Evaluating Attack and Defense Strategies for Smartphone PIN Shoulder Surfing

Hassan Khan, Urs Hengartner, Daniel Vogel
Cheriton School of Computer Science
University of Waterloo
{h37khan, urs.hengartner, dvogel}@uwaterloo.ca

ABSTRACT

We evaluate the efficacy of shoulder surfing defenses for PIN-based authentication systems. We find tilting the device away from the observer, a widely adopted defense strategy, provides limited protection. We also evaluate a recently proposed defense incorporating an "invisible pressure component" into PIN entry. Contrary to earlier claims, our results show this provides little defense against malicious insider attacks. Observations during the study uncover successful attacker strategies for reconstructing a victim's PIN when faced with a tilt defense. Our evaluations identify common misconceptions regarding shoulder surfing defenses, and highlight the need to educate users on how to safeguard their credentials from these attacks.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—Input devices and strategies; K.6.5 Computing Milieux: Security and Protection—Authentication

Author Keywords

Shoulder Surfing, Authentication, Mobile Devices

INTRODUCTION

Traditional smartphone unlocking mechanisms rely on something that the user knows such as PINs, passwords, or Android's pattern lock. These are widely available and widely used [11, 12]. Despite the increasing availability of biometric options like fingerprint and facial recognition, traditional mechanisms remain the fallback authentication in case the biometric approach fails. Shoulder surfing attacks on traditional methods are particularly devastating since the attacker does not require any special equipment or skill. Therefore, significant effort has focused on securing traditional authentication mechanisms against such attacks [1, 5, 6, 16, 21, 29].

Although several defenses had been proposed, evidence supporting the prevalence of shoulder surfing in the wild was missing until recently [7, 12]. Harbach et al. [12] conducted an online survey (n=260), and a field study (n=52), to show that 35% of their participants were concerned that someone may observe them during smartphone unlocking and steal their credentials. During the field study, participants indicated that shoulder surfing was a possibility for 17% of the sampled device usage sessions. When shoulder surfing was possible in non-public settings, the potential attacker was most often a malicious insider, like a co-worker or a family member. Furthermore, the most widely used shoulder surfing defense reported was tilting the device screen away from the observer (28% of the participants). Other research found that the choice of this defense was motivated by users' inclination to react in a subtle way when the observer was an acquaintance [7, 19].

In addition to user-initiated defense strategies, several shoulder surfing resistant variations of traditional authentication mechanisms have been proposed by researchers [5, 6, 22]. Most require significant changes to the contemporary PIN entry interface [22] or add new out-of-band communication [6]. One recent promising approach requires no changes to the PIN user interface, but leverages capacitive touchscreens that sense pressure [1, 16]. Here, a user can explicitly apply more pressure to a subset of PIN key touches as a part of their secret. Lab experiments demonstrate that this "invisible pressure component" [16] prevents shoulder surfing [16].

We subject the defense strategies of tilting the screen to enter a 4-digit PIN ("PIN") and entering pressure-sensitive PINs ("ForcePINs") to shoulder surfing attacks. We record videos of 30 victims entering a PIN and a ForcePIN from two unobstructed views (top and side of the device) and a side view where the device screen is tilted away from the camera. We recruit 30 attackers to mount over 1,000 shoulder surfing attacks by watching videos of victims. We also elicit feedback from attackers on the strategies they used against both defenses.

Our results show that tilting the device screen provides limited protection. For 45% of attacks on this tilting defense, the attackers were able to correctly guess the complete PIN by observing their victim authenticate an average of three times. For another 50% of the attacks, the attackers were able to partially guess the PIN digits. In terms of attacker strategies, attackers who paid attention to the pattern of relative finger movement (i.e., movement in direction and distance relative to the previous tap) performed significantly better than the attackers who guessed only based on the current position of the finger and the layout of the keypad.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

© 2018 ACM. ISBN 978-1-4503-5620-6/18/04…$15.00
DOI: https://doi.org/10.1145/3173574.3173738

ForcePINs failed to provide any advantage over PINs against shoulder surfing attacks. All attackers identified and exploited the timing side channel in ForcePINs, where entering digits with high pressure took longer than entering digits with normal pressure. Consequently, for 94% of the attacks, the attackers were able to completely guess the PIN along with the pressure digits. In all but two attacks on ForcePIN, all pressure digits were guessed correctly.

The main contributions of this paper are:

- Examination of effects of different viewing angles and viewing distances on shoulder surfing.
- Identification of a common misconception, namely that tilting the device screen away from an observer protects the PIN.
- Evidence that the "invisible pressure component" assumption [16] does not hold for ForcePINs due to the presence of a timing side channel.
- Identification of (successful) strategies adopted by attackers to circumvent shoulder surfing defenses.

RELATED WORK

We discuss only literature related to shoulder surfing attacks on smartphones.

Shoulder Surfing in the Wild & Users' Countermeasures

Only a few studies have empirically evaluated the susceptibility of knowledge-based authentication mechanisms to shoulder surfing attacks. Schaub et al. [20] conducted lab experiments to investigate the susceptibility of password entry on different smartphone platforms to shoulder surfing attacks. They demonstrated significant differences for shoulder surfing success for different keyboards and showed that keyboards with low usability were less susceptible. Von Zezschwitz et al. [23] evaluated the shoulder surfing susceptibility of Android's pattern lock and demonstrated that it is highly influenced by pattern length, line visibility, number of overlaps and number of intersections. Ye et al. [28] developed a smartphone app that can be used by an attacker to record and crack (using computer vision techniques) a user drawing Android's pattern lock in five attempts or less. In their ongoing research, Davin et al. [3] outlined a method to measure and compare the efficacy of shoulder surfing attacks on smartphone authentication mechanisms. However, they offered no conclusive suggestions from their preliminary findings and indicated that extensive experiments are a part of their ongoing work.

Researchers have also measured users' perceptions of the threat posed by shoulder surfing and in-the-wild shoulder surfing experiences. Harbach et al. [12] conducted an online survey and a field study with 260 and 52 participants, respectively, to understand smartphone unlocking behaviour and the threat posed by shoulder surfing. Their online survey indicated that 35% of users were concerned that someone may observe their secret when they were entering it. During the field study, through in-situ feedbacks, users indicated that shoulder surfing was a possibility for 17% of their device usage sessions. The top three defenses used by the participants against shoulder surfing included: tilt screen away (27.7% of the participants), wait a moment (16.2%), and turn around (11.2%).

Eiband et al. [7] surveyed 174 participants to elicit stories about shoulder surfing incidents (not restricted to authentication) from both victims and observers. 48% of the respondents admitted to shoulder surfing and 33% of the respondents mentioned catching someone in the act. Victims included strangers (74% of the victims), acquaintances (20%) and other (6%). In 6% of the incidents, authentication data was shoulder surfed. In terms of the top three defenses, respondents reported: turn display/body away (43% of the respondents), put device down (13%), and turn device off (13%).

The efficacy of the most widely adopted defense (tilting the device screen) has not been established. Therefore, it is important to understand the efficacy of tilting as a defense.

Defenses Against Shoulder Surfing Attacks

Several shoulder surfing resistant solutions have been proposed for knowledge-based authentication systems. Harbach et al. [12] broadly classified them into four categories: (i) indirect input systems; (ii) additional layer of implicit biometric systems; (iii) input obfuscation systems; and (iv) non-observable channel systems. We provide a brief overview and canonical examples for each system.

Indirect input systems display a challenge on one interface and require a response on another interface (such as back-of-device [5, 6]) or device (such as Google Glass [26]). The requirement of an additional interface or device is an inherent limitation of these systems.

Researchers have proposed systems that provide an additional layer of defense by also considering the touch input behaviour of the device user during authentication [4, 14, 21]. However, in earlier work, we demonstrated that touch input behaviour can be mimicked through shoulder surfing attacks [15].

Input obfuscation systems transform the secret entry interface to prevent shoulder surfing [10, 17, 22, 27, 29]. However, most proposals are complex and reduce usability [12]. For instance, SwiPIN [22] transforms the interface by presenting PIN digits in red and yellow coloured layouts. It also presents the PIN entry fields in the same coloured fields. The user inputs digits through the swiping or tapping gesture in the entry field that corresponds to the layout colour of their digit. After each digit input, the layouts are randomized again. Wiese and Roth [25] demonstrated that the increased complexity of SwiPIN (and similar schemes) can be broken through 6–11 observations by a human observer followed by computer-based simulations.

Systems that establish non-observable channels include approaches that use the pressure dimension [1, 16, 18] and tactile feedback [2]. The former approaches exploit the observation that smartphone touchscreens are capable of distinguishing different pressure levels. Krombholz et al. [16] proposed a scheme where a user adds explicit pressure to a subset of the entered digits. Introducing a binary state for pressure (normal/high) increases the password space of a 4-digit PIN from 10^4 to 20^4. They conducted two experiments to demonstrate that their scheme is resilient to shoulder surfing attacks. In

the first experiment, one of the authors acted as a shoulder surfer. In the second experiment, one of the authors entered ForcePINs collected from the participants on camera and two unrelated observers mounted 50 attacks in total. The authors acknowledge the limited nature of their evaluation. Arif et al. [1] proposed a similar scheme and argued that: *"...in theory it should be more difficult to guess the amount of pressure applied on a key just by observing the user"*.

ForcePINs are the most promising defense since they: (i) do not change the PIN interface so they are easy to deploy and understand; (ii) require only a small change to how PINs are entered so they are easy to use; and (iii) do not requite any new hardware [16]. However, a rigorous evaluation of shoulder surfing attacks on ForcePINs has not been performed.

STUDY

Threat Model

The attacker is in close vicinity of the victim during authentication and may be a stranger or a malicious insider. The former may have observed the user entering their PIN only once in a public place (e.g., a coffee shop) while the latter (a spouse or a co-worker) may have repeatedly observed the user. An attacker may also use their smartphone to secretly record the user authenticating thereby enabling them to repeatedly observe the victim. Other influential factors include the viewing angle (e.g., from the side vs. the top when the victim is sitting and the adversary is passing by) and the distance from the victim. Attackers may gain access to the victim's device when it gets lost or stolen or is left unattended. The device may limit incorrect attempts, thereby preventing bruteforcing.

The tilting defense requires that the victim chooses a tilt angle such that the attacker cannot see the device screen or the location of the finger tip used for PIN entry on the device screen. However, in practice, the tilt angle is subjective, and different users may tilt the screen of their devices differently. In order to objectively measure the efficacy of tilting, we attempt to capture the natural behaviour of users when they are tilting their device away from the observer.

For the ForcePIN defense, we use the same threat model as Krombholz et al. The adversary is able to *"clearly observe all sensitive information"* and *"observe the typing behaviour"*. Therefore, we assume that the adversary can clearly observe the ForcePIN that is entered and is also able to watch the ForcePIN entry behaviour. Since device tilting is a common defense strategy adopted by users, we also investigate the scenario where the tilting and ForcePIN defense strategies are used in combination.

Design

We conduct a lab study to investigate the efficacy of the two defense mechanisms. We consider four conditions: (i) PIN: no defense during standard PIN entry; (ii) Tilt: tilting the phone to hide PIN entry from attackers; (iii) ForcePIN: selectively adding pressure during PIN entry to a subset of digits; and (iv) ForcePIN-Tilt: a combination of ForcePIN and Tilt. The no defense baseline is introduced to compare the attackers' success for the two defenses against an unsuspecting victim who is entering a standard PIN.

Independent variables: We treat the viewing angle of the attacker as an independent variable. We investigate two viewing angles: a top view and a side view (Figure 1). The viewing angle is of interest for PIN and ForcePIN. Since Tilt requires tilting the device screen away from the observer, testing from multiple observer angles is not meaningful. A related aspect for tilt is the tilt angle and we investigate it in the "Discussion" section.

Dependent variables: The dependent variables include: (i) attacker success rate; (ii) number of observations required; and (iii) number of incorrect guesses. The attackers' success is measured in terms of correct (all four digits) or partially correct guesses (between 1–3 digits in their correct order). For ForcePIN, the correct guess of the pressure digit(s) is also important and we report it separately to quantify the protection offered by this feature. The other variables, number of observations required and the number of incorrect guesses, capture the effort required by an adversary to bypass a defense.

DATA COLLECTION

To conduct attacks under the investigated variables, we required videos of users entering PINs, entering ForcePINs, and entering PINs or ForcePINs while tilting their device screen away from the observer. We now describe the data collection setup and descriptive statistics of the gathered data. We received approval from our university's IRB for all experiments involving human participants.

Apparatus

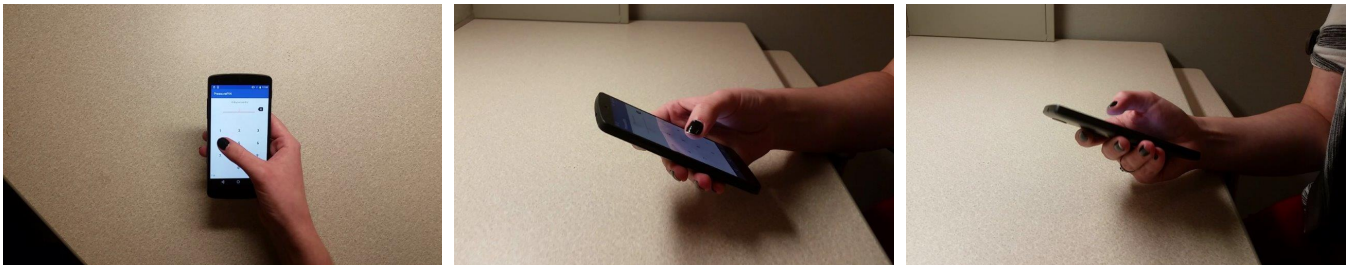
We developed an Android app to configure and enter PINs and ForcePINs. Our app presents a PIN keypad to users that is identical (in terms of layout and key sizes) to the standard PIN keypad of Android 6.0. In addition to recording PIN entry time, every time a user enters a digit, our app logs the raw pressure value (returned between 0 and 1 by Android API's `MotionEvent.getPressure()`) and the keyhold interval. All experiments were conducted on a Nexus 5 device.

For ForcePINs, we performed a pilot experiment (n = 3) to determine the threshold for the high pressure level. Subjects were asked to apply normal and high pressure on the screen surface when entering PINs. Our experiment showed that subjects reliably produced a high pressure level over the value of 0.7 (i.e., no normal pressure input accidentally exceeded this value). We, therefore, use this value as the threshold.

While Arif et al. do not use haptic feedback for input with high pressure, Krombholz et al. use vibration-based feedback. During the pilot study, we found that haptic feedback was helpful as it enabled users to reliably register high pressure input. Therefore, we used vibration-based haptic feedback. To ensure that haptic feedback was not used as a side channel by the attackers, we muted the audio in the footage.

Shoulder Surfing Videos

For our experiments, we need authentication videos of users against four conditions—PIN, ForcePIN, Tilt, and ForcePIN-Tilt. For PIN and ForcePIN, we captured videos from two viewing angles: a top view and a side view. Each angle provided an unobstructed view of the participant's finger on the



(a) View from top (b) View from side (c) View with device tilted

Figure 1. The shoulder surfing views evaluated in this work (cropped images).

touchscreen and the numeric pad (see Figure 1). For Tilt and ForcePIN-Tilt, users were asked to tilt their device away from the camera as a defense against shoulder surfing. We did not explicitly tell the participants how much to tilt the device. This effectively captured the subjective nature of the tilt angle.

All videos were shot in 1080p format (1920x1080 pixels) with a frame rate of 30 FPS. The camera was held between 30–45 centimeters away from the smartphone such that the smartphone occupied 5–10% of the video frame. We assume an "over-the-shoulder" attack scenario: the adversary was seated 60 centimeters apart from the screen displaying a victim's device in its real-life size.

Collection Procedure

We recruited participants through local advertisement websites (Craigslist and Kijiji), our university's mailing list and word-of-mouth advertising. The recruited participants were first asked to configure a 4-digit PIN and then enter it ten times. They were asked to enter it three more times to facilitate recording of videos in the aforementioned angles. Videos were shot in the following order: the side view, the top view, and the side view with the device screen tilted.

Participants were then introduced to ForcePINs. They were not informed about ForcePINs earlier to avoid any change in their input behaviour. Participants repeated the aforementioned steps for ForcePINs, where they were instructed to configure a ForcePIN with at least one pressure digit. Participants were permitted to reuse the PIN that they chose for the previous task. Before entering their ForcePIN ten times, participants were given the opportunity to get familiarized with it by entering it as many times as they wanted on an interface that provided visual feedback (pressure digits in bold typeface).

Data Statistics

The data collection part of the experiment was completed by 30 participants. 60% of the participants were male. 56% of the participants were between the ages of 18–25 years, and the rest were between 26–30 years old. All participants were graduate students. In terms of authentication preferences on their smartphones, 33% used PIN, 33% used Android's pattern lock, and the rest used fingerprint. For PIN and ForcePIN, we shot 60 videos each (30 for each of the top and side views). For Tilt and ForcePIN-Tilt, we shot 30 videos each from the side view. The tilted videos of two participants were excluded as their screen was visible (tilt angles $< 10°$). Raw data against 300 PIN and ForcePIN entries was logged.

The majority of participants (19/30) chose a ForcePIN where only one digit was entered with high pressure. From now on we call such a digit a "pressure digit". In terms of the distribution of the pressure digit among the four digits, we observe that eight, six, five and two participants chose the first, second, third and fourth digit, respectively. Eight participants chose a ForcePIN with two pressure digits while only three participants chose three pressure digits. Finally, eleven participants chose the same sequence of digits for their PIN and ForcePIN.

SHOULDER SURFING ATTACKS

Experiment Protocol

We recruited 30 participants using the same method as in the data collection part of the experiment. 53% of the participants were male. 60% of the participants were between the ages of 18–25 years, 27% were between 26–30 years and the rest were over 30 years old. 87% of the participants were graduate students and the rest were professionals. In terms of authentication preferences on their smartphones, 47% used PIN, 30% used Android's pattern lock, and the rest used fingerprint. A subset of 22 participants had also participated in the data collection. (Participants never attacked their own videos.) Participants who had not participated in the data collection were first asked to perform the same steps as the other participants during data collection (i.e., configuring and entering a PIN and ForcePIN) but their video was not recorded. This step was taken to ensure that all attackers were primed in the same way with the knowledge of defenses.

Participants were informed that they would observe shoulder surfing footage to identify PIN digits (in correct order) and pressure digits for the ForcePIN condition. Each participant attacked 24 unique victims against the following conditions: (i) PIN: four PIN entries each from top and side views; (ii) ForcePIN: four ForcePIN entries each from top and side views; (iii) Tilt: four PIN entries with tilting defense; and (iv) ForcePIN-Tilt: four ForcePIN entries with the tilting defense. These conditions were counterbalanced across participants.

For each video, the researcher played it once and then asked participants whether they would like to guess the PIN (or ForcePIN) by telling it to the researcher or watch the video again. Participants were allowed to watch a video as many times as they liked. If they guessed an incorrect response, they

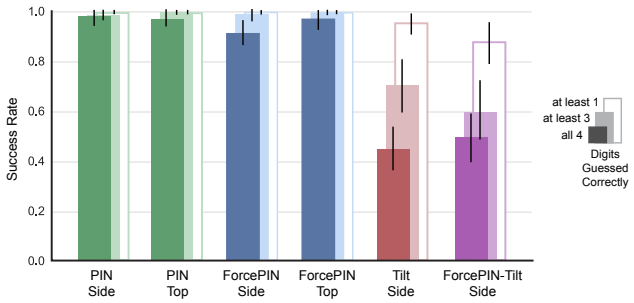
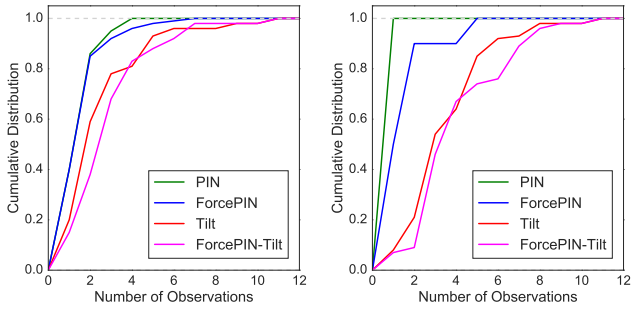


Figure 2. Success rates by condition and viewing angle (95% CI error bars).



(a) Correct guesses. (b) Partially correct guesses.

Figure 3. Number of observations required for correct or partially correct guesses.

were only told that the response was incorrect. They were not told if it was partially correct or whether the mistake was made for the pressure digit(s). Participants were given the opportunity to make up to three guesses before continuing to the next victim. Participants were given pen and paper to take notes and a smartphone that they could use to recall the keypad layout during the experiment. Once participants completed the shoulder surfing task, they were asked about their strategy for guessing the secret. Participants were paid $10.

Results

We now report how successful attackers were at guessing the victim's secret and the amount of effort required. For test statistics, we use a paired t-test when comparing between two conditions (e.g., side view vs. top view for PIN) and a one-way ANOVA when comparing more than three conditions (e.g., PIN vs. ForcePIN vs. Tilt vs. ForcePIN-Tilt). Post hoc comparisons using multiple pairwise t-tests are used only if the ANOVA test is significant. For data that is not normally distributed, we use a Kruskal-Wallis test and perform post hoc comparisons using multiple pairwise Mann-Whitney rank tests. In all cases, a $p < 0.05$ critical value is used for statistical significance. For multiple comparisons of the same data category, we apply Bonferroni correction to p-values (and set the significance cut-off at α/n, where n is the number of multiple comparisons [13]).

Attacker success

A correct guess is when the attacker identifies the digit (and pressure setting for ForcePIN) in the correct sequence position. We report the success rate for completely correct guesses. We also report the success rate for partially correct guesses, where one or more digits (including pressure setting for ForcePIN) are correct. Success rates for partially correct guesses suggest a trend and a vulnerability. Guessing one or more digits suggests an attacker could make a completely correct guess with more observations. Guessing all but one digit means the attacker could mount a bruteforce attack to find the remaining digit. Recall that attackers were allowed up to three guesses; we report the guess closest to the victim's secret.

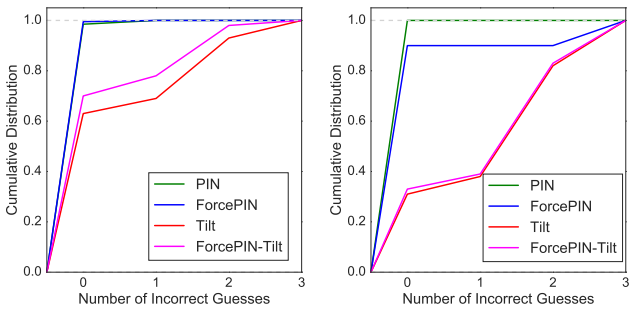
In Figure 2, we report the attack success rate by defense method (including PIN, which is no defense) and different viewing angles when applicable. Attackers were able to successfully guess the complete secret for PIN and ForcePIN in over 97% of the attacks, and they were able to partially guess the secret for the remaining attacks. An alarming result is that for ForcePIN, in all but two attacks, all pressure digits were guessed correctly. For Tilt, attackers were able to correctly determine the PIN for 43% of the attacks. For another 23%, attackers guessed three digits, and for another 22%, they guessed one or two digits correctly. For ForcePIN-Tilt, the success rates were comparable to Tilt.

We also measure the effect of the viewing angle on the correct guess success rate for PIN and ForcePIN. A t-test comparing side and top views for PIN indicates no statistically significant difference *($t = 1.0$, $p = 0.64$)*. Similarly, a t-test comparing side and top views for ForcePIN indicates no statistically significant difference *($t = 2.04$, $p = 0.10$)*. For the remaining results, we focus on the side view since it is common to all defense conditions including Tilt.

Of particular interest is whether ForcePIN, Tilt, or the combined ForcePIN-Tilt made a significant reduction in attacker success rates. A one way ANOVA indicates a significant effect of defense type on attackers' correct guess success rate ($F_{3,30} = 63.15, p < 0.001$). Post hoc comparisons using the Tukey HSD test showed no significant difference between PIN (M=0.98; SD=0.06) and ForcePIN (M=0.91; SD=0.12), and between Tilt (M=0.45; SD=0.23) and ForcePIN-Tilt (M=0.5; SD=0.27) defenses (all $p > 0.07$). There were significant differences between the group including PIN and ForcePIN and the group including Tilt and ForcePIN-Tilt (all $p < 0.001$). Our results show that tilting provides some increased protection compared to no defense, but ForcePIN does not.

Attacker effort

We use two metrics to capture the amount of effort an attacker required for correct or partially correct guesses. First, we report the number of observations an attacker required for a correct or partially correct guess. Second, we report the number of incorrect guesses an attacker made during the shoulder surfing attack. This is critical since a victim's device may lock out the attacker after a certain number of incorrect authentication attempts. Note that our setup is similar to the real-world scenario in that the attacker is only provided accept or reject feedback and no other information about their guess. Since, in the real world, attackers may make multiple observations



(a) Correct guesses. (b) Partially correct guesses.

Figure 4. Number of incorrect guesses before a correct or partially correct guess is made.

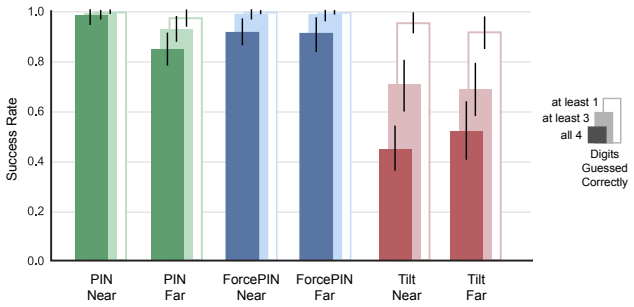


Figure 5. Success rates against different condition from near (∼60 centimeters) and far (∼5.5 meters) distances (95% CI error bars).

before gaining access to the victim's device, these metrics effectively capture the attacker effort.

Figure 3 shows the number of observations required for a correct (Figure 3(a)) and partially correct (Figure 3(b)) guess for the PIN, ForcePIN, Tilt and ForcePIN-Tilt conditions. Figure 3(a) shows that 40% and about 85% of participants required one or at most two observations only, respectively, to correctly guess the secret for both PIN and ForcePIN. Furthermore, over 90% of the attackers required at most three observations to correctly guess the secret for PIN and ForcePIN. For Tilt and ForcePIN-Tilt, 20% and 15% of the participants, respectively, required one observation to make a correct guess. At most four observations were required by 80% of the attackers to make a correct guess for Tilt and ForcePIN-Tilt.

Figure 3(b) shows the number of observations required for partially correct guesses. Since most attacks on PIN resulted in correct guesses, we do not have enough data to report against it. For partially correct guesses on ForcePIN, 50% of attackers required only one observation while 90% required at most two observations. For Tilt and ForcePIN-Tilt, only 8% and 7% of the participants, respectively, required only one observation to make a partially correct guess. At most five or at most seven observations for Tilt and ForcePIN-Tilt, respectively, were required to make a partially correct guess. While Tilt requires more observations, given that an average user unlocks their smartphone 47 times a day [12], it would likely be easy for an insider to repeatedly observe a victim.

Figure 4 shows the number of incorrect guesses made before a correct (Figure 4(a)) and a partially correct (Figure 4(b)) guess is made. Figure 4(a) shows that for both PIN and ForcePIN only two attackers made an incorrect guess before correctly guessing the secret. For Tilt and ForcePIN-Tilt, 63% and 70% of attackers required one incorrect guess before making a correct guess, respectively. Only at most two incorrect guesses were made by over 90% of the attackers before making a correct guess for Tilt and ForcePIN-Tilt.

Figure 4(b) shows the number of incorrect guesses made before a partially correct guess. Since most attacks on PIN and ForcePIN resulted in correct guesses, we do not have enough data to report against this metric for them. For Tilt and ForcePIN-Tilt, only 31% and 33% of the participants, respectively, made one incorrect guess before making a partially correct guess. Over 80% of the attackers made at most two incorrect guesses for Tilt and ForcePIN-Tilt, respectively, to make a partially correct guess. These results show that while tilting is relatively more resilient, ForcePIN fails to provide any advantage over PIN against shoulder surfing attacks.

Effect of the Viewing Distance on Attackers' Success

The results reported above are based on an "over-the-shoulder" attack scenario, where the adversary was 60 centimeters from the victim's device. This is obviously a desirable scenario for the attacker, but it is also easily achievable in real settings [7] and has been the focus of previous studies [16]. Regardless, it is informative to see if attacker efficacy degrades when observing the victim from a distance, such as across a room.

To examine this, we conducted a small experiment with ten participants attacking ten victims against each of the PIN, ForcePIN and Tilt conditions from a distance. We did not evaluate the ForcePIN-Tilt condition because our previous results show that the attackers can defeat Tilt in most cases, so ForcePIN-Tilt unlikely offers more protection when tilted. The protocol was exactly the same as our main experiment, but this time the attacker viewed videos on a television placed 5.5 meters away (the average length of a living room in an average new house [8]). The video player frame size was adjusted so the size of the smartphone matched its real-life size. This simulating enabled us to re-use videos from the main experiment and we were concerned that shooting videos from across the room would make protecting the anonymity of participants difficult.

In Figure 5, we show the success rate of attackers against PIN, ForcePIN and Tilt when the victim is near and far. All near attack data is from the main 30-participant experiment. For the far distance, the average success rate for correct guesses is 86%, 91% and 52% for PIN, ForcePIN and Tilt, respectively. Furthermore, for 94%, 99% and 70% of the attacks for the far distance, attackers were able to correctly guess at least three digits (including the pressure digit for ForcePIN) for PIN, ForcePIN and Tilt conditions, respectively.

We compare the correct guess success rate of attackers between near and far cases using independent samples t-tests. There is a significant difference between near and far cases for PIN *($t = 5.01$, $p < 0.001$)*. The success rate for PIN is 12% lower when attacked from far away. This matches our intuition; it should

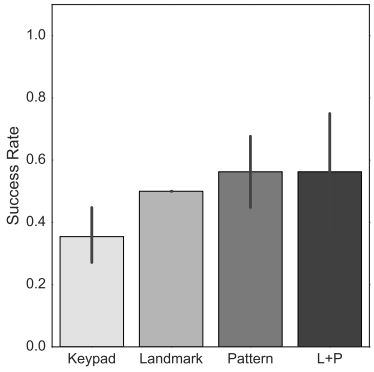


Figure 6. Success rates against different attack strategies (95% CI error bars; L+P = Landmark+Pattern).

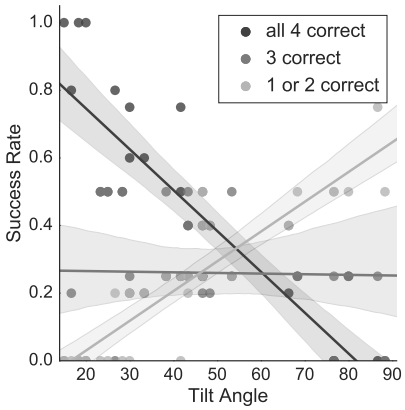


Figure 7. Full and partial success rates for different tilt angles.

be more difficult to shoulder surf from far away. Surprisingly, there are no significant differences between near and far case for ForcePIN or Tilt.

The average ForcePIN success rates in our measured data are very similar, including distributions. Without additional experiments, we can only speculate that the act of providing pressure digits is somehow compensating for any loss of efficacy at increased viewing distance. For Tilt, we believe the reason is that attacks on tilting are already difficult given the low success and high variance, so a small increase in difficulty when attacking at a distance may not be detected.

We also compare the number of observations required to make a correct guess between near and far cases using independent samples t-tests. A t-test between near (M=1.99; SD=1.2) and far (M=7.85; SD=4.08) cases for Tilt indicates significant differences *($t = 12.41$, $p < 0.001$)*. No significant differences were found between near and far cases for PIN and ForcePIN (all $p > 0.45$).

DISCUSSION

We note that a surprisingly large number of attacks against Tilt were successful. Furthermore, since we allowed the attackers to make at most three guesses, the tilting defense may not be as effective against determined insiders using repeated observations combined with bruteforcing or for strangers recording the victim authenticating. Another influencing factor is the coping strategy of users against insiders. Users may adjust their strategy depending on their relationship with the observer, for example avoiding using the device instead of tilting to avoid signs of mistrust [7].

In this section, we first investigate different strategies adopted by attackers to bypass Tilt and the related secondary factor of tilt angle. We then report attackers' strategies against ForcePIN and the effect of attackers' viewing distance. We ground our discussion in additional analysis of our attack data.

Attackers' Strategies Against Tilt Defense

We asked attackers what they watched for to guess the secret digits for the Tilt defense. For qualitative analysis of their responses, three researchers coded all participant responses using the grounded theory approach [9]. An inter-rater agreement between the three researchers was almost perfect (Fleiss's $\kappa = 87\%$). The discrepancies were resolved by picking the majority code (there were no ties). The strategies of attackers were categorized under the following labels:

Keypad: Attacker watches for finger intersection with keys in the keypad. For example, *"where the thumb was... could be within a set of keys"*.
Landmark: Attacker watches the absolute position in a coordinate frame defined by fixed element on the phone or user interface. For example, *"how their thumb was moving from the start point... for start point how far from sides [corners] and enter key"*
Pattern: Attacker watches the pattern of relative finger movement in terms of direction and distance from the last tap location. For example, *"movements of their thumb and how far apart or which direction they were from the last one"*

Figure 6 shows the correct guess success rate against each strategy. The Keypad strategy was adopted by twelve attackers who had an average success rate of 36%. The Landmark strategy was used only by two attackers and they achieved a success rate of 50%. Twelve participants used the Pattern strategy to achieve a success rate of 55%. Four participants used a combination of Landmark and Pattern strategies. Their responses included: *"relative distance between keys... corners of the screen and enter key location was helpful"*. These participants achieved a success rate of 55%.

A one way ANOVA indicates a significant effect of strategy on attacker success ($F_{3,25} = 4.04, p = 0.01$). Post hoc comparisons using the Tukey HSD test showed only a significant difference between Keypad (M=2.83; SD=1.33) and Pattern (M=4.81; SD=1.40) strategies. These results show that the attackers who adopt a strategy beyond a simple visual hit test on a target have a better chance of success.

Effect of Tilt Angle on Attackers' Success

Recall that we placed no restriction on how victims adjust the tilt angle of their device for the Tilt condition. This enabled us to capture the natural behaviour of participants, but it also enables us to investigate the influence of tilt angle based on these natural variances. To measure tilt angle, three people independently viewed all 28 videos (two videos where the

screen was visible were excluded) used for the Tilt condition and estimated the tilt angle in 5° intervals for each. Estimated tilt angles for each video differed by 10° on average (SD=7°), 11 videos differed by 5° or less, 12 videos differed by 10° or 15°, 5 videos differed by 20° or 25°. For each video, we use the mean of the three estimated angles given by the raters. The overall mean estimated tilt angle was 45° (SD=22°) with a minimum angle of 15° and maximum angle of 88°.

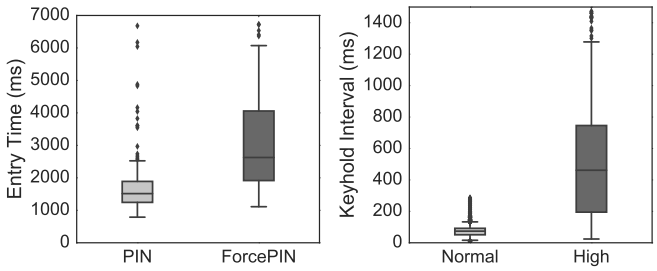
Figure 6.1 shows the correlation between success rates for guesses and tilt angle in victim video. An increase in the tilt angle results in a corresponding decrease in the correct guess success rate, and a corresponding increase in the success rate for partial guesses of 1 or 2 digits. The success rate for partial guesses of 3 digits remains about the same. This balanced trade-off follows from the relationship between completely correct and partially correct guesses. As expected, these trends suggest that making a correct guess becomes increasingly difficult for higher tilt angles.

Figure 6.1 also shows that no attacker was able to correctly guess the PIN when a tilt angle of 70° or higher was used. However, attackers were still able to partially guess the PIN when a tilt angle close to 90° was used. This was possible due to the Landmark+Pattern attack strategy, where some attackers paid attention to the thumb movements near the corner of the device screen to determine whether they pressed digits in the first column (i.e., 1 or 4 or 7).

We calculated linear regressions to see if success rates can be predicted using tilt angle. The linear relationship with correct guess success rate is significant ($F_{1,26}$=78.62, $p < 0.0001$, $R^2 = .75$) and can be predicted using: $0.9881 - 0.01208 * tilt\_angle$. The linear relationship with partially correct guesses of 1 or 2 digits is also significant ($F_{1,26}$=106.3, $p < 0.0001$, $R^2 = .80$) and can be predicted using: $0.1461 + 0.0088 * tilt\_angle$. No significant linear relationship was found for partial guesses of 3 digits ($F_{1,26}$=0.01, p = 0.9, $R^2 = 0.0005$).

Attackers' Strategy Against ForcePIN Defense

We asked attackers about the strategy they used to guess ForcePIN. All 30 attackers indicated that they determined pressure digits through timing differences between pressure digits and non-pressure digits. We analyze the logged data to confirm this. First we compare the entry times between PIN and ForcePIN. We use the same definition as Krombholz et al. and define entry time as the duration between the key down event for the first digit and the key release event for the enter key. Entry times for PIN and ForcePIN are provided in Figure 8(a). On average, users take 1.7*s* (M=1.4*s*; SD=0.92*s*) to enter PINs whereas ForcePIN entry takes 3.2*s* (M=2.5*s*; SD=2.1*s*). An independent samples t-test indicates this 1.5*s* difference between PIN and ForcePIN is significant *(t = -10.8, p < 0.05)*. This corroborates the findings of Krombholz et al. and Arif et al. — both found statistically significant differences for entry times between PIN and ForcePIN. Krombholz et al. also reported ForcePIN took 1.3*s* longer to enter, very similar to the difference we found.



(a) Authentication times. (b) Keyhold intervals.

Figure 8. Timing and keyhold interval differences between PIN and ForcePIN and between normal and pressure digits.

To investigate the timing channel more closely, we examine the keyhold interval: the duration between pressing and releasing a key. Using the logs from the data collection, we measure keyhold intervals when the victim taps on keys with normal pressure and with high pressure in the ForcePIN defense.

Figure 8(b) illustrates that the average keyhold interval for normal pressure keys is 88*ms* (M=83*ms*; SD=38*ms*), but for keys pressed with higher pressure, it is 566*ms* (M=486*ms*; SD=419*ms*). An independent samples t-test indicates this 483*ms* difference is significant *(t = -58.6, p < 0.05)*. Detecting when a victim takes almost six times longer to enter a digit would be very easy, especially considering the less than 100*ms* interval of normal pressure digits. Our analysis provides the empirical evidence behind the timing side channel. We suspect that the attackers in Krombholz et al.'s experiment failed to identify this side channel due to their limited evaluation: only two attackers unrelated to the experimenters mounted 50 attacks in total. In our experiments, we had 30 observers mount 360 attacks on ForcePIN.

LIMITATIONS

Our study has reasonable limitations due to the inclusion of human subjects: the scope is limited to people willing to participate and it contains self-reported views. Since these are unavoidable, we discuss limitations specific to our study.

For the Tilt condition, victim participants may not have been as motivated to protect their secret from shoulder surfing attacks because they were using a temporary secret for the purpose of this study. This may have resulted in over reporting of the success rates. While this limitation is difficult to avoid, it is representative of the insider attack scenario where the user desires to be subtle. Furthermore, our analysis shows that participants used a range of tilting angles, which enabled us to report protection offered against different angles.

While we provided victim participants the opportunity to practice their chosen PINs as many times as they wanted, they were not well trained. This resulted in an optimal condition for the attackers. However, this limitation was unavoidable for a time-constrained, lab-based experiment.

The 250*ms* haptic feedback for the pressure digit(s) of ForcePIN may have influenced the entry times. However, this

potential influence on the entry times does not impact the presence of a side channel. This is evident from the longer entry times reported by Arif et al. for their experiment (without the haptic feedback).

Attacker participants did not receive a performance-based reward, so they may have had less motivation and this may have affected their performance. However, our results still provides a lower baseline on attackers' performance.

Wiese and Roth suggested that it is preferable to conduct shoulder surfing attacks on live users [24]. In our study, each attacker attacked 24 unique victims. To conduct the same study on live users would be extremely difficult in terms of participant recruiting and experiment session time. Moreover, using videos enabled us to exercise much more control over the attacker's distance and viewing angle, it eliminated potential side channels like audio, and it guaranteed that the environment context was consistent.

CONCLUSION

We conducted experiments using 30 subjects to understand shoulder surfing defenses and attack strategies on smartphone PINs. Our experiments show that it is quite easy to correctly guess PINs with two observations, on average. Furthermore, attackers were surprisingly effective in shoulder surfing PINs from across the room. We also subject the most commonly used defense, tilting the device screen away from the observer, to shoulder surfing attacks and show its limited efficacy. We show that while tilting the device screen away from the attacker with an angle of 70° or higher prevents complete guessing of PINs, smart attackers look for other clues (such as the proximity of the finger to the corner of the smartphone) to partially guess the PIN. Finally, we conduct experiments and gather empirical evidence to show that ForcePIN has an inherent timing side channel, which renders it completely ineffective against shoulder surfing attacks.

Our work calls attention to educating users about the threat of shoulder surfing and common misconceptions. First, it shows that smartphone users need to be careful even when the observer is located across the room. Second, while tilting the device screen may hide its contents, it does not prevent shoulder surfing attacks on PINs. Therefore, it is important to educate smartphone users on the inefficacy of this defense.

ACKNOWLEDGEMENT

We gratefully acknowledge the support of NSERC for grants RGPIN-2014-05499 and RGPIN-402467-2013.

REFERENCES

1. Ahmed Sabbir Arif, Ali Mazalek, and Wolfgang Stuerzlinger. 2014. The use of pseudo pressure in authenticating smartphone users. In *11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST. DOI: http://dx.doi.org/10.4108/icst.mobiquitous.2014.257919

2. Andrea Bianchi, Ian Oakley, Vassilis Kostakos, and Dong Soo Kwon. 2011. The phone lock: audio and haptic shoulder-surfing resistant PIN entry methods for mobile devices. In *5th International Conference on Tangible, Embedded, and Embodied Interaction*. ACM. DOI: http://dx.doi.org/10.1145/1935701.1935740

3. John T. Davin, Adam J. Aviv, Flynn Wolf, and Ravi Kuber. 2017. Baseline Measurements of Shoulder Surfing Analysis and Comparability for Smartphone Unlock Authentication. In *CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/3027063.3053221

4. Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. 2012. Touch me once and i know it's you!: implicit authentication based on touch screen patterns. In *30th Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI:http://dx.doi.org/10.1145/2207676.2208544

5. Alexander De Luca, Marian Harbach, Emanuel von Zezschwitz, Max-Emanuel Maurer, Bernhard Ewald Slawik, Heinrich Hussmann, and Matthew Smith. 2014. Now you see me, now you don't: protecting smartphone authentication from shoulder surfers. In *32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/2556288.2557097

6. Alexander De Luca, Emanuel Von Zezschwitz, Ngo Dieu Huong Nguyen, Max-Emanuel Maurer, Elisa Rubegni, Marcello Paolo Scipioni, and Marc Langheinrich. 2013. Back-of-device authentication on smartphones. In *31st Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/2470654.2481330

7. Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. Understanding shoulder surfing in the wild: Stories from users and observers. In *35th Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/3025453.3025636

8. Paul Emarath. 2013. Spaces in New Homes, National Association of Home Builders. (2013). https://www.nahb.org/en/research/housing-economics/special-studies/spaces-in-new-homes-2013.aspx

9. Barney G Glaser and Anselm L Strauss. 2009. *The discovery of grounded theory: Strategies for qualitative research*. Transaction publishers.

10. Jan Gugenheimer, Alexander De Luca, Hayato Hess, Stefan Karg, Dennis Wolf, and Enrico Rukzio. 2015. Colorsnakes: Using colored decoys to secure authentication in sensitive contexts. In *17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM. DOI: http://dx.doi.org/10.1145/2785830.2785834

11. Marian Harbach, Alexander De Luca, and Serge Egelman. 2016. The Anatomy of Smartphone Unlocking: A Field Study of Android Lock Screens. In *34th Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI:http://dx.doi.org/10.1145/2858036.2858267

12. Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *10th Symposium on Usable Privacy and Security*. https://www.usenix.org/system/files/conference/soups2014/soups14-paper-harbach.pdf

13. Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* (1979), 65–70.

14. Hassan Khan, Aaron Atwater, and Urs Hengartner. 2014. Itus: an implicit authentication framework for Android. In *20th Annual International Conference on Mobile Computing and Networking*. ACM. DOI: http://dx.doi.org/10.1145/2639108.2639141

15. Hassan Khan, Urs Hengartner, and Daniel Vogel. 2016. Targeted Mimicry Attacks on Touch Input Based Implicit Authentication Schemes. In *14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM. DOI: http://dx.doi.org/10.1145/2906388.2906404

16. Katharina Krombholz, Thomas Hupperich, and Thorsten Holz. 2016. Use the Force: Evaluating Force-Sensitive Authentication for Mobile Devices. In *12th Symposium on Usable Privacy and Security*. https://www.usenix.org/system/files/conference/soups2016/soups2016-paper-krombholz.pdf

17. Anindya Maiti, Kirsten Crager, Murtuza Jadliwala, Jibo He, Kevin Kwiat, and Charles Kamhoua. 2017. Randompad: Usability of randomized mobile keypads for defeating inference attacks. In *IEEE EuroS&P Workshop on Innovations in Mobile Privacy & Security*. IEEE.

18. Behzad Malek, Mauricio Orozco, and Abdulmotaleb El Saddik. 2006. Novel shoulder-surfing resistant haptic-based graphical password. In *EuroHaptics*.

19. Ildar Muslukhov, Yazan Boshmaf, Cynthia Kuo, Jonathan Lester, and Konstantin Beznosov. 2013. Know your enemy: the risk of unauthorized access in smartphones by insiders. In *15th International Conference on Human-computer Interaction with Mobile Devices and Services*. ACM. DOI: http://dx.doi.org/10.1145/2493190.2493223

20. Florian Schaub, Ruben Deyhle, and Michael Weber. 2012. Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In *11th International Conference on Mobile and Ubiquitous Multimedia*. ACM. DOI: http://dx.doi.org/10.1145/2406367.2406384

21. Muhammad Shahzad, Alex X Liu, and Arjmand Samuel. 2013. Secure unlocking of mobile touch screen devices by simple gestures: you can see it but you can not do it. In *19th Annual International Conference on Mobile Computing & Networking*. ACM. DOI: http://dx.doi.org/10.1145/2500423.2500434

22. Emanuel Von Zezschwitz, Alexander De Luca, Bruno Brunkow, and Heinrich Hussmann. 2015a. SwiPIN: Fast and secure pin-entry on smartphones. In *33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/2702123.2702212

23. Emanuel Von Zezschwitz, Alexander De Luca, Philipp Janssen, and Heinrich Hussmann. 2015b. Easy to draw, but hard to trace?: On the observability of grid-based (un) lock patterns. In *33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/2702123.2702202

24. Oliver Wiese and Volker Roth. 2015. Pitfalls of Shoulder Surfing Studies. In *NDSS Workshop on Usable Security*. https://www.internetsociety.org/doc/pitfalls-shoulder-surfing-studies

25. Oliver Wiese and Volker Roth. 2016. See you next time: a model for modern shoulder surfers. In *18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM. DOI:http://dx.doi.org/10.1145/2935334.2935388

26. Christian Winkler, Jan Gugenheimer, Alexander De Luca, Gabriel Haas, Philipp Speidel, David Dobbelstein, and Enrico Rukzio. 2015. Glass Unlock: Enhancing Security of Smartphone Unlocking Through Leveraging a Private Near-eye Display. In *33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/2702123.2702316

27. Qiang Yan, Jin Han, Yingjiu Li, Jianying Zhou, and Robert H Deng. 2013. Designing leakage-resilient password entry on touchscreen mobile devices. In *8th ACM SIGSAC Symposium on Information, Computer and Communications Security*. ACM. DOI: http://dx.doi.org/10.1145/2484313.2484318

28. Guixin Ye, Zhanyong Tang, Dingyi Fang, Xiaojiang Chen, Kwang In Kim, Ben Taylor, and Zheng Wang. 2017. Cracking Android pattern lock in five attempts. In *Network and Distributed System Security Symposium*. https://www.internetsociety.org/doc/cracking-android-pattern-lock-five-attempts

29. Nur Haryani Zakaria, David Griffiths, Sacha Brostoff, and Jeff Yan. 2011. Shoulder surfing defence for recall-based graphical passwords. In *7th Symposium on Usable Privacy and Security*. ACM. DOI: http://dx.doi.org/10.1145/2078827.2078835