Understanding Visual Perceptions of Usability and Security of Android's Graphical Password Pattern *

Adam J. Aviv
U.S. Naval Academy
Annapolis, MD
aviv@usna.edu

Dane Fichter
Swarthmore College
Swarthmore, PA
dficht1@swarthmore.edu

ABSTRACT

This paper reports the results of a user study of the Android graphical password system using an alternative survey methodology, pairwise preferences, that requests participants to select between pairs of patterns indicating either a security or usability preference. By carefully selecting password pairs to isolate a visual feature, a visual perception of usability and security of different features can be measured. We conducted a large IRB-approved survey using pairwise preferences which attracted 384 participants on Amazon Mechanical Turk. Analyzing the results, we find that visual features that can be attributed to complexity indicated a stronger perception of security, while spatial features, such as shifts up/down or left/right are not strong indicators for security or usability. We extended and applied the survey data by building logistic models to *predict* perception preferences by training on features used in the survey and other features proposed in related work. The logistic model accurately predicted preferences above 70%, twice the rate of random guessing, and the strongest feature in classification is *password distance*, the total length of all lines in the pattern, a feature *not* used in the online survey. This result provides insight into the *internal visual calculus* of users when comparing choices and selecting visual passwords, and the ultimate goal of this work is to leverage the visual calculus to design systems where inherent perceptions for usability coincides with a known metric of security.

1. INTRODUCTION

Graphical passwords, credentials based on users recalling or drawing shapes or images, have become more common place with the increase in touch oriented devices, such as smartphones and tablets. Perhaps the most prevalent graphical password system in use is Android's graphical password pattern which comes standard on Android devices. Like any password system, graphical passwords suffer from many of the same issues as text based passwords: users select weak and insecure passwords. We know about text-based password from large studies [10, 14, 19, 20, 13] of publicly available, real-world passwords databases [2] and through collaboration with industry [9] and academic institutions [19].

Studies of the Android password pattern have demonstrated that users are similarly inclined to make poor password choices [27, 3]. Unlike studies in the text-based password space which are based on large-scale analysis of leaked or real-world password data, the results for user selection of graphical passwords are generally based on in-lab surveys where participants are asked to provide a password that meets some criteria, such as being memorable or easy or secure. The findings are informative and persuasive regarding user-choice in this domain – users are inclined to choose weak passwords as compared to the total entropy of the space – but these results also belie larger challenges of studying graphical passwords in the same maner as text-based passwords. Due to the unavailability of large graphical password corpora, standard metrics, such as guessability probability [18], and large-scale study of actually used passwords are not possible.

Motivated by the challenges of large-scale studies of graphical passwords, we explored the use of a new survey methodology in this space that is based on *pairwise preferences* that attempts to measure the visual perceptions that influence secure and usable choices. The pairwise preference survey consists of users selecting between two passwords, indicating a preference for one password in the pair that meets some criterion, such as perceived security or usability compared to the *other* password. By carefully selecting the password pairs, visual features of the passwords can be isolated and the impact of that feature on users' perception of security and usability can be measured.

It is important to note that pairwise preference provides *different* and *complementary* insights into user choice as compared to related in-lab user studies. Pairwise preference data instead attempts to assess the visual processes that indicate *perceived* security or usability under some personal/user-based understanding and calculation of those terms. Perceptions influence choice, especially for passwords where users are consciously attempting to select password that adhere to a security and usability criteria that may not be well defined. The goal of this research is to better understand the visual process that drive security and usability perceptions that in turn can inform new designs of security systems where user perceptions actually match some known metrics of security.

Using the pairwise preference methodology we conducted a large IRB-approved survey that attracted 384 participants on Amazon Mechanical Turk. The resulting dataset averages 19.6 preference selections for over 1,100 password pairs that focus on 6 visual features, including password length, crosses, and left/right and up/down shifting. Additionally, users provided individual security and usability ratings of over 2,000 passwords, averaging 4.6 ratings per

*This research was partially supported by ONR grant N001614WX30023, Swarthmore Faculty Research Grant, and the Swarthmore STEM Student Summer Support. Material is based upon work supported by the Maryland Procurement Office under contract H98230-14-C-0127. Additional acknowledgement to Aashish Srinvas for preliminary work on this research.

ACSAC '14 New Orleans, LA USA
Copyright 2014 ACM ACM 978-1-4503-3005-3/14/12 ...$15.00
http://dx.doi.org/10.1145/2664243.2664253.

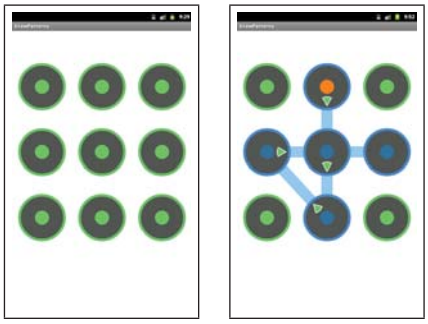

Figure 1: A sample pattern entry screen and a sample pattern as appearing in the survey. The orange marker indicates the start point with shaded lines and arrows indicating the direction of the pattern.

password. In analysis, we find that *length*, the number of contact points used in the password, is the strongest separator of usable or security preferences —passwords with more contact points are perceived as more secure as compared to those patterns with fewer contact points. Other visually complex features, such as crosses, also affect preferences, and when comparing complexity features, users were more inclined to perceive crossing patterns as the most secure. Spatial features, such as patterns shifted to the right/left or up/down, have little to no effect on user perceptions. When users individually rated passwords, they noted that perceived symmetry of the password negatively correlated with security, as in users find symmetric patterns to have less security, while perceived symmetry has little to no effect on perceived usability.

The survey data was expanded and applied to *predict* user preference of password pairs to model perceptions. Using the standard machine learning technique of logistic regression, we constructed a predictive model trained on features from the online survey as well as additional visual features described in related work [4, 21, 27, 3]. The model is able to predict the preferred secure patterns at an accuracy of 70% and preferred usable patterns at an accuracy of 65%. Surprisingly, of the training features *pattern distance*, the total length of all the lines in the password (not the total number of contact points) was the strongest predictor of a preference, a feature *not* used in the online survey.

2. BACKGROUND

The Android password pattern is a recall based graphical password scheme based loosely on Pass-Go [24] and Draw-A-Secret [17][1]. The user is required to select a pattern by traversing a grid of 3x3 points. A pattern must contain at least four points, cannot use a point twice, and all points along a path must be included (no skipping points). Figure 1 provides a reference visual for the input of the pattern (*left*) as well as a sample pattern (*right*) where the orange dot indicates the start point in the pattern with arrows describing the direction of the traversals. Note that in the '+' shaped pattern, the middle point is not associated with the horizontal traversal since it was previously selected during the vertical traversal. In our survey, we used this visualization of the password pattern because it mimics the pattern entry screen on Android 2.x versions, which is still widely used.

Analysis of the Android password pattern has shown that there exist 389,112 possible passwords [6]. Despite the relative complexity of the scheme, the total number of passwords is actually quite small compared to a general ordering of points, which would allow for over a million passwords. The entropy of the password space is also quite small, on the order of a 3-digit PIN [27]. Other studies have also suggested that the actual number of patterns that are likely usable, that is, can be entered accurately without repetition, is probably much smaller than 389,112. In a study of accelerometer readings, the authors reported many users complaining about certain patterns being "hard to enter" [7].

[1]For a larger review of the graphical password landscape, we refer to the reader to two surveys [8, 23].

3. RELATED WORK

Password Study Methodologies .

Research in text-based password selection is very well developed. Early work in the area by Morris and Thompson in 1979 showed that humans are bad at choosing passwords [20]. More recent studies break down into two main categories: (1) the study of large leaks of passwords, and (2) survey based studies where users either self-report password statistics or researchers have users provide a password in the context of a system.

The study of large password leaks [10, 14, 19] is a fairly standard method, but it is not clear how well these data sets represent broader password choice because they tend to only contain the passwords that were cracked through conventional brute force methods [1]. This tends to bias results towards a study of weak passwords and not necessarily passwords in general use. In the graphical password arena, the likelihood of such a leaked password set is extremely low because graphical passwords are not used for remote authentication, yet.

Another methodology is to perform surveys where participants self-report properties about their passwords [22, 30]. These surveys are quite helpful in understanding user preferences, but the demographics of these surveys are often limited to university settings. Online surveys provide a much better option because it can reach a broader set of participants with limited expense, and have become a common methodology [18, 28]. Online surveys try and gauge user choice by having participants select passwords to protect an important account, like a bank account or class web pages [29], and then researchers analyze the results [16].

The graphical password studies, due to a lack of large data sets, have relied primarily on in-lab experimentation. For example, they either bring participants into a lab to use the graphical password scheme [24, 17] or perform pen-and-paper experiments [27]. Unfortunately, the collected data for these experiments can be inconsistent [12], particularly when asking participants to provide passwords [17].

3.1 Studies of Passwords on Mobile Devices

Zhao *et al.* investigated the security of the picture password scheme [29] used by Microsoft tablets, which requires users to select a password by circling points and drawing lines over a chosen image. The user then must recall the specific drawings in order to unlock the device. Picture password schemes often suffer from hotspots [26]. Zhao *et al.* collected sample picture passwords by securing a course website with the picture password scheme and analyzing the chosen passwords, including an analysis of images chosen and how circles and lines are drawn upon those images (hotspots). They concluded that for many images, users are likely to circle and draw on certain spots which can be identified using image processing techniques.

Andriotis *et al.* conducted a survey to advance smudge attacks [6] on Android password patterns by training a predictive model with user surveyed data [3]. The goal was to improve the performance of the smudge attack with information about likely user patterns. Androitis *et al.* surveyed 22 participants requesting them to provide a

"secure" password and an "easy" password to help train their predictor. From these results, they found that users are inclined to select certain shapes, like the "L"- or ¬-shape, as well as a strong bias to start in the upper left contact point. We were able to obtain the user choices from this study, and we reevaluated 8 of those choices, selected at random, from the set of 22. The goal is to see if user choice of security and usability is consistent for in-lab settings. We refer to this as the *Bristol* study because it was conducted at the University of Bristol, UK.

In recent related work, Uellenbeck *et al.* developed a method for quantifying the entropy of Android passwords [27] using partial guessing entropy [9]. They found that the security of the password pattern is on the same order as a 3-digit pin, low entropy. They also conducted a number of pen-and-paper surveys of patterns and identified a number of key n-grams, or sequences of swipes, that users are most inclined to chose. For example, the various forms of "L" were very common, and they also identified a strong bias for users to initiate their patterns in the upper-left corner. The work of Uellenbeck *et al.* is very informative and persuasive about the processes of user choice, and these results compliment our own. Unfortunately, we did not have access to the Uellenbeck results until after our online surveys completed, and were not able to integrate their findings with our survey[2]. We believe, however, that there are interesting places to expand prior results by investigating the preferences of security and usability of the identified n-grams. We do note that at least one of the n-grams was in our survey a ¬-like shape that individually rated as usable but not secure.

4. SURVEY METHODOLOGY

The survey was conducted in two rounds on Amazon Mechanical Turk (MTurk), running within a HIT iFrame (Human Intelligence Task). The survey was hosted on a server at our institution, and all data collected was initially stored at Amazon and now resides at a secure server onsite. Survey participants were limited to those residing in the United States over the age of 18, as per the requirements of the IRB. We did not collect additional demographic information, and instead relied on MTurk to properly check the qualifications of the participants.

The online survey is organized into five parts:

- *Pre-Survey Questionnaire*, which inquires about ownership of smartphone and the use of graphical passwords;
- *Secure Pairwise Preference*, which requires participants to select between a pair of pasword patterns based on a security preference;
- *Password Rating Section*, which requires participants to answer questions about individual passwords;
- *Usable Pairwise Preference*, which requires participants to select between a pair of patterns based on a usability preference;
- *Post Survey Questionnaire*, which asks participants to reflect on the survey.

The survey is designed such that participants cannot easily bypass sections by simply clicking through, and if they do so, such actions are easily detected. For a visual reference to the survey parts, refer to Figure 2. On the left is a sample of a preference selection and on the right is post selection. Not pictured is the individual password rating survey which displayed a single pattern with a series of questions.

[2]We do, however, use the start and end points of a pattern as a feature in the predictive models. Like other spatial features, such as up/down and left/right shifting, the start and end point had little to no impact on predicting preferences. See Section 6 for details.

The online survey is interactive and user-responsive, written in a combination of PHP and Javascript. The pairwise survey requires participants to highlight their selections and confirm the selection before proceedings. We also included attention tests that require participants to recognize that a pair of patterns were identical and thus select the same choice, and we performed post survey analysis to measure the structure of the responses for any indication dishonest intentions during survey taking, for example always selecting left or right, or back and forth. A failure of either test could result in a rejection of the data from the study and/or a rejection of the user within the MTurk ecosystem.

4.1 Survey Questions and Formats

Pre-Survey Questions.

The pre-survey questionnaire was designed to gather some baseline data on the participants and their familiarity with smartphones and Android in particular. We asked the following questions:

- *Do you currently own an Android device?*
- *If you own an Android device, do you use the graphical password feature?*
- *How frequently do you have to re-enter your graphical password because you entered it incorrectly?*
- *Do you find that your graphical password sufficiently secures your Android device?*
- *Describe your handedness?*

We found that a large number of the survey participants owned or used an Android device: 354 of the 384, or 92%. Of the Android owners, 257 of the 354 use graphical passwords, or 72%, and 249 of the 257 graphical password users found that their password was "sufficient" or 97% qualified respondents. Due to the informal definition in the questionnaire, the participant is open to define "sufficiently secure" as he/she sees fit. There is likely a range of interpretation, for example the participant may interpret the question as his/her pattern is *sufficiently secure* for the purpose of preventing casual, unintended use of the device, such as snooping by a friend or family member. The term could also imply a stronger sense of security, such as preventing access by a determined actor like law enforcement. In either case, the high percentage of qualified respondents describing their password as sufficient does suggest a level of comfort with their password choice based on some personal security criteria.

With regards to how frequently users must reenter their passwords due to incorrect entry, a potential measure of usability: 11 reported that they must do so "Very Frequently", 44 reported "Frequently," and 153 reported that he/she "Rarely" must reenter their pattern due to incorrect entry. When combined with the previous responses about security, most participants are Android users who feel comfortable with their password choice with respect to a personal security criteria and ease of password entry. Finally, and unsurprisingly, the majority of participants are right handed (336) as compared to ambidextrous (15) and left handed (33).

Pairwise Preference Surveys.

In both pairwise preference surveys, for security and usability, the same survey format was used. Due to the similarity between the surveys, the password rating survey was placed between the two pairwise surveys to break up the work flow. In both pairwise surveys, each participant indicated a security and usability preference on 48 pairs of patterns. The same set of patterns was used in both the secure and usable sections, but the order of the survey as well as the order within the pair (*i.e.*, left vs. right) was randomized

Figure 2: Survey screenshots: (*left*) Secure comparison prior to selection and (*right*) secure comparison after selection before confirmation.

within each survey. The selection of the 48 password pairs was at random from a carefully selected pool of 1,109 possible password pairs. The pair selection methodology is discussed in the following section. An additional 2 pairs were added to the survey for attention tests bringing the total to 50 pairs used in the preference surveys. A visual of the pairwise preference survey is presented in Figure 2, showing both the comparison state, prior to selection, and the confirmation state, post selection.

Depending on the survey part, in the comparison state, the participant was instructed: "Your job is to determine which of the passwords is the most secure" or "most usable." Recall, that there are two pairwise preference surveys, one for indicating secure preferences and one for indicating usable preferences. Since both surveys are visually similar, an additional prompt was presented prior to starting the second pairwise survey on usability: "The next section looks like the first password section you completed, but it is not. Instead of choosing the most secure of the two passwords, your task is to choose the most usable." This information and the fact the two pairwise surveys are separated by the individually rating survey (discussed next), we feel this properly informed and directed the participant.

We left the definition of security and usability intentionally vague in the survey prompts. Participants applied their own definitions to these terms, and it is possible that many interpreted "secure" as "not usable" or "usable" as "not secure" or in some combination. The goal of the survey is not to precisely identify a user preference for security/usability based on a predetermined definition, but rather to assess the collective visual stimulants that would influence a preference for security/usability. Furthermore, individual preferences have little effect on the result, and the combined preference as a fraction of participants selecting a pattern in the pair was considered.

In addition to indicating a preference for a password in the pair, the participant may also indicate "same" if he/she felt that both passwords exhibit the same security and usability. Once a password is selected, it is highlighted, and confirmation state is entered, where the participant must confirm their selection by pressing a "submit" button or select "cancel" to re-select. Similarly, if "same" is selected, the confirmation state is entered with the same choice, submit or cancel.

Finally, each participant was presented with four attention tests, two in the security survey and two in the usability survey. The attention test consisted of showing the participant the same password on the left and right, and, thus, the participant should select "same." Failing three of four attention tests and/or the entropy test, resulted in rejecting the results of the user. In the end, we accepted data from 384 participants and rejected 54. This resulted in an average of 19.6 total preference classifications per password pair, with 11 usable choices and 11 secure choices per pair, on average.

Password Rating Survey.

The password survey differed from the pairwise comparison in that participants were only directed to assess a single password pattern at a time. Each participant rated 25 patterns chosen at random from the 2,214 patterns in the pool of password pairs. The survey presented the participant with a single password surrounded by the following prompts:

- *Does the above password appear to be symmetrical?*
- *Describe any symbols or geometric shapes which appear in the password.*
- *Rate the security of this password.*
- *Describe the usability of this password.*

Questions of symmetry is a binary, "Yes" or "No", response; the question of shapes is an free text, optional answer; and, the usability and security questions provided four responses between "Highly Unusable", "Unusable", "Usable", and "Highly Usable", with "usable" replaced with "secure" appropriately for the security question. Overall, all 2,114 password patterns received ratings averaging 4.6 ratings per password.

Post-Survey Questionnaire.

At the end of the survey, we asked the participant to answer a brief post survey questionnaire.

- *If you use Android's graphical password, how has this survey changed your opinion of your own password?*
  - My password is less secure than I believed.
  - My password is more secure than I believed.
  - My opinion of my password is unchanged.
  - I do not use Android's graphical password.
- *Which statement about the relationship between security and usability of passwords is most true?*
  - The passwords which look the most secure look more difficult to remember.
  - The passwords which look the most secure look easier to remember.
  - There is no clear relationship between usability and security.

The goal of the post-survey questionnaire is to assess if the passwords viewed affected the participant's security preference of his/her own password and if there was an inverse view of security and usability.

Unsurprisingly, in response to the question about the relation-

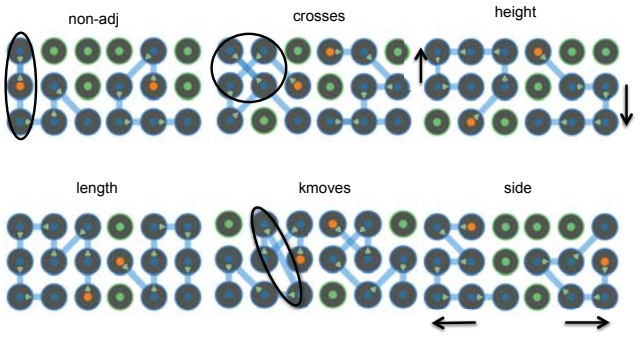

Figure 3: Features tested in the survey

ship between security and usability, 92% of respondents reported that the relationship is inversely related. This is also present in the pairwise data results presented in the following sections. The response to the question about participants who use a graphical password indicates that after viewing many of the passwords selected for the survey, 45% of participants felt that his/her current password was less secure and only 4% felt that it was more secure. *This suggests that a mechanisms that simply shows users a random graphical password before making a selection could improve user choice.* Developing an experiment to test this hypothesis is a focus of future work.

4.2 Selecting Password Pairs

Password pairs were selected to isolate a single feature of a password, such that one password exerts that feature and all other features in the set are similar within some bound. For example, if a password pair was selected for the *length* feature, then one password would be long, connecting more contact points, than the other, but all other features would be similar within a threshold, such as the number of crosses.

Six features were considered in the survey, and examples of those features and pairs are presented in Figure 3. The features are described below with the shorthand inset:

- Length (*length*): The total number of contacts points used in the pattern. This does not consider the total length of the swipes, but in Section 6 the total length of the swipes is characterized.
- Crosses (*crosses*): The total number of times the patterns cross over each other. A subcategory of crosses is *exes*, which is a 90 degree cross.
- Non-Adjacent (*non-adj*): The total number of non-adjacent swipes which occur when the pattern double-backs on itself by tracing over a previously contacted point. For example, starting in the middle contact point, swiping right, then doubling back to the contact point to the left of the initial point.
- Knight-Moves (*kmoves*): The total number of knight-moves which occur when a contact point is connected to another point that is two spaces in one direction and then one space over in another direction, like how a knight moves in chess. These are also referred to as 30 degree swipes [6] in related work.
- Height (*height*): The amount the pattern is shifted towards the upper contact points or the lower contact points.
- Side (*side*): The amount the pattern is shifted towards the left or right contact points.

The goal is to select pattern pairs randomly from all possible patterns, but, unfortunately, many patterns are visually complex at a level where participants could not directly distinguish the patterns and the tested features easily. To counteract this effect, passwords were first classified based on the features above using a simple normalized metric. This rated the prevalence of each feature on the range of [0,1], where 0 indicated the feature was *not* present and 1 indicated that the feature was prevalent at its most extreme. For example, a pattern can have at most 14 crosses in it, so by counting the total number of crosses in a pattern and dividing by 14 provided a simple normalized metric is computed.

Feature	Total Pairs	Strict Pairs
Length	114	53
Crosses	135	35
Non-Adj	151	43
Kmoves	132	44
Height	134	51
Side	139	60
Crosses vs. Non-Adj	100	-
Crosses vs. Kmoves	100	-
Non-Adj vs. Kmoves	100	-
Bristol	8	-

Table 1: Number of Feature Pairs Used in the Survey

To limit the overall visual complexity of the pattern, only patterns whose features were under the 0.7 normalized threshold were considered, except for the features *length*, *height*, or *side* since extremes in those categories are necessary for completeness, for example, to have a pattern completely left or right shifted or of length 8 or 9. Additionally, we limited *kmoves* such that at least one of the intermediary contact points, between which the knight move traverses, must be selected prior. This was argued to increase the usability of the pattern [6, 7]. With these limitations, we selected patterns by randomly sampling from all possible patterns until coverage across features was achieved. In addition to the single feature pairs, we selected a set of pairs that pitted two features against each other for the combination of *kmoves*, *non-adj*, and *crosses* to better assess which features have stronger visual preferences.

Table 1 describes the total number of pairs selected in each category. Due to randomization, we do not have equal coverage across pairs. Additionally, each feature has a strict characterization as well as a general characterization. This is because we wished to identify pairs that generally characterized a difference in the features, for example, a password pair where one pattern has more *crosses* but the two passwords also differed in length by a single contact point. Strict pairs contained passwords that were nearly identical in all features except for the tested feature, and we found that the results are consistent across general and strict pairs.

We acknowledge that the limitation on password selection for the pairs can impact the results; however, the goal of the experiment is to assess the features that inform perceptions of security and usability. Participants who are challenged to quickly recognize patterns due to visual complexity are unlikely to produce informed data, and we felt that decreasing the visual complexity was necessary to this end. Patterns are still selected through a random process, and it should be noted that a number of password pairs contained highly complex patterns which are not likely to be password used in the wild[3].

However, prior studies do provide an opportunity to reevaluate findings in a different setting. In particular, the study by Andriotis *et al.* asked users to first select a password that is secure and one that is usable. We obtained the selected passwords from the re-

[3]One participant did report in non-survey feedback that his/her pattern was part of the dataset.

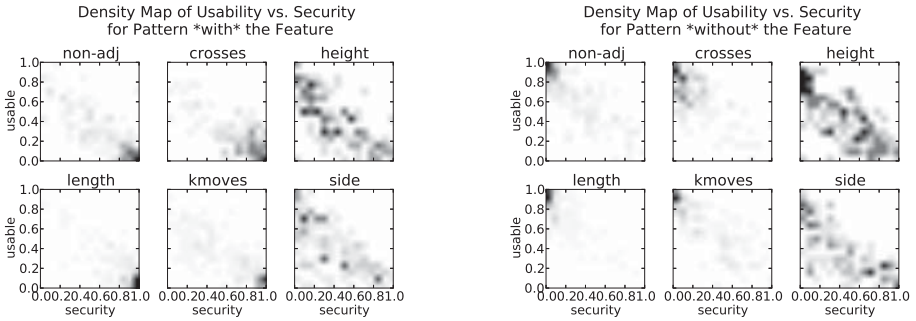

Figure 4: Density maps between security (x-axis) and usability (y-axis) for each features: *left*, considering the pattern in the comparison pair that had the feature; *right*, considering the pattern in the comparison that did have the feature.

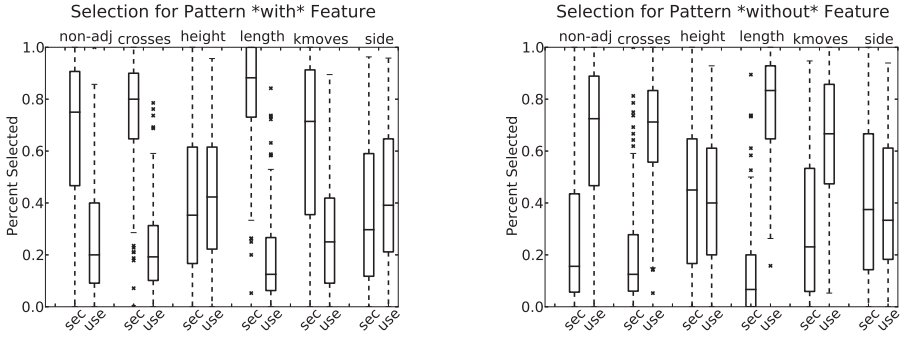

Figure 5: Confidence intervals for percent chosen secure or usable for each pattern that (*left*) has the feature and (*right*) dos not have the feature.

searchers and used a subset of those selections in the survey. Since the study was conducted at the University of Bristol in the UK, we describe these pairs as *Bristol* pairs. We were unable to include password features and n-grams identified by Uellenbeck *et al.* because prior work was not available at the time of the survey.

4.3 Compensation and Reducing Harm

The survey was posted as a MTurk HIT (Human Intelligence Test) twice over the summer of 2013. In the first instance, we compensated participants at a rate of $2.50 for completing the survey without failing the attention and entropy tests. We reject users through MTurk who failed the attention test, which resulted in a number of complaints, including to our IRB. Upon further review, we determined that, yes, these participants should be rejected, but the stigma of MTurk rejections was undue harm to the participant. After re-reviewing the survey with the IRB, we decided to accept all participant through MTurk and release compensation but to remove their data from the data set.

As a result of the complaints, in the second round survey, we changed the compensation to $1.50 for completing the survey — all participants, regardless of the quality of the data received the $1.50 compensation and were accepted within the MTurk ecosystem. We then further reviewed the data, and if a participant passed the attention tests, we provided a bonus of $1.00, equalling the $2.50 compensation from the previous survey. We found that this methodology reduced harm to participants within the MTurk ecosystem.

4.4 Survey Limitations

The described surveys are designed to provide insight into the perceptions of security and usability as queued by different visual features. We believe that this data is informative about user choice habits during password selection; however, there are limitations to this method, and in particular, the fact that perceptions should not be interpreted as a metric for security or usability. While perceptions may correlated with security metrics, unfortunately we are currently unable to directly compare perceptional metrics with those of a true security metric which is in part a consequence of the system being analyzed. There are currently no standard metrics for password strength for the Android password pattern. The closest proposal is described by Uellenbeck *et al.* [27], but its effectiveness remains untested in the field and is based on data unavailable to the authors. In the text based passwords space there is more opportunity as standard metrics are known, and the pairwise preference methodologies of measuring perceptions could be fruitfully applied there, an area of future work. For graphical passwords, however, we see perceptional metrics playing a role in designing selection systems; for example, as discussed in Section 6, generating a ranking of passwords based on perceptions where a user first chooses a password they deem usable, and then are suggested passwords with the same usability perceptions but have some stronger security property. A user will have a password that they perceive as usable but meets a security criteria (perhaps from [27]), and thus will be more likely to use the stronger password.

5. SURVEY RESULTS

In the survey, over 1,100 pairs were rated with a preference assigned to one pasword in the pair for security and usability or same usability/security. Each pair received 19.6 ratings on average, and the combined the overall preferences for a single password is based on the fraction of participants preferring the password. Each password is assigned four preference statistics:

- *sec*, the fraction of participants preferring the pattern more

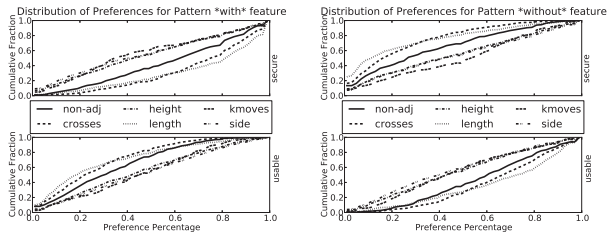

Figure 6: Cumulative fraction distribution of percent selected secure (top) and usable (bottom) for the pattern with the feature (left) and without the feature (right).

secure in the pair;

- *use*, the fraction of participants preferring the pattern more usable;
- *samesec*, the fraction of participants preferring neither pattern as more secure;
- *sameuse*, the fraction of participants preferring neither pattern as more usable.

Note, that for a given password pair, (p_1, p_2):

$$sec(p_1) + sec(p_2) + samesec(p_1, p_2) = 1$$
$$use(p_1) + use(p_2) + sameuse(p_1, p_2) = 1$$

Each password pair is selected to test a visual feature, and one password has the feature, while one does not. This distinction is indicated with the keywords *with*, indicating results for the password *with* the feature, and *without*, indicating results for the password in the pair *without* the feature.

5.1 Pairwise Preference

The most straight forward analysis of pairwise preferences is to measure the relationship between usability and security preferences for each password pair based on the tested feature. This comes in two flavors: an assessment of usability and security preferences for the passwords *with* the feature and a complimentary assessment for the passwords *without* the feature. Note that *with* and *without* are not inverses because participants can indicate *same* preference if neither password is preferred under the criteria.

Figure 4 presents a density image map of the distribution of security versus usability for each visual feature tested, both for passwords *with* the feature (*left*) and for passwords *without* the feature (*right*). Density maps with higher concentrations in the lower right corner indicate a stronger preference for security, while density maps with a higher concentration in the upper left corner indicate a stronger preference for usability.

Of the tested features, *non-adj*, *crosses*, *kmoves* and *length*, the passwords *with* the feature demonstrate a clear preference for security while passwords *without* the feature demonstrate clear preference for usability. For these features, there is a strong inverse relationship between usability and security. The spatial features, *height* and *side*, however, do not exhibit an inverse relationship and have scattered densities along the diagonal.

These observations continue in Figure 5, which is a box-and-whisker graph of the distribution of preferences for passwords *with* and *without* the feature. Here, the inverse relationship of *non-adj*, *crosses*, *kmoves*, and *length* is even more prevalent. For example, the difference in the median between usability and security for passwords *with* the *length* feature is extreme at 0.75, and, at the same time, the difference in the median preference for *side* and *height* are nearly identical for usability and security both *with* and *without* the feature, varying at most 0.09.

Feature	With-Pref	Feature	Without-Pref
length	24.36	height	3.22
kmoves	22.12	side	-2.16
crosses	21.75	length	-17.28
non-adj	20.56	non-adj	-22.07
height	-2.27	crosses	-22.60
side	-5.55	kmoves	-23.80

Table 2: Security versus usability Preferences for patterns with and without the feature as measured by the area between the cumulative fraction distributions: Left the increased security preference, and right, the negative security preference.

5.2 Preference Ranking Metric

We wished to develop a ranking metric to better understand the usability and security preferences of the feature. To do so, we first generated the cumulative fraction distribution for each feature, which considers the fraction of passwords that receive the preference, or higher. The cumulative fraction graphs are presented in Figure 6. Distributions that are shifted towards the lower right corner, or *cupped*, suggest a stronger usability/security preference because a higher fraction of passwords receive stronger ratings in the category. Similarly, distributions that are shifted towards the upper left corner, or *capped*, suggest a weaker usability/security preference since a larger fraction of passwords received weaker preferences in the category.

In the left graphs of Figure 6, we generate the cumulative fraction graphs (cfg) for the passwords *with* the feature, the security cfg on top and the usability cfg on the bottom. The ranking metric for the features is based on measuring the area between the usability cfg and the security cfg: the more strongly capped the usability cfg is and the more strongly cupped the security cfg is, the greater the security preferences are for that feature. Table 2 presents the results of that measure. For the *with* passwords, the most strongly preferred secure feature is *length*, the number of contact points used in the pattern, and all major non-spatial features also exhibit a strong security preference.

Interestingly, for the *without* passwords, this metric can be seen as how much the *lack of* the feature reduce the security preference for the password. We describe this as the *negative security preference* since it negatively impacts the perceived security of the password in the pair. The ordering of negative security preferences are not consistent with the prior result. The *kmoves* feature exerts the strongest negative security preference, and *length* has the weakest negative security preference. This implies that of the features tested increasing the length of the password increases the security preferences but a lack of length does not strongly decrease security preference as compared to other features, like a lack of *kmoves*. In the Section 6, a broader set of features is explored in relation to user preference, and confirms that *length* is a strong indicator of perceived security and usability, but an additional feature, the password distance, or the length of the individual strokes, is an even stronger indicator than *length* alone.

5.3 Comparing Feature Preferences

The previous results considered the tested features in isolation, that is, a single pattern in the pair has the feature while the other does not. Now, we consider features in comparison to each other where within a pair one pattern has a feature and the other has a different feature with all other aspects being equal.

To start, the effect of the *length* feature is compared to other tested features. Recall that the data set contains both strict pairs and general pairs. A strict pair is a pure comparison with only one feature changed. A general pair may have all features being the

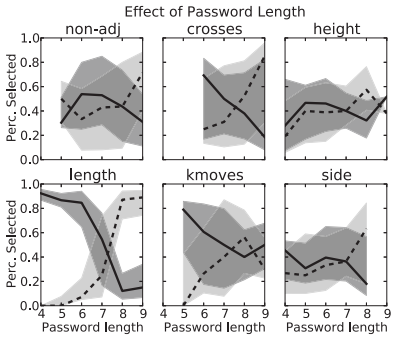

Figure 7: Length Correlation: Effects of usability (dashed line) and security (solid line) for increasing length. The shaded region is the first and third quartile.

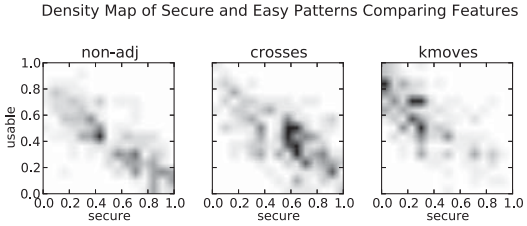

Figure 8: Density distribution of security and usability preferences for pairs of patterns that compared two features.

same, but have one feature vary slightly under the threshold, such as *length*. We investigated all general pairs that varied in length to determine if *length* was a singular driver in preference. These results are presented in Figure 7. The y-axis in each subplot is the length of the pattern; the solid line is the median security preference; the dashed line is the median usability preference; and the shaded region is the first and third quartile.

For some of the features, a clear inverse relationship between length and usability/security preference is present. Most demonstrable is the *length* feature itself; however, the strength of this relationship is not present for all features. Notably, both *non-adj* and *kmoves* do not have as strong an inverse relationship — as length increases, security increases and usability decreases — and the spatial features *height* and *side* have almost no relationship to length, which is aligned with prior findings. This suggests that while the length of the password is a major contributor, participants are sensitive to the other tested feature.

We further explored the relationship between features by pitting the non-spatial features, excluding *length*, in a pairwise preference comparison. The density map of those results are presented in Figure 8. From these results, we can see that *non-adj* is weakly shifted towards usability when compared to *crosses* and *kmoves*, while *crosses* are considered more secure in comparison. The *kmoves* feature, interestingly, was the usable preferred of the features tested, clearly shifted towards the upper left.

5.4 Comparison to Bristol Studies

Included in the pairwise features was passwords that were selected by participants from a related study by Andriotis *et al.* at the University of Bristol, UK [3]. In that study, participants were asked to provide one password that is "secure" and one password that is "easy." We wished to test if user generated passwords purposely selected in a similar usable/secure criteria would be preferred in the same way within our study. A set of 8 password pairs from the *Bristol* study were included in the pairwise preference study and those results are presented in Figure 9. There does exist a secure preference for the user selected secure choice, as well as usable

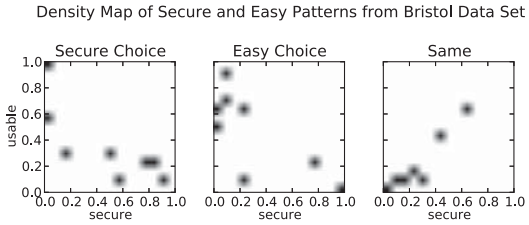

Figure 9: Bristol Data: Usability and Security preferences of data collected in the Bristol study.

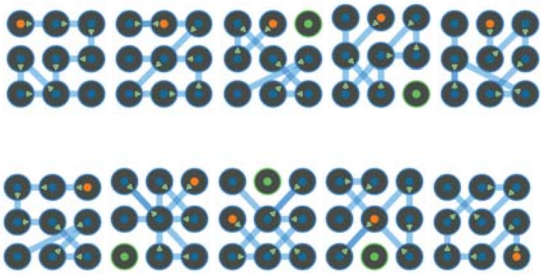

Figure 10: Top 10 highest rated passwords for most "Highly Usable" and "Highly Secure." The highest rated is in the top left, and moving left-to-right by row, the tenth, highest rated is the lower right

preference for the user selected easy choice. There are significant outliers, however, at the opposite extremes. This experiment does not draw direct conclusions on the validity of the Bristol study, but does suggest that user provided data for usability studies can contain inconsistencies. The pairwise preference survey could provide a way to validate prior results and expand on those studies.

5.5 Individual Password Ratings

Finally, we report on the results of the password rating survey. Recall that participants were directed to provide a rating for individual passwords in addition to pairs of passwords, indicating the password's symmetry, any apparent shapes, a usability assessment, and a security assessment. Of most interest to this study is the rating of security and usability, which we can place on a range [0,3], where 0 indicates "Highly Insecure" or "Highly Unusable" and 3 indicates "Highly Secure" and "Highly Usable." A rating of 1.5 would be neutral with respect to usability and security. The rating was compiled by taking the average response from all ratings by participants for a given password. We only included passwords that received at least 5 ratings in these results.

Participants identified a large portion of passwords used in the study as at least "secure" and "usable" as indicated by an average score of at least 2 in both categories. To further understand this result, the top 10 highest rated secure *and* usable patterns in our data set of approximately 2,000 passwords were identified and presented in Figure 10, with the number one most secure and usable password in the upper left and the number ten in the lower right, moving left-to-right by row. The set of features in these top ten patterns runs the gamut, but all have a length of at least 7, again demonstrating the impact of length on perception of security. While these patterns apear complex, they all received positive usability scores, suggesting that *users would be open to using a wider variety of pattern forms than they might use if selecting independently.* A mechanism that simply shows users randomly selected passwords may increase the overall security of user choice, as these results and the exit survey demonstrates.

5.6 Perceptions of Symmetry

Finally, the relationship between perceived password symmetry and usability/security is measured. Recall that participants were directed to rate "Yes" or "No" if the pattern "appeared to be symmetric." We looked at the correlation of the average rating, on the scale for [0,1] where 1 indicate symmetry and 0 is the lack thereof, with the security and usability rating. We applied the standard Pearson product-moment correlation coefficient, and found that symmetry and security correlation have an $r = -.249$ while symmetry and usability have an $r = 0.058$. This result indicates that users perception of security is negatively impacted by perceived symmetry, but symmetry does not have a strong effect on perceived usability. In some ways, this is contrary to studies that suggest symmetry is a strong indicator of memory in psychology [5, 15] and usability studies for the Draw-a-Secret graphical password scheme [25].

These results could be explained by how participants associated symmetry with guessability of a passwords and usability with ease-of-entry. For example, these results might imply that users see symmetric password as something that an attacker would likely guess before an asymmetric password, thus decreasing the perceived security. When interpreting usability as ease-of-entry, it is also inline that symmetry has little impact on usability because, in the end, the user must still enter the password and even symmetric passwords can be hard to enter. This is consistent with related work on symmetry and mnemonics [5, 15]; a password could be easy to remember but still hard to enter.

6. PREDICTING PREFERENCES

In this section, we move beyond presenting the direct empirical results of the survey and instead broaden the focus by expanding the set of passwords features that may affect preference. The following experiments involves constructing a logistic regression model based on a set of extended features to measure their predictive power in anticipating a user's security or usability preference.

6.1 Extended Features

For these experiments, the set of features were extended beyond those surveyed to include metrics presented in prior studies [3, 4, 27]. This larger set of features were used to train a logistic regression model on training data randomly sampled, 900 pairs from the 1,100 collected, and the remaining data was withheld as testing. Below, the additional features and their characteristics are explained, as well as where the feature was used in prior work.

- Ex-Crosses (*exes*): the number of exes within corners of the patterns, not just general crosses.
- Starting Point: (*start*) : the start contact point (as used in [3, 27])
- Ending Point (*end*): the end contact point (as used in [3, 27])
- Distance (*dist*): total geometric distance of the strokes in the pattern, not just the length, as calculated by the total points used [4].
- Point Entropy (*pfreq*): the entropy of the points used in the password as calculated from the total distribution of point found in all patterns [4].
- Stroke Entropy (*sfreq*): the entropy of pairwise points in the pattern (or strokes) as calculated from the total distribution of strokes found in all patterns [4].
- Vertical Symmetry (*vsym*): the vertical symmetry of the pattern with respect to mirroring the pattern top to bottom including both points and strokes [4, 21].
- Horizontal Symmetry (*hsym*): the horizontal symmetry of the pattern with respect to mirroring the pattern left to right including both points and strokes [4, 21].

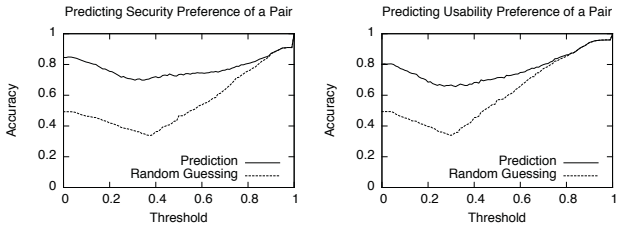

Figure 11: Predicting rates for selecting the preferred password in a pair above a threshold as compared to random guessing.

Each password in the pair received values based on the features above as well as the normalized feature values used in pair selection for the survey (*length*, *kmoves*, *crosses*, *non-adj*, *height*, *side*), totalling 13 different features. Features were assigned to pairs based on the difference between the individual features of the two passwords, and labeling occurred per pair (see below).

6.2 Preference Labeling

Each pair received a label from three values -1, 0, 1, where a -1 indicated that the left item in the pair was preferred, 0 indicating that neither was preferred, or 1 indicating the right item was preferred. Labels were assigned based on a threshold difference in preference: if $sec(p_1) - sec(p_2) > \tau$ a label of -1 is assigned to indicate a preference for p_1, or if $sec(p_2) - sec(p_1) > \tau$ a label of 1 is assigned to indicate a preference for p_2, and if neither is true, then a label of 0 is applied to indicate no strong preference.

There is a range of choices for the threshold with tradeoffs, and to assess these tradeoffs, we ran a five-fold cross-validation over the training set with the thresholds ranging from 0 to 1. The results of the experiment is presented in Figure 11 with a comparison to random guessing (i.e., always selecting the most common label). At a threshold of approximatively 0.33, random guessing is at a minimun because the number of samples with each label is roughly equal. We use this threshold going forward because this represents the most challenging predictive scenario. Additionally, the 0.33 threshold is a natural choice because it requires that one password in the pair have a preference at least twice as great as the other to receive a preference label other than 0.

6.3 Prediction Results

At the 0.33 threshold rating, the logistic model can accurately separate the data for a security preference at a rate of 70% and usable preference at a rate of 66%, which is twice the rate of random guessing in both instances. With lower thresholds, such as threshold of 0.0 which would label the data with -1 or 1 if there is any difference in preference between the patterns, the model can accurately separate the data at a rate of 80% or higher for both usability and security which is significantly greater than random guessing, which is just below 50%.

The final consideration is to analyze which of the features are the strongest predictors of preference. To measure this, we used the 0.33 threshold value and constructed separate models using only a single feature from the set. The results of that experiment is presented in Figure 12. One might expect, as before, that the features used in the survey would be the strongest indicator as these were the features used when selecting pairs. As indicated, the *length* feature continues to be a strong predictor of user preference for both usability and security, but the *dist* feature is by far the strongest feature followed by *crosses* and then *length*. In some ways, this is intuitive. The total distance of a pattern, which is a measure of

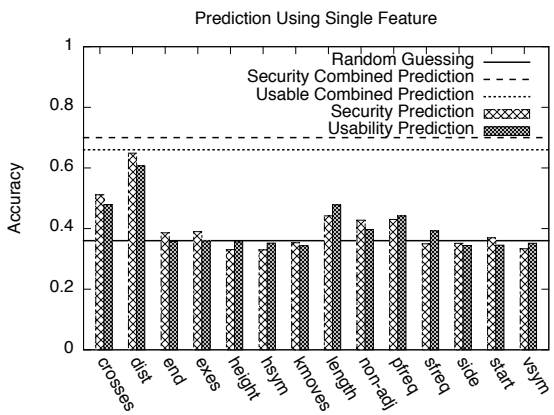

Figure 12: Predictions using single feature from the set.

the length of all the lines, is a good indicator of visual complexity because to add any of the other tested features, such as *non-adj*, *kmove*, *crosses* and *length*, requires increasing the distance of the pattern, by definition, and thus increased pattern distance results in increased visual complexity. This result is also important because it is indicative of the visual calculus that users apply when determining a preference for usability and security, and it can be leveraged in designing new visual security systems as well as aides, like password meters, to help user selection of patterns.

Other findings of this experiment are interesting in that they indicate that the more theoretic metrics used in prior work [4] have little impact on pairwise preferences. This includes metrics involving symmetry; neither horizontal nor vertical symmetry can predict a preference above random guessing. Information theoretic metrics were more capable, including *pfreq* and *sfreq*, but primarily the entropy of contact point distributions seems to be a more reasonable predictor of preference. Finally, spatial metrics, again, fail to strongly correlate with preference, including the start and end point, which were highlighted in prior studies [3, 27] as having strong biases in user-selected passwords.

7. CONCLUSION

We presented the results of a large user study of pairwise preferences for usability and security of the Android password pattern which provides insights into user perceptions that inform choice. Pairs of patterns were selected based on six visual features, and we concluded that spatial features, such as shifting to the side or up/down, had little impact on a preferred pattern in the pair. More visually striking features have a stronger impact, with the length of the pattern being the strongest indicator of preference. These results were extended and applied by constructing a predictive model with a broader set of features from related work, and we found that the distance feature, the total length of all the lines in a pattern, is the strongest predictor of preference. Our findings provide insight into users' visual calculus when assessing a password, and this information could be leveraged to develop new password systems or user selection tools, like password meters. Moreover, with a good predictive model of user preference, it can be applied to a broader set of passwords, including those not used in the survey. Ranking data based on learned pairwise preferences is an active research area in machine learning [11], and the resulting rankings metric over all potential patterns in the space would be greatly beneficial to the community. It could enable new password selection procedures where users are helped in identifying a preferred usable password that also meets a security requirement.

References

[1] John the ripper. `http://www.openwall.com/john/8`.

[2] The Password Project. `http://thepasswordproject.com`.

[3] Panagiotis Andriotis, Theo Tryfonas, George Oikonomou, and Can Yildiz. A pilot study on the security of pattern screen-lock methods and soft side channel attacks. In *Proceedings ACM Conference on Security and privacy in Wireless and Mobile Networks*, WiSec, 2013.

[4] Majid Arianezhad, Douglas Stebila, and Behzad Mozaffari. Usability and security of gaze-based graphical grid passwords. In *Financial Cryptography and Data Security*, pages 17–33, 2013.

[5] Fred Attneave. Symmetry, information, and memory for patterns. *The American journal of psychology*, 68(2):209–222, 1955.

[6] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith. Smudge attacks on smartphone touch screens. In *Proceedings of the 4th USENIX Workshop On Offensive Technologies*, WOOT'10, 2010.

[7] Adam J. Aviv, Ben Sapp, Matt Blaze, and Jonathan M. Smith. Practicality of accelerometer side-channels on smartphones. In *Annual Computer Security Applications Conference (ACSAC)*, 2012.

[8] Robert Biddle, Sonia Chiasson, and Paul C Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys*, 44(4):19, 2012.

[9] Joseph Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *IEEE Symposium on Security and Privacy (SP)*, 2012.

[10] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A birthday present every eleven wallets? the security of customer-chosen banking pins. In *Financial Cryptography and Data Security*, pages 25–40. Springer, 2012.

[11] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of International Conference on Web Search and Data Mining*, pages 193–202, 2013.

[12] Sonia Chiasson, Robert Biddle, and Paul C van Oorschot. A second look at the usability of click-based graphical passwords. In *Proceedings of the Symposium on Usable Privacy and Security*, SOUPs, pages 1–12, 2007.

[13] Matteo Dell'Amico, Pietro Michiardi, and Yves Roudier. Password strength: an empirical analysis. In *Proceedings of IEEE INFOCOM*, pages 1–9, 2010.

[14] Serge Egelman, Joseph Bonneau, Sonia Chiasson, David Dittrich, and Stuart Schechter. It's not stealing if you need it: A panel on the ethics of performing research using public data of illicit origin. In *Financial Cryptography and Data Security*, pages 124–132. Springer, 2012.

[15] Robert Stanton French. Identification of dot patterns from memory as a function of complexity. *Journal of Experimental Psychology*, 47(1):22, 1954.

[16] Kirsi Helkala and Nils Kalstad Svendsen. The security and memorability of passwords generated by using an association element and a personal factor. In *Information Security Technology for Applications*, pages 114–130. Springer, 2012.

[17] Ian Jermyn, Alain Mayer, Fabian Monrose, Michael K Reiter, Aviel D Rubin, et al. The design and analysis of graphical passwords. In *Proceedings of the 8th USENIX Security Symposium*, pages 1–14. Washington DC, 1999.

[18] Patrick Gage Kelley, Saranga Komanduri, Michelle L Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *IEEE Symposium on Security and Privacy (SP)*, pages 523–537, 2012.

[19] Michelle L Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring password guessability for an entire university. In *Proceedings of the Conference on Computer & Communications Security (CCS)*, 2013.

[20] Robert Morris and Ken Thompson. Password security: A case history. *Communications of the ACM*, 22(11):594–597, 1979.

[21] PC van Oorschot and Julie Thorpe. On predictive models and user-drawn graphical passwords. *Transactions on Information and System Security*, 10(4):5, 2008.

[22] Jeffrey M Stanton, Kathryn R Stam, Paul Mastrangelo, and Jeffrey Jolton. Analysis of end user security behaviors. *Computers & Security*, 24(2):124–133, 2005.

[23] Xiaoyuan Suo, Ying Zhu, and G Scott Owen. Graphical passwords: A survey. In *Annual Computer Security Applications Conference (ACSAC)*, 2005.

[24] Hai Tao and Carlisle Adams. Pass-go: A proposal to improve the usability of graphical passwords. *IJ Network Security*, 7(2):273–292, 2008.

[25] Julie Thorpe and Paul C van Oorschot. Graphical dictionaries and the memorable space of graphical passwords. In *USENIX Security Symposium*, 2004.

[26] Julie Thorpe and Paul C van Oorschot. Human-seeded attacks and exploiting hot-spots in graphical passwords. In *USENIX Security Symposium*, 2007.

[27] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. Quantifying the security of graphical passwords: the case of Android unlock patterns. In *Conference on Computer & Communications Security (CCS)*, 2013.

[28] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, and Lujo Bauer. How does your password measure up? the effect of strength meters on password creation. In *Proc. USENIX Security*, 2012.

[29] Ziming Zhao, Gail-Joon Ahn, Jeong-Jin Seo, and Hongxin Hu. On the security of picture gesture authentication. In *USENIX Security Sympsosium*, 2013.

[30] Moshe Zviran and William J Haga. Password security: an empirical study. *Journal of Management Information Systems*, 15:161–186, 1999.