Comparisons of Data Collection Methods for Android Graphical Pattern Unlock

Adam J. Aviv
aviv@usna.edu
United States Naval Academy

Jeanne Luning Prak
jlprak@mimech.us
Broadneck High School

ABSTRACT

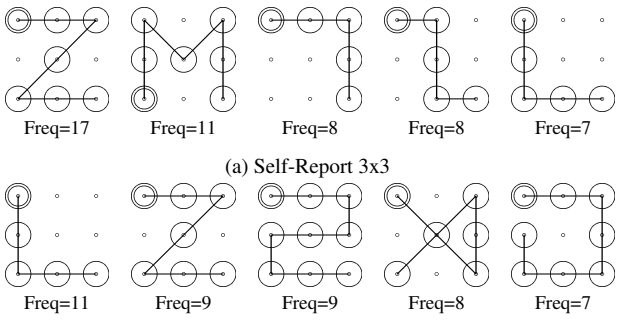
We conducted two methodologically different studies to measure differences between Android graphical patterns collected in-lab via pen-and-paper surveys and those collected on-line via self-reporting on the users own mobile device. We find that there exist subtle but potentially significant differences between data collected in-lab using pen-and-paper and data collected on-the-device via self-reporting. In particular, the guessability of self-reported and pen-and-paper patterns diverge at the tail ends: the more common/least-secure patterns reported via self-report are much more easily guessed than pen-and-paper patterns, but less common/more-secure self-reported patterns are much harder to guess than pen-and-paper patterns. With respect to visual features, the self-reported patterns contained statistically significantly more crosses and exes than pen-and-paper patterns, and self-reported patterns also tend to shift towards the top of the grid space while pen-and-paper patterns shift toward the bottom of the grid space. These results suggest that while in-lab surveys for Android graphical passwords using pen-and-paper are a reasonable substitute for real/in-the-wild data, there are likely subtle ecological differences that need some accounting. Unfortunately, overall, the scope of human-generated passwords for this authentication scheme in both collection methods remain weak, on the order of a random 3-digit PIN, confirm prior results in this space.

1. INTRODUCTION

Studying graphical passwords in the same manner as text-based passwords is challenging because graphical passwords are not used for remote authentication, and are thus unlikely to be hacked and leaked to the public where researchers can analyze them. As a result, large corpora of real world graphical passwords do not exist for study. To compensate, researchers have conducted studies within the lab [1, 3] to collect and analyze data. In this poster-abstract, we seek to test the ecological validity of in-lab methodologies by comparing an in-lab survey to an on-line one where users self-report their graphical password on their own device. We do not claim that prior work is invalid; to the contrary, we wish to better understand the vagaries of this space and, in fact, *confirm prior results herein.*

Our analysis focuses on Android's graphical password pattern unlock, which is perhaps the most commonly used graphical password system today and is available on all Android devices. Password patterns are selected by connecting a series of contact points in a 3x3 grid without lifting, repetition, or avoidance.

This research focuses on two large IRB approved studies. The first study consisted of an in-lab, pen-and-paper based survey closely following the methodology of Uellenbeck et. al [3] where participants "draw" a set of patterns to select their own passwords and attempt to guess the passwords of others. The second study was conducted entirely online using Amazon Mechanical Turk and requested that participants optionally self-report their graphical passwords using their mobile device. In total, we collected 491 pen-and-paper/in-lab passwords and 440 on-line/self-reported passwords.



(a) Self-Report 3x3

(b) Pen-Paper 3x3

Figure 1: Most Common Patterns

Comparing the two data sets, we find that there are a number of strong similarities. In both cases, the set of user generated passwords is highly predictable and much less varied than that of the total allowable passwords (389,112 total password patterns). Many visual features also followed similar distributions. However, overall, using a modified version of the guessing algorithm from [3], we find that the partial guessing entropy of self-reported passwords is higher (up to 1-bit higher) than that of in-lab studies for partial-guessing fractions greater than 0.2. Unfortunately, the entropy is still much lower than guessing a set of random patterns (6-bits lower), and thus, we can continue to conclude that user-generated passwords for Android's graphical password system are significantly weaker on whole than the allowable set of passwords.

2. METHODOLOGY

The survey for the in-lab study used a pen-and-paper model where participants selected their own personal patterns by drawing them on a grid on the paper survey form, and then also attempted to guess the patterns of others in their session by drawing additional patterns. The participates were incentivized with winning an edible treat for both recalling their own patterns and guessing others. Each participant selected 3 patterns of their own and were asked to guess 10 patterns of others. We analyzed the combined set, totaling 491 patterns with 38 participants taking part during three sessions.

The second survey was conducted using Amazon Mechanical Turk with a payment of $0.75 for participation. Participants were required to take the survey on an Android device (likely their own personal device), and were then asked to either self-report their cur-

	Self-Report		Pen-and-Paper			
	freq.	$\bar{x}$	freq.	$\bar{x}$	t-test	χ^2-test
Crosses	34/440	0.141	16/491	0.049	$p < 0.01$	$p < 0.005$
Knight-Moves	24/440	0.077	14/491	0.045	$p < 0.1$	$p < 0.1$
Non-Adjacent	64/440	0.164	54/491	0.116	$p < 0.05$	$p < 0.15$
X's	17/440	0.043	5/491	0.010	$p < 0.005$	$p < 0.01$
Length	–	6.055	–	6.283	$p < 0.05$	–
Stroke Length	–	5.823	–	5.919	$p < 0.25$	
Side	–	-0.036	–	-0.094	$p < 0.25$	–
Height	–	0.134	–	-0.136	$p < 0.0005$	–
Repeats	203/440	–	245/491	–	–	$p < 0.5$
Sym.	336/440	–	398/491	–	–	$p < 0.01$
Uniq Sym.	133/440	–	153/338	–	–	$p < 0.1$

Table 1: Composition Comparison and Significance Testing: frequency of occurrences, average occurrence per pattern, and p-scores for t-test and χ^2-test as appropriate.

rent Android password or alternatively report statistics about their password. Standard attention tests were used to ensure accurate data, such as requiring users to repeat entries during the survey and attest that they took the survey honestly. In total, 750 participants took the survey and 440 of those self-reported their password and passed the attention tests.

3. RESULTS

Standard Features. An initial analysis of the visual features of the patterns shows that they are highly similar (Table 1). Of the features considered, only the length, height (the shifting toward the top or botom of the gird space), number of crosses (line segments that cross over another line segment), and number of exes (line segments that cross and form an 'x' shape) are significantly different. The differences in height and crosses are particularly notable: 8% of self-reported patterns contain crosses whereas only 3% of pen-and-paper patterns contain crosses. The two sets of data were virtually opposite in with respect to height: self-reported patterns tend toward the top of the grid space where as pen-and-paper patterns tends towards the bottom of the grid space.

All other visual features, including knight-moves (moving to a dot one over and two up), connections of non-adjacent dots, stroke length, and side the pattern tended towards, are not significantly different. The number of repeated patterns and symmetric patterns are also similar, with the two sets being within 5% of each other (Table 1).

Information Theoretic. To measure the guessability of the data, we followed the guessing algorithm as described by Uellenbeck et. al which first applies a naive guess based on repetitions founding in the training set followed by using probability estimates from a Markhov chain. We amended this algorithm slightly to include rotations and flips of the training sets in the naive guessing and for seeding the transition matrix with initial probabilities. Figure 3 shows the results of the average of 10 runs of a 5-fold cross-validation using 400 randomly selected patterns from each set. At first, patterns from the pen-and-paper study are harder to guess, but after the initial guessing rounds, the self-reported passwords are harder to guess overall. This is likely due to the fact that there are also more outliers in the self-reported data than in the pen-and-paper data that are dissimilar from the rest of the set which makes it more difficult to guess using the Markhov model.

The increased difficulty of guessing the self-reported passwords as compared to the pen-and-paper passwords is apparent in the partial-guessing entropy [2] which was calculated using the same formulation as Uellenbeck et. al. We find similar entropy values as reported in prior work, and for $\alpha = 0.1$ (the fraction of passwords

(a) Self-Report 3x3 (b) Pen-Paper 3x3

Figure 2: Frequency of Pattern Start and End Points (in percent)

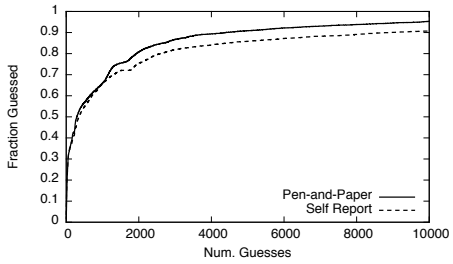


Figure 3: Guessability

Distribution	α =0.1	α =0.2	α =0.5	α =0.7	α =1.0
Pen-and-Paper	6.83	6.95	8.96	10.17	12.53
Self-Reporting	5.82	6.57	9.19	10.32	13.95
Random 4-Digit PIN	13.28	13.28	13.28	13.28	13.28
Random 3-Digit PIN	9.97	9.97	9.97	9.97	9.97
Random 2-Digit PIN	6.64	6.64	6.64	6.64	6.64
Random 3x3 Pattern	18.57	18.57	18.57	18.57	18.57

Table 2: Comparing Partial Guessing Entropy

an attacker wishes is to crack), the entropy for pen-and-paper is higher than that of self-reported passwords; however this changes for larger α fractions. At $\alpha = 1.0$ (guessing all the patterns) the entropy of the self-reporting is 1-bit higher and is roughly as secure as a 4-digit PIN, whereas the entropy of the pen-and-paper is more than 1-bit lower in entropy.

4. CONCLUSION

We presented the analysis of two methods for collecting graphical passwords, focusing on Android's graphical password pattern. We find that there exist consistencies between in-lab/pen-and-paper methods and on-line/self-reporting methods, and, in fact, confirm the results of prior work. However, there are differences that researchers should be aware of; most prominent, the partial-guessing entropy for self-reported passwords is higher than that of pen-and-paper ones. Unfortunately, overall, the results confirm that user-generated passwords are on the whole much worse (more guessable) than it should be considering the overall size of the password space.

5. ACKNOWLEDGMENT

Ravi Kuber and Devon Budzitowski for contributions to this poster-abstract. This work was partially supported by the National Security Agency and the Office of Naval Research.

6. REFERENCES

[1] P. Andriotis, T. Tryfonas, G. Oikonomou, and C. Yildiz. A pilot study on the security of pattern screen-lock methods and soft side channel attacks. In *WiSec'13*, 2013.

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

[3] S. Uellenbeck, M. Dürmuth, C. Wolf, and T. Holz. Quantifying the security of graphical passwords: The case of android unlock patterns. In *CCS'13*, 2013.