Evaluating the Effectiveness of Using Hints for Autobiographical Authentication: A Field Study

Yusuf Albayram
Department of Computer Science & Engineering
University of Connecticut, Storrs, CT, USA
yusuf.albayram@uconn.edu

Mohammad Maifi Hasan Khan
Department of Computer Science & Engineering
University of Connecticut, Storrs, CT, USA
maifi.khan@engr.uconn.edu

ABSTRACT

To address the limitations of static challenge question based authentication mechanism, recently smartphone-based autobiographical authentication mechanisms are being explored where challenge questions are generated using users' day-to-day activities captured by smartphones dynamically. However, users' poor recall rate in such systems is still a significant problem that negatively affects the usability of such systems. To address this challenge, this paper investigates the possibility of using hints that may help users to recall recent day-to-day events more easily and explores various design alternatives for generating hints. Specifically, in this paper, we generate challenge questions and hints for three different kinds of autobiographical data (e.g., call logs, SMS logs, and location logs), and evaluate the effect of different question types and hint types on user performance by conducting a real-life study with 24 users over a 30 day period. To test whether hints are useful/harmful for adversaries' response accuracy, we simulate various kinds of adversaries (e.g., naive and knowledgeable) by recruiting volunteers in pairs (e.g., close friends, significant others). In our study, we observed that, for legitimate users, hint was effective for all different question types. Interestingly, we found that hint has negative effect on strong adversarial users and no significant effect on performance for naive adversarial users.

1. INTRODUCTION

Several recent research efforts have investigated the idea of autobiographical authentication leveraging smartphone usage data (e.g., call log, SMS log) and highlighted the advantages of these systems over static challenge question based approaches due to the dynamicity offered by such smartphone-based solutions [4, 13]. However, as recall of information relies solely on users' memory in such systems, which are often imperfect and unreliable [19], error rates while answering autobiographical questions generated based on day-to-day activity logs (e.g., call log, SMS log) are often high [13], negatively affecting the usability of such systems [4]. Interestingly, while providing cues/hints may act as a "memory trigger" to jog one's memory and can be an effective way to improve the recall rate and reduce the error rate, only a small number of prior efforts looked into the possibility of using hints to facilitate recall and mostly focus on static password based systems [21, 32]. To address this void and complement prior efforts, in this paper, we focus on generating hints for autobiographical authentication systems where hints are generated along with the challenge questions dynamically. Specifically, for hint and challenge question generation, the presented system leverages events that are related to *episodic memory* [11] which refers to memories related to everyday experiences (e.g., call logs) instead of events that are related to *semantic memory* which refers to memories related to facts and general knowledge (e.g., important events such as graduation) [24]. This is due to the fact that, in case of autobiographical authentication, events that are related to *episodic memory* is more suitable and relevant for hint and challenge question generation as these are the types of memories that are formed every day and have a short shelf life (i.e., they are forgotten within days or weeks), offering the required dynamicity compared to relatively less dynamic semantic memory.

Based on these observations, in our study, we investigate hint generation algorithms that consider frequency of events along with historical patterns of a user to generate customized hints that may help a user to recall near past events without revealing the actual information. Specifically, we consider three different kinds of day-to-day events for generating challenge questions and hints (e.g., phone call send and receive events, SMS message send and receive events, location-visit events). As hints may reveal secrets and help adversaries to learn about a user if not designed carefully, over multiple iterations, we finalized the design of hints for specific event types that present information in a non-revealing way. To study the effect of hints on users' performance, we recruited 24 users and conducted a real-life study for over a month. To test whether hints are useful/harmful for adversaries' accuracies, we simulate two different kinds of adversaries (e.g., naive vs. knowledgeable) by recruiting volunteers in pairs (e.g., close friends). Over the course of the study, each user is periodically presented with three sets of challenge questions. The first set is generated based on users' own data. The second set is generated based on a randomly selected user's data. Finally, each user is presented with a third set of questions which is generated based on user's friend's data.

Symposium on Usable Privacy and Security (SOUPS) 2015, July 22–24, 2015, Ottawa, Canada.

In our study, we observed that, for legitimate users, hints were effective for all different question types. We also found that hints had no significant effect on response correctness of naive adversaries and had negative effect on strong adversaries. Finally, users' exit interview data suggests that legitimate users found hints to be helpful and indeed improve the usability of such systems. To summarize, this paper makes the following key contributions:

1. Investigate the effect of hints on legitimate users' response correctness for different types of autobiographical challenge questions.
2. Simulate different kinds of adversarial users (e.g., naive vs. knowledgeable) by recruiting participants in pairs (e.g., close friends), and investigate the effect of hints on adversarial users' response correctness for different types of autobiographical challenge questions.
3. Finally, investigate the effect of hints on usability of such systems through performance change and qualitative feedback collected using an exit survey.

The rest of the paper is organized as follows. Section 2 presents prior work that are related to our current study and discusses the basic concepts underlying our analysis. Section 3 explains the study design and describes how the autobiographical questions and hints are generated. Key findings along with detailed analysis are presented in Section 4. Limitations of our study along with possible future directions are discussed in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORK

To facilitate resetting of passwords or provide an extra layer of security for authentication, various fallback authentication mechanisms are being investigated. One of the most widely used approaches is known as knowledge-based authentication (KBA) (e.g., static challenge question based technique) [7–9, 35]. KBA can be further divided into two categories, namely, static KBA and dynamic KBA. In static KBA, the questions are often predetermined (e.g., "What is the name of your first pet?"), and are often susceptible to various forms of vulnerabilities such as easy predictability and poor recall rate [6, 22, 29–31, 33–35, 41]. Furthermore, static questions are becoming weaker due to improved information retrieval techniques and increase in online content [31]. For instance, by mining online sources (e.g., social networking sites, public records), an attacker often can obtain the details about one's personal information to answer many of the challenge questions commonly used for backup authentication.

To address the limitations of static KBA schemes, dynamic KBA schemes are being investigated recently where challenge questions are generated on the fly based on user's recent activities such as browsing history [5], Facebook activity [12], calendar events [27], user's email history [26], electronic personal history [28], or financial activity (e.g., several major American credit bureaus authenticate users by generating questions based on past financial transactions).

More recently, Gupta et al. [16] investigated the memorability of users' smartphone usage behavior (e.g., emails, calendar events, calls) and attempted to leverage that to authenticate users. One of the main limitations of this work is that the challenge questions are generated based on a user's routine (e.g., who do you call the most?) rather than day-to-day activities which are more dynamic. Similarly, another work used email activities to generate challenge questions (e.g., who sent you the most emails?) [40]. Das et al. [13], Albayram et al. [4], and Hang et al. [18] presented authentication frameworks that exploit smartphone usage data (e.g., phone call history, location traces, app usage) to generate dynamic challenge questions.

While these recent efforts on smartphone based autobiographical authentication mechanisms report encouraging results, they rely solely on users' memory for recall which is often imperfect and thereby unreliable [19], causing poor recall rate [4, 13]. Although providing cues or hints can be an effective way to jog one's memory and improve recall rate, none of these works has investigated the possibility of using hints to facilitate recall. Interestingly, a number of prior efforts have looked into this possibility in different contexts. For example, Wagenaar [38] studied the recall performance of a person's daily life events, and pointed out the effectiveness of providing cues on memory retention. He also highlighted that providing different forms of cues (e.g., who, when, where) increase the chance of recalling an event. Vemuri et al. [37] noted that one needs only a trigger rather than original content in order to remember.

There is a limited number of works in the literature that looked into the possibility of using hints in authentication settings. For example, Hertzum [21] proposed using minimal-feedback hints where users select certain characters of their passwords that will be revealed during password entry in order to jog users' memory. He conducted a user study with 14 users and found that, while hints aid users' to recall their passwords, the selected passwords were weak. Similar findings were found by Lu et al. [25]. Renaud et al. [32] investigated the use of some abstract images (e.g., Cueblot) as password cues. However, they found that the presence of abstract images did not have a positive effect on users' performance in password-based authentication.

Based on prior studies, we identify that, in case of autobiographical based systems, events related to *episodic memory* is more suitable for question and hint generation as these are the types of memories that are formed every day and have a short shelf life (i.e., they are forgotten within days or weeks). Specifically, Conway [10] conducted a study in which participants were asked to list as many specific memories as they can remember from yesterday, from two days back, from three days back, and so on. They found that users can recall a good number of events that are one day old compared to events that are older. Beyond a 3-day retention interval, memories appear to be much more concerned with routines and schema than with specific episodic memories. Further, Kristo et al. [24] examined several factors that may influence recall rate of recent autobiographical events using an Internet-based diary study. They found that the content and the time of the events were remembered better compared to the details of the events. Also, among the time elements, time of the day was remembered better. Events that occur less frequently were also remembered better compared to events that occur frequently.

Motivated by these prior efforts, in this work, we looked into the challenge of generating effective hints for smartphone-based dynamic authentication mechanism by identifying events that are more likely to be remembered by a user. While our

work is inspired by prior efforts, our work differs from prior work in several aspects. First, to the best of our knowledge, we are the first to investigate the challenge of generating hints dynamically for smartphone based autobiographical authentication systems. Second, we conducted a real-life study that investigates the strengths and weaknesses of providing hints for different categories of questions and users (e.g., legitimate, naive adversary, strong adversary). Finally, we evaluate the effect of hints on usability aspect of such systems through an interview style exit survey. The details of our work is presented in the following sections.

3. METHODOLOGY

In this section, we describe the smartphone application that was developed for collecting and analyzing autobiographical data along with the algorithms for question and hint generation. We then present the design of the study. The details are below.

3.1 Autobiographical Data Collection Application

We developed an android application for devices running Android 2.3 or higher to collect and analyze autobiographical user data. The application collects the communication history and the location traces of a user while running in the background. It then generates challenge questions and hints using the collected data. Table 1 lists the details of the data that are collected in our study.

Data	Details of collected data
Call	Type (outgoing, incoming), Duration, Name of the person, Time
SMS	Type (sent, received), Receiver/Sender Name, Length of SMS message, Time
Location	Latitude, Longitude, Time, Accuracy (i.e., the expected error bound)

Table 1: Details of the collected data.

In order to obtain the location information with minimal energy overhead, we utilize the latest Google Fused Location API [1] along with Android's activity recognition API [2]. Specifically, Android's activity recognition API provides an easy way to detect if a user is moving or not (e.g., walking, biking, or in a vehicle). The application software leverages the Android's activity recognition API to decide whether to track location or not. For example, when a user is not moving, the app does not track location at all. However, if a user is walking, biking, or in a moving vehicle, the app starts logging location data. Once the data items are collected, the question generation component generates challenge questions as follows.

3.2 Autobiographical Question Generation Component

Using the aforementioned data items, the application generates 5 different types of questions as listed in Table 2. Details about each type of question are below.

Questions Based on Communication Activity

Communication questions are generated based on a user's recent communication history (e.g., SMS history that includes both sent and received messages, and call history that includes both incoming and outgoing phone calls). This category of questions asks a user to recall the name of the person he/she called or SMS messaged, or the name of the person who called him/her or SMS messaged him/her at a certain time. Examples of communication questions are shown in Figure 1(a) and Figure 1(b). For this type of question, a user is asked to enter the answer (i.e., person name) into a textbox. To enhance the usability, we utilize the "autocomplete" feature which suggests possible entries as a user types in the textbox. This is especially helpful as "autocomplete" feature reduces potential errors due to possible misspellings.

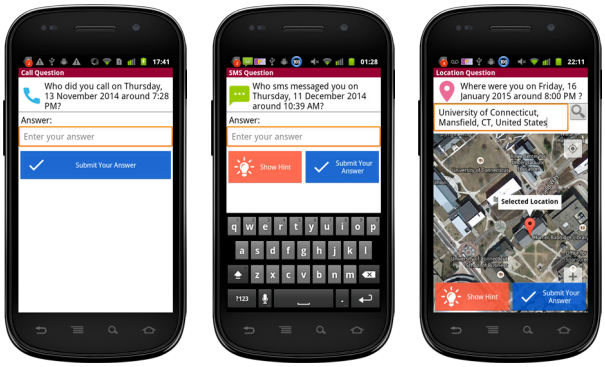
Question Type	Question
Phone call (Incoming)	Who called you at <time>?
Phone call (Outgoing)	Who did you call at <time>?
SMS (Received)	Who sent you the SMS message at <time>?
SMS (Sent)	Who did you send the SMS message at <time>?
Location	Where were you at <time>?

Table 2: List of the question types.

Questions Based on Location Information

Location questions are generated based on a user's recent location traces tracked by the application. The collected location data is composed of a sequence of coordinates with latitude, longitude, and the relevant temporal information (i.e., time stamp). To avoid considering each geographical coordinate as a unique physical location, in our work, we use a clustering algorithm that groups geographical coordinates based on their distance in order to infer user's locations. However, as a user may visit new places over time and we do not know a priori the total number of places a user may visit, we chose to employ a density-based clustering approach that can incrementally adapts the number of clusters (i.e., the number of distinct physical locations). Specifically, we use the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [14] that is based on the notion of density reachability. Briefly, the DBSCAN algorithm requires two parameters: ϵ distance threshold (e.g., 75 meters) and min_{pts} minimum number of points within a cluster. The algorithm starts by assigning a random point in a cluster and expands it with neighborhoods of at least min_{pts} points that are within a distance ϵ from it. In our work, to calculate the distance between two geographical coordinates, we use Haversine Distance, though the algorithm can work with any distance function. Based on the distance of examined coordinates and ϵ distance threshold, the DBSCAN algorithm either creates new clusters or expands/updates the existing clusters. As new coordinates arrive, new coordinates are first examined to determine whether they can be assigned to any of the existing clusters. If not, the new coordinates are given as input to the DBSCAN algorithm to regenerate the clusters including the new locations. The output of this algorithm is a set of clusters that is used to generate questions. Please note that this algorithm only runs whenever the application needs to generate location questions for a user.

For location-based questions, a user is presented an interactive map and is asked to select the location that he/she had visited during a certain time window of a specific day. The interactive map was implemented leveraging Google Maps Android API [3] where the initial zoom level was set to 1 to make the most of the world visible. The rationale behind this choice is to avoid influencing users to select locations



(a) Call (b) SMS (c) Location

Figure 1: Screenshots showing different types of questions.

from a certain geographic area, which may reduce the overall security of the system [36]. In order to select a location on the map, the minimum required zoom level is set to 16, which gives reasonable details and higher security since an adversary has to guess a location at a finer resolution. A user needs to long press on the map to pin his/her location and set a marker at the selected location (e.g., like the one in Figure 1(c)). Instead of zoom in/out manually, user may also use a search box to zoom-in on the right area/location very quickly.

Algorithm for Generating Challenge Questions

As a user often makes/receives a large number of phone calls and/or sends/receives a large number of SMS messages, and/or visits many different places in a given day, it is non-trivial to pick the specific instance of an event that may be used to generate the question. Ideally, the system should pick an event that is easy for a legitimate user to recall but hard for an adversary to guess.

To address this challenge, we develop an algorithm that gives preference to rare events compared to more predictable events. Intuitively, if a person rarely receives a phone call from person X, it is more likely that he/she will remember that event. To implement the algorithm, we represent a user's history H as a sequence of certain type of events (e.g., phone call). In H, each event is represented as a triplet of the form $e_i = (a_i, d_i, t_i)$, where a_i represents an activity (e.g., making a phone call), d_i is the duration of the event, and t_i is the time-stamp of the event. Assuming that n activities were recorded for a user in a given time frame, the history for that time frame will be represented as a time ordered sequence of triplets, and will be denoted as $H = \{e_1, ..., e_n\}$. Subsequently, we convert the history H into a Time Window-Event Matrix as shown in Table 3 by splitting each day into a set of m time windows $W = \{w_0, ..., w_m\}$ of fixed size (e.g., 1 hour). Next, each event is assigned to a specific time window W_i based on the event's time-stamp t_i.

Once events are assigned in specific time windows, the system computes an "interestingness" weight for each event based on statistical measure of randomness, and attempts to pick events for generating questions by giving preference to more infrequent events in a user's schedule. To identify the infrequent events for a given *Time Window-Event Matrix* (e.g., as shown in Table 3), the algorithm analyzes daily and weekly activity patterns of a user and calculates the weight for an event as follows.

1. First, the algorithm calculates $P(e_i)$ which denotes the probability of an event e_i. For example, probability of calling John in the last 30 days based on call log data.

2. Next, calculate $P(e_i|w_m)$ which denotes the probability of event e_i for a specific time interval w_m. For example, the probability of calling John between 10:00 am and 11:00 am in the last 30 days. This probability is calculated to identify daily patterns.

3. Next, calculate $P(e_i|w_m, dow_k)$ where dow_k denotes the "day of the week" from the set $DOW = \{dow_1, ..., dow_7\}$ where $dow_1 = Monday, \ldots, dow_7 = Sunday$.
$P(e_i|w_m, dow_k)$ denotes the probability of an event e_i for a time interval w_m on day dow_k of the week. For example, the probability of calling John between 10:00 am and 11:00 am on Mondays in the last 30 days. This probability is calculated to identify weekly patterns.

4. Finally, to give priority to long lasting events which are more likely to be remembered by a user easily, the algorithm calculates T_e^i which denotes the sum of the duration of event e_i in the history H and subsequently, multiply with d_i (duration of the event). For example, multiply a recent phone call duration made to John with the sum of the duration of phone calls that made to John in the last 30 days based on call log data. The main intuition behind this multiplication is that we want to give priority to the latest events that lasted longer compared to other events of the same type.

5. Based on the above probabilities, we compute the *weight* of an event as follows:

$$Weight = \frac{P(e_i)\, P(e_i|w_m)\, P(e_i|w_m, dow_k)}{T_e^i \times d_i}$$

Once weight for individual events are calculated, the algorithm sorts all events based on *weight* and pick according to that order whenever the system needs to generate challenge questions for a particular data type. Please note that the lower the weight, the higher the chance of that event to be selected by the algorithm.

Due to the above scheme, higher weight questions that are relatively easy to guess because of "regularity" are filtered out and the preference is given to more infrequent events which are more likely to be harder to guess but easier to recall by legitimate users.

Please note that the above scheme can be applied for any data types such as call log, SMS log, and location log. However, necessary changes may need to be made based on data types. For instance, for SMS log, there is no duration for SMS messages, and thus duration needs to be ignored or may be replaced with the length of SMS messages. For location log, to avoid considering each geographical coordinate as a unique physical location, geographical coordinates need to be clustered first.

3.3 Hint Generation

Since human memory is fallible, when it comes to autobiographical authentication where the challenge questions

Window \Day	Nov 27	Nov 28	...	Dec 27
$00:00-00:59$	–	–	...	$\{Received call from-Jeff, 55sec\}$
$01:00-01:59$	–	–	...	–
⋮	⋮	⋮	⋮⋮⋮	⋮
$14:00-14:59$	$\{Called-Alice, 55sec\}, \{Called-Bob, 32sec\}$	$\{Called-Bob, 89sec\}$	...	$\{Called-Bob, 17sec\}$
$15:00-15:59$	$\{Called-John, 300sec\}$	$\{Received call from-Jeff, 42sec\}$	...	–
$16:00-16:59$	$\{Called-John, 14sec\}$	$\{Called-Bob, 20sec\}$	...	$\{Called-Bob, 89sec\}$
$17:00-17:59$	$\{Called-Bob, 27sec\}$	–	...	–
⋮	⋮	⋮	⋮⋮⋮	⋮
$23:00-23:59$	–	$\{Received call from-Bob, 14sec\}$	...	$\{Called-Mike, 14sec\}$

Table 3: Time Window-Event Matrix of a user's phone call log history. Numbers right next to a person's name indicates the duration of the phone calls in seconds.

are generated using users' everyday interactions with their smartphones, a user may not always remember who he/she called or texted at a specific time. However, providing some auxiliary information about a certain event as hints may help users to recall that particular event. For example, a user may not remember whom he texted at 3 pm today, as the user may have texted more than one person around that time, making it a bit difficult to guess which person the question is referring to. However, if the user is provided some auxiliary information such as "the same person texted at 11:47pm yesterday", the user is more likely to remember the answer. Please note that such hints do not reveal any privacy sensitive information other than the fact that he texted someone at 11:47 pm yesterday.
While hints might be helpful for users, generating hints is a nontrivial problem and ideally should satisfy the following properties.

1. **Efficacy:** A hint should be useful for legitimate users.
2. **Ambiguity:** A hint should not be useful for anyone else other than the legitimate user.
3. **Privacy:** A hint should not leak/reveal user's privacy sensitive information (i.e., preserve the privacy of the legitimate user).

To ensure the above properties, in this work, we designed different kinds of hints as follows.
For communication questions (i.e., Call and SMS), hints are generated based on recent communication events that involves the same receiver/sender that the challenge question is asking about. Intuitively, knowing that the user talked (or messaged) to the same person a few hours or days earlier can jog the user's memory. For example, if the question is "Who did you call at 2 pm on Wednesday?" and the correct answer is "John", information regarding phone call(s) made to "John" or received from "John" within the last few days (e.g., 2 days) can be used as hints (e.g., sample hint: you called the same person on Wednesday at 11:25 am). Similarly, information regarding SMS message(s) received from "John" or sent to "John" may be used as hints as well (e.g., sample hint: you sent a SMS message to the same person on Wednesday at 1 pm). For location questions, hints are generated based on recent historical location information (e.g., sample hint: you visited the same place on Monday at 11 am).
Furthermore, in the absence of recent historical information relevant to the event, we generate hints in a negative format.

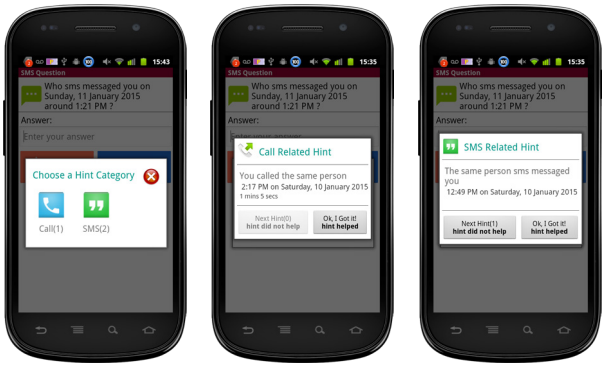


Figure 2: Screenshots showing a SMS question along with the provided hints.

For example, we may generate hints such as "You did not call this person within the last 2 days". All different kinds of hints that may be generated for different types of questions are listed in Table 4.
In case there are multiple hints in different categories, users are provided with an option to choose from any of the available hint categories they want (e.g., Call, SMS). Finally, once a user uses a hint during the session, the user is asked whether the viewed hint was helpful or not. A sample question along with the provided hints are shown in Figure 2.

3.4 User Score Calculation

As users may make different types of mistakes for different question types while answering the challenge questions, we need to calculate the score differently for different question types. For example, in case of communication questions (i.e., Call and SMS), a user can either pick the answer from a list of suggestions that are populated using an "auto-complete" functionality, or a user may type in his/her answer instead of selecting from the list of names suggested by the auto-complete feature. However, while typing, a user may make spelling mistakes. For example, a common first name "Adrianna" may be spelled as "Adrianne" or "Adrienne". Thus, instead of scoring the answer based on an exact match (where the correct answer and the user's answer are matched 100%), in the current implementation, there is an error tolerance to accommodate typing errors (e.g., 85% similarity score between two strings). Specifically, if the Jaro-Winkler distance between the entered text and the correct answer is greater than 85% similarity score, the answer is considered to be cor-

Question Type	Hint Type and Hints
Call (e.g., Who did you call on <time>?)	1) Call/Outgoing/Positive: You called the same person on <time>, and the duration of the phone call was <duration>. 2) Call/Incoming/Positive: The same person called you on <time>, and the duration of the phone call was <duration>. 3) Call/Outgoing/Negative: You did not call this person within the last <♯ of days>. 4) Call/Incoming/Negative: This person did not call you within the last <♯ of days>. 5) SMS/Incoming/Positive: The same person SMS messaged you on <time>. 6) SMS/Outgoing/Positive: You SMS messaged the same person on <time>. 7) SMS/Outgoing/Negative: You did not SMS messaged this person within the last <♯ of days>. 8) SMS/Incoming/Negative: This person did not SMS message you within the last <♯ of days>.
SMS (e.g., Who did you SMS message on <time>?)	1) SMS/Outgoing/Positive: You SMS messaged the same person on <time>. 2) SMS/Incoming/Positive: The same person SMS messaged you on <time>. 3) SMS/Outgoing/Negative: You did not SMS messaged this person within the last <♯ of days>. 4) SMS/Incoming/Negative: This person did not SMS message you within the last <♯ of days>. 5) Call/Outgoing/Positive: You called the same person on <time>, and the duration of the phone call was <duration>. 6) Call/Incoming/Positive: The same person called you on <time>, and the duration of the phone call was <duration>. 7) Call/Outgoing/Negative: You did not call this person within the last <♯ of days>. 8) Call/Incoming/Negative: This person did not call you within the last <♯ of days>.
Location (e.g., Where were you on <time>?)	1) Location/Positive: You were in the same location on <time>, and you stayed there <duration>. 2) Location/Negative: You did not visit this location within the last <♯ of days>.

Table 4: List of possible hints for different types of questions.

rect. Otherwise, the score is set to 0. Please note that the Jaro-Winkler distance metric is best suited for comparing short strings such as names [39].

In case of location questions, as users may not place the marker on exactly the same location coordinates estimated and identified by the system, there is an error tolerance (e.g., 75 meters great circle distance) in our system, which is calculated based on the Haversine distance formula[1] between the selected coordinates and the estimated location. If the distance between the selected geographical location and the estimated location is greater than 75 meters, the answer is considered to be incorrect and the score is set to 0.

3.5 Study Design

To evaluate our system, we recruited 24 participants from the college campus through the university email list server. To simulate strong adversaries, we recruited participants in pairs (e.g., close friends, significant others). The social relationships between the pairs of participants are shown in Table 5.

Over the course of the experiment, each participant was presented with three sets of questions multiple times each week. The first set of question was generated based on participant's own data. For example, a participant would receive a phone call question in the following format: "Who did you call at 11:25am on Wednesday?". The second set of question was generated based on participant's pair's (e.g., close friend or couple) data. In this case, the role of a strong adversary is played by the pair of each participant. For example, the participant would receive a phone call question about his/her partner in the following format: "Who did your partner call at 4:20pm on Friday?". The third set of question was generated based on a randomly selected participant's data whose identity was not revealed to the participant who answered the question. In this case, participants played the role of a naive adversary. For example, the participant would receive a phone call question about a stranger in the following format: "Who did a stranger call at 2:51pm on Monday?".

In order to evaluate the effectiveness of using hints for autobiographical authentication, we devised a within-subject user study with two conditions. In order to have consistent comparisons between questions with hint(s) and with no hint, a question was asked twice. In the first condition, a question was presented without hint and subsequently, in the second condition, the same question was presented with hint. In all cases, participants were not given any feedback regarding his/her performance throughout the study.

Each participant was compensated with a $25 Amazon gift card for two weeks of participation. The study was approved by the University's Institutional Review Board (IRB).

Pair#	Relationship	Closeness Rate	Live Together
Pair-1	Friends	4	No
Pair-2	Friends	5	No
Pair-3	Friends	4	No
Pair-4	Friends	5	No
Pair-5	Friends	4	No
Pair-6	Friends	4	No
Pair-7	Friends (Roommates)	5	Yes
Pair-8	Friends (Roommates)	5	Yes
Pair-9	Boyfriend/Girlfriend	5	No
Pair-10	Boyfriend/Girlfriend	5	No
Pair-11	Boyfriend/Girlfriend	5	No
Pair-12	Boyfriend/Girlfriend	5	No

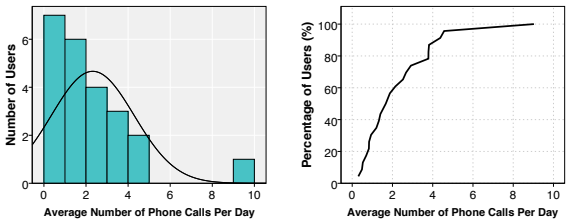
Table 5: The social relationships between the pairs of participants and their ratings on how well they know each other on a Likert-scale of 1 (Very little) to 5 (Pretty well). The last column shows whether participants live together or not.

4. FINDINGS

During a period of 30 days, from 24 participants (12 paired participants), we collected a total of 3296 question-answer responses where the questions were presented with no hints and 3296 question-answer responses where the questions were presented with hints. Out of 3296 questions, participants used hints in 832 questions. One of the participants withdrew from the study after two weeks of participation. All participants (10 female, 14 male) were undergraduate students from a broad range of degree programs. The age of participants ranged from 18 to 23 years with an average age of 19.33 years (Median=19 years with SD=1.28).

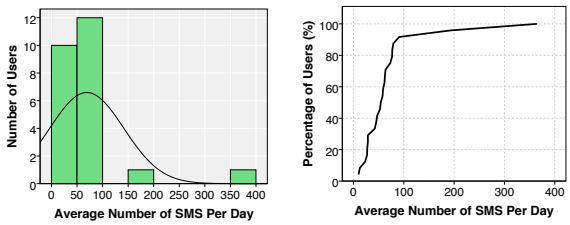
Figure 3 shows the statistics for phone call data including the number of phone calls made and received. As shown in the histogram in Figure 3(a), the plot appears to be right-skewed as most participants make 1 - 4 phone calls per day. Figure 3(b) shows the cumulative distribution of participants with respect to the average number of phone calls per day. 80% of

[1] `http://en.wikipedia.org/wiki/Haversine_formula`



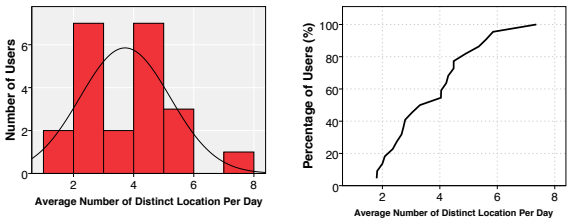
(a) (b)

Figure 3: (a) Histogram of average number of phone calls per day across all users, (b) Cumulative distribution of users with respect to the average number of phone calls per day.



(a) (b)

Figure 4: (a) Histogram of average number of SMS messages per day across all users, (b) Cumulative distribution of users with respect to the average number of SMS messages per day.



(a) (b)

Figure 5: (a) Histogram of average number of distinct locations per day across all users, (b) Cumulative distribution of users with respect to the average number of distinct locations per day.

the participants made less than 4 phone calls per day during the study period (i.e., 1 month). Figure 4 shows the statistics for SMS data including the number of sent and received text messages. The histogram for the SMS data appears to be right-skewed as most participants (91%) sent and received 50-100 SMS messages per day. One of the participants received and sent 150-200 SMS messages per day and another participant received and sent 350-400 SMS messages per day. Figure 5 shows the number of distinct locations visited by users during the study period. The histogram appears to be centered and closer to normal as shown in Figure 5(a). The cumulative distribution of participants with respect to the average number of distinct locations per day is shown in 5(b). Most participants visited 2-6 unique locations per day.

Intuitively, as users send/receive a large number of SMS messages per day, we expected that SMS related questions will be the hardest to answer, and the call and location-based questions will be comparatively easier to answer. Also, we expected hints to improve performance of legitimate users while having minimal/no effect on adversarial users' accuracy.

To explore the effects of different factors (e.g., question type, hint vs. no hint) on response correctness for different categories of users (e.g., legitimate vs. adversarial users), we used a Mixed-effect logistic regression model [17] to analyze the data. The Mixed-effect logistic regression model contains fixed effects and random effects. As we have repeated measurements from the same individual, a *user* was included as a random-effect variable to account for multiple measurements (i.e., multiple responses) within users. Thus, each user has his/her baseline likelihood for answering a question correctly. All other independent variables were included as fixed-effect variables. The coefficients listed in Table 6 show the relationships between the dependent variable (i.e. response correctness) and independent variables (e.g., question type, time taken to answer). The categorical variables are designated using their baselines. The coefficients marked with the "*" represent the variables that have statistically significant (p <0.05) effect on response correctness. The log odds was used as a measure of association between the dependent variable and independent variables and their influencing factors. The coefficients listed in Table 6 represent the change in response correctness when the coefficient

Features	Coefficients			Baseline
	Legitimate	Strong Adversary	Naive Adversary	
Age	0.0297317	-0.3907673	-0.0404524	
Gender	0.1075227	-0.1329869	0.6497308	Female
Time to answer (seconds)	0.0014976	0.0066606	-0.0030984	
Question Type: Call	1.379083 *	0.7397752 *	1.354868	Question Type: SMS
Question Type: Location	0.6789546 *	1.503114 *	2.878599 *	Question Type: SMS
Hint Used	0.2568961 *	-0.8211419 *	-0.5674617	Hint Not Used
Confidence	0.6430592 *	0.5549432 *	0.112277	

Table 6: Coefficients for the Mixed-Effect Logistic Regression model for the user study. The coefficients show whether listed features had a statistically significant effect on response correctness. Significant features are designated by a * next to their coefficients.

is increased by one-unit while controlling all other numerical variables at their mean values and categorical variables at their baseline. In addition, a positive coefficient implies that an increase in the independent variable affect response correctness positively. A negative coefficient suggests the opposite. We discuss the details regarding our findings below.

• Effect of Question Type on Accuracy Score.
In our evaluation, we condition a question type on its baseline as shown in Table 6. This indicates that the coefficient for one question type (e.g., Location questions) significantly differs from the baseline (i.e., SMS message questions). The response correctness of phone calls and location questions significantly differ from the SMS message questions for legitimate and strong adversarial users. For naive adversarial users, only location questions appeared to significantly differ from the SMS message questions. This may be due to the fact that naive adversarial users had a higher chance of guessing where a random person was at any given time as they knew that the other participants were from the same campus/locality, which is less likely to be the case in real-life. Hence, in real-life, the success rate for naive adversary is more likely to be lower for location-based questions. However, it was very difficult for a naive adversary to guess who the person texted or called at any given time.
We found that phone call questions were answered more often than SMS and location based questions. When it comes to guessing, location based questions were guessed more easily than questions about communications (i.e., phone call and SMS message).

• Effect of Hints on Accuracy Score.
As one of the main purposes of this paper is to evaluate the effectiveness of using hints for autobiographical authentication, we devised a within-subject user study with two conditions. In order to have consistent comparisons between questions with hint(s) and with no hint, a question was asked twice. In the first condition, a question was presented without hint and subsequently, in the second condition, the same question was presented with hint. In all cases, no feedback was given to participants to avoid biasing them.
In our evaluation, we compared different users' accuracy to evaluate whether hints help users to recall the answer. We found a significant difference between questions with hint and without hint in terms of response correctness. More interestingly, when legitimate users used hints, their response correctness improved significantly, whereas when strong adversarial users used hints, their response correctness reduced significantly, and hints had no significant effect on response correctness for naive adversarial users. While such negative effect on strong adversarial users could be due to increased ambiguities caused by hints, further investigation focusing on this particular aspect of our finding is needed to identify the underlying reasons behind such effect.
Furthermore, in our study, we found that strong adversaries used hints for 35 call, 100 SMS, and 73 location questions out of 260 call, 476 SMS, and 440 location questions respectively. Usage of hints reduced response accuracy for 84.6% adversarial users for call questions, 94% adversarial users for SMS questions, and 87.5% adversarial users for location questions. In contrast, in case of legitimate users, it improved response accuracy for 84.2% users for call question, 91% users for SMS questions, and 77.5% users for location questions. Figure 6 shows the effect of hints on overall average response correctness across three different question types (i.e., Call, SMS, and Location) for three different types of users (i.e., Legitimate, Strong, and Naive adversarial users).

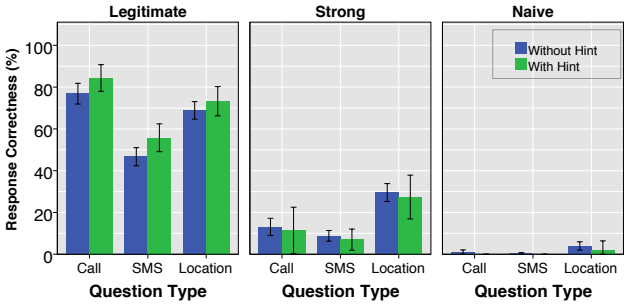


Figure 6: Comparison of response correctness rate of hint and non-hint conditions across three different question types (i.e., Call, SMS and Location) for three different user types (i.e., Legitimate, Strong and Naive adversarial users). 95% confidence intervals are included.

• Effect of Hints on Users' Confidence.
To understand the effect of hints on users' level of confidence, which may passively indicate whether users find hint to be useful or not, after answering a question, users (i.e., both legitimate and adversarial users) were asked to rate their level of confidence in their answer on a 5-point Likert scale where 1 means "Not confident at all" and 5 means "Very confident". To analyze the effect of hints on user's level of confidence, we used Wilcoxon signed-rank test to evaluate the difference in confidence level for their answers. The Wilcoxon test is similar to t-tests, except it does not make any assumption regarding the distributions of the compared samples, which is appropriate for our analysis. Our analysis shows that the use of hints significantly increases users' confidence on their answer for questions about communication (i.e., Phone Call and SMS message). Specifically, for phone call and SMS message based questions, we found that legitimate users were significantly more confident in the correctness of their answers when hints are used with ($Z = -5.986$, $p < 0.01$) and ($Z = -3.313$, $p = 0.01$) respectively. However, in the case of location based questions, legitimate users were not statistically more confident when they used hints ($Z = -0.741$, $p = 0.459$). Figure 7 shows the effect of hints on legitimate users' level of confidence for 3 different question types (i.e., Call, SMS, and Location) when they used hints.

• Relation Between "Time taken to Answer" and Accuracy Score.
"Time taken to Answer" indicates the amount of time that was taken by a user to answer a question. The effect of time on accuracy score was insignificant for legitimate and adversarial users. Furthermore, we observed that adversarial users (i.e., strong and naive adversaries) took less time on average to answer the questions compared to legitimate users. Also, when users used hints, the amount of time to answer the questions were longer as expected. Table 7 summarizes the time taken by legitimate users to answer different types of questions. Specifically, legitimate users took on average 15.50 seconds with a median of 11 seconds to answer phone call questions with no hint compared to 22.45 seconds with

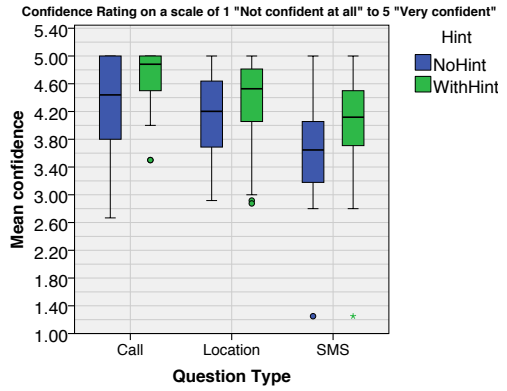


Figure 7: Impact of Hints on Legitimate User's Confidence rating while answering questions with hint and without hint.

a median of 18 seconds with hints, 18.15 seconds with a median of 12 seconds to answer SMS based questions with no hint compared to 29.34 seconds with a median of 25 seconds with hints, and 28.67 seconds with a median of 12 seconds to answer location based questions with no hint compared to 30.21 seconds with a median of 25 seconds with hints. The mean time taken to answer phone call and SMS based questions varied significantly when users used hints. However, the differences were not significant when users used hints for location based questions. Part of the reason may be due to the fact that the number of hints generated for location based questions were less compared to call and SMS based questions. Hence, users did not have to spend significant amount of time to go over the hints.

	Without Hint			With Hint		
	Mean	Median	SD	Mean	Median	SD
Call	15.50	11	13.24	22.45	18	15.18
SMS	18.15	12	18.03	29.34	25	17.90
Location	28.67	23	19.12	30.21	25	17.57

Table 7: Time taken for legitimate users to answer different types of questions.

• **Effect of Age and Gender on Accuracy Score.** While due to the limited size of the study, the effect of age and gender on response correctness cannot be claimed with high confidence, in our study, gender and age had no significant effect in predicting response correctness. Part of the reason may be because participants were undergraduate students with similar age (between 18-23) and we also had an almost equally balanced male and female population. A large-scale study is needed to verify the effect of age and gender on response accuracy.

4.1 Accuracy of Model-based Authentication

As each individual user is different and may perform differently, we attempt to account for this variations by building models for each user based on individual response patterns, and subsequently leverage that model to identify legitimate users. In our work, we first present a simple threshold based scheme, and next compare that against the performance of a more sophisticated Bayesian classifier based model which is inspired based on prior work [13].

4.1.1 Classification Performance Evaluation Metrics

ROC (Receiver operating characteristics) plots are commonly used for evaluating classification performance, in which TPR (true positives rate) on Y-axis is plotted as a function of FPR (false positives rate) on X-axis, and shows the trade-off between TPR and FPR for all possible thresholds (i.e., cut-off points) [15]. In this context, the true positive rate (TPR) corresponds to the success rate of legitimate users, while the false positive rate (FPR) denotes the success rate of adversaries. We also use the area under the ROC curve (AUC) to measure the performance of the test. Note that the performance of the test can be quantified with a single value by calculating AUC value [20] which is an important indicator of the classification performance. AUC = 0.5 represents a test performed at chance for binary classification (i.e., the model performs no better than a coin flip), while AUC = 1 means a perfect test where all legitimate users succeed and all adversaries failed to enter the system. Hence, the larger the AUC value, the better the model/test. Please note that, while evaluating performance of both threshold and Bayesian classifier based model, we use AUC value for different attack scenarios and vary the number of questions using the data collected from our field study where users are presented hints. The details are below.

4.1.2 Classification Accuracy of Threshold Based Scheme

As a single question may not be enough for reliably authenticating a user, we assume that multiple questions may be asked in a single session. Hence, in this scheme, we calculate the score of a user by taking average accuracy over multiple challenge questions in a session.

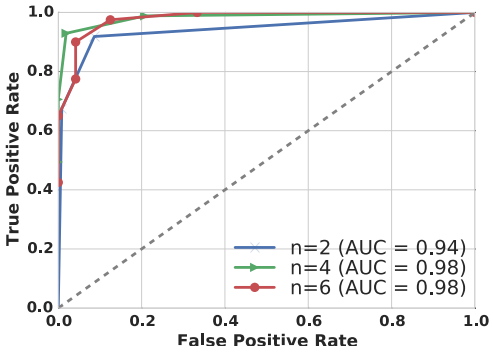
We generate ROC curves for the threshold-based scheme to show how the number of questions would affect TPR and FPR in identifying users for two different attack scenarios. In particular, in the first scenario, we assume the existence of only strong adversaries in the system (i.e., all attackers are strong adversaries), while in the second scenario, we assume the existence of only naive adversaries in the system (i.e., all attackers are naive adversaries).

The three curves in each plot in Figure 8 are generated for different number of questions (n). For brevity, we only show the ROC curves for $n = 2$, $n = 4$, and $n = 6$. From these figures, it can be seen that the AUC values are 0.94 for $n = 2$, 0.98 for $n = 4$, 0.98 for $n = 6$ when modeled against naive adversaries. In contrast, the AUC values are 0.80 for $n = 2$, 0.81 for $n = 4$, and 0.81 for $n = 6$ when modeled against strong adversaries.

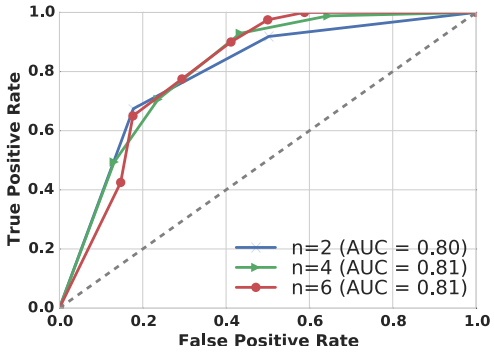
Although the performance is better when modeled against naive adversary compared to strong adversary (Figure 8(a)), the performance of threshold based scheme against strong adversary is not impressive, even when the number of questions in a session increases. This motivates us to explore the Bayesian classifier based scheme which is explained next.

4.1.3 Classification Accuracy of Bayesian Classifier Based Scheme

As different user's performance vary significantly, instead of relying solely on user's accuracy score (e.g., threshold-based scheme), one possible alternative is to learn a user's response pattern and subsequently leverage the response patterns along with accuracy score to authenticate a user.

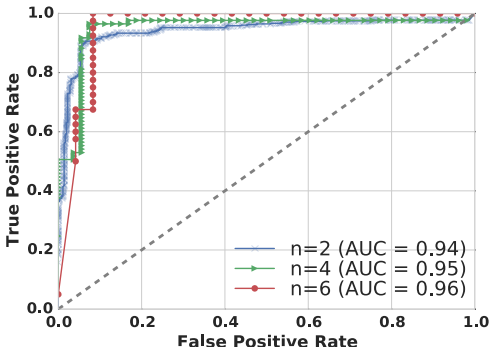


(a) Against Naive Adversary

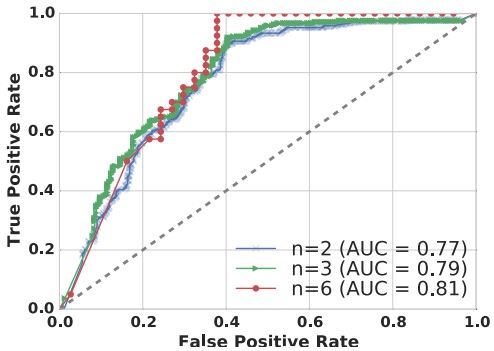


(b) Against Strong Adversary

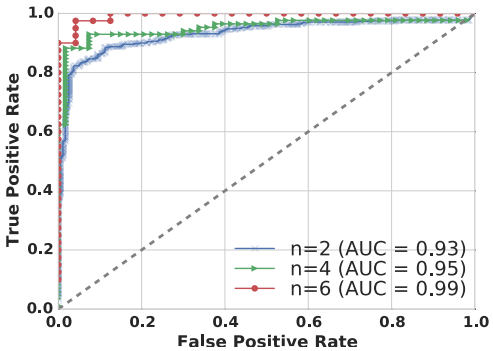
Figure 8: Receiver Operating Characteristic (ROC) curves for threshold based scheme when modeled against strong and naive adversary for different number of questions ($n = 2$, $n = 4$ and $n = 6$).



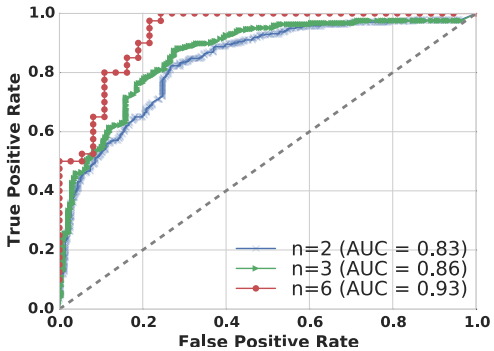
(a) **Modeled:**Against Naive Adversary **Tested:**Legit vs. Naive Adversary



(b) **Modeled:**Against Naive Adversary **Tested:**Legit vs. Strong



(c) **Modeled:**Against Strong Adversary **Tested:**Legit vs. Naive



(d) **Modeled:**Against Strong Adversary **Tested:**Legit vs. Strong

Figure 9: Receiver Operating Characteristic (ROC) curves for Bayesian classifier based scheme when modeled against strong and naive adversary for different number of questions (e.g., $n = 2$, $n = 4$ and $n = 6$). ROC curves are plotted for different user types (legitimate user, strong adversary, and naive adversary) and attack scenarios.

For example, a user who usually answers call and location questions correctly but SMS questions incorrectly is more likely to answer call and location questions correctly and SMS questions incorrectly in future attempts (i.e., repeat a similar pattern). Using this scheme, even if an adversary can somehow observe and learn a user's daily activities and answers all the questions correctly, the adversary will require to closely imitate the response errors and behavior of a legitimate user to gain access to the system. Based on this observation, in our work, we use the Bayesian classifier based model from Das et.al paper [13] to authenticate users reliably. Please note that, in [13], authors applied the

model only with two features which are categorical variables (i.e., question type and answer selection method (e.g., multiple choice vs. open ended)). We extended this prior approach and applied this model using different features such as question type, hint used, amount of time to answer, and user's level of confidence regarding the correctness of their answers. In our study, question type and hint used are categorical variables, and the amount of time to answer and user's confidence regrading their answers are continuous features. Here, if the response feature is a categorical variable such as questions type, we compute its probability by using its contingency table. If the response feature is a continuous variable such as the amount of time taken to answer a question, we compute its probability using probability density function assuming that the continues variable is distributed according to Gaussian distribution [23] as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (1)$$

where μ is the mean of the samples and σ is the variance of the samples. Given the μ and σ, equation 1 returns the probability of a particular sample belonging to the particular distribution.

Briefly, the Bayesian classifier based model [13] takes in a sequence of responses along with an "adversary model" as input (for simplicity, we pick a specific adversary type (e.g., strong, naive adversary)), and outputs a confidence rating ranging from 0 to 100 which represents how confident the system is in identifying the user as a legitimate user based on the user's response pattern.

Model Evaluation

To explore how the confidence ratings of the Bayesian classifier based model vary for different types of users and for different attack scenarios, we used the data collected from our field study where three different user types (i.e., legitimate, strong, and naive adversarial users) were simulated in each session. We then split each of these three different users' responses into n folds where n denotes the number of sessions. Subsequently, we use data from $(n-1)$ sessions to train the model and use the remaining session for testing. We repeat the process n times where each time we use a different session for testing.

To test the system, we try two different attack scenarios (against strong adversaries and against naive adversaries) as follows.

In the first case, we construct two models: one model represents the legitimate user which is trained using the legitimate user's data excluding the present session's data. The second model represents the strong adversary which is trained using the corresponding legitimate user's strong adversary data excluding the present session's data. For testing, we use the corresponding legitimate user's naive adversary data excluding the present session's data. Once the models are constructed, we calculate confidence rating for this scenario by using the remaining session for testing for three different user types.

Likewise, in the second case, we construct two models one using the legitimate user's data and the other using the corresponding legitimate user's naive adversary data. For testing, the corresponding legitimate user's strong adversary data is used. As before, the present session's data is excluded from the training data. We vary the number of challenge questions that we consider in each session. We also utilized combinations of different response features and we observed different confidence ratings for different response features after answering n questions aggregated across all sessions. For brevity, we only show plots for $n = 2$, $n = 4$, and $n = 6$ where we obtain the highest classification accuracy by using two features–question type and hint used.

Figure 9 shows the ROC curves for the Bayesian classifier-based scheme for two different attack scenarios. Specifically, Figure 9(a) and Figure 9(b) show the ROC curves when modeled against naive adversary (i.e., two models are trained–one for the legitimate user and one for the naive adversarial user) and tested using legitimate, naive, and strong adversarial user's data respectively. Similarly, Figure 9(c) and Figure 9(d) show the ROC curves for naive and strong adversarial users when modeled against strong adversary. As before, performance of each scenario/test is provided using the AUC value for $n = 2$, $n = 4$, and $n = 6$.

From these figures, for all attack scenarios, we observe that, as users answer more questions, regardless of the modeled adversary, the confidence estimate increases along with the AUC values. In other words, the system becomes more confident in identifying the actual user as the number of questions answered in a session increases. Also, we see that the classification performance varied greatly depending on the modeled adversary. For example, the highest classification performance was obtained against naive adversary. The AUC values are 0.96 and 0.99 for $n = 6$ when modeling against naive and strong adversary respectively. On the other hand, the classifier offers a much more conservative classification performance when tested against strong adversary. Specifically, AUC values are 0.93 and 0.81 after answering 6 questions when tested against strong adversary and modeled against naive and strong adversary respectively. Intuitively, as a strong adversary has significant knowledge regarding a user's schedule (e.g., girlfriend), strong adversarial users are more likely to gain access to the system by answering questions more accurately compared to naive adversarial users.

Please note that, while the Bayesian classifier based model achieves high accuracy, the model requires training data for both legitimate user and his/her adversarial users, which may not be available in real-life. Investigating alternative models that can be trained using a group of adversarial users' data that do not include any specific adversary, which is more likely to be available, is one of our future work.

4.2 Usability Aspect Regarding Autobiographical Authentication and Usage of Hints: A User's Perspective

To understand the impact of providing hints on usability of such systems, at the end of the study, the participants were asked to complete an exit survey for an additional $10 Amazon Gift card. We asked the participants to answer several questions to understand their perception regarding *Autobiographical Authentication* and the effectiveness of using hints. We used a five-point Likert-scale where 1 indicates strong disagreement and 5 indicates strong agreement with the given statement. Table 8 summarizes the survey results. As the survey responses are ordinal data, it is appropriate to employ median and mode rather than mean and standard deviation. Also, in case where multiple modes exist, the smallest value is shown in Table 8.

As per the survey data, most of the participants found phone

	Call		SMS		Location	
Question	**Mode**	**Median**	**Mode**	**Median**	**Mode**	**Median**
Q_1: It was easy for me to recall	4	4	4	3.5	5	4
Q_2: It was easy for my close friends to guess	3	3	3	3	3	3
Q_3: It was easy for me to guess my close friends' questions	3	2.5	1	2	3	3
Q_4: It was easy for a stranger to guess	1	1	1	1	1	1
Q_5: It was easy for me to guess stranger's questions	1	1	1	1	1	1
Q_6: I found hints to be useful while answering	4	3.5	4	3	4	3.5
Q_7: I found hints to be useful while guessing my close friend's questions	1	2	1	1.5	1	2
Q_8: I found hints to be useful while guessing a stranger's questions	1	1	1	1	1	1
Q_9: I do not think hints generated based on my data can leak my privacy sensitive information if shown to my close friend	4	4	5	4	4	4
Q_{10}: I do not think hints generated based on my data can leak my privacy sensitive information if shown to a stranger	5	5	5	5	5	4

Table 8: User Feedback on Autobiographical Authentication scheme and effectiveness of using hints. A five-point Likert-scale was used on a scale of 1 (strong disagreement) to 5 (strong agreement) with the given statement.

call and location based questions to be easier to recall (mode 4, median 4 and mode 5, median 4 respectively) compared to SMS based questions (mode 4, median 3.5 in Q_1). When it comes to guessability, most of the participants disagreed that guessing the answers of their close friend's question would be easy. They reported that questions about communications (i.e., Call and SMS) would be harder to guess compared to location based questions (with mode 1 and median 2 for SMS based questions, mode 3 and median 2.5 for call questions, mode 3 and median 3 for location questions in Q_3). A majority of the participants strongly agreed that a stranger would not be able to guess their questions and vice versa (mean and median values are all 1 for these cases). When it comes to the effectiveness of using hints, the majority of participants found hints to be useful while answering communication and location based questions (for call and location based questions mode 4, median 3.5, for SMS based questions mode 4, median 3 in Q_6). Moreover, a majority of the participants disagreed that hints would be helpful while guessing (mean and median values are all less than 2 for these cases). Finally, most of the participants did not think that hints generated based on their data can leak their privacy sensitive information if shown to their close friends (response mean and median values are all greater than or equal to 4 in Q_9), or if shown to a stranger (response mean and median values are 5, except for location question with a mode 5 and median 4 in Q_{10}).

5. DISCUSSION

In this paper, we investigated the strengths and weaknesses of providing hints for different categories of autobiographical questions and users (e.g., legitimate, strong, and naive adversarial users). We also presented a Bayesian classifier based model to account for differences in individual user's response pattern and subsequently leveraged that to identify legitimate users with high accuracy while reducing the success rate of adversaries. Based on our findings, we strongly believe that the proposed authentication framework can significantly improve the overall system security by adding another layer of easy-to-use authentication mechanism. For instance, the presented scheme can be used in addition to password based mechanism to prevent attacks launched from remote locations using stolen passwords, or can be used to facilitate resetting of passwords.

While we found that providing hints improved legitimate users' response accuracy while having no significant effect on naive adversarial users' performance, interestingly, hints appear to have negative effect on response correctness for strong adversarial users. While such negative effect could be due to increased ambiguities caused by hints, further investigation focusing on this particular aspect of our finding is needed to identify the underlying reasons behind such effect.

Finally, we would like to point out that, in our study, all of the participants ($n = 24$) were undergraduate students with similar age (range is 18-23). Due to the limited size of the study and skewed age distributions of the participants, the effect of age and gender on response correctness cannot be claimed with high confidence, and needs further investigation. Also, the effect of smartphone usage behavior (e.g., low vs. heavy smartphone users) on response accuracy cannot be verified due to limited size of the study, and needs further investigation.

6. CONCLUSION

This paper investigates the possibility of using hints to improve users' recall rate for autobiographical based authentication systems. To evaluate the effect of hints on legitimate and adversarial users' performance, a real-life user study is conducted with 24 users over a period of 30 days. The findings suggest that hint indeed has a significant positive effect on legitimate users' response correctness while negative effect on strong adversarial users' response correctness, and no significant effect on response correctness for naive adversarial users. Finally, based on exit survey data, it was noted that users feel positively about using hints while answering challenge questions.

7. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No. CNS-1251962. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency. We thank the anonymous reviewers for their valuable suggestions and comments.

8. REFERENCES

[1] Android fused location provider api. `https://developers.google.com/android/reference/com/google/android/gms/location/FusedLocationProviderApi` (Accessed: 06/11/2015).

[2] Android's activity recognition api: Recognizing the user's current activity. `https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionApi` (Accessed: 06/11/2015).

[3] Google maps android api. `https://developers.google.com/maps/documentation/android/` (Accessed: 06/11/2015).

[4] Y. Albayram, M. M. H. Khan, A. Bamis, S. Kentros, N. Nguyen, and R. Jiang. A location-based authentication system leveraging smartphones. In *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*, volume 1, pages 83–88. IEEE, 2014.

[5] F. Asgharpour and M. Jakobsson. Adaptive challenge questions algorithm in password reset/recovery. *First International Workshop on Security for Spontaneous Interaction (IWIISI '07), Innsbruck, Austria, (2007)*, 7:6 pages, 2007.

[6] 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*, WWW '15, pages 141–150, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[7] Y. Chen and D. Liginlal. Bayesian networks for knowledge-based authentication. *IEEE Trans. on Knowl. and Data Eng.*, 19(5):695–710, May 2007.

[8] Y. Chen and D. Liginlal. A maximum entropy approach to feature selection in knowledge-based authentication. *Decision support systems*, 46(1):388–398, 2008.

[9] S. Chokhani. Knowledge based authentication (kba) metrics. In *KBA Symposium-Knowledge Based Authentication: Is It Quantifiable*, 2004.

[10] M. A. Conway. Episodic memories. *Neuropsychologia*, 47(11):2305–2313, 2009.

[11] M. A. Conway, D. C. Rubin, A. Collins, S. Gathercole, M. Conway, and P. Morris. The structure of autobiographical memory. *Theories of memory*, pages 103–137, 1993.

[12] S. K. Dandapat, S. Pradhan, B. Mitra, R. Roy Choudhury, and N. Ganguly. Activpass: Your daily activity is your password. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2325–2334, New York, NY, USA, 2015. ACM.

[13] 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*, UbiComp '13, pages 211–220, New York, NY, USA, 2013. ACM.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[15] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[16] P. Gupta, T. K. Wee, N. Ramasubbu, D. Lo, D. Gao, and R. K. Balan. Human: Creating memorable fingerprints of mobile users. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 479–482. IEEE, 2012.

[17] L. C. Hamilton. *Statistics with STATA: Version 12*. Duxbury Press, Boston, MA, USA, 8 edition, 4 2012.

[18] A. Hang, A. De Luca, and H. Hussmann. I know what you did last week! do you?: Dynamic security questions for fallback authentication on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1383–1392. ACM, 2015.

[19] E. Hassan. Recall bias can be a threat to retrospective and prospective research designs. *The Internet Journal of Epidemiology*, 3(2):339–412, 2006.

[20] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[21] M. Hertzum. Minimal-feedback hints for remembering passwords. *interactions*, 13(3):38–40, 2006.

[22] M. Jakobsson. *The death of the Internet*. John Wiley & Sons, 2012.

[23] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[24] G. Kristo, S. M. Janssen, and J. M. Murre. Retention of autobiographical memories: An internet-based diary study. *Memory*, 17(8):816–829, 2009.

[25] B. Lu and M. B. Twidale. Managing multiple passwords and multiple logins: Mifa minimal-feedback hints for remote authentication. In *IFIP INTERACT 2003 Conference*, pages 821–824, 2003.

[26] M. Nishigaki and M. Koike. A user authentication based on personal history-a user authentication system using e-mail history. *The Journal on Systemics, Cybernetics and Informatics*, 5(2):18–23, 2007.

[27] A. Nosseir, R. Connor, and M. Dunlop. Internet authentication based on personal history-a feasibility test. 2005.

[28] A. Nosseir and S. Terzis. A study in authentication via electronic personal history questions. In *ICEIS 2010 - Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 5, HCI, Funchal, Madeira, Portugal, June 8 - 12, 2010*, volume 5, pages 63–70, 2010.

[29] L. O'Gorman, A. Bagga, and J. Bentley. Call center customer verification by query-directed passwords. In *Financial Cryptography*, pages 54–67. Springer, 2004.

[30] J. Podd, J. Bunnell, and R. Henderson. Cost-effective computer security: Cognitive and associative passwords. In *Computer-Human Interaction, 1996. Proceedings., Sixth Australian Conference on*, pages

304–305. IEEE, 1996.

[31] A. Rabkin. Personal knowledge questions for fallback authentication: Security questions in the era of facebook. In *Proceedings of the 4th Symposium on Usable Privacy and Security*, SOUPS '08, pages 13–23, New York, NY, USA, 2008. ACM.

[32] K. Renaud, T. McBryan, and P. Siebert. Password cueing with cue (ink) blots.

[33] S. Schechter, A. B. Brush, and S. Egelman. It's no secret. measuring the security and reliability of authentication via "secret" questions. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, pages 375–390, Washington, DC, USA, 2009. IEEE Computer Society.

[34] S. Schechter, S. Egelman, and R. W. Reeder. It's not what you know, but who you know: A social approach to last-resort authentication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1983–1992. ACM, 2009.

[35] S. Schechter and R. W. Reeder. 1 + 1 = you: Measuring the comprehensibility of metaphors for configuring backup authentication. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, page 9, New York, NY, USA, 2009. ACM.

[36] J. Thorpe, B. MacRae, and A. Salehi-Abari. Usability and security evaluation of geopass: A geographic location-password scheme. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, page 14, New York, NY, USA, 2013. ACM.

[37] S. Vemuri and W. Bender. Next-generation personal memory aids. *BT Technology Journal*, 22(4):125–138, 2004.

[38] W. A. Wagenaar. My memory: A study of autobiographical memory over six years. *Cognitive psychology*, 18(2):225–252, 1986.

[39] W. E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. 1990.

[40] K. Xu, D. Yao, M. A. Pérez-Quinones, C. Link, and E. Scott Geller. Role-playing game for studying user behaviors in security: A case study on email secrecy. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2014 International Conference on*, CollaborateCom '14, pages 18–26. IEEE, 2014.

[41] M. Zviran and W. J. Haga. User authentication by cognitive passwords: An empirical assessment. In *Information Technology, 1990.'Next Decade in Information Technology', Proceedings of the 5th Jerusalem Conference on (Cat. No. 90TH0326-9)*, JCIT, pages 137–144, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.