I Know What You Did Last Week! Do You? Dynamic Security Questions for Fallback Authentication on Smartphones

Alina Hang[1], **Alexander De Luca**[1,2,3], **Heinrich Hussmann**[1]
[1]Media Informatics Group,University of Munich (LMU), Munich, Germany
[2]DFKI GmbH, Saarbrücken, Germany, [3]Fraunhofer FKIE, Bonn, Germany
(alina.hang, alexander.de.luca, heinrich.hussmann)@ifi.lmu.de

ABSTRACT
In this paper, we present the design and evaluation of dynamic security questions for fallback authentication. In case users lose access to their device, the system asks questions about their usage behavior (e.g. calls, text messages or app usage). We performed two consecutive user studies with real users and real adversaries to identify questions that work well in the sense that they are easy to answer for the genuine user, but hard to guess for an adversary. The results show that app installations and communication are the most promising categories of questions. Using three questions from the evaluated categories was sufficient to get an accuracy of 95.5% - 100%.

Author Keywords
Fallback Authentication; Dynamic Security Questions;

ACM Classification Keywords
H.5.1. Information Interfaces and Presentation (e.g. HCI): Evaluation/methodology

INTRODUCTION
Mobile devices employ a variety of authentication schemes like PINs, alphanumeric passwords or graphical passwords to protect sensitive data (e.g. photos, contact details) on a user's device. However, fallback solutions are required when these systems fail (e.g. when users enter their PINs/passwords incorrectly for too many times). Personal Unblocking Codes (PUCs) are often used in a mobile context to enable users to regain access to their device [16]. However, they are inconvenient since users do not (and should not) carry them around all the time. Retrieving a PUC in a situation where the user is left alone, maybe in a foreign country, with the locked device, is difficult in practice. Also the fallback solutions provided by the most popular operating systems for mobile devices are very difficult to use under such circumstances, like using a specific online account for Android or connecting Apple devices to a computer running iTunes [1].

Mobile devices are often used on the go and should offer fallback solutions that can be used locally without Internet connection and which are immediately available when needed. In turn, the time needed to complete the fallback authentication process is not as critical as for primary authentication.

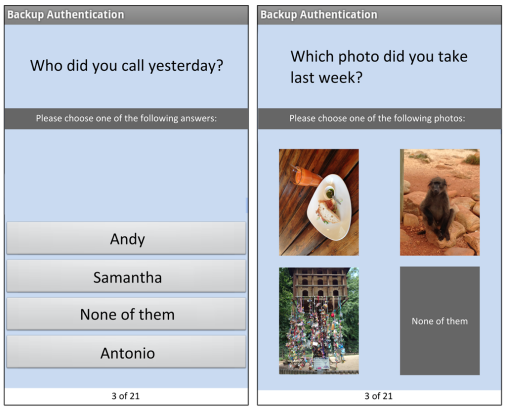

Figure 1. Screenshot of exemplary security question with four possible answers using text (left) or images (right). Translated from German.

In this work, we propose to take advantage of the enormous amount of data that is already stored on mobile devices to create dynamic security questions. This approach is immediate, does not require any additional tokens and is individual for each user. Since the questions are based on personal information and behavior, users should be able to answer them. In turn, persons who want to gain unauthorized access should not have enough knowledge to provide the correct answers.

The main contributions of this paper are the iterative design and evaluation of dynamic security questions in due consideration of different types of human adversaries (close and acquainted). Our studies indicate that privacy implications of certain questions play an important role in the design of dynamic security questions. Though certain questions were easy to remember, they turned out to be less accepted by users. Thus, different dimensions have to be taken into account when designing dynamic security questions. The study results show that the right combination of questions can yield up to 100% accuracy. In particular, questions about app usage and app installations are the most promising ones. They offered the best trade-off between usability (i.e. memorability) and security. While participants were better in answering questions about more recent activities, the performance of adversaries seem not to be influenced by the timespans used.

RELATED WORK
The most common solutions for fallback authentication are designed for web services and are often based on email or security questions.

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea

ACM 978-1-4503-3145-6/15/04$15.00.
http://dx.doi.org/10.1145/2702123.2702131

Communication-based Password Reset

Email-based password reset is a popular approach for fallback authentication. In case of password loss, the (new) password or a link to reset the password is sent to the corresponding email account. According to Garfinkel [5], this approach works well, but comes with certain shortcomings (i.e. lack of email encryption). In a mobile context, email-based password resets do not work well as the user's only email access may be through the smartphone. Alternatives for fallback authentication have been proposed in various work. Schechter et al. [18] suggest social authentication for password reset. However, their results show that users also had memorability problems with this approach.

Security Questions

Numerous web services use security questions for password reset [4]. In order to use them, users have to answer a number of questions at set-up time, which have to be recalled during fallback authentication.

Most security questions are predefined and are based on facts (e.g. "What is your mother's maiden name?"). Though these kinds of questions seem to be easy to answer, several studies have shown that these traditional security questions are inadequate and have a bad performance in terms of usability and security (e.g. [6], [8], [15], [17]).

Since fallback authentication does not occur frequently (in the best case, never), users are often not able to recall the answers they have given due to the ambiguity of possible answers or the inapplicability of the questions. Therefore, users prefer to select easier questions, which in turn are also easier to guess for others (e.g. [15]). Haga et al. [8] and Schechter et al. [17] have shown that such questions are often correctly guessed by family members, friends or acquaintances. In particular, in times of social networks, answers to questions are often researchable by adversaries (e.g. [6], [7], [14]). Also, most security questions can be guessed by choosing the most popular answers [17].

To overcome the shortcomings of traditional security questions, some web services enable users to define their own questions [10]. However, users often lack creativity, forget the answers to their own questions or generate questions that are not secure enough [11]. Other solutions are based on the users' preferences (e.g. [9]). However, preferences might change over time and are vulnerable to insider threats: threats by adversaries that know the user well.

Dynamic Security Questions

The previous section has shown that the design of security questions is challenging. Users have to state the answer during enrolment so that the answer might not be up-to-date at the time when the actual fallback authentication is performed.

Thus, Babic et al. [2] propose the design of dynamic security questions that are based on the Internet activities of the user. Since the answers are derived automatically, no enrolment is needed and the answers are always up-to-date.

Category	Question + Timespan
SMS (out)	Who did you text [Y | LW | LM]?
SMS (in)	Who texted you [Y | LW | LM]?
Call (out)	Who did you call [Y | LW | LM]?
Call (in)	Who called you [Y | LW | LM]?
App	Which App did you use [Y | LW | LM]?
Music	Which artist did you listen to [Y | LW | LM]?
Photos	Which photo did you take [Y | LW | LM]?

Y=Yesterday; LW=Last Week; LM = Last Month

Table 1. Overview of 21 security questions used for the pre-study. One question for each category combined with three different timespans.

For a mobile context, Das et al. [3] presented an autobiographical authentication approach based on different categories (e.g. communication, technology usage, etc.). They created 13 different security questions and evaluated them in a user study. Participants performed best in answering questions about communication and were not as good in answering questions about app usage.

In order to test the security level of their security questions, Das et al. [3] introduce a theoretical adversary model that considers different types of adversaries, e.g. adversaries that know nothing, adversaries that know everything, etc. This is an interesting approach considering that studies with real adversaries are complex, time consuming and often produce high dropout rates. A theoretical model can provide hints at the security level of security questions.

Nonetheless, the model does not consider all circumstances that might influence the performance of real adversaries. Our work is similar to the work by Das et al. [3] in the sense that we analyze different types of dynamic security questions based on personal information on mobile devices. However, it is significantly different from it as we evaluate the security of those questions with human adversaries. We distinguish between close adversaries (e.g. family, friends) and acquaintances. This way, we were able to evaluate the actual security of the approach rather than a theoretical approach. For instance, with respect to true positive and true negative rates, we identified app usage as one of the best categories.

THREAT MODEL

We assume an adversary that is in possession of the smartphone. The adversary does not have the PIN or password necessary to unlock the device. Thus, the device gets blocked completely. The adversary can now use the fallback authentication mechanism to try to get access to the device. Assuming an equal distribution for each answer option, the chances for an adversary without any knowledge about the device owner to get access to the device is $(\frac{1}{x})^n$ with x being the number of possible answers for each question and n being the number of questions that have to be correctly answered.

To simulate a worst-case scenario, we consider adversaries with advanced knowledge about the user. This way, breaching the device is not plain luck. According to [13] these adversaries are very likely and thus, very interesting. The goal is to identify questions that are easily answered by users and that are, at the same time, harder to guess for these adversaries.

BRAINSTORMING

We conducted a brainstorming with four smartphone users (one female) to identify interesting information categories for the design of dynamic security questions. The participants were recruited over different channels like social networks, bulletin boards or personal communication. The participants were aged between 23-25 years (average: 24 years). All of them were students and had a technical background. Participants were invited to our lab. After a brief introduction, in which we explained the idea of dynamic security questions, they were asked to name and discuss different information categories that they thought to be suitable for the design of dynamic security questions. Participants received gift vouchers for their participation.

As a result, we identified seven categories: SMS (outgoing and incoming), call (outgoing and incoming), app, music and photos. These categories mostly agree with the results by Das et al. [3].

PRE-STUDY

In order to evaluate the seven categories, we created one question for each category. Additionally, each question was combined with the timespans *yesterday*, *last week* and *last month*. We did this to see, how far we can go back in time for the dynamic security questions. Table 1 gives an overview of all the security questions used in the pre-study.

Prototype

The prototypes consisted of two Android apps:

Logging Application

The logging app collected usage information on the user's device running in the background. Therefore, the participants did not have to cope with any changes to their smartphones. We collected information about incoming/outgoing text messages, incoming/outgoing calls (e.g. contact number), app usage (e.g. name of app) and music (e.g. artist). To ensure that our research complies with federal privacy laws, we paid particular attention to respect the participants' privacy when implementing the logging application and conducting the user studies. None of the information that were collected ever left the devices. If requested, we showed the participants a list of all the information that was logged.

Questions Application

The second app generated the actual security questions based on the logged information. The questions appeared in random order. For each of the 21 questions, 4 possible answers were provided, one of them being the option *none of them*. The incorrect answer options were generated from the logged data (excluding the correct answer). Figure 1 shows two examples on how the answer possibilities looked like. For all information categories, possible answers were provided as text (see figure 1, left). Only for the questions about photos, pictures were used instead (see figure 1, right). In case there was no logged information available to generate a question (e.g. when a user did not listen to music during a specific timespan), the corresponding question was skipped. To avoid that the questions might be too easy to guess for adversaries, we excluded popular apps like Facebook or WhatsApp for the questions "Which app did you use [yesterday, last week, last month]?" The apps that were excluded are based on the data from a survey by [20], which lists the 20 most used apps in 2012 (the year in which the pre-study was conducted).

Category	Question + Timespan
SMS (out)	Who did you text [Y | LW]?
SMS (in)	Who texted you [Y | LW]?
Call (out)	Who did you call [Y | LW]?
Call (in)	Who called you [Y | LW]?
App	Which App did you use [Y | LW]?
App Install	Which app did you install/update [Y | LW]?
Photos	Which photo did you take [Y | LW?

Y=Yesterday; LW=Last Week

Table 2. Overview of 14 security questions used for the main study. One question for each category combined with two different timespans.

Study Design

For the pre-study, we used a within-subject design with the independent variables *category* (7 levels) and *timespan* (3 levels). This resulted in 7x3=21 different security questions (see table 1).

In order to participate in the pre-study, participants had to own an Android Smartphone. The pre-study consisted of two phases: a logging phase and a phase in which the participants had to answer questions. For the second phase, participants were asked to bring a person who knew them very well to the study. This person was asked to act as an adversary. We gave participants examples of close persons (e.g. partners and family). During the study, we also asked participants/adversaries to rate as how close they would describe their relationship.

Procedure

In the first phase of the pre-study, we sent instructions by email to our participants that informed them about the general procedure and showed them how to install the logging application that we also made available via email. The logging phase lasted for four weeks, after which we contacted the participants again and invited them to our lab.

The goal of the second phase was to let participants and adversaries answer the security questions. We installed the question app on the participant's device. Then the adversary was asked to leave the room and to wait outside. In the meantime, the participant had to answer all the questions from the question app. The questions were created in random order. This was followed by a questionnaire to gather subjective assessments. Once the participant had finished, the adversary was invited back in and had to complete the same tasks with the participant's smartphone. Participants and adversaries received gift vouchers as incentives.

Participants

Altogether, 38 persons took part in our pre-study: 19 smartphone owners and 19 persons that the smartphone owners brought along. In the remainder of the paper, smartphone owners will be referred to as participants and accompanying persons will be referred to as adversaries.

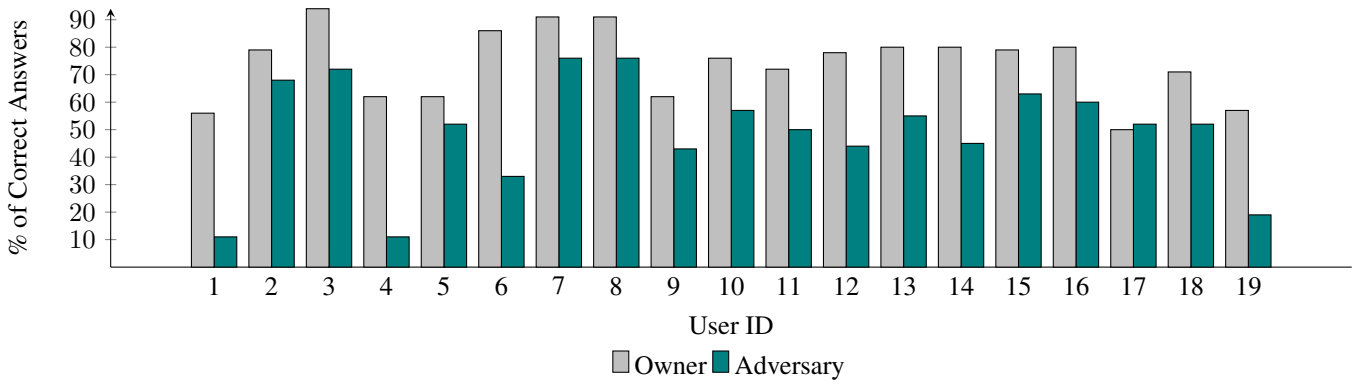

Figure 2. Overview of number of questions asked per participant and the number of correct answers by each participant as well as their corresponding adversary in the pre-study.

We recruited 19 participants (7 female) for the pre-study using mailing lists and public bulletin boards. They were aged between 20-31 years (average: 26 years). 14 of them had a college degree, 5 had a high school degree. Most of our participants had a technical background. All of them owned an Android smartphone, which was a prerequisite to participate in the study.

All participants brought one person to the study to act as an adversary. The 19 adversaries (10 female) were aged between 20-30 years (average: 24 years). 12 had a college degree, 7 a high school degree.

The relationship between participants and adversaries varied among the participants. One brought her spouse, 7 brought their partner, 8 brought a friend and three brought some acquaintances. Participant and adversary were asked to name the relationship with the respective other person in the questionnaire. Only in three cases, the relationships did not match. In those cases, the adversary considered the participant as friend, while the participant considered the adversary as an acquaintance, or the other way around. However, it is hard to draw the line between friend and acquaintances. Altogether, we can say that there is a good agreement on the relationships between participants and adversaries.

Results

We collected a total of 372 answers to dynamic security questions from participants as well as 372 answers from the adversaries. On average 20 answers were gathered per participant and per adversary.

Participants provided 275 (74%) correct answers. This is a higher percentage than can be found in similar work [3]. Adversaries gave 191 (51.3%) correct answers. The participant that performed worst only got 50% of the questions right. The best participant gave a correct answer in 94% of the cases. The best adversary reached 76% of correct answers, while the worst adversary had only 11% of correct answers. Figure 2 gives an overview of the number of correct answers for participants and their respective adversaries.

Users 3, 7 and 8 achieved over 90% of correct answers. However, their corresponding adversaries also answered over 70% of the questions correctly. Participants 7 and 8 did not only participate in our study as normal users, but they also acted as adversaries for each other. They were best friends who liked to communicate with each other using their smartphone. They told us that, in some cases, they were the answer to each other's questions. For example, for the question"Who did you call yesterday?", the adversary herself was the answer to the question. In contrast to this, there are participants who knew surprisingly little about themselves (ID 1, 17, 19). They answered less than 60% of their own questions.

Participants performed particularly well in giving correct answers to questions about text messaging: 92.8% for SMS (out) and 79% for SMS (in). However, adversaries achieved high percentages as well: 65% for SMS (out), 61.4% for SMS (in). Having a look at the other categories, adversaries achieved less than 50% of correct answers for questions about app usage, music, photos and calls (in). For the categories music and photos, participants could only answer 62.1% and 56%, respectively.

Questions about app usage showed the best trade-off between usability and security. While participants were able to answer over 70% of questions about app usage, their adversaries only achieved around 35%.

Participants tended to overestimate their performance. The difference between the number of answers they thought they had answered correctly and the actual number of correct answers was on average 3 (min=0; max=8). For adversaries the average difference between self-assessment and real performance was 4 (min=0; max=17).

Timespans

We asked participants to rate each question with respect to ease on a 5-point Likert scale, ranging from very difficult (1) to very easy (5). There were some questions that almost all participants found easy or very easy to answer (min=17 participants; max=19 participants). Interestingly, all these questions were about the timespan *yesterday*. This subjective assessment corresponds to their actual performance.

Participants had more correct answers for the timespan *yesterday* (87.3%) than for questions with the timespan *last week*

(71%). The timespans *last week* yielded more correct answers than questions about *last month* (62.3%). No tendencies were observed for adversaries (around 50% correct answers).

Perceived Security

In the questionnaire, we asked the participants to rate the perceived security of a question by assessing how easy they think certain questions could be guessed by an adversary.

Questions about music and app usage were considered as safe categories. There is no clear tendency for the other categories. We also did not find any particular timespans that participants found less or more secure.

User Stories and Comments

In general, the participants found our approach entertaining and fun. One participant in particular enjoyed the fact that she could learn more about her own usage behavior. Though some participants were concerned about the time it takes to answer all the questions, most of them found it convenient.

With respect to security, most participants thought the approach was secure against strangers, but they also assume that close persons or even acquaintances will be able to answer some of the questions.

Despite the positive feedback, there were participants who were concerned about their private data. After the study, one participant and his girlfriend jokingly said: "We just broke up - just kidding, but your study could destroy relationships. You are revealing information that you actually want to protect". They were referring to photos and communication details that we revealed as answer options during the study.

Another interesting comment was stated by two participants whose birthday was during the logging phase. One of them said: "I would have been better in answering questions about last month, if it wasn't my birthday. I received so many text messages". Thus, special events might influence the performance of a user.

MAIN STUDY

Based on the results of the pre-study, we improved the security questions and evaluated them further.

The information category music was replaced by the category app installation. This app-related category was included, since the category app usage had the best trade-off between usability and security in the pre-study. Other questions were not modified. Though questions about communication were easy to attack in the pre-study, we kept them to evaluate them in the main study with different types of adversaries (close adversary and acquainted adversary).

With respect to the timespan, we removed *last month* and kept only the timespan *yesterday* and *last week*. To enhance privacy, photos were blurred for the main study.

Study Design

The study design was similar to the pre-study. We used a within-subject design with two independent variables: *category* (7 levels) and *timespan* (two levels). Altogether, we evaluated 7x2=14 security questions (see table 2).

The study consisted of two phases: a logging phase and a phase, in which the participants had to answer security questions in random order. Again, the prerequisite to participate in the study was to own an Android smartphone.

Study Procedure

The study procedure was identical to the first study. The only difference was that we asked participants to bring two persons with them, one person that they know very well and one person they are only acquainted with. This way we were able to analyze our questions in terms of security with respect to different types of adversaries. Again, we provided participants with examples for close as well as acquainted persons and also asked participants/adversaries to rate the closeness of their relationship to each other.

Participants

Participants who took part in the pre-study were not allowed to participate again. Participants were recruited over bulletin boards, mailing list and social networks. Altogether, we recruited 18 participants for the main study. We ended the study with 11 participants (+11 close adversaries + 11 acquainted adversaries). There were different reasons for the high dropout rate, one of them being the vacation season where participants and adversaries did not show up for the second phase. Other participants had to be excluded, since the second application could not be installed on their smartphones.

The 11 participants (5 female) were aged between 19-33 years (average: 24 years). Two of them had a college degree, while 9 had a high school degree. The 11 close adversaries (three female) were aged between 19-33 years (average: 23 years). 6 of them had a college degree. 5 had a high school degree. Acquainted adversaries (two female) were aged between 19-58 years (average: 27 years). 5 of them had a college degree. 6 of them had a high school degree.

The relationship to the close adversary was manifold. Two participants brought their spouse, three brought their partner, 5 brought a very close friend and one brought her brother. There were no contradictions in the relationship between participants and their close adversary. They also all stated that they knew each other well to very well.

Eight participants brought a person they knew from college, school or work as acquainted adversaries. One of them brought a friend, two brought a friend of a close friend. Acquaintances and participants named the same relationship. There was no contradiction. Most adversary-participant pairs did not know each other well.

Results

Correct Answers to Security Questions

We created 154 dynamic security questions (14 per participant). Each of those questions had to be answered by the participants as well as their adversaries. Participants gave 135 (87.7%) correct answers. Close adversaries had 84 (54.5%) correct answers, while acquainted ones had 67 (43.5%).

The best participant had 92.9% correct answers. The worst participant achieved 78.6%. The best close adversary had

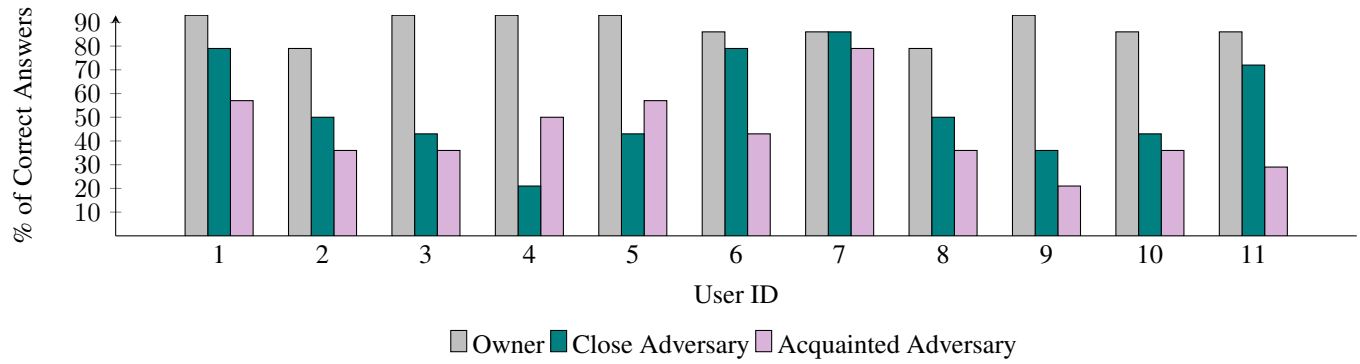

Figure 3. Overview of number of questions asked per participant and the number of correct answers by each participant as well as their corresponding adversary in the main study.

85.7%, the worst had 21.4%. The same observations can be made for acquainted adversaries. The worst had 21.4%, while the best yielded 78.6%.

An overview of the number of correct answers is shown in figure 3. Five participants had over 90% of correct answers, 4 got over 85% and 2 had over 75%. Participants were very good in answering their questions. They had only between 1-3 of 14 questions incorrect.

The results show that there are some participants that are easier to attack than others. In particular, for participants 1, 6, 7 and 11, close adversaries were good in answering their security questions. The acquainted was not as good as the close one. Participants that were hard to attack had the ID 3, 4, 9, 10. While they had a high percentage of correct answer, their adversaries, close and acquainted, yielded only a low percentage of correct answers.

The participants were good in assessing their performance. The average difference between estimation and actual performance was only 1 question (min=0; max=2;). The difference for adversaries was on average 3.

Timespans

As mentioned before, participants were good in answering questions. They made only 1-3 errors. However, they made at most one error for questions about *yesterday*. This error was always made for the question about app usage.

In general, participants achieved high percentage of correct answers for *yesterday* (94.8%) and were better in answering them than answering questions on *last week* (80.5%). Adversaries were better in answering yesterday questions than questions on *last week*. However, the differences are minimal. Close adversaries had 55.8% of correct answers for *yesterday* questions and 53.3% for questions about *last week*. In turn, acquainted adversaries achieved a lower number of correct answers than the close adversaries (44.2% for yesterday, and 42.9% for last week).

We conducted a mixed design ANOVA with the between-factor user type (user, close or acquainted adversary) and the within-factor timespan (*yesterday* and *last week*). The analysis showed no significant main effect for timespan, but a highly significant main effect for user type ($F(2,30) = 24.114$; $p < .001$). The test did not show any interaction effects between the factors timespan and user type.

The post-hoc test showed highly significant differences between user and close adversary ($p < 0.001$; *Bonferroni corrected*) and also between user and acquainted adversary ($p < 0.001$; *Bonferroni corrected*). No significant differences were found between close and acquainted adversaries.

Perceived Security

The majority of our participants found questions about communication activities and photos to be insecure. Only the information category app installation (with the timespan *yesterday*) was rated to be appropriate. The opinions for the other categories were ambiguous.

Accuracy of Security Question

Accuracy calculations are a good indicator on how well an authentication system performs. In our case, we used accuracy to calculate how well each security question worked. True positives (TP) are successful authentication attempts by legitimate users, while true negatives (TN) are unsuccessful attacks of potential adversaries. In turn, false negatives (FN) are counted when a legitimate user is accidentally rejected, while false positives (FP) depict the number of successful authentication attempts by potential adversaries. The formula for the calculation looks like this:

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum FP + \sum TN + \sum FN}$$

We will use this formula for the forthcoming analyses. The formula returns a value between 0 and 1 (or 0 and 100 in percentage). A value of 0 is bad and a rate of 1 (or 100) means that all attacks failed and the user always succeeded. Thus, a very desirable result.

Since we used two different timespans and two types of adversaries, there are four values for accuracy. This way, we can compare if the accuracy of a system depends on the timespan and type of adversary who tries to attack our questions.

Table 3 gives an overview of the calculated accuracy values that only consider the answers by users and attempted attacks by close adversaries for calculation. Table 4 shows the accuracy values that only take into account answers by users and acquainted adversaries. With respect to the close adversary,

	Close Adversary									
	Yesterday					Last Week				
Question / Category	TP	TN	FP	FN	ACC	TP	TN	FP	FN	ACC
Who did you text? / SMS (out)	11	4	7	0	68.2	9	5	6	2	63.6
Who texted you? / SMS (in)	11	5	6	0	72.7	11	2	9	0	59.1
Who did you call? / Call (out)	11	5	6	0	72.7	8	7	4	3	68.2
Who called you? / Call (in)	11	4	7	0	68.2	7	3	8	4	45.5
Which app did you install/update? / App Install)	11	7	4	0	81.8	9	6	5	2	68.2
Which app did you use? / App)	7	6	5	4	59.1	9	5	6	2	63.6
Which photo did you take?) / Photos	11	3	8	0	63.6	9	8	3	2	77.3

Table 3. True positives (TP), true negatives (TN), false positives (FP) , false negatives (FN) and accuracy (ACC; in %) of each question and the timespans yesterday, last week, taking into account answers by close adversaries.

	Acquainted Adversary									
	Yesterday					Last Week				
Question / Category	TP	TN	FP	FN	ACC	TP	TN	FP	FN	ACC
Who did you text? / SMS (out)	11	8	3	0	86.4	9	6	5	2	68.2
Who texted you? / SMS (in)	11	4	7	0	68.2	11	4	7	0	68.2
Who did you call? / Call (out)	11	8	3	0	86.2	8	6	5	3	63.6
Who called you? / Call (in)	11	8	3	0	86.4	7	6	5	4	59.1
Which app did you install/update? / App Install)	11	7	4	0	81.8	9	8	3	2	77.3
Which app did you use? / App)	7	4	7	4	50	9	6	5	2	68.2
Which photo did you take?) / Photos	11	4	7	0	68.2	9	8	3	2	77.3

Table 4. True positives (TP), true negatives (TN), false positives (FP), false negatives (FN) and accuracy (ACC, in %) of each question and the timespans yesterday, last week, taking into account answers by acquainted adversaries.

the highest accuracy was achieved for app install (*yesterday*), with over 80% (11 TP, 7 TN, 4 FP and 0 FN).

When considering the attacks by acquainted adversaries only, the highest accuracies are achieved for yesterday's SMS (out) and Call (in and out), with 86.4% (each with 11 TP, 8 TN, 3 FP and 0 FN). This is followed by App Install (yesterday) with 81.8% accuracy. This category also achieves high accuracy when combined with the timespan last week 77.3% (9 TP, 8 TN, 3 FP and 2 FN).

Best Combination

So far, we analyzed the accuracy of each individual question. However, in a fallback authentication system multiple security questions are combined to provide enhanced security. Thus, we calculated the accuracy values for all possible combinations of the 14 security questions. With n = 14, there are $\sum_{k=1}^{n} \binom{n}{k} = 16383$ possible combinations. The best combination in terms of accuracy was found for the questions: "Which app did you install yesterday?", "Who did you call yesterday?" and "Who did you text yesterday?" Table 5 gives an overview of the accuracy values for this combination. The threshold in the table depicts the number of correctly answered questions that are required to authenticate successfully. If only the attacks by close adversaries are considered, this combination can yield an accuracy of over 95%. All legitimate users were authenticated successfully (11 TP, 0 FN). Only one adversary got unauthorized access (1 FP, 10 TN). If we consider the same combination with only attacks provided by acquainted adversaries, an accuracy of 100% is achieved (11 TP, 11 TN, 0 FP, 0 FN).

Close Adversaries					
Threshold	TP	TN	FP	FN	ACC
1	11	2	9	0	59.1
2	11	4	7	0	68.2
3	11	10	1	0	95.5
Acquainted Adversaries					
Threshold	TP	TN	FP	FN	ACC
1	11	5	6	0	72.7
2	11	7	4	0	81.8
3	11	11	0	0	100

Table 5. True positives (TN), true negatives (TN), false positives (FP), false negatives (FN) and accuracy (ACC; in %) of the best combination of questions, taking into account attacks from close (top) and acquainted (bottom) adversaries, respectively. Threshold depicts the number of questions that must be answered correctly.

To achieve the reported accuracy values, users have to answer only three questions correctly (threshold = 3). There are also several other combinations with more questions that yield the same accuracy values. This is important to prevent random adversaries to succeed by chance. With only three questions (and four answer options), the chances for a random adversary is $(\frac{1}{4})^3 = 0,016$ (1.6% chance). In turn, using 7 questions will reduce the chances to 0.01% ($(\frac{1}{4})^7 = 0,000061$). For comparison: The chances of a 4-Digit PIN to be guessed by a random adversary are $\frac{1}{10000} = 0.0001$ (0.01%).

A combination with 7 questions and 95.5%/100% accuracy for close/acquainted adversaries consists of the questions "Which app did you install yesterday/last week?", "Which app did you use last week?", "Who did you call yesterday?", "Who called you last week?" and "Who texted you yesterday/last week".

DISCUSSION

Related Adversary Attack

The actual performance varied from participant to participant. There were participants who knew a lot about themselves, but we also had adversaries who were quite familiar with their usage patterns and interestingly, also made similar mistakes. This confirms our assumption that the biggest threats are persons who are very close to the user (e.g. jealous partners). We argue that such attacks are very likely [13] and thus have to be considered when designing a fallback authentication system.

Limited Device Usage

We also had participants who knew very little about themselves, but at the same time, were hard to attack. We assume that those users have very low smartphone usage with particular usage patterns (e.g. using special apps), which makes it more difficult to guess their answers. In a real-world deployment, it might be difficult to generate enough security questions for those users, since only few data is available. The same problem may arise for users who seldom install new apps.

Alternative solutions need to be provided in those cases. One might think of designing more questions within a category for which the user has enough data on the device or switch to another fallback solution. The latter approach is particularly important in case the device has been idle for some time (i.e. no usage data is available) in order to prevent adversaries to always select the option *none of them* as the correct answer.

Limiting Data Exposure

Dynamic security questions could be exploited to circumvent the primary authentication. Adversaries do not even have to get into the user's device to spy on information as some information can be leaked through dynamic questions based on usage patterns. That is why we analyzed different types of questions to gain insights on the acceptance of each question. We learned that certain questions are more accepted by participants than others.

For example, the pre-study showed that using photos was a bad idea. That is why we introduced blurred photos to evaluate whether this would reduce the privacy concerns that users have. However, the comments of the participants showed that still, they would prefer not answering photo questions at all. When designing systems with dynamic security questions, it is important to focus on data types that are personal, but that users do not mind if exposed to others. As shown by our results, different types of data on the phone have different levels of sensitivity.

Choosing Question Categories

The information categories that were related to active communication were the categories that our participants found easiest to answer. Since active communication is something that is initiated by the users (e.g. calling someone), it is probably easier to remember when being asked for it. However, these categories alone are not suitable for the design of dynamic security questions, since related adversaries are also good in answering those questions.

The categories about app usage and app installations seem to be the most promising ones. They had the best trade-off between usability and security. While participants were good in answering those questions, adversaries performed much worse for them.

In addition to this, the participants rated these categories among the safest, together with the category music. In comparison to the other categories, questions about app usage or music listening habits are less personal. Participants are probably not so reluctant to share which artist they listened to or which app they used. However, information with whom they texted or what photos they have on the device are much more sensitive. In particular photos can reveal information that better should be protected.

Though the category music was considered as safe by our participants, it is not suitable for the design of dynamic security questions due to memorability issues. Music is something one does on the go while doing other things. It is a passive activity, so that it is hard to actually remember which music one had listened to at a specific point in time. Furthermore, music is increasingly consumed through streaming services, where it is difficult to log the data.

Choosing Timespans

With respect to the timespans, the participants felt that questions about *yesterday* were easier to answer. Their actual performance for the timespan *yesterday* was also better than for other timespans. This is quite reasonable since activities that just happened a day ago, are more present in one's memory.

We got interesting insights when taking a closer look at the categories with respect to the included timespans. Some questions seem to work better with longer timespans, while others are easier to answer when they happened just recently.

For example, app installations are something that the user does consciously and thus, it will be more likely to remember the answer if it did not happen too long ago. On the other hand, using an app (if it is not newly installed) can be included in the users' routine (checking mails on the go, etc.). Though these kinds of things are done actively, they are part of an automated routine. Thus, it is easier for users to answer if the used an app during the week, instead of answering if they used it the day before.

Design for Special Events

Participants had difficulties remembering the correct answers when special events occurred (e.g. one's birthday). Two of them received too many text messages so that they could not remember who of the persons texted them. Unusual events should be taken into account when designing dynamic security questions. The identification of unusual events could be based on outlier detection. In case an event (e.g. incoming message) occurs more frequently than usual, the corresponding questions should not be asked to avoid memorability issues identified in the study. Another consideration in this case could be to invert the question (e.g. "Who didn't call you yesterday?"). However, this could cause additional social tension by highlighting that some expected event did not occur.

Adversaries and Security

Das et al. [3] showed with their theoretical model that naïve and observing adversaries performed worst in answering security questions. As examples for naïve and observing adversaries, they mentioned close family members. These are exactly the adversaries that we had in our study.

However, our results suggest that adversaries can perform quite differently. We had adversaries that performed well, but we also had adversaries who performed not as good. Altogether, the adversaries were always worse in answering the questions than the participants. How well an adversary can answer certain questions depended on how well they knew their victim. It also depended on the usage patterns of the user. In particular, participants that seldom used their smartphones were harder to attack.

Participants were better in answering the security questions in the main study. We achieved an increase of 13% of correct answers compared to the pre-study. The redesign of the security questions by removing unsuitable questions and adding promising categories seems to have positively influenced the results. The percentage of correct answers for adversaries did not change much (around 50%). Close adversaries were better than acquainted ones, which is reasonable, since close adversaries are supposed to know more than acquaintances.

Another threat by close adversaries is that they are more likely to be able to observe the behavior of the user (e.g. partner). Excluding the most popular answers from a question, is a first step to make observation more difficult, since adversaries are probably more likely to observe more regular patterns (e.g. someone using Facebook frequently). However, how easy it is for close adversaries to perform observation attacks needs to be addressed in future studies.

Real and Estimated Performance

In the first study, we found that participants overestimated their actual performance. This means that users are confident in their answers, making the overall authentication process more enjoyable and less frustrating. At the same time, being rejected even though one believes that the answers were correct might cause frustration too. Thus, it is desirable to provide a close to real experience.

Participants in the main study were very good about their estimation. Improving the security questions most likely removed error prone questions (e.g. music), reducing wrong estimates. Close adversaries and acquaintances had a good self-assessment about their performance. This could be an indicator that adversaries know quite well which answers they are able to answer and which they are not. This is a danger, in particular for close adversaries who would like to spy on information of their friend or partner. Exploiting the fact that they know which questions are missing, they might try to observe their "victim" in order to get the missing pieces.

Best Combination of Questions

The individual accuracy for each security question helped us to identify promising and problematic questions. However, the accuracy that can be achieved with one question is not sufficient due to the high number of FP and FN. The accuracy can be optimized by combining multiple questions. Using a combination of 7 questions increases security drastically. The chance for a random adversary to successfully authenticate is almost zero. Thus, our presented approach can yield a similar theoretical security level as a 4-digit PIN that is often used as primary authentication method on smartphones. This is important, since a fallback authentication mechanism needs to be at least as secure as the primary authentication. Also, the use of 7 questions seems appropriate in terms of time in the context of fallback authentication.

The best combinations contained questions about the user's communication history that were identified as privacy-intruding during the studies. However, this does not mean that one is not allowed to use these question categories. Instead, one should think of ways on how to use them, but preserve the user's privacy at the same time. For example, this could be removing the last name of the contact or even try to only give away the initials or the first three letters of a name.

Number of Attempts

The best combinations presented in this paper are very strict, meaning that users have to answer all questions correctly (within three questions) or are only allowed to make one mistake (within seven questions). While this makes it difficult for adversaries to attack, it also puts a high pressure on the user.

In case the user makes an error, one could think of an incremental approach, where with each error, more questions are asked (like in [19]). For this, the same combinations of questions can be used again, but with different answers (e.g. there may be more than one app that the user installed yesterday). However, it has to be taken into account that an incremental approach also comes with new risks (e.g. hinting at adversaries which questions have been answered incorrectly). Thus, it is important to set a limit on the incremental approach after which the user has to perform another fallback authentication (with another approach) or the authentication fails completely (similar to PUC).

LIMITATIONS AND IMPROVEMENTS

Though we collected a lot of data in our study, there was some information we could not access. For example, there are many different apps for text messaging like WhatsApp, Viber, etc. for which it was not possible to log text messaging information. This is suboptimal since those apps might have revealed further usage patterns and performances.

We had a high dropout rate for the main study, which was mostly due to the fact that participants found it hard to fulfill the requirements of bringing two people with specific properties with them for the lab study. Thus, a clear limitation of the studies is the number and age distributions of the participants. Participants were also rather tech-savvy which could have an influence on the higher app usage and thus, also limits the representativeness of our sample. We encourage additional studies with a larger and more diverse user population that take into account the different types of users, for example, in

a between-groups experiment (i.e. high vs. low device usage). The study of less familiar adversaries would also be an interesting aspect to be considered in future research.

CONCLUSION

In this paper, we presented an iterative design process for dynamic security questions that are based on available (usage) data of mobile devices. We identified suitable and unsuitable categories of information conducting two user studies and tested them under the worst possible circumstances with different types of human adversaries. We identified significant differences between users and adversaries when answering dynamic security questions.

Most importantly, we showed that the design of dynamic security questions is a challenging task that does not only involve security and usability aspects, but also privacy concerns. Thus, the most usable options (i.e. categories that the user can answer best) are not necessarily the best solutions. Instead, we identified app usage and app installation as promising categories, since they have the best trade-off between usability and security. Nonetheless, they work best in combination with respect to the user's communication history, yielding over 95% to 100% accuracy.

Dynamic security questions have been proposed in the past (e.g. [2]) and are actually used in the real world (e.g. [12]). However, their privacy implications have been rarely discussed. We hope that the insights gained in this paper will inspire further discussion and research in this area. In particular, if dynamic security question are used, how can we prevent the revelation of too much personal information when their design actually relies on the personal information?

ACKNOWLEDGMENTS

We would like to thank Stephan Thalhammer and Philipp Hauptmann for their help with conducting the user studies.

REFERENCES

1. Apple Support. ios: Forgotten passcode or device disabled after entering wrong passcode. `http://support.apple.com/kb/ht1212` (Accessed: 03/03/2014).
2. Babic, A., Xiong, H., Yao, D., and Iftode, L. Building robust authentication systems with activity-based personal questions. In *Proc. SafeConfig 2009*, ACM Press (2009), 19–24.
3. Das, S., Hayashi, E., and Hong, J. I. Exploring capturable everyday memory for autobiographical authentication. In *Proc. UbiComp 2013*, ACM Press (2013), 211–220.
4. Furnell, S. An assessment of website password practices. *Computers & Security 26*, 7-8 (2007), 445 – 451.
5. Garfinkel, S. L. Email-based identification and authentication: An alternative to pki? *IEEE Security & Privacy 1*, 6 (2003), 20–26.
6. Griffith, V., and Jakobsson, M. Messin' with texas deriving mother's maiden names using public records. In *Proc. ACNS 2014*. Springer (2005), 91–103.
7. Gupta, P., Gottipati, S., Jiang, J., and Gao, D. Your love is public now: Questioning the use of personal information in authentication. In *Proc. ASIA CCS 2013*, ACM Press (2013), 49–60.
8. Haga, W. J., and Zviran, M. Question-and-answer passwords: An empirical evaluation. *Information Systems 16*, 3 (1991), 335 – 343.
9. Jakobsson, M., Stolterman, E., Wetzel, S., and Yang, L. Love and authentication. In *Proc. CHI 2008*, ACM Press (2008), 197–200.
10. Just, M. Designing and evaluating challenge-question systems. *IEEE Security & Privacy 2*, 5 (2004), 32–39.
11. Just, M., and Aspinall, D. Personal choice and challenge questions: A security and usability assessment. In *Proc. SOUPS 2009*, ACM Press (2009), 8:1–8:11.
12. Microsoft. Microsoft live mail. `https://login.live.com/` (Accessed: 04/08/2014).
13. Muslukhov, I., Boshmaf, Y., Kuo, C., Lester, J., and Beznosov, K. Know your enemy: The risk of unauthorized access in smartphones by insiders. In *Proc. MobileHCI 2013*, ACM Press (2013), 271–280.
14. Polakis, I., Lancini, M., Kontaxis, G., Maggi, F., Ioannidis, S., Keromytis, A. D., and Zanero, S. All your face are belong to us: Breaking facebook's social authentication. In *Proc. ACSAC 2012*, ACM Press (2012), 399–408.
15. Rabkin, A. Personal knowledge questions for fallback authentication: Security questions in the era of facebook. In *Proc. SOUPS 2008*, ACM Press (New York, NY, USA, 2008), 13–23.
16. Rysgaard, B. A method for protecting user data stored in memory of a mobile communication device, particularly a mobile phone, 2001. *European Patent No. EP 1107627* (2001).
17. Schechter, S., Brush, A. J. B., and Egelman, S. It's no secret: Measuring the security and reliability of authentication via 'secret' questions. In *Proc. SOUPS 2009*, ACM Press (2009), 40:1–40:1.
18. Schechter, S., Egelman, S., and Reeder, R. W. It's not what you know, but who you know: A social approach to last-resort authentication. In *Proc.CHI 2009*, ACM Press (2009), 1983–1992.
19. Schechter, S., and Reeder, R. W. 1 + 1 = you: Measuring the comprehensibility of metaphors for configuring backup authentication. In *Proc. Soups 2009*, ACM Press (2009), 9:1–9:31.
20. Statista: The Statistics Portal. Top 20 smartphone-apps in deutschland 2012. `http://de.statista.com/statistik/daten/studie/239434/umfrage/nutzeranteile-der-top-20-smartphone-apps-in-deutschland` (Accessed: 03/03/2014).