additional term to the model: the bit-string similarity of the actual answer correctness vector and the expected answer correctness:

$$(5)\ S(r_1,\dots,r_n|u) = \frac{n - |E(r_1,\dots,r_n|u) - corr(r_1,\dots,r_n)|}{n}$$

The expected answer correctness vector is generated by thresholding the empirical probability distribution for the user for each response at 0.5. In other words, if the user is at least 50% likely to get a response with the given features correct, we expect the user to answer correctly. Otherwise, we expect the user to answer incorrectly. The bit-string similarity term, $S(r_1,\dots r_n \mid u)$, will equal 1 if there is no difference between the expected correctness vector and the actual vector, and will equal 0 if there is no agreement between the two vectors (i.e., the actual response is incorrect whenever we expect it to be correct and vice versa). Multiplying (4) and (5), our final equation to calculate the confidence rating of an attempted authentication becomes:

$$(6)\ C(u|r_1,\dots,r_n,\breve{u}) = P(u|r_1,\dots,r_n,\breve{u}) * S(r_1,\dots,r_n|u)$$

In summary, equation (6) takes in a set of responses and an adversary model as input, and yields a high confidence rating only if the responses are unlikely to be observed given our adversary model *and* likely to be observed given our user model. The theoretical range of this rating spans 0 to 100, but in practice a "high" confidence rating should be much lower than 100. Indeed, a rating of 100 requires that there should be a 0% chance that the response sequence comes from the adversary and a 100% similarity between the expected and actual correctness vectors—both possible, but unlikely in practice.

Adversaries

We simulated five different adversaries, each based on plausible real-life counterparts.

Simple Adversaries

The **naive adversary** guesses an answer at random from a set of 10 answers, one of which is correct. This adversary represents "chance" in a 10-answer recognition question. For recall questions, this adversary is an overestimate in the likely case where there are more than 10 possible answers.

The **observing adversary** guesses an answer selected uniformly at random from all answers that were correct for the same question type in the past. In other words, the adversary has compiled a list of plausible answers to every question type, based on the target user's previous responses. This adversary represents a close friend or family member who might know, for example, that the user only text messages one of a small set of people (e.g., her mother, her brother, and the friend himself), but not the exact answer to a specific question.

As the name implies, the **always-correct adversary** always answers a challenge question correctly. This adversary could represent malware logging software that chronicles everything that the phone records. It could also model an adversary who steals the phone and can retrieve the knowledge base directly.

Empirical Adversaries

The **empirical-observing adversary** theorizes that user response patterns are more alike than different. She has collected question-answer responses for all question types from a separate population of users, from which she constructed an empirical probability distribution—the population mean—that captures the likelihood that the average user might get a response correct. Like the observing adversary, the empirical-observing adversary also has pre-compiled a list of plausible answers to every question for the victim user with one additional detail: she has the probability distributions of the answers, as well. Thus, she knows which of the plausible answers are more likely. To put this into context, consider the fact that a user might text his mother, girlfriend and brother, but that the rates at which he texts these contacts are likely different.

The empirical-observing adversary might represent a technically proficient friend or stalker who knows the user's habits well enough to narrow down plausible answers and also tries to model the user's errors by generalizing from other users. However, emulating the population mean given a set of plausible answers can get complex: The adversary must be careful not to get the answer correct *too* often as to over-perform, and thus answers with their best guess

selectively if their best guess might outperform the population mean.

The **empirical-knows-correct adversary** is the strongest we consider, a combination of the empirical-observing and always-correct adversaries. The adversary has not only has access to an population mean, but also *knows* the correct answer to every question asked. The empirical-knows-correct adversary might represent a strong cracker who not only has malware logging software on the user's system, but also an empirical probability distribution modeling response error. This adversary intentionally answers questions correctly or incorrectly to best emulate the population mean.

Evaluation

We ran simulations using data we collected from our field study to observe how confidence ratings varied with the attempting authenticator and the modeled adversary. We simulated 6 attempting authenticators—the actual users and each of the five adversaries trying to impersonate the users. We calculated confidence estimates for responses within *sessions*—sequences of questions answered within a few minutes of each other. As a result of our study design, most sessions comprised of 5 questions answered, though some went as far as 13.

We used two features in constructing the empirical probability that a user, P_u, or a population, P_{pop}, would answer a question correctly: the question type and the answer method. For example, the empirical probability that a user will get a *fact-based* recall question about app usage correct is the rate she got other *fact-based* recall questions about app usage correct in the training data.

The training data used to construct the empirical probability for the user, P_u, included all of a user's data excluding the data from the questions in the session being evaluated. For example, consider a user who recorded 24 question-answer response sessions over the course of the study. If we are calculating the confidence rating for session 1, we construct P_u from sessions 2-24 of the user's data. Likewise, we use all of the user's data but data from the present session to construct the list of plausible answers used by the observing and empirical-observing adversaries. The training data used to calculate the population mean, P_{pop}, that is used by the empirical-observing and empirical-knows-correct adversary is all of the data for every user but the victim.

In Figure 2, we show how confidence rating varies with the number of questions answered when the attempting authenticator is the actual user. For example, if we modeled againt a naïve adversary, Figure 2 suggests that a user should generally obtain a confidence rating between 71 and 75 after answering 5 questions.

There are several encouraging points to glean from Figure 2. No matter the adversary, the confidence rating increases with the number of questions answered. In other words, when the attempting authenticator is the user, himself, our framework becomes more confident that it is the user attempting to authenticate as more questions are answered.

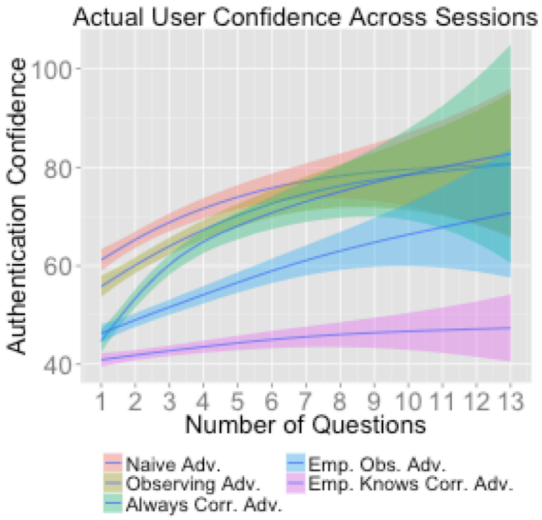

Figure 2. Confidence score estimates for users, aggregated across all sessions. Each line represents an adversary modeled against, and the translucent shades around the lines represent the 95% confidence intervals.

Expectedly, the confidence estimate varies with the modeled adversary. Our framework estimates the highest confidence rating for the user when modeling against a naïve, observing, or always-correct adversary, garnering ratings just over 70 after five questions. This result is encouraging, because these adversaries are by far the most likely. Users must answer many more questions to achieve a comparable confidence estimate when modeled against the empirical-observing adversary. Assuming an empirical-knows-correct adversary, however, our model offers a much more conservative estimate—around 45 after 5 questions—but still increases in confidence as more questions are answered. These results make sense: the empirical-observing and empirical-knows-correct adversaries have more resources than the others.

We also simulated all of the adversaries trying to impersonate the user, plotting the results in Figure 3. For brevity, we only show the plots for three of the five adversaries modeled against. The results, again, are encouraging. Relative to impersonators, the user always obtains a comparatively high rating as the attempting authenticator, no matter the adversary. Furthermore, regardless of the modeled adversary, users were able to obtain higher ratings after answering more questions.

Impersonators' performances varied greatly depending on the adversary being modeled against. For example, the always-correct impersonator obtained high confidence estimates—often higher than the user—when modeled against any adversary except itself. When modeled against itself, however, an impersonating always-correct adversary obtains a confidence estimate close to 0. In fact, this pattern

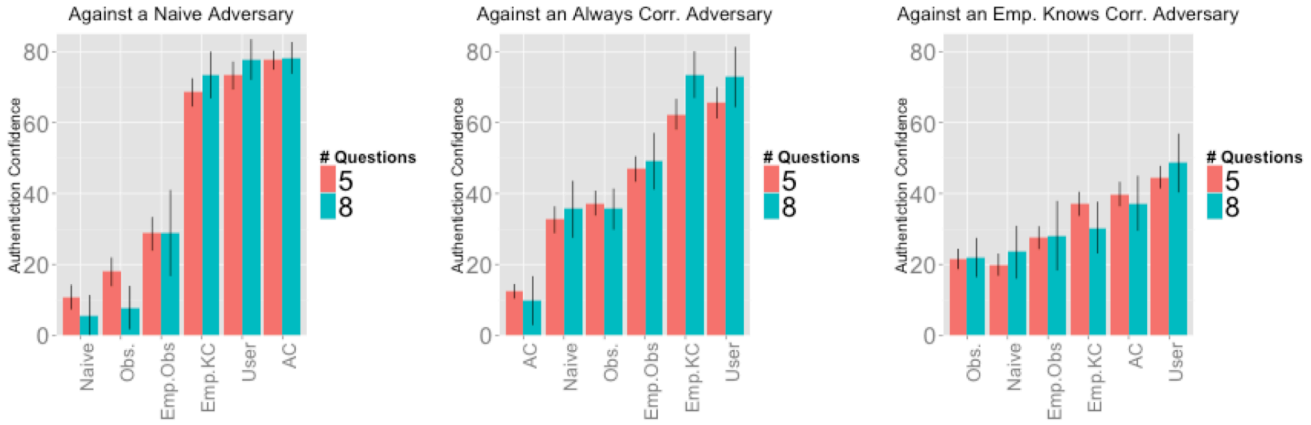

Figure 3. Each graph represents the mean confidence rating after answering 5 and 8 questions, aggregated across all users and all sessions. Confidence ratings are plotted for all six attempting authenticators—the user and the five impersonating adversaries—modeled against the naïve, always correct and empirical-knows-correct adversaries. 95% confidence intervals are included.

is relatively constant across all impersonators: when correctly modeled against, the impersonator obtains much lower confidence estimates.

The most likely adversaries—naïve and observing—do not perform well no matter the modeled adversary, with mean confidence ratings around 15 or lower after five questions. Even the more realistic of the empirical adversaries, the empirical-observing adversary, performs badly relative to the actual user, generally obtaining confidence ratings around 25-30 after five questions whereas the user obtains ratings closer to 70. The empirical-knows-correct proves more robust and effective, performing well no matter the modeled adversary—at times indistinguishably from the user. On the other hand, when modeled against itself, the empirical-knows-correct impersonator obtains increasingly lower confidence estimates as more questions are answered, while the user obtains increasingly higher estimates. Modeling against an empirical-knows-correct adversary, however, produces much more conservative ratings overall.

These results are encouraging. Employing the correct adversary model is key to mitigating false positives, but we can see that no matter the modeled adversary, the actual users obtain high confidence ratings when authenticating. Furthermore, the most likely adversaries—the naïve, the observing and the always-correct—do not achieve very high confidence ratings in impersonating the user, or are easy to identify and model against.

DISCUSSION

In this paper, we explored the feasibility of a new type of authentication that leverages the unique context-sensing and logging affordances of smartphones [6]. We offered the first empirical model of capturable everyday memory and constructed a framework that estimates a continuous confidence rating that an attempting authenticator is the user based on his answers to a series of autobiographical questions. Our evaluation shows that autobiographical authentication shows promise: by accounting for systematic response error, our framework generally estimates high confidence ratings for actual users and low confidence ratings for even sophisticated impersonators.

However, there is a lot to be done before autobiographical authentication is practical. The most glaring limitation is that autobiographical authentication is slow. On average, it took users 22 seconds to answer each question, even for recognition questions. Thus, a five-question session would take about a minute and a half—substantially more time than simply entering a password. One reason for the delay is the cognitive overhead involved with each question. With a password, the challenge and the response are known ahead of time, and procedural memory can help with its input. With autobiographical authentication, the question must be parsed and the answer recalled. Based on our exit survey, users were not thrilled with the idea of answering autobiographical questions instead of using a password. However, their sentiment rested on the assumption that they every question should be answered correctly.

Another limitation is that autobiographical authentication requires constant device usage to replenish its knowledge base. Constant usage may be a reasonable assumption for smartphones, but not for other digital devices and services (e.g., a social networking site). One solution is to integrate the knowledge bases for all of a user's devices on the cloud so that her laptop can use information from her phone for authentication. Future research will be needed to investigate this and other solutions to overcome this limitation.

While perhaps not a replacement for passwords, autobiographical authentication can shine in risky situations or situations where safe use of passwords is unusable—for example, in accessing sensitive information from an unknown location. Furthermore, with the continuous confidence estimate, authentication can be tiered instead of binary. For example, read access to non-sensitive information could require a low score—perhaps 40—while write access to security settings should require a higher

score—perhaps 70. In addition, autobiographical authentication can adopt different adversary models in different contexts [6]. For example, when at home, the most likely adversary is the observing adversary that represents a family member. If the phone is at an unknown location, we might adopt a stronger adversary model such as the empirical-knows-correct adversary.

There remain a number of interesting, open questions. For example, it would be pertinent to find heuristics to detect which type of adversary we should model against. We should also explore other questions that can be answered correctly more often and those that one user answers differently than the general population. These are the questions that will stump the empirical adversaries. Finally, we should optimize the question generation algorithm to minimize the number of questions asked to achieve a stable confidence estimate.

CONCLUSION

In summary, we made two contributions: (1) a model of capturable everyday memory—ephemeral, event-specific memories captured by smartphones and remembered by users; and, (2) a framework for and evaluation of autobiographical authentication—an authentication scheme based on users answering questions about capturable everyday memory. We found that users answer autobiographical questions predictably. By accounting for systematic response error in answering questions, we derived a formula for computing a confidence rating that the attempted authenticator is the user from a sequence of question-answer responses. We tested our formula against five simulated adversaries based on plausible real-life counterparts. Our simulations indicate that our model of autobiographical authentication performs well in assigning high confidence estimates to the user and low confidence estimates to impersonating adversaries. While at an early stage, this work represents an important step in enhancing traditional authentication for personal mobile devices.

ACKNOWLEDGMENTS

This research was funded in part by the National Defense Science and Engineering Graduate Fellowship. Special thanks go to Ian Oakley and the Portugal ICTI.

REFERENCES

1. Adams, A. and Sasse, M.A. Users are not the enemy. CACM 42, 12 (1999), 40–46.
2. Allan, A. Passwords are near the breaking point. Gartner Research Note, December 2004 (2004).
3. Conway, M.A. and Pleydell-Pearce, C.W. The construction of autobiographical memories in the self-memory system. Psych. Review 107, 2 (2000), 261–288.
4. Crovitz, H.F. and Schiffman, H. Frequency of Episodic Memories as a Function of Their Age. Bltn. of the Psychonomic Soc. 4, (1974), 517–518.
5. Galton, F. Psychometric experiments. In Brain. 1879, 149–162.
6. Hayashi, E., Das, S., Amini, S., et al. CASA: A Framework for Context-Aware Scalable Authentication. Proc. SOUPS'13, (2013).
7. Herley, C. So long, and no thanks for the externalities: the rational rejection of security advice by users. Proc. NSPW'09, (2009).
8. Jackson, K.M. and Trochim, W.M.K. Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses. Org. Rsrch Meth. 5, 4 (2002), 307–336.
9. Jakobsson, M., Stolterman, E., Wetzel, S., and Yang, L. Love and authentication. Proc. CHI'08, ACM (2008), 197–200.
10. Just, M. and Aspinall, D. Personal choice and challenge questions: a security and usability assessment. Proc. SOUPS'09, ACM (2009), 8.
11. Komanduri, S., Shay, R., Kelley, P.G., et al. Of passwords and people: measuring the effect of password-composition policies. Proc. CHI'11, ACM (2011), 2595–2604.
12. Kristo, G. and Janssen, S. Retention of autobiographical memories: An Internet-based diary study. Memory 17, 8 (2009), 816–829.
13. Rabkin, A. Personal knowledge questions for fallback authentication: Security questions in the era of Facebook. Proc. SOUPS'08, (2008).
14. Raudenbush, S.W. and Bryk, A.S. Hierarchical Linear Models in Social and Behavioral Research: Applications and Data-Analysis Methods. Sage Publications, Thousand Oaks, (2002).
15. Rice, A. A Continued Commitment to Security. The Facebook Blog, (2011).
16. Rubin, D.C., Schrauf, R.W., and Greenberg, D.L. Belief and recollection of autobiographical memories. Memory & cognition 31, 6 (2003), 887–901.
17. Schechter, S., Brush, A., and Egelman, S. It's no Secret: Measuring the Security and Reliability of authentication via 'secret' questions. Proc. S&P'09, Ieee (2009), 375–390.
18. Shay, R., Komanduri, S., Kelley, P.G., et al. Encountering stronger password requirements: user attitudes and behaviors. Proc. SOUPS'10, ACM (2010).
19. Suo, X., Zhu, Y., and Owen, G.S. Graphical passwords: A survey. Proc. ACSAC'05, IEEE (2005).
20. West, R.L., Crook, T.H., and Barron, K.L. Everyday memory performance across the life span: Effects of age and noncognitive individual differences. Psychology and aging 7, 1 (1992), 72–82.