Usable, Secure and Deployable Graphical Passwords

by
Paul Dunphy

A thesis submitted in partial fulfilment of the
degree of Doctor of Philosophy
in Computing Science

School of Computing Science
Newcastle University

November 2012

Abstract

Evaluations of the usability and security of alphanumeric passwords and Personal Identification Numbers (PINs) have shown that users cannot remember credentials considered to be secure. However, the continued reliance upon these methods of user authentication has placed end-users and system designers in a coevolutionary struggle, with each defending competing concerns of usability and security. Graphical passwords have been proposed as an alternative, and their use is supported by cognitive theories such as the picture superiority effect which suggest that pictures, rather than words or numbers, could provide a stronger foundation upon which to design usable and secure knowledge-based authentication. Indeed, early usability studies of novel systems harnessing this effect appear to show promise, however, the uptake of graphical passwords in real-world systems is low. This inertia is likely related to uncertainty regarding the challenges that novel systems might bring to the already delicate interplay between usability and security; particularly the new challenges faced in scaffolding user behaviours that comply with context-specific security policies, uncertainty regarding the nature of new socio-technical attacks, and the impact of images themselves upon usability and security.

In this thesis we present a number of case studies incorporating new designs, empirical methods and results, that begin to explore these aspects of representative graphical password systems. Specifically, we explore: (i) how we can implicitly support security-focused behaviours such as choosing high entropy graphical passwords and defending against observation attack; (ii) how to capture the likely extent of insecure behaviour in the social domain such as graphical password sharing and observation attack; and (iii) how through the selection of appropriate properties of the images themselves we can provide security and usability benefits. In doing so, we generate new insights into the potential of graphical passwords to provide usable, secure and deployable user authentication.

Acknowledgements

Thanks must firstly goto Microsoft Research who funded this project through their PhD scholarship programme, their support has been much appreciated. During my PhD I received lots of great advice from some very smart people, but special thanks must go to my advisor Patrick Olivier for his support throughout my time in the Culture Lab, and to Jeff Yan for support in the early stages of my studies. I benefitted from their expertise and experience immeasurably. Feedback from Aad van Moorsel and Lizzie Coles-Kemp during my thesis defence also much improved this work.

During the time I have spent in Newcastle I have met a number of great people who made the years spent pursuing the PhD very enjoyable. Anja Thieme, Jonathan Hook and Stephen Lindsay especially helped to periodically remind me that more important things in life exist than sitting in front of a computer, and actually, that a number of those things can be found in the nearest pub. Culture Lab has a working atmosphere unlike any other I have experienced, and the camaraderie that exists is undoubtedly at the core of its success. A workplace where people are as much friends as colleagues is a rare place indeed. Everybody I have encountered in the lab both past and present has contributed to a really great few years.

One of the great pleasures that accompanies the study of a PhD is travel; I was made very welcome on a number of trips by great hosts all over the world. I had some great times visiting Alexander de Luca, Max-Emanuel Maurer and Emanuel (Emario) von Zezschwitz in Munich; Stuart Taylor was a great host at Microsoft Research Cambridge; and I had a great internship with Andreas P. Heiner and N.Asokan at the Nokia Research Centre in Helsinki.

Finally, thanks must goto my parents who supported me to goto University in the first place. Hopefully, finishing the PhD will go some way to convince them that I've actually been doing something useful for the past few years.

Collaborations

The work described in this thesis is the result of a number of fruitful collaborations: my supervisors Patrick Olivier and Jeff Yan; the numerous students who over the years I have had the pleasure of supervising; and researchers within the Culture Lab, and Nokia Research Centre (Helsinki).

More specifically, the implementation of the authentication mechanisms on the multi-touch surface described in Chapter 4 was led by David Kim who also participated in the generation of ideas that informed their design, as did John Nicholson, James Nicholson, Jonathan Hook, and Pam Briggs. The authentication mechanisms described in Chapter 5, were the result of numerous discussions with Andreas P. Heiner and N. Asokan at the Nokia Research Centre in Helsinki. The graphical password description study described in Chapter 6 was greatly assisted by James Nicholson who undertook much of the study recruitment, helped with the running of the study, and contributed ideas to its design.

These contributions were essential to the work carried out in this thesis, and while I am very grateful to all these willing collaborators I fully accept all responsibility for errors, and omissions in the work presented.

Contents

Abstract **1**

Acknowledgements **2**

Collaborations **3**

1 Introduction **15**
1.1 Overview . . . 15
1.2 Research Context . . . 17
1.2.1 Password Security and Usability . . . 17
1.2.2 User-Centred Security . . . 18
1.2.3 Graphical Passwords . . . 18
1.3 Research Questions and Contributions . . . 19
1.4 Research Approach . . . 21
1.5 Thesis Structure . . . 22
1.6 Prior Publications . . . 24

2 User-Centred Security and Graphical Passwords **25**
2.1 User-centred Security . . . 25
2.2 User Authentication . . . 28
2.2.1 Summary . . . 29
2.3 Alphanumeric Passwords . . . 29
2.3.1 Security . . . 30
2.3.2 Scaffolding Choice of Secure Passwords . . . 32
2.3.3 Studying the Usage Context . . . 36
2.3.4 Summary . . . 38
2.4 Graphical Passwords . . . 38
2.4.1 Recall-based Graphical Passwords . . . 39
2.4.2 Recognition-based Graphical Passwords . . . 43
2.4.3 Cued recall-based Graphical Passwords . . . 50
2.5 Thesis Overview . . . 55

3 Scaffolding Choice of High Entropy Graphical Passwords **61**
3.1 Threat Model . . . 62
3.2 Background Draw a Secret (BDAS) . . . 62

3.3 User Study . . . 63
3.3.1 Method . . . 63
3.3.2 Study Materials . . . 64
3.3.3 Results . . . 65
3.4 Discussion . . . 71
3.4.1 Security Implications of BDAS . . . 72
3.5 Study Limitations . . . 73
3.6 Conclusion . . . 74

4 Observation Resistant Graphical Passwords through Multi-Touch Interaction 76
4.1 Threat Model . . . 77
4.2 A Framework for Observation Resistant Interactions . . . 78
4.3 Observation Resistant User Authentication on Multi-Touch Surfaces . 80
4.3.1 Exploiting Intuitive Physical Gestures . . . 80
4.3.2 Exploiting Multiple Concurrent Touches . . . 80
4.3.3 Exploiting the Invisibility of Pressure Change . . . 82
4.3.4 Reflection . . . 83
4.4 Observation Attack User Study . . . 84
4.4.1 Method . . . 85
4.4.2 Study Materials . . . 86
4.4.3 Results . . . 87
4.5 Discussion . . . 89
4.6 Study Limitations . . . 90
4.7 Conclusion . . . 91

5 Observation Resistant Graphical Passwords through Image Portfolios 92
5.1 Threat Model . . . 93
5.2 Intersection Attack vs. Observation Attack . . . 94
5.3 Simulated Observation Attacks . . . 95
5.4 Implementation . . . 97
5.4.1 High Entropy Graphical Passwords . . . 98
5.4.2 Low Entropy Graphical Passwords . . . 98
5.5 User Study . . . 98
5.5.1 Method . . . 100
5.5.2 Study Materials . . . 101
5.5.3 Results . . . 101
5.5.4 Login Durations . . . 103
5.6 Observation Attack User Study . . . 103
5.6.1 Results . . . 104

5.6.2 Questionnaire . . . 105
5.7 Discussion . . . 107
5.8 Study Limitations . . . 108
5.9 Conclusion . . . 109

6 Graphical Password Sharing and Image Similarity 111
6.1 Threat Model . . . 112
6.2 Graphical Password *Description* . . . 112
6.3 User Study 1 - Description Collection . . . 114
6.3.1 Method . . . 114
6.3.2 Study Materials . . . 114
6.3.3 Results . . . 116
6.4 User Study 2 - Passfaces Authentication using Description . . . 116
6.4.1 Method . . . 117
6.4.2 Study Materials . . . 120
6.4.3 Results . . . 122
6.4.4 Gender Differences . . . 124
6.5 Discussion . . . 125
6.6 Study Limitations . . . 126
6.7 Conclusion . . . 126

7 Towards Automated Detection of Image Similarity 128
7.1 Usable and Secure Graphical Passwords and Image Filtering . . . 129
7.1.1 Similarity in the Recognition-based Graphical Password Login 131
7.2 User Study 1 - Human Consensus of Image Similarity . . . 132
7.2.1 Method . . . 132
7.2.2 Results . . . 133
7.2.3 Choosing an Automated Similarity Metric . . . 133
7.3 User Study 2 - Recall Test Using Automatically Selected Image . . . 136
7.3.1 Method . . . 136
7.3.2 Results . . . 138
7.4 Discussion . . . 140
7.4.1 Security Implications . . . 142
7.5 Study Limitations . . . 144
7.6 Conclusion . . . 144

8 Conclusion 146
8.1 Research Questions . . . 147
8.1.1 How can interaction design scaffold secure behaviour in graphical password schemes? . . . 147
8.1.2 How can strategic selection of images provide scaffolding in graphical password schemes? . . . 148

8.1.3 What is the impact of graphical passwords upon socio-technical threats? . . . 150
8.2 Summary of Contributions . . . 151
8.3 Future Work . . . 152
8.3.1 Scaffolding User Choice of High Entropy Graphical Passwords 152
8.3.2 Mass Longitudinal Comparisons of Representative Graphical Password Systems . . . 153
8.3.3 Experience-centred Security . . . 153

Appendix:

A Draw a Secret (DAS) Information sheet 169

B Background Draw a Secret (BDAS) Enrolment and Login Sheet 171

C Example log file from two weeks of usage for the mobile-based graphical password system 173

D Examples of collected descriptions 182

E Example enrolment screen for *Mechanical Turk* user study 187

List of Figures

Figure 1 A visual description of the thesis structure. 22

Figure 2 Illustration of self-reported data where users were asked to remember the cause of their last password problem, and the regularity at which they used that password (as reported by Sasse et al. [109]). 37

Figure 3 An example of a DAS [68] graphical password that has a single stroke and length of seven. The raw encoding of the drawing is (2,2) (2,3) (3,3) (3,2) (2,2) (1,2) (5,5). 40

Figure 4 Examples of lines that cross fuzzy boundaries in the DAS system. Such lines would difficult for users to recreate according to the rules of DAS due to proximity to the cell boundaries. 40

Figure 5 The *Pass-Go* [127] system requires users select cell intersections to assemble a graphical password. 42

Figure 6 A mono-grid recognition-based graphical password system [28]. In this configuration the image positions are unchanged between selections and logins. Red squares represent example selections. 44

Figure 7 An example of a multi-grid recognition based graphical password system [98] in a one-key-per-screen configuration. Red squares represent example selections. 44

Figure 8 An example of a multi-grid recognition-based graphical password system where key images are randomly spread across grids. Red squares represent example selections. 45

Figure 9 The five conditions from the study carried out by Everitt et al. [40] exploring the usability of remembering multiple Passfaces graphical passwords. Rows represent the 5 study weeks (Tuesday, Wednesday, Thursday, Friday) and each distinct character represents a distinct graphical password. The recurrence of those characters depicts how frequently users had to login using that graphical password. 47

Figure 10 Example visualisation of 5 chosen points in an image as required by the Passpoints system, taken from [143]. 51

Figure 11 The cued-recall graphical password scheme proposed by Weinshall [139]. Users begin in the top left corner, and traverse the grid by moving down if they see a key image, or right otherwise. The user must then enter the number they arrive to at the edge of the grid. . . 52

Figure 12 Dirik et al. [33] used image processing techniques such as corner detection to predict likely sequences of Passpoints graphical passwords. (left) an overlay of click points collected from a corpus of users; (right) overlay of click points predicted using their image processing methods. 54
Figure 13 Illustration of the sub-domains of graphical passwords encountered in each chapter of the thesis. 57
Figure 14 Example of a DAS drawing grid (left); example of a BDAS grid (right). 62
Figure 15 An image with few hotspots (left) and a large number of hotspots (right). 63
Figure 16 The background images selected for the user study, BDAS participants were able to choose one upon which to draw their BDAS graphical password. 65
Figure 17 Examples of complex drawings that participants were able to remember: (left) Basketball and backboard created using DAS (7 strokes, length 39) (right) Persian name written with BDAS (9 strokes, length 27). 66
Figure 18 Examples of simple drawings created in the study that had a low stroke count, (left) from a BDAS participant (7 strokes, length 7), (right) from a DAS participant (4 strokes, length 4). 66
Figure 19 Entropy of the drawings (in bits) created by participants in the BDAS group. Calculations are made from the tables provided by Thorpe and van Oorschot [131]; off chart calculations are estimated. . 67
Figure 20 Entropy of the drawings (in bits) created by participants in the DAS group. Calculations are made from the tables provided by Thorpe and van Oorschot [131]; off chart calculations are estimated. 68
Figure 21 The BDAS drawing that one participant could not recall in the five minute recall test. The blue dots indicate erroneous attempts to remember the starting point of the circle that forms the head of the person (4 strokes, length 18). 69
Figure 22 The BDAS drawing that a participant could not recreate correctly in the one week recall test (12 strokes, length 34). 70
Figure 23 The drawing incorrectly repeated by the DAS participant in the one week recall test: (left) the original drawing; (right) the drawing as created by the participant one week later (7 strokes, length 24). . . . 70
Figure 24 Based upon a password space of at most 40 bits, The resulting impact of the location of a hotspot upon the password space using the assumption that a user will visit a hotspot in every stroke: (left) the hotspot is a corner cell (middle) the hotspot is a border cell (right) the hotspot is a central cell. 73

Figure 25 *Exploiting Intuitive Physical Gestures* - ShieldPIN screenshot with added example interaction (left), in situ (right): the PIN keypad only appears once the shielding gesture is detected in the green zone which serves to force the user to enter the PIN and visually cover the keypad. 81
Figure 26 *Exploiting Multiple Concurrent Touches* - Colour Rings screenshot with added example interaction (left), in situ (right): The user drags coloured rings to select key icons amongst decoys, one ring lassos a key image of the user, the rest represent decoy selections to create confusion for an attacker. 82
Figure 27 *Exploiting the Invisibility of Pressure Change* - Two images of a hand resting upon a multi-touch surface (based upon Frustrated Total Internal Reflection (FTIR) [57]) where different levels of pressure upon each finger are illustrated by different sized circles upon the interface. 83
Figure 28 PressureFaces screenshot with added example interaction (top), photo (bottom). The user increases pressure on one finger per hand in the coloured pressure zones to communicate an (x, y) coordinate and select an object. 84
Figure 29 The custom-built FTIR table used for the evaluation (49x95x105cm) and the user study context. 86
Figure 30 Percentage of observers able to replicate the inputters credentials (by authentication method). 87
Figure 31 The distribution of successful login durations recorded for each system. 88
Figure 32 The impact of the Pressure Grid on the login durations of participants using *Faces* and PIN. 90
Figure 33 Top: mean number of observations required by an attacker to learn a sufficient number of key images to gain a single successful login; (bottom): mean number of observations required by an attacker to learn the entire key image portfolio, useful for unrestricted future access. Both represent averages of 10,000 simulated sessions. 96
Figure 34 Pareto chart that illustrates the outcomes of 10,000 simulated observation attacks and the likelihood of various contexts that provide an attacker with a successful login. 97
Figure 35 Screenshots of the *high entropy* and *low entropy* systems. In the high entropy system, images were displayed in a 3x3 grid, and in the low entropy version displayed in a 2x2 grid. 99
Figure 36 Example log entry for a single authentication attempt. The sequence of numbers refers uniquely to images presented at login. . . 101

Figure 37 Success rates per day per system; participants were reminded to use the system less frequently during the second week and this affected success rates in the high entropy group. 102
Figure 38 The distributions of login durations recorded across both conditions across the user study. 104
Figure 39 The context of the observation attack component of our user study, participants could sit or stand. 105
Figure 40 Histograms of the login durations collected from legitimate users and those posing as observers. There is a clear difference between the login durations of legitimate users and imposter users. This could be used to inform design of a login timeout. 106

Figure 41 The classification of educational backgrounds of participants recruited to the description collection study. 115
Figure 42 The sequence of faces given to participants in the Description Collection study. 115
Figure 43 Randomly assembled grid: decoys are selected at random within the set of faces of the same gender (target face highlighted in red). . . 118
Figure 44 Visual groups: a grid of faces grouped by visual similarity (the target face is highlighted). 119
Figure 45 Illustrative verbal grouping procedure: Example tabulation of raw data for a cross-section of possible facial features. For face 2, all face numbers occurring on the same row as face 2, for any feature, are candidate decoys. 120
Figure 46 Verbal groups: a grid of faces grouped by verbal similarity (the target face is highlighted). 121
Figure 47 Example Passfaces grid in the description study. Participants were required to select the face to which an audio description refers. The interface widget below the image grid is the audio control panel. 121
Figure 48 Random vs Visual vs Verbal groups: A breakdown of the scores achieved in each condition. 123
Figure 49 Example faces: participants showed diverse performance depending on the image. 123
Figure 50 Comparison of male and female performance in the login task, 5/5 indicates a successful login. 125

Figure 51 Extremes of decoy image selection for the same key image: (left) decoys are semantically different; (centre) semantic and visual similarity to key image; (right) decoys are semantically similar yet different from the key image. 129

Figure 52 Points at which to consider image similarity across an example login. Red indicates a per grid requirement, and blue indicates a per login consideration. 132
Figure 53 Visualisation of the number of strong matches identified per image in the first study. A strong match is determined by a threshold upon the number of participants that must have classed an image pair as being similar. Overall: 1 participant=4424 strong matches; 2=2462 strong matches; 3=1425 strong matches, 14=148 strong matches. . . . 133
Figure 54 The key images chosen for the study, these were the same in all study conditions. 137
Figure 55 Example image grids assembled for key image #4 using similar, middle, and dissimilar decoy image criteria. 138
Figure 56 The number of login errors made per key image and per experimental condition. 139
Figure 57 Length of successful logins in each condition: top left) similar; top right) middle; bottom left) dissimilar; bottom right) overall. . . . 141
Figure 58 *Similarity matrix* that illustrates the pairwise EMD distance d between images in a single 3x3 image grid. This grid has been assembled with key relative filtering [102]. 143
Figure 59 (Left) grids assembled using minimum distance approach where $d = 2$; (right) similarity intervals where $4 > d > 0$. 144

List of Tables

Table 1 A collation of the user authentication quality criteria proposed by Renaud [100]. 29
Table 2 The race affiliation effect noted by Davis et al. [27], white females chose faces to comprise their graphical password within their own race 50% of the time. 48
Table 3 A collection of evaluations of recognition-based graphical passwords, with characteristics highlighted that appear relevant to the discussion of the future deployability of graphical passwords. 58
Table 4 A collection of evaluations of recall-based graphical passwords, with characteristics highlighted that appear relevant to the discussion of the future deployability of graphical passwords. 59
Table 5 A collection of evaluations of cued recall-based graphical passwords, with characteristics highlighted that appear relevant to the discussion of the future deployability of graphical passwords. 60

Table 6 Complexity of drawings created by participants in both experimental groups. 67
Table 7 Recall results for both experimental groups at the five minute recall test. 68
Table 8 Complexity of successfully recalled drawings in the five minute recall test. 69
Table 9 Recall results for each experimental group at the 1 week recall test. 69
Table 10 Complexity of successfully recalled drawings in the one week recall test. 71

Table 11 Principles for designing observation resistant user authentication. 79
Table 12 Observation attack resistance techniques used in methods proposed in this chapter and in related work (* = primary; + = supporting). 85
Table 13 Rounded percentage of logins where participants guessed a particular number of authentication components (138 attempts collected across all systems). 88
Table 14 Perceptions of PIN, Faces and Pressure Grid on a 1-5 likert scale where 1 is not very confident and 5 is very confident. 89

Table 15 Success rates and attempts recorded for each mechanism during week one where participants were requested to login twice per working day. 102
Table 16 Success rates and attempts recorded for each mechanism during week two, where participants were requested to login twice, every two days. 102
Table 17 Questionnaire results at the end of the two week study. 106

Table 18 Number of successful logins in the different experimental conditions. 122
Table 19 Examples of descriptions collected that provided interesting authentication behaviours. These descriptions refer to images in Figure 49. 124
Table 20 Female vs male performance in each experimental condition. . . 124
Table 21 Gender combinations of describer and listener and the number of successful logins that resulted with that combination (e.g. M:M = male listener using male description). 125

Table 22 Results from filtering procedure on an image set with 800 photographs, resized to 384x286. 135
Table 23 The number of key images correctly identified (out of four) in study two. Success is 4/4. 139
Table 24 Significant differences noted in user performance across experiment conditions on a per-image basis. 140

Chapter 1

Introduction

1.1 Overview

From the very dawn of computing, knowledge-based authentication (KBA) has been, and remains to be, the predominant technique by which a user's secure access to computer systems is realised. One study reports that the average user has approximately 25 online accounts that require passwords, and must enter an average of eight passwords per day [46]. In the face of the cognitive demands that result, users are known to respond by adopting strategies of credential selection and usage that can ultimately weaken the security of the systems they use [74]. Attempts to correct this subversive behaviour are based upon methods to *scaffold* [146] more desirable user authentication behaviours. That is, to provide advice or system features that direct user behaviour along a path that is thought to produce an outcome with security benefits; for example, implementing strict credential selection guidelines [65] proactive password checking [148] or password expiry [152]. However, research is beginning to show that these scaffolds can have either positive or negative effects upon usability and security, as designers seek to support or suppress particular user behaviours. Indeed, where interventions simply force unreasonable constraints upon user behaviour, users may become ever more likely to adopt new insecure workarounds to reclaim the missing usability [99]. It is well known that users are often not minded to follow the strict security guidelines relating to KBA that might be prescribed [155]. This depicts a scenario where system administrators and users are placed in a coevolutionary struggle; where the security and usability of the authentication system is paradoxically threatened – yet sustained – by the ability of its users to appropriate interactions with secure systems into a personal framework of coping with the cognitive load, which may inevitably have insecure components. This has led some to speculate that the current forms of KBA provide a poor fit to the socio-technical challenges inherent in contemporary user authentication, and will abandoned in the future [22].

Recently, the implications of research in human-computer interaction (HCI) [104], particularly relating to usability, have started to impact the computer security com-

munity, as security researchers proclaim their desire to design secure systems that people can use [25]. This still nascent field of *user-centred* security [1] advocates engagement with users in the design and evaluation of security systems considered to be socio-technical. One design response of the this community is the graphical password [124, 5], an alternative method of KBA that was designed from the outset with human cognitive limitations in mind and aims to provide a much-needed balance between usability and security. Graphical passwords are underpinned by the cognitive theory of the picture superiority effect [117] which suggests that a concept is more likely to be remembered experientially if presented as a picture rather than as words or numbers. Numerous such schemes have emerged, and end-user evaluations measuring memory retention of these new types of credential have showed considerable promise [11, 28, 143].

However, even where the underlying computer infrastructure is flexible (such as web-based interfaces) the uptake of graphical passwords is low. Inertia is likely, in part, caused by the extent to which existing methods of KBA are entrenched into our digital infrastructure, but also the gaps that remain in our understanding of how unanticipated positive and negative scaffolds placed upon user behaviour might mutate the socio-technical challenges currently experienced with existing methods of KBA. The increasing diversity of platforms and contexts that require user authentication: ranging from mobile devices, to desktop computers and touchscreen displays, also increases the difficulty to transfer insights from one platform to another. A myriad of other reasons could be posited, however an innovative empirical research approach is required to begin to explore the space of security and usability issues, threats, requirements and value that should be accommodated by contemporary methods of user authentication. Indeed, Sasse et al. [109] comment that security engineers should adopt a more holistic approach to the design and evaluation of security mechanisms that includes the context, the user, and the technology.

In this thesis we present system designs, empirical methods and results that explore methods of scaffolding authentication behaviours in exemplar graphical password systems, and evaluate the impact upon usability and security. We take a case study approach that explores the impact of interaction design and image choice; we also consider a number of previously unexplored contextual phenomena including insecure behaviour such as graphical password sharing and use of graphical passwords in (increasingly prevalent) shared public interfaces such as digital tabletops (where users are particularly vulnerable to observation attack). The aim of the research is to contribute to the set of empirical methods and results pertaining to the field of graphical password-based KBA, and to begin to provide a more holistic account of how graphical passwords can be a usable, secure and deployable method of user authentication.

[1] The field of user-centred security is sometimes interchangeably referred to as usable security, security usability, and others. In this thesis we typically use the name *user-centred security*

1.2 Research Context

This thesis contributes to the research areas of human-computer interaction and computer security. Relevant sub-domains include password security and usability, user-centred security, and graphical passwords.

1.2.1 Password Security and Usability

User authentication is usually the first stage of interaction had by users with a security sensitive system. Alphanumeric passwords are ubiquitous for this purpose; however, over time their adversarial model has changed significantly from providing a means of *security through obscurity* in the early days of computing when computers were not widely networked, to providing a source of entropy [115] to resist automated remote guessing attacks. However, early in this transition it became clear that the theoretical security benefits offered by passwords were not being born out in everyday system deployments. Morris and Thompson [87] collected a sample of 3289 passwords, where 86% of them were categorised as being shorter in length than six characters or appearing in a common word dictionary; their findings suggested that user-chosen passwords could be relatively easy to guess if an individual were inclined to try. Klein [74] did try, and carried out an experimental attack on a corpus of 15,000 UNIX passwords and discovered that using a typical word dictionary containing 62,727 words he was able to guess 25% of that password collection. The first real-world exploitation of these weaknesses was by the infamous *Morris Worm* [41, 122], malware which used a small attack dictionary of 432 words to seed password guessing to great effect, infecting 10% of computers on the Internet at that time.

To secure passwords for this threat, organisations have attempted to scaffold [146] behaviour that results in stronger (less guessable) passwords. Unfortunately, research has suggested that characteristics of strong passwords inhibit their memorability [155]. Indeed, while the research and institutional focus is placed upon the need to impose stronger passwords upon users, users in parallel develop practices that enable them to cope with those more complicated passwords e.g. writing down and reusing passwords. Such workarounds are difficult to regulate centrally and often prioritise usability at the expense of security which contributes to the adage that users are the weakest link in the security chain [109]. This scenario is perhaps the most prominent exemplar of the conflict between usability and security. After decades of KBA deployments, the security community is beginning to acknowledge the extent of these workarounds and the importance of holistic study of user authentication systems [109]. At present it can be said that password security and usability is at an impasse [60], due to the need for users to behave ever more securely in terms of password choice and management, while usability is already overexerted and contingent upon insecure coping practices.

1.2.2 User-Centred Security

The field of computer security has its roots in military systems where the adversarial model is such that users themselves are treated principally as threats to system security [153]. This meant that strategies to design secure systems were focused principally upon its conceptualisation as a technical problem. Recently, the study of HCI has diffused through the computer security community, with the aim to design secure systems that people can use [25]. This field of usable security advocates engagement with users in the design and evaluation of security facing mechanisms, with the motivation that user-facing security systems should be more openly considered as socio-technical systems. Indeed, in recent years Bruce Schneier, a prominent cryptographer was quoted in one of his books to acknowledge the wider concerns at play in security: *"If you think technology can solve your security problems, then you don't understand the problems and you don't understand the technology"* [111].

A number of seminal papers motivated the importance of changing the user-hostile mentality in the design of security systems. Zurko and Simon [153] first used the term user-centred security, Whitten and Tygar [141] carried out a usability study of the email encryption software Pretty Good Privacy (PGP) 5.0 with damning results. Adams and Sasse [1] argued that the need to know principle inherited by organisations from the military roots of the field should be abandoned, due to its detrimental effect upon user motivation to comply with security policies. This entry point for computer security into 1990s HCI resulted in a conceptualisation of the user as an information processing machine, their capabilities reasoned over using cognitive strengths and limitations, leading to abstract modelling of user decision making and behaviour e.g. Norman's seven stages of action [93]. The importance of considering the role of users in the security of digital systems has manifested in a the form of a dedicated ACM symposium: *Symposium on Usable Privacy and Security* (SOUPS) and increasingly, conferences traditionally focused upon systems and network security solicit work in this domain.

1.2.3 Graphical Passwords

The field of user authentication has arguably received the most research attention from the user-centred security community; this is due to the wide applicability of the problems associated with remembering securely composed passwords. The hope is that a user-centred approach to the design of KBA can alleviate the need for users to resort to insecure workarounds to cope with the cognitive load of remembering strong credentials. This has led to the exploration of new approaches to user authentication including that of graphical passwords [5, 124]; systems of this class are designed to complement rather than conflict with human cognitive ability due to their harnessing of a *picture superiority effect* [117] observed in cognitive psychology research. This has stimulated research into the design of systems that harness visual stimuli as an

authentication credential, first proposed in a patent by Blonder [7] in 1996 in which users were challenged to remember previously chosen areas of an image. Systems can be specifically designed to provide differing levels of entropy: PIN-level, password-level and crypto-level [5]. Exemplar systems have emerged and include the commercial Passfaces [96], Draw a Secret [68] and Passpoints [143]. A prominent research approach encompasses the theoretical analysis of proposed systems complemented by empirical studies to validate the memorability of the credentials. Seminal results include a field study of the Passfaces system across 12 weeks, where users of Passfaces made one third of the errors made by users of alphanumeric passwords [11].

While numerous studies of memorability have shown promise, fewer studies have explored the likely need for graphical password systems to incorporate mechanisms for positive and negative scaffolds to support or restrict context specific user behaviours; this reflects the reality that different deployment contexts will have different security and usability requirements. Renaud [101] considered the impact of methods of creating image content upon the memorability of the resulting images. Everitt et al. [42] carried out a longitudinal study of users asked to remember multiple Passfaces [96] graphical passwords. Hlywa et al. [62] explore the impact of different image types upon the usability of recognition-based graphical passwords. Such studies are promising, as they question the key assumptions regarding the extent that perceived benefits of graphical passwords would be born out under the constraints of everyday deployments. Such deployments could bring about novel attacks [54, 31, 97, 27], but also entice users to reformulate attacks and coping strategies trained upon alphanumeric passwords and PINs. One specific threat to which graphical passwords are perceived to be particularly vulnerable is that of observation attack: using simple observation techniques to capture the credentials of the user at the point of login. This is thought to be a particularly potent attack vector due to the difficulty to mask the image stimulus onscreen combined with its perceived memorability, which may make the credentials also memorable to collocated attackers. Such an attack bypasses the security gains of encouraging users to choose strong passwords. A number of graphical password protocols have been designed specifically to defend against this threat e.g. [144]. However, providing a balance between usability and security in this context is an open problem, due to the tendency for designed defences to force indirection into user interactions that are complicated for a user to perform, and an observer to capture.

1.3 Research Questions and Contributions

The research question for this thesis is *how graphical password systems can provide secure, usable and deployable knowledge-based authentication?* This can be decomposed into a number of sub-questions:

- How can interaction design scaffold secure behaviour in graphical password schemes?
- How can strategic selection of images provide scaffolding in graphical password schemes?
- What is the impact of graphical passwords upon socio-technical threats?

In the course of responding to these questions, in each chapter we make contributions to the field of graphical passwords and KBA more generally through novel designs, empirical methods and discussion. The main contributions of this thesis are as follows:

1. We highlighted the importance of the role of deployability when considering the usability and security of graphical passwords. Our conceptualisation of deployability comprised the explicit recognition of the problem of providing context-specific scaffolding to desired user authentication behaviours, and the impact of interventions upon usability and security. Our empirical studies explored how interventions could positively and negatively shape user behaviours.

2. We proposed a framework for supporting designers to scaffold user behaviours to defend against observation attack. By applying this framework to the contemporary problem of authentication on shared public multi-touch interfaces, we were able to articulate a design space of observation resistant augmentation to a standard graphical password.

3. We identified *description* as a threat to recognition-based graphical passwords and performed an empirical study to explore the ability of users to verbally share Passfaces [96] graphical passwords. We discovered that through the manual manipulation of similarity of decoy images and key images we could directly impact on users ability to identify key images from audio descriptions.

4. We proposed novel empirical methods to facilitate the study of socio-technical threats in laboratory-based studies, and the usability of systems outside of a laboratory environment. We measured user performance in matching audio descriptions to face images as an estimator for the prevalence of *description*. Also we refined a method proposed elsewhere to explore the observation resistance of graphical passwords on shared displays and mobile devices, asking users to perform as attackers and observe live authentication sessions. Finally we performed the first field study of graphical passwords on mobile devices, where our system was installed on the personal devices of participants.

5. We performed the first attempt to harness image processing techniques to identify instances of image similarity that could be detrimental to usability if included in a recognition-based graphical password grid. The need for a human

to perform this task is a severe limitation in terms of the deployability and scalability of this genre of graphical password. We conducted a usability evaluation of the best method we tested with over 300 participants recruited using Amazon Mechanical Turk where login challenges were assembled to include differing levels of similarity. We found that manipulation of the similarity levels allowed us to provide both positive and negative scaffolding of user authentication behaviours behaviours, highlighting the importance of image choice in the usability and security of graphical passwords.

1.4 Research Approach

To explore the research questions we adopt a case study approach, this follows from our position that overarching principles and insights for KBA are our ultimate goal, and the best route to these is often the detailed unpicking of concrete examples of authentication schemes (and their assumptions). These case studies were chosen to represent a cross section of issues faced with KBA in the past, but to also be mindful of the new challenges graphical passwords may pose: (i) the problem of users choosing strong credentials; (ii) observation attack; (iii) password sharing; (iv) image filtering. The latter came about as it was noted to be a particular challenge throughout the thesis, and thus a likely deployment challenge to future graphical password instantiations. To explore each phenomenon we chose a existing graphical password system that we believed would provide the greatest insight, due to the lack of a consensus around which system is the most promising. The sum of these case studies amounts to our consideration of the challenges of usability, security and deployment challenges facing graphical passwords (a visual overview is presented in Figure 1). Our methodology is primarily empirical and quantitative, and lends from techniques prominent in the HCI community: *contextual inquiry* [104]: A method of data gathering for system requirements that involves a combination of discussion and observations made with the potential users of a system; *heuristic evaluation* [92]: the evaluation of a system by experts based upon its adherence to particular usability standards or guidelines; *human subject experiments*: principles of experiment design are applied to perform hypothesis testing regarding the usability of system interventions. In some cases we were interested to capture the experiences of participants using the systems we designed, in those situations we would use questionnaires and informal discussions to capture this feedback. The length of each study was chosen to appropriately capture the phenomenon under study. Participants were recruited opportunistically in a university or organisational setting through word of mouth and other message distribution lists where the research was being conducted, as a result this means people not present in those settings (such as older people) are under-represented in this research.

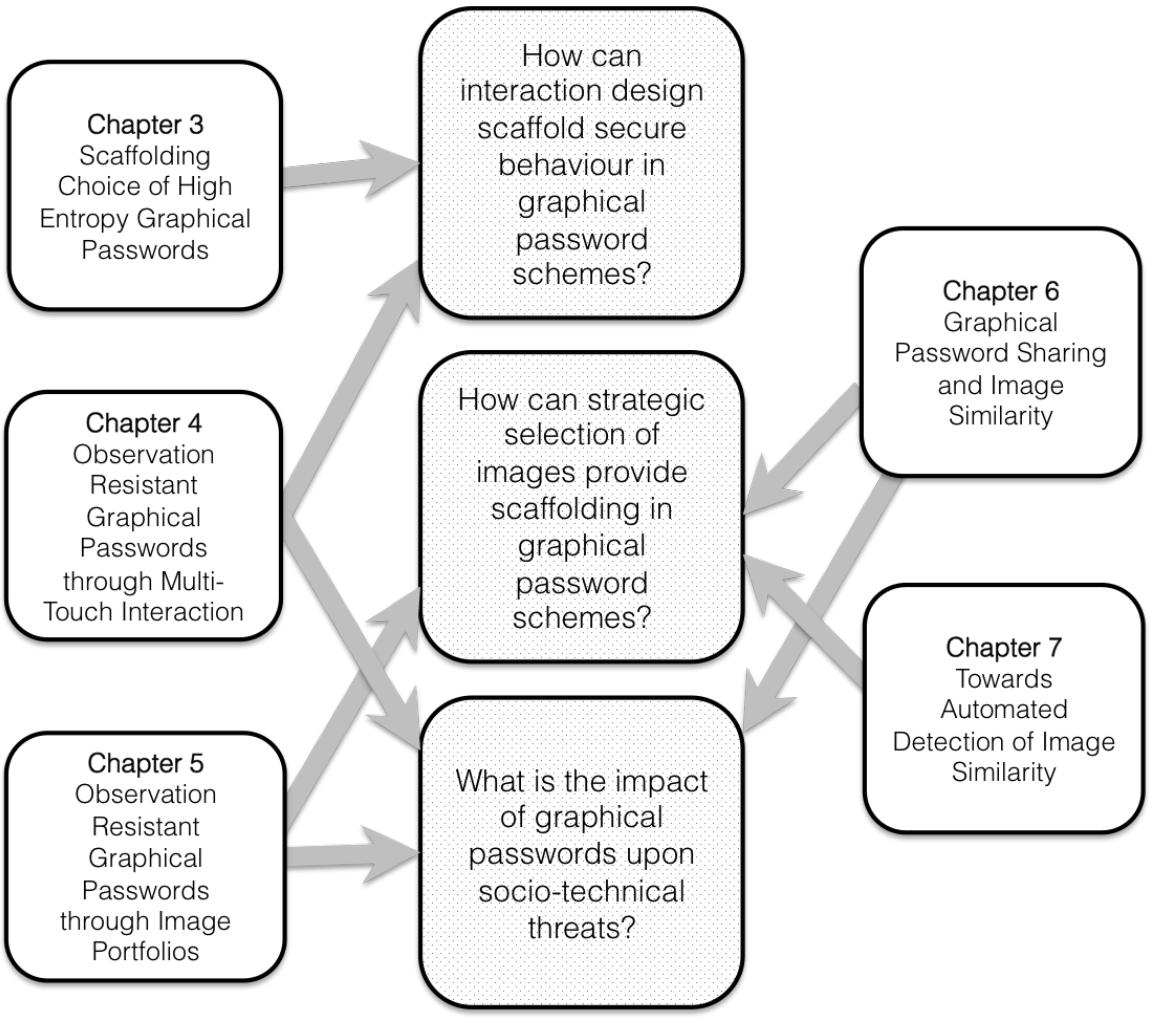


Figure 1: A visual description of the thesis structure.

1.5 Thesis Structure

The thesis is organised as follows: in Chapter 2 we present a literature review that encompasses relevant aspects of user-centred security, the security and usability of alphanumeric passwords, and the current state of the art of graphical passwords. The focus of our review of graphical passwords places a particular emphasis on the nature of empirical studies conducted with representative schemes, and analysis of usability and security properties.

Chapter 3 comprises the first case study into how interaction design can be used to scaffold the choosing of strong graphical passwords. Draw a Secret (DAS) [68] is a graphical password scheme designed to enable usable generation of memorable credentials that have high levels of entropy [115]; credentials that, for example, can be used in authentication or to seed the generation of encryption keys. Previous research has highlighted the fact that theoretical estimates of DAS security may not be born out in practice as users are likely to select weak drawings that exhibit predictable (i.e. guessable) characteristics [130, 131]. To address this we proposed and evaluated a novel DAS variant called Background Draw a Secret designed to encourage users to

avoid the user choice biases predicted for DAS, through the use of added visual cues in the drawing grid. Our empirical evaluation sheds light on the resulting patterns of user choice encountered in both systems.

Chapter 4 describes a different case study for interaction design being used to scaffold secure user behaviours, in which we focused upon how harnessing (and designing for) the affordances of particular technologies can facilitate secure entry of KBA credentials in the face of observation attack. The class of KBA system we considered is less concerned with the generation of high levels of entropy, however inherits a vulnerability to observation attack due to both, the simple interactions required, and the memorable nature of the visual stimulus presented. Traditionally, the responsibility to secure such systems against observation attack is placed on the user, through behaviours that are often difficult to maintain in the face of pressure to adhere to social norms. In this chapter we propose a framework for observation resistant interactions for KBA in general and explore the design space for authentication on inherently social shared multi-touch interfaces, a technology of contemporary interest that has been explicitly designed to support collocated users.

In Chapter 5 our focus shifts slightly to question of how manipulation of the frequency that images are presented to users can scaffold desirable security behaviours. We again concern ourselves with observation attack and recognition-based graphical passwords, but consider the problem of the observation resistance provided by an image portfolio-based system [28, 31], in which users are challenged to identify only a random subset of their secret images at every login. Through an empirical study across two weeks we evaluated the usability of this approach and showed its resistance to intersection attack [31].

In Chapter 6 we identify a new threat to recognition-based graphical password systems, which stems from factors relating to the visual and verbal image similarity of the constituent images, and to elaborate this we report our exploration of users ability to share Passfaces [96] graphical passwords. The results from our novel empirical study suggest a relationship between the level of similarity and the associated difficulty of sharing graphical passwords by description (i.e. verbal description).

Chapter 7 considers a question that gradually emerged as we progressed through the project as a whole: how to systematically assemble image grids that resemble a usable and secure login? In answering this question we explored the role that image processing can play to optimise image selection (i.e. the systems choice of key and decoy images) with respect to the detection of image similarity. In our user study we identified colour histograms as being useful image signatures, and assembled login challenges based upon judicious manipulation of thresholds to affect image similarity. In an empirical study we measured user recall performance using image sets assembled in an automated manner, and explored its impact on both usability and security. Finally, in Chapter 8 we revisit our contributions, consider the findings of our case studies in relation to each other, and propose some promising and pressing directions

for future work in this domain.

1.6 Prior Publications

Significant portions of this PhD research have been previously presented in peer-reviewed academic conferences:

1. Dunphy, P. and Yan, J. 2007. Do background images improve "draw a secret" graphical passwords?. In Proceedings of the 14th ACM Conference on Computer and Communications Security. CCS '07. ACM, New York, NY, 36-47.

2. Dunphy, P. and Yan, J. 2007. Is FacePIN secure and usable?. In Proceedings of the 3rd Symposium on Usable Privacy and Security. SOUPS '07, ACM, New York, NY, 165-166.

3. Lin, D., Dunphy, P., Olivier, P., and Yan, J. 2007. Graphical passwords & qualitative spatial relations. In Proceedings of the 3rd Symposium on Usable Privacy and Security. SOUPS '07, ACM, New York, NY, 161-162.

4. Dunphy, P., Fitch, A., and Olivier, P. 2008. Gaze Contingent Graphical Passwords at the ATM. In Proceedings of the 4th COGAIN Annual Conference on Communication by Gaze Interaction. COGAIN '08.

5. Dunphy, P., Nicholson, J., and Olivier, P. 2008. Securing passfaces for description. In Proceedings of the 4th Symposium on Usable Privacy and Security. SOUPS '08. ACM, New York, NY, 24-35.

6. Dunphy,P., Heiner, A.P., N. Asokan. 2010. A closer look at recognition-based graphical passwords on mobile devices. In Proceedings of the Sixth Symposium on Usable Privacy and Security. SOUPS '10. ACM, New York, NY, USA.

7. Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J., Nicholson, J., and Olivier, P. 2010. Multi-touch authentication on tabletops. In Proceedings of the 28th international Conference on Human Factors in Computing Systems. CHI '10. ACM, New York, NY, 1093-110.

8. Nicholson, J., Dunphy, P., Coventry, L., Briggs, P., Olivier, P. 2012. A security assessment of tiles: a new portfolio-based graphical authentication system. In Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts (CHI EA '12). ACM, New York, NY, USA, 1967-1972.

9. Dunphy, P., Olivier, P. On Automated Image Choice for Secure and Usable Graphical Passwords. In Proceedings of the 28th Annual Computer Security Applications Conference. ACSAC 12. ACM, New York, NY, USA.

Chapter 2

User-Centred Security and Graphical Passwords

Given the scope of social, psychological and technical concerns that must be embraced in the development of usable and secure user authentication mechanisms, our research inevitably adopts an inter-disciplinary approach. To this end, this chapter reviews a number of research sub-domains that relate to our proposed research approach: the emergence of user-centred security; user authentication; and, the state of the art of graphical passwords.

2.1 User-centred Security

We presented an overview of the research domain of user-centred security in Section 1.2.2, which briefly highlights key learnings from the field; in this section we expand upon that description to include more detail of seminal research results that shaped thinking in this nascent field.

The contention that the constraints placed upon end-users should be considered in the design of secure systems was born out of a retaliation to the user-hostile roots of computer security in the military sector. Indeed, the perspective of security engineers towards users has traditionally been that of the user-as-the-problem; one example of this perspective is provided by Herschberg:

> *"They range all the way from very occasional users, contacting their systems once or twice a day almost incidentally, up to very professional users sitting at their terminals all day. Still, these users appear to have one characteristic in common: they are all, men and women, rather devoid of imagination"* [61] (p. 1).

The foundations of the research community that challenged this hostile perspective were laid by Zurko and Simon [153] who called for a user-centred approach to the design of security models, systems, and software. Their motivation was to highlight

the importance of usability in the design of secure systems across a wide range of user stakeholders: end-users, application programmers, system administrators and even social units. This recognised that poor usability of system features can unintentionally undermine the security of a system through the resulting impact upon work practices and social norms. However, prior to the formal assembly of a purpose-fit research community, a number of researchers in more traditional fields of security had already highlighted the need to design secure systems usable by humans. Wood suggested that *"One of the fundamental trade-offs found in many areas of the computer security field is ease of use versus security"* [145] (p. 3). Also, as early as 1975 Saltzer and Shroeder coined the term *psychological acceptability* to describe how users should be able to interact with security mechanisms:

> *"It is essential that the human interface be designed for ease of use, so that users routinely and automatically apply the protection mechanisms correctly. Also, to the extent that the user's mental image of his protection goals matches the mechanisms he must use, mistakes will be minimised. If he must translate his image of his protection needs into a radically different specification language, he will make errors."* [108] (p. 6).

Saltzer and Shroeder also identified design principles that have remained a fixture in discourse on computer security, including: the principle of least privilege, open design, and fail-safe defaults. In 2005 Bishop [6] argued that three areas of computer security particularly violated psychological acceptability: passwords, patching and configuration. He suggested that while reaching the goal of psychological acceptability might be impossible, there is still scope to improve the way that security systems are designed and deployed. Indeed, Adams and Sasse [1] claimed that the lack of understanding between security administrators and users contributed to unsatisfactory outcomes with respect to security. They propose that to improve, the culture of *need to know* inherited from the military sector should be abandoned in favour of practices that motivate and encourage security awareness amongst users. Research methods proposed to meet this challenge were drawn from the field of human-computer interaction (HCI) and included contextual design, lab-based experimentation involving users, and contextual enquiry [104]. In 2005 the field came into formal existence with the release of Cranor and Garfinkel's edited collection of papers on usable security [25], and the first Symposium on Usable Privacy and Security (SOUPS) which has taken place every year since.

One source of relief for the security community in the face of this new wave of insight was the assumption that the use of cryptography was one system feature that could not be undermined by users; however, this assumption was soon under attack. Whitten and Tygar [141] present a seminal usability evaluation of the Pretty Good Privacy (PGP) email encryption software, which demonstrates that users could make potentially dangerous errors during typical usage. The creators of PGP had claimed

that it was usable by novice users, however: three (of a total of 12) participants unintentionally sent emails in plain text, seven sent email using the wrong encryption key, and one participant was unable to encrypt at all. The results of this study prompted Whitten and Tygar to identify properties of secure systems that make the design of usable and secure systems difficult:

- *The unmotivated user property*: user interaction with security is usually a secondary task. Designers should not assume that users will be motivated to learn the required skills to use a system correctly.
- *The abstraction property*: interaction with security-facing systems usually requires interaction with abstract concepts such as choosing rules for access control. Such procedures are intuitive to engineers but not to inexpert users.
- *The lack of feedback property*: the current state of a security system can be complicated to articulate or visualise in feedback to an inexpert user.
- *The barn door property*: once a user has made a mistake in configuring a security system, recovery can be difficult (if not impossible) where data may have been leaked.
- *The weakest link property*: a system is only as secure as its weakest link, which is often the user.

In the context of Public Key Infrastructure (PKI), Davis [26] highlighted that much of the promised security and organisational benefits of a PKI were obtained by shifting responsibility for critical operations onto end-users, who could not in practice be relied upon to execute them correctly. Davis highlighted the existence of what he termed *compliance defects*, rules of operation that are both difficult to follow and unenforceable, and recommended that non-experts should not interact directly with a PKI. In a similar vein, Anderson [2] provided concrete examples of how user errors or insider involvement led to security breaches in the banking industry, where the use of cryptography was a principal security feature. These cases served as warning to the security community that a reliance purely upon cryptography did not provide a guarantee of a secure system.

The early discourse on user-centred security provides critique of the relationship between users and security, and highlighted the difficulty of designing systems that were both theoretically and practically secure. One domain that particularly received research attention to ground such critique is that of user authentication, due to it being a task universally encountered by users, and the notable usability and security issues at play.

2.2 User Authentication

User Authentication [120] is usually the first security procedure encountered by users of a security-focused system; during this process the system challenges the user to prove their right of access. The process begins with *identification*, where the user makes a claim of identity, and is followed by *authorisation*, where the user provides credentials that prove the claimed identity [100]. The credentials are usually based upon one of three *factors* of authentication: something you *have*; something that *is*; or something you *know* [120, 100]. *Something you have* authentication is typically based upon the possession of a digital authentication token, which is particularly common in the context of enterprise security. One user study of smart cards and USB tokens highlights usability issues relating to their mobility in terms of the diverse range of form factors, and the associated need to install drivers on each new machine that should accept the tokens [98]. Biometrics measure human physiological or behavioural characteristics and are often assumed to be the best solution, as they present the perception that user authentication can be achieved purely through technical innovation. However, in practice biometrics present their own set of usability and security issues, in particular, the stability and salience of the measured characteristic across diverse user groups [24]. *Something you know* [1] authentication typically requires the user to remember a an alphanumeric secret, and prove knowledge of this secret to access a system. This approach is ubiquitous on the Internet as it is convenient to deploy, and the credentials are easy to replace. However, the increase in systems that harness this method of authentication creates a high cognitive load for users who may forget the association of passwords to user accounts, or even the password itself. Attempts to cope with the cognitive load include users choosing passwords that are easily guessable [74], or writing them them down.

In addition to choosing appropriate factors of authentication, there are various system architectures that impact system design: local authentication, direct authentication, indirect authentication and offline authentication [120]: *local authentication* is where the authentication and access control mechanism both reside within a single security perimeter (e.g. a PIN on a mobile device); *direct authentication* is where users authenticate directly to a remote service, but communicate with the service remotely via a system in a different security perimeter (e.g. logging into a website); *indirect authentication* is where there are many points of service and a single point of authentication (e.g. Kerberos [89]). *Offline authentication* is where authentication can be completed without an active connection to a service (e.g. PKI).

[1] Often referred to as knowledge-based authentication

Category	Features
Accessibility	Special Hardware or Software
	Convenience
	Inclusivity
Memorability	Retrieval Strategy
	Meaningfulness
	Depth of Processing
Security	Predictability
	Abundance
	Disclosure
	Confidentiality
	Privacy
	Crackability
Cost	Hardware
	Software
	Key Replacement
	Maintenance

Table 1: A collation of the user authentication quality criteria proposed by Renaud [100].

2.2.1 Summary

User authentication is a fundamental process required to secure computer systems for unauthorised access. The process of selecting the appropriate factor and architecture has been considered by Renaud who proposed a framework to aid such decision making [100] (a summary of this framework is presented in Table 1). However, despite widespread technical innovation and the diversity of potential solutions, the bottom line for most organisations is the financial cost of a solution; this serves to partly explain why alphanumeric passwords and PINs are widely used despite shortcomings in terms of both usability and security. Not only are alphanumeric passwords and PINs easy to integrate in a system, but users themselves are well accustomed to their usage, and their limitations are perceived to be well understood by organisations. In the following section we further explore the security and usability considerations brought about by the ubiquitous deployment of knowledge-based authentication.

2.3 Alphanumeric Passwords

Alphanumeric passwords (herein passwords) have been widely used to secure computers for unauthorised access since the time of the *Compatible Time Sharing System* (CTSS) at the Massachusetts Institute of Technology. Today, the need to authenticate users in a familiar manner across a diverse range of service providers has firmly placed the password as the authentication method of choice. The study of passwords in research literature can broadly be separated into studies of security, the impact of security upon usability, and contextual studies of password usage.

2.3.1 Security

Passwords can serve two functions in the security of computer systems [86]: *authentication*: to validate the identity of a user based upon knowledge of a shared secret; and *key generation*: to provide a source of entropy (as termed by Shannon [115]) to seed the generation of cryptographic keys. Today, it has become increasingly important that the same password is usable and secure in both modes of operation. However, in any context of use passwords are vulnerable to a range of attacks:

- Replay Attack: the password of a legitimate user is reused by an unauthorised person.
- Social Engineering: an attacker persuades a legitimate user to reveal a password [84].
- Observation Attack: an attacker captures a password using simple observation techniques at the point of password entry.
- Phishing: the password is captured via an interface that is maliciously designed to masquerade as a trusted entity (this usually culminates in a replay attack) [66].
- Brute Force Attack: an automated guessing attack where every possible valid password is tried until the correct one is discovered. This can take place online, where guesses are mediated by the system under attack (against which certain automated defences can be realised) or offline where the attacker has obtained a copy of an encrypted password database and can make an unrestricted number of guesses. This is guaranteed to succeed, but could take a prohibitively long time.
- Dictionary Attack: an optimised version of a brute force attack that uses the assumption that password content is likely to contain a word found in a typical word dictionary [74]. This can also be executed online or offline.

From the earliest days of password usage, automated guessing has been prioritised as the most potent attack. Automated tools can be configured to make many thousand password guesses per second, whereas a human may be able to make only one or two. The threat of automated attacks meant that without stipulation over the complexity of a password, little confidence could be had in the identity of the individual who would present it to a system. In order to reason over this required complexity, measures of password strength are computed in terms of entropy based upon the seminal work of Shannon [115]. Where each character of a password is chosen randomly, entropy H (measured in bits) can be calculated by $H = log_2(a^l)$. Where a represents the size of the alphabet from which each character is drawn (e.g. $a = 26$ where the password is comprised of lowercase alphabetic characters), and l is the length of the

password. However, if it were demonstrated that passwords were chosen in a non-random manner, this computation of entropy would overestimate password strength in practice.

Morris and Thompson [87] present early reflections upon password security through the analysis of the content of over 3000 UNIX passwords. Strikingly, they observed that users were very likely to choose predictable passwords, as many were very short in length, with 15 comprised of only a single ASCII character. Also, as a significant proportion of the passwords were comprised of pronounceable words, they propose that a dictionary attack could be a particular threat, and tenuously conclude that *"The use of encrypted passwords appears reasonably secure in the absence of serious attention of experts in the field"* (p. 4). This concedes that the security of passwords was based purely upon the hope that nobody would try to exploit their clear weaknesses. In an experiment, Klein [74] carried out such a dictionary attack on a corpus of collected UNIX password files and discovered that using a dictionary of only 62,727 words he was able to guess 25% of 15,000 passwords. Arguably, the first real-world execution of a dictionary attack was by the infamous Morris Worm [122], which infected over 6000 machines [41] using an attack dictionary of just 432 words as seeds for guesses.

Measuring password entropy in the presence of such user-imposed bias relies upon an understanding of the particular biases that might manifest in a user population. A report on password choice by the National Institute of Standards and Technology (NIST) [12] bases estimates of the entropy of user-chosen passwords upon Shannon's analysis of the entropy of English text [116]. This analysis suggests that the first character in a password provides 4 bits of entropy; the next 7 characters provide 2 bits each; characters 9 to 20 provide 1.5 bits each; and beyond 21 each character has 1 bit. For example, a strong eight character password where characters are drawn randomly from the full ASCII character set could have 53 bits of entropy, but a user-chosen eight character password may only have 18 bits. While the NIST analysis has a certain illustrative value, estimates could be refined to account for the idiosyncrasies of a particular user sample.

The typical attack scenario is that the attacker may wish to compromise the password of a particular user in order to abuse their associated system access privileges. In a batch guessing attack [48] the attacker has no specific user target, and so chooses a password considered to be common amongst the user population and uses it to guess the password of a sequence of known users (user identifiers, as compared to passwords, are more readily accessible in many systems). Single Sign On (SSO) is a class of software tool designed to store all user passwords, and provides functionality to provide the correct password to the correct online service. This potentially removes issues of memorability, although a user must instead remember a single master password to access the SSO system. Chiasson et al. [20] carried out a usability evaluation of two SSO systems but found that users reported usability issues, which included the lack of visibility of when the tools were working. Other disadvantages concern portability, as

SSO software must usually be installed on each system that it is used; also users tend to have difficulty to trust SSO software with their credentials, as this is a convenient instantiation for malware to steal large numbers of passwords.

The literature on password security amply demonstrates that when users are unconstrained in password choice, they are unlikely to create passwords desirable from a security perspective. Indeed, that security is contingent upon behaviour that users are unlikely to display, is a fundamental limitation in the design of a system, and a clear example what Davis termed a compliance defect [26]. However, rather than simply abandon password authentication at this point as being inadequate, the research focus changed to explore how user behaviour could be scaffolded (within the limits of existing infrastructure) to prefer secure passwords. That is, to develop interventions that support users to choose from a class of more desirable passwords.

2.3.2 Scaffolding Choice of Secure Passwords

The observed tension between usability and security in relation to alphanumeric passwords has led to the proposal of a number of interventions that attempt to find an appropriate balance. Such proposals range from security guidelines, technological interventions and self-help strategies; all of which aim to add diversity to user choice of passwords with minimal changes to existing infrastructure.

2.3.2.1 Password Guidelines & Proactive Password Checking

One early Federal Information Processing Standard (FIPS) provided guidance on the level of password entropy that should be enforced for contexts that require low, medium and high levels of security [44]. NIST [12] proposed that password composition rules enforced by *proactive password checking* [74, 148] can help to increase the entropy of chosen passwords, and should be a solution of choice for businesses seeking to protect their data from unauthorised access. Proactive password checking is a process by which a system prevents users choosing passwords judged to be weak at the point of enrolment, before acceptance into a system. The most common way to judge weakness in this context is a dictionary attack, where the user chosen password is searched for within a list of passwords judged to be predictable (i.e. those containing common pronounceable words). However, there is a trade off between the extensiveness of the checking and the requirement that, from the point of view of the user creating the password, the checks are completed within a reasonable period of time. Yan [148] claimed that the use of the dictionary attack as a key determinant of password weakness may still miss passwords that could be considered predictable; in response he advocates an approach based upon analysis of widely used determinants of password entropy (i.e. password length, use of upper and lower case characters etc.). Proctor et al. [99] carried out a study of the impact of imposing these compositional restrictions on the security of the resulting passwords chosen by users. In a user

study they discovered that increasing the length restriction on the passwords users were asked to create from five to eight characters reduced the percentage of passwords they were able to guess by 42% when using the *John the Ripper* [94] password cracking tool. Also they observed that their intervention increased the time it took for users to choose a password that satisfied the prescribed criteria. Finally, a qualitative analysis of the passwords they collected shows that users tended to satisfy the need to add complexity to passwords in predictable ways which could be exploited by an attacker (e.g. adding numbers to the end of a password to increase its length). Komanduri et al. [76] also carried out a user study considering the entropy of passwords produced when users were placed under various password choosing constraints, and discovered that the most strict password composition policies did not produce the highest entropy passwords. In their study, a group of users constrained only by a minimum length requirement of 16 characters in the password, chose higher entropy passwords (44.67 bits) than those constrained to an 8 character password where characters were drawn from the full ASCII set (34.3 bits). Both estimates were made to incorporate possible biases using Shannon's estimates of the entropy of English text [116]. Despite this interesting finding, they did however note a relationship between the increasing entropy of the password a user would choose, and the increased likelihood of the user writing it down. Both studies indicate that password composition guidelines focusing upon a minimum length requirement alone, appears to provide the most usable means to encourage users to increase password strength. However, both provided warnings of that the nature of coping techniques that users may adopt to cope with the added cognitive load may undermine security.

2.3.2.2 Mnemonic Passwords

Mnemonic passwords are a self-help technique often advocated as a usable approach to support users to choose high entropy passwords; the deployment benefit attached is that no significant modifications to existing system infrastructure is required. Mnemonic passwords are based upon a phrase, for example: a bird in the hand is worth two in the bush which can serve as a mnemonic for the password abithisw2itb!. Kuo et al. [77] examined the security of user choice of mnemonic passwords using the John the Ripper password cracking tool, and were able to guess 4% of 144 passwords compared to 11% for the passwords created in a traditional manner by a control group. Unsurprisingly, their analysis also shows that many of the base phrases used to generate mnemonic passwords were taken from widely known sources, such as song lyrics or well known lines from movies, potentially leaving them vulnerable to appropriately configured guessing attacks. Consequently, they propose that a good phrase should not be based upon the text from such sources. Yan et al. [147] also carried out a study with 288 participants across a university term to evaluate the memorability and security of user-chosen mnemonic passwords, contrasting these with both pass-

words based upon standard selection advice and randomly generated passwords. A dictionary attack against all collected passwords was able to compromise over 30% of user selected passwords based upon standard advice, but only 14% for the combined random and mnemonic groups. As in previous studies, users described randomly generated passwords as being significantly more difficult to remember, thereby lending further support for the more widespread utilisation of advice that encourages mnemonic-based password use. In addition the authors noted a significant number of occurrences (10% of participants) where participants did not comply with the password selection advice they were given; another clear reminder that it is necessary to study how password systems are used in practice and that systems should take account of compliance issues [26].

The base phrase chosen to seed a password mnemonic has itself been proposed as a secure and usable *passphrase* [78]. This is based upon the observation that the very length of an entered phrase mitigates against guessing attacks as the password space increases exponentially with each additional character. Porter [145] proposed that passwords should have at least 64 bits of entropy and based upon a passphrase of up to 80 characters. Of course, remembering and typing long passwords is inevitably associated with usability issues. In a field study that explored the use of passphrases, Keith et al. [70] found that the use of passphrases caused a significant increase in the number of login errors when compared to a group of users authenticated by conventional passwords. In addition, they report that the source of most failed authentication attempts could be attributed to typography errors, rather than users failing to remember the passphrase correctly. When ignoring typography errors, the success rate of those using passphrases increased by 14%. This adjustment placed the success rate of passphrase users above that of participants using passwords under typical password choosing constraints.

2.3.2.3 Cognitive Passwords

The research presented so far documents methods to improve the security of passwords that leave existing password infrastructure undisturbed; other approaches have considered how reconception of the password protocol could serve a usability purpose. *Cognitive passwords* require the user to answer some form of question to gain access to a system. Smith [145] considers the use of word association (or associative passwords) where the user password is comprised of a series of single word responses to a sequence of stimulus words e.g. where black is the stimulus, white could be a candidate user response. More recently Jakobsson et al. [67] propose a protocol based upon the user recalling preferences expressed at enrolment (e.g. do you prefer cats or dogs?); the design rationale being that preferences have been shown to be relatively stable in psychological studies. In a user evaluation they explored the entropy of the responses gathered from particular questions and reported low error rates when the stability of

the preferences was tested upto 14 days later. Zviran and Haga [154] carried out a questionnaire-based empirical study of 106 users to explore the usability of cognitive passwords, in which users are provided with a cue for passwords using a question that relates to some aspect of their personal experience or personal history, for example, what is the name of your first school? The result suggested that participants better retained cognitive passwords than two modes of alphanumeric password (self-chosen and randomly assigned).

Zviran and Haga [155] conducted a 103 participant user study to compare each of the aforementioned forms of password: cognitive passwords (comprised of 20 items), associative passwords (comprised of 20 items), passphrases, self-chosen passwords and system generated passwords. In their within-subject study, each participant enrolled with each password type on the first day; after a three-month delay participants returned and were tested on memory retention for all assigned credentials. The results show that 27% of participants could remember their self-generated password, compared to 13% for a system assigned password and 21% using a passphrase. Unfortunately, subjects were assigned multiple cognitive password items but only a single alphanumeric password, which raises concerns of how comparable these results might be in practice. However, on average participants recalled 15/20 of their cognitive password items, and recalled on average 14/20 of associative password items. It is interesting to note that in a post-study questionnaire subjects expressed a preference for the method that did not correspond with predications made based upon their measured performance; participants ranked self-generated passwords their favoured approach despite the overall success rate (i.e. 27%) being significantly lower than for other approaches (i.e. cognitive passwords).

While cognitive passwords and associative passwords promise usability benefits, they may potentially be more vulnerable to guessing attacks through social engineering, and are more likely to result in low entropy credentials due to the password response bring constrained by a question. Although widely used in commercial systems, such approaches are generally only deployed as an additional layer of security in relation to password reset protocols (and password hint provision). Indeed, recent research has suggested the information used in cognitive questions is not as secret as may be perceived, and that even unsophisticated attackers can be effective at guessing the answers to those questions [110].

The search for a method to support user choice of secure passwords has produced a number of candidate solutions that attempt to balance usability and security. Renaud [100] provides a user-centred framework to assist designers in making a suitable trade-off between accessibility, memorability, security and cost. However, in order to apply such a framework it is important that system designers understand the trade-offs being made by users themselves in response to contextual factors of everyday life, such as the existing cognitive load imposed by other systems (e.g. the number of systems for which they need unique passwords, and the relative infrequency of use

of some systems), and organisational factors (such as forced password expiry) that impact the password management problem. So far, the studies we have presented have involved the study of systems in isolation from this surrounding phenomena, however an interesting branch of password security and usability research has focused upon documenting the everyday constraints placed upon users in everyday interactions with password systems.

2.3.3 Studying the Usage Context

Our review of alphanumeric passwords has shown that a widely explored approach to finding a balance between the usability and security of passwords is to develop methods to facilitate (or force) users to choose strong passwords. However, many of these studies of password security have been conducted without taking account of the experiences of users and the realities of everyday deployments that can constrain usability and security. Sasse et al. [109] conducted a qualitative study that explored socio-technical issues centred upon passwords in a large organisation. The findings shed light on a diverse and conflicting array of concerns at play between an organisation and its users; their analysis is framed within four domains: (i) technology; (ii) user; (iii) goals and tasks; and (iv) context. Participants in the study had an average of 16 passwords to manage and they discovered that users experienced the most password resets related to rarely used passwords (see Figure 2). In a related investigation, Inglesant and Sasse [65] conducted a diary study and follow-up interviews to explore the impact of password policies upon people working in an organisation. They suggest that time spent on actions such as devising passwords, responding to password expiry, and password reset is not emphasised sufficiently as time employees spend unable to do their jobs. In conclusion they suggest that the institutional focus upon password strength and frequently changing passwords is counterproductive, and highlight the importance of studying the context of use when designing security policies to ensure compatibility with work practices.

The abundance of computer systems with which typical users engage (in their work and non-work activities) and the corresponding high number of secure passwords they are required to manage, is a key source of pressure that leads to insecure password practices. Florencio and Herley [46] report online user password habits recorded via a software component bundled with a version of *Windows Live Toolbar*. They found that an average web user had 6.5 passwords and 25 accounts that require password authentication (these findings were from a study conducted in 2007 and these figures are likely to be an underestimate for 2012). While proposals for just how many passwords a user should be asked to remember have hovered around seven (as argued by Miller in the context of human information processing [83]), there is no widely accepted figure. Inevitably, users will find workarounds to the near impossible task of managing such large numbers of passwords; Chiasson et al. [20] reports that 26 of

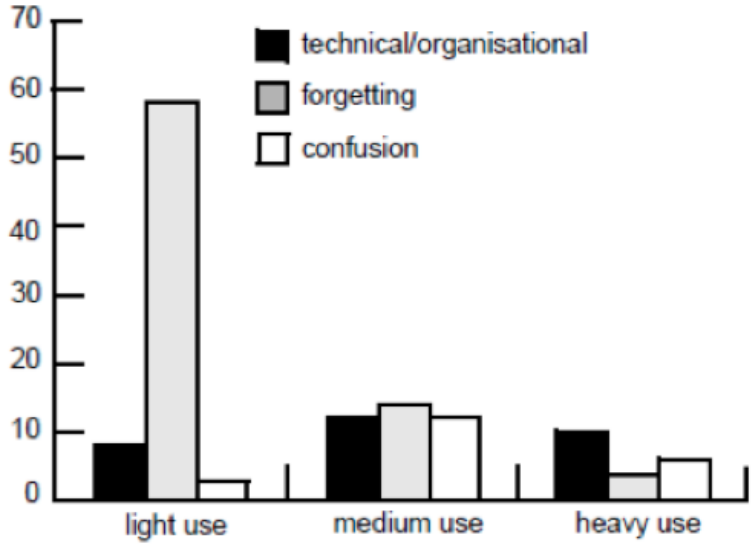


Figure 2: Illustration of self-reported data where users were asked to remember the cause of their last password problem, and the regularity at which they used that password (as reported by Sasse et al. [109]).

27 participants in their study claimed to reuse passwords across websites. Research also documents other workarounds such as sharing credentials [28, 69], writing down passwords [147, 154] and basing new passwords upon old passwords to aid memorability [1]. Florencio and Herley [47] explored the nature and enforcement of password security policies across a corpus of websites. They discovered that those websites with the most users, were the least likely to impose onerous password selection guidelines on its users. This suggests that those sites with more active users (usually large companies) were more concerned to ensure customers had easy access to the site rather than less widely known sites, who appeared more likely to impose strict password selection guidelines upon its users.

Singh et al. [118] describe password usage in scenarios where authentication credentials were shared amongst diverse social groups. In a striking account of the practices of a group of indigenous Australian islanders, they describe how banking transactions were delegated to a single person who would board a flight to visit the bank on the mainland. Such examples serve to show how sharing of secret security credentials can be appropriated as an act of trust, and also be necessary for survival. While such behaviour may seem extraordinary to designers, it provides an important glimpse of how social imperatives can encourage non-standard security behaviour.

Studies of contexts that have password deployment as a key feature use have yielded insights into the nature of password use that could not have been uncovered by lab-based studies and mathematical analyses alone. As we have seen, how security mechanisms are supposed to be used and how they are used in practice can often be very different [147].

2.3.4 Summary

Computer users have more passwords than ever before and approaches to securing alphanumeric passwords for threats related to users' lack of motivation to comply with security policies have been explored. We have highlighted that the security of the password is based upon the assumption that users comply with password selection guidelines. Studies have highlighted that password selection biases can be exploited where a bias towards pronounceable words exists; methods to scaffold secure behaviour, such as proactive checking have been seen to create a space for new workarounds. However, while passwords have a number of widely understood limitations, they remain to be cheap and convenient for security administrators and easy to understand for users. Our review of the principle proposals to scaffold secure behaviour in the user groups of passwords does not contain a panacea, as technical interventions may not be trusted [20] and self-help techniques may be ignored [147]. Rather than simply scaffolding the use of standard alphanumeric passwords where usability is already overexerted, an alternative is to design new usable knowledge-based authentication methods that have memory requirements that complement our actual cognitive abilities.

2.4 Graphical Passwords

Cognitive psychology literature can readily explain the biases exhibited by users in selection of alphanumeric passwords. Hulme et al. [63] report that memory span for words is significantly better than for non-words, and suggest that words hold a more prominent place in long term memory. The same research community has also demonstrated other interesting memory capabilities had by humans, for example, it has been proposed that there exists a *picture superiority effect*: that concepts are more likely to be remembered experientially if presented as pictures rather than words. Shepard [117] demonstrated that human performance on a recognition task was better with pictures than with words (selected to be both common and rare in everyday usage), and sentences. This effect is thought to be explained by Paivio's *dual coding theory* [95]: that memory is comprised of both visual and verbal components that provide a complementary encoding for an item in memory. Such results have motivated research that seeks to re-envision knowledge-based authentication for the contemporary challenges faced when authenticating users. *Graphical passwords* [5, 124] are knowledge-based authentication protocols proposed to enable users to remember authentication credentials in a usable and secure fashion, and reduce the inclination for users to adopt insecure behaviours that may undermine security. The most common taxonomy of graphical password systems categorises approaches in terms of the memory task placed upon the user: recall; cued-recall; recognition; Tables 3,4,5 (positioned at the end of this chapter) provide a grouping of prominent systems that emphasises

system and evaluation features relevant to the discussion of security, usability, and deployability. A related tabulation is provided by Chiasson [13] which focuses upon the results of empirical studies and security characteristics. In the following sections we describe the state of the art in each of the three genres of the graphical password.

2.4.1 Recall-based Graphical Passwords

Recall-based graphical passwords require users to remember credentials in the absence of a memory cue (thus deployments of alphanumeric passwords are also usually recall-based). This genre of graphical password was proposed with the aim of generating credentials with crypto level [5] entropy and as such can be deployed in the modes of both local and direct authentication. Draw a Secret (DAS) [68] is the exemplar recall-based system and was proposed to exploit the stylus capability of Personal Digital Assistants (PDAs). At the enrolment phase the user is presented with an nxn grid and is asked to draw something memorable within its boundaries; the drawing is encoded internally as a sequence of cells (x, y) crossed by the pen of the user punctuated by pen lift events which are represented in the as $(n + 1, n + 1)$ (see Figure 3). The benefit of such a raw encoding, when contrasted with more sophisticated methods of pattern matching, is that the raw encoding is exactly reproducible by the user without the need for an exact visual correspondence between the enrolled drawing and the authentication attempt. As a result, the encoded drawing can be used as an encryption key and can be securely stored using a one-way function [32]: a function that is easy to compute but prohibitively difficult to invert. In order to authenticate using the enrolled drawing, the user must re-sketch the drawing ensuring the same grid cells are crossed in the same order, and the pen is lifted at the same points in the sequence. However, one usability constraint is that users are not permitted to draw into so-called *fuzzy boundaries*, areas of the grid that could lead to complexity being added to the drawing that is difficult to replicate later (e.g. lines cutting cell intersections or cell borders) as illustrated in Figure 4.

An authentication mechanism that represents a constrained instantiation of DAS is available on the Google Android platform for local authentication to touchscreen devices. This system presents a 3x3 grid, and users are restricted to the creation of patterns comprised of straight lines. In addition this variant restricts users to a single stroke and only allows users to visit each of the nine cells once. The formula used to calculate the DAS password space [68] can be modified to calculate the password space of the Android patterns; the full password space is 18.6 bits. Such a small password space is still suitable for local authentication when a restriction is enforced upon the number of permitted incorrect attempts to safeguard against guessing attacks.

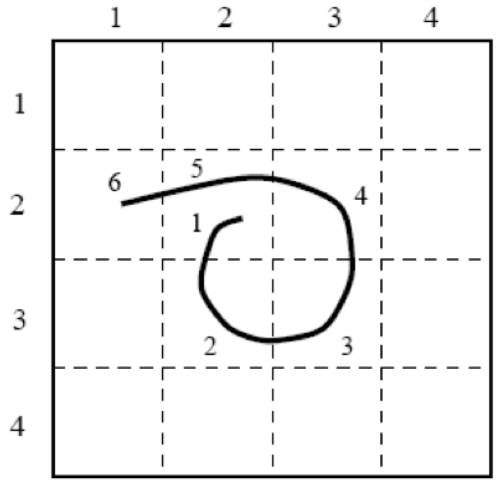


Figure 3: An example of a DAS [68] graphical password that has a single stroke and length of seven. The raw encoding of the drawing is (2,2) (2,3) (3,3) (3,2) (2,2) (1,2) (5,5).

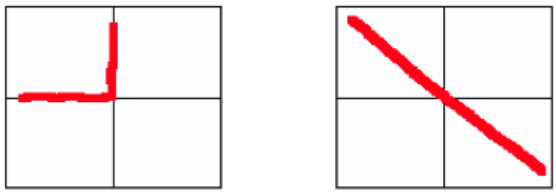


Figure 4: Examples of lines that cross fuzzy boundaries in the DAS system. Such lines would difficult for users to recreate according to the rules of DAS due to proximity to the cell boundaries.

2.4.1.1 Security of Recall-based Graphical Passwords

DAS is the canonical recall-based graphical password system; studies have been conducted into its theoretical security in terms of its password space, however relatively few conducted with respect to other attack vectors; a summary of the possible threats is as follows:

- Replay Attack: the credentials can be used by an unauthorised person.
- Dictionary Attack: where user choice is allowed, there may be a vulnerability to guessing attacks.
- Brute Force Attack: where users choose credentials from a predictable set, it may be possible to mount a guessing attack using all possible graphical passwords passwords.
- Phishing: users may be tricked to provide their graphical password to unauthorised services.
- Observation attack: observation of a single login may reveal enough information to facilitate a replay attack [151].

The theoretical password space for DAS is large (or as large as user motivation allows). The key determinants of the security of a DAS graphical password are: (i) length: the number of cells crossed in the drawing; and (ii) stroke count: the number of separate lines, captured as the number of pen-up events; a drawing that crosses 11 cells has 53 bits of entropy, which is greater than an eight character password drawn from the full ASCII set. As with alphanumeric passwords, the threat of automated guessing attacks has been prioritised as the most significant threat. Thorpe and Van Oorschot [131] studied the password space of DAS and determined that the number of strokes present in a drawing had a greater security impact than the length of a drawing; this led to the recommendation that for a drawing of length L the stroke count should be at least $L/2$. In later work [130], the same authors proposed that the theoretical password space of a knowledge-based authentication system has little significance if users are shown to choose credentials from a much smaller *memorable* password space (as observed with alphanumeric passwords). To illustrate this point in a graphical password context, they constrained their analysis to drawings of maximum length 12 and approximated the memorable password space for DAS using the set of possible mirror symmetric drawings. In fact, their smallest attack dictionary assumed mirror symmetry about the centre horizontal and vertical axes, which reduced the theoretical password space from 57 bits to 42 bits; although, this is still a large attack dictionary. Such results are striking but the predicted biases were not born out of empirical evidence with users of DAS. Nali and Thorpe [88] report the results of a small empirical study of DAS which used a 6x6 grid in which 45% of drawings collected were symmetric and 80% of drawings had a small number of strokes (1-3) which provided preliminary evidence that earlier theoretical predictions may be born out in practice.

When gathering empirical data from the usage of graphical password mechanisms, the most common way to reason over usability is to measure the memory accuracy of the user who has been tasked to remember authentication credentials across a period of time. The end measurement is usually referred to as the success rate S which can typically be calculated in two ways: a function of the number of successful and unsuccessful login attempts recorded across all study participants S_a; or the fraction of users who were able to authenticate successfully using a particular system (irrespective of the number of attempts required) S_p.

$$S_a = \frac{|attempts_s|}{|attempts|} \tag{1}$$

$$S_p = \frac{|users_s|}{|users|} \tag{2}$$

Equation 1 is most commonly used[2], where *attempts* refers to all authentication attempts (successful and unsuccessful) recorded across all users of a system, and

[2]The success rates reported are based upon Equation 1 unless otherwise stated

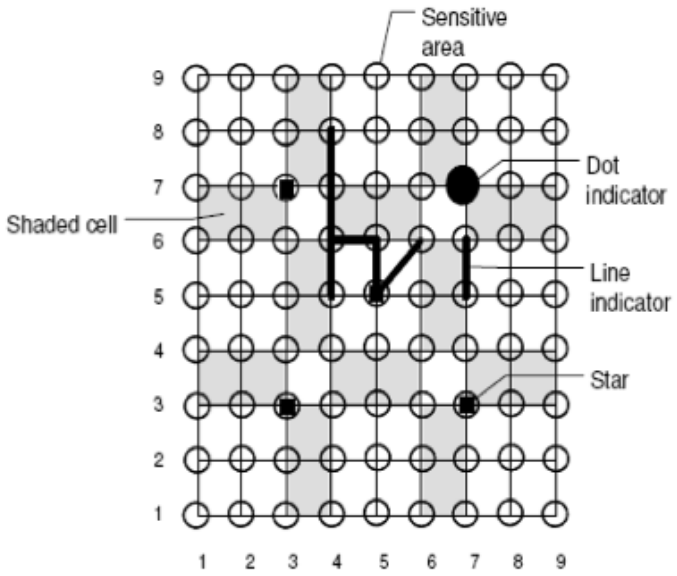


Figure 5: The *Pass-Go* [127] system requires users select cell intersections to assemble a graphical password.

$attempts_s$ refers to the subset of those attempts that resulted in successful authentication. In Equation 2 $users$ refers to the set of users who were asked to use a system and $users_s$ refers to the set of users who were able to authenticate successfully.

Tao and Adams [127] propose Pass-Go, a system based upon a Chinese board game; the system is designed to provide a larger password space than DAS, and overcome the likely usability issues caused by fuzzy boundaries (see Figure 5). In a user study with 167 computer science students across 13 weeks the success rate was 78%.

Due to the difficulty of executing automated guessing attacks upon credentials perceived to be strong, it is likely that much less sophisticated attacks such as observation attack become more attractive. Zakaria et al. [151] explored how interaction design could help to secure DAS for observation attack. They empirically evaluated a number of defence techniques: *disappearing strokes*: where each stroke would disappear after it was completed; *decoy strokes*: where randomly positioned and shaped strokes would appear on screen for obfuscation; and *line snaking*: where the tail of the stroke fades away whilst the user is still in the process of drawing. In an empirical study they identified the disappearing stroke technique as providing the best balance between usability and security, and that the decoy stroke technique provided little protection.

2.4.1.2 Summary

DAS is a promising candidate to enable users to choose and remember high entropy authentication credentials, as proposed attack dictionaries still appear prohibitively large. However, research to-date has focused upon theoretical analysis of the password

space and the prediction of biases that could reduce this space in practice [130]. Such predicted weakness has been observed to some extent in a small study [88], however, no formal empirical studies have been conducted to explore likely user behaviour with this system to determine the types of drawings users are likely to create, if such drawings are memorable, or if scaffolding can be provided to support the choice of drawings that are difficult to predict. Such research would be important to determine whether in practice DAS and potential variants would be usable and secure in practice.

2.4.2 Recognition-based Graphical Passwords

Recognition-based graphical passwords harness the reliability of human memory to recognise events or stimuli that have been previously seen to authenticate users. During an enrolment phase a user becomes acquainted with a sequence of k key images; at login, users must perform a visual search to identify these key images amongst d decoy images which provide the authentication challenge. If the user can demonstrate knowledge of the sequence of key images, they are considered to be the legitimate user. This genre of system is often proposed for contexts requiring password or PIN level entropy [5], and as a suitable approach to local authentication, or as a second authentication factor for direct authentication [96]. Representative schemes can be broadly classified as either being *mono-grid* [28] or *multi-grid*.

Mono-grid schemes provide an intuitive mapping from the standard PIN entry, as a user is presented with a single grid of images and is challenged to identify key images in a specific order. The associated entropy can be calculated as $log_2(\frac{(k+d)!}{((k+d)-k)!})$, which in the configuration illustrated in Figure 6, is less than the entropy offered by a 4 digit PIN (where $k = 4$ and $d = 6$) since each image can only be selected once. In this configuration images are usually static after each user selection, and across future authentication sessions, although the positions of the key and decoy images may be shuffled.

Multi-grid systems have emerged as a method to scaffold higher entropy graphical passwords through a modification to the original mono-grid design, where images presented during the login phase are distributed across a sequence of image grids of equal dimensions. Key images can be distributed across grids using one of two approaches: *one-key-per-screen* [96] (see Figure 7): a single key image appears in each displayed grid, and where g is the number of grids, password entropy is higher than that of a mono-grid system system for the same number of key images: $log_2((k+d)^g)$; or *random* [59, 125] (see Figure 8): images randomly spread across grids as zero, one, or more than one key image can appear in a single grid which provides a positive impact on password entropy: $log_2(\binom{(k+d)}{k})$. Finally, key images can be presented in one of two modes: *monolithic*: where the user identifies all known key images; *portfolio* [56][31]: users are challenged to identify a random subset of their known key images. These two modes do not have a direct impact upon password entropy, but can serve to

Figure 6: A mono-grid recognition-based graphical password system [28]. In this configuration the image positions are unchanged between selections and logins. Red squares represent example selections.

complicate attacks to capture and reuse key images (e.g. through observation attack).

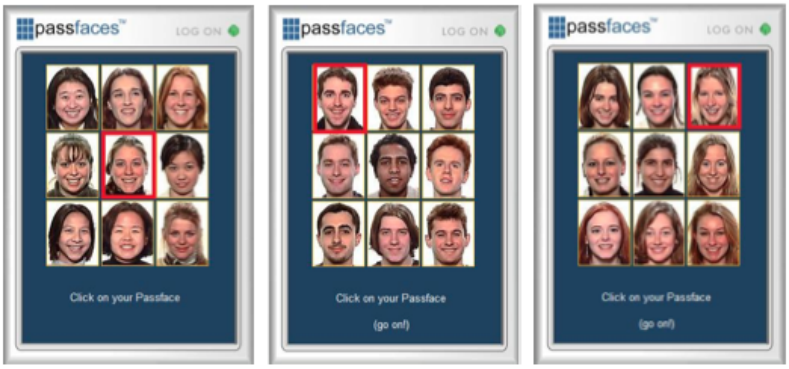


Figure 7: An example of a multi-grid recognition based graphical password system [98] in a one-key-per-screen configuration. Red squares represent example selections.

2.4.2.1 Usability Studies

Early usability research in this domain has focused upon the memorability of diverse types of image stimulus, mainly centred upon different types of photographs. Brostoff & Sasse [11] conducted a field trial of the commercial Passfaces [96] system over a three month period with 34 participants where the system secured a university coursework submission system. At the end of the study they report that users of Passfaces experienced a 5% login failure rate, compared to 15% for those using alphanumeric passwords. The study also showed that those using Passfaces accessed the system less frequently than those using alphanumeric passwords; the mean number of logins was 34 ($\sigma = 20$) for alphanumeric passwords, and 12 ($\sigma = 7$) for Passfaces users. This reduction in usage could have been the result of a number of factors, for example perceived effort associated with Passfaces; a subtle artefact of experimental design;

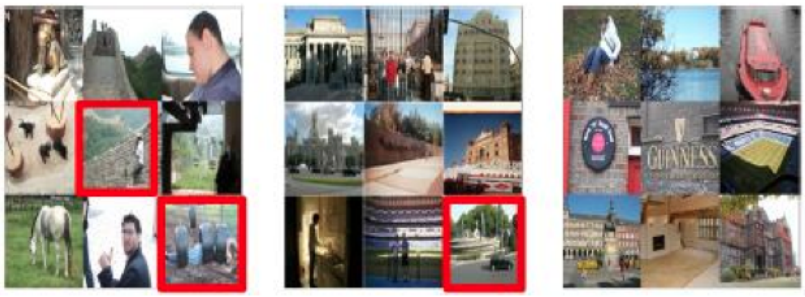


Figure 8: An example of a multi-grid recognition-based graphical password system where key images are randomly spread across grids. Red squares represent example selections.

or the contextual motivations of the students. However, it is notable that the lower login failure rate for Passfaces was apparent despite this reduced usage.

De Angeli et al. [28] conducted a usability analysis of a photograph-based *Visual Identification Protocol* (VIP) in three forms: VIP1 requires users to identify 4 key images in a fixed order from the onscreen set of 10; VIP2 is the same as VIP1 with the exception that images appear in random positions at each login; VIP3 assigns users 8 key images, and at login requires the user to identify 4 from the onscreen image set of 12. In a one-week recall study with 61 participants they found that the VIP3 system had the largest error rate of over 10%. Elsewhere, there were no significant differences between those using VIP1, VIP2 and and those using standard PINs in a control group. The key source of errors in the mono-grid style systems (i.e. the VIP1 and VIP2 conditions) resulted from users forgetting the specific ordering of the key images. These results lend further weight to an observation by Davis et al. [27] that problems of remembering image order are significant (for Davis et al. this was related to their Story system). However, in the VIP3 condition the key source of error came from simply making the wrong selection, likely exacerbated by the variation in key images between logins (i.e. the portfolio mode of operation). Despite this detailed analysis of errors, all participants were able to provide credentials within three attempts.

Whereas De Angeli et al. [28] evaluated publicly available sets of photographs, Tullis and Tedesco [133] explored the usability of personal photographs for user authentication. Across three studies they found that participants were most accurate in an authentication procedure when the key images were personal photographs as opposed to generic stock images, even where decoy images were handpicked to be semantically similar to the key image. In a remarkable follow-up study six years later, Tullis et al. [134] found that nine out of thirteen participants were able to successfully authenticate using credentials comprised of personal images. To strengthen these results, there does appear to be an intuitive logic to the utility of personal photographs in this context, that users are more likely to establish meaningful associations with images drawn from personal collections due to having an increased familiarity with

the content. Renaud [101] provided further support for this hypothesis in a study that explored the impact upon usability of differing levels of user involvement in creation of images for authentication. Users with a greater involvement in the creation of the image content achieved higher subsequent success rates in the authentication challenge; the group with the highest accuracy was a group that produced handwritten doodles, followed by those who added their own photographs from a camera.

Dhamija and Perrig [31] conducted an evaluation of the Déjà vu system, which explored the use of fractal images as authentication credentials. The use of fractals in this context means that, unlike in the other studies, images could be automatically generated to comprise a login which is potentially a useful property from the perspective of deployment. In a one-week recall test they discovered that those assigned photographs or fractals had a higher login success rate than those using PINs and passwords. No statistical tests were performed to determine the significance of this result, although the small differences might suggest no significant differences existed over this short time scale.

2.4.2.2 Memory Interference

One important question concerning graphical passwords is the extent to which remembering multiple credentials impacts usability. In a study of 172 participants over 4 weeks, Moncur and Leplâtre [85] carried out the first study of the usability of multiple mono-grid recognition-based graphical passwords. Participants were distributed across 5 conditions, one of which was a control group using PIN; they assigned 5 credentials to each user, and tested recall at two-week intervals. Despite the need to recall the order of images due to the mono-grid configuration of the authentication challenge, participants assigned five graphical passwords demonstrated significantly higher login accuracy than those remembering five PINs [3]. Everitt et al. [42] carried out a five week study of a Passfaces-based system with 100 participants. The experimenters controlled both the number of graphical passwords given to users and the frequency with which participants accessed their experimental system (Figure 9 depicts and describes the experimental conditions). Their findings can be summarised as follows:

1. Frequency: participants using a particular graphical password every day (Group 5, week 1) had a lower failure rate (0%) than those using the same graphical password once per week (Group 1) (1.96% not statistically significant).

2. Training: of the participants assigned four graphical passwords, those who enrolled with all four at the same time (Group 4) performed significantly worse than those who were gradually enrolled with new graphical passwords (Group 5).

[3] Note that the unusually high dropout rate of 65% makes it difficult to establish the general applicability of this result.

3. Interference: participants assigned four different graphical passwords had the worst failure rate overall at 15% (Group 4). Participants with the best performance (1.45%) used one graphical password once per week (Group 1).

4. Long Term Recall: in a long-term recall test, conducted four months after the original experiment, of the participants who had multiple graphical passwords, and were tested on a different one every day, Group 4 had the highest failure rate (14%) and Group 1 and Group 5 (who were assigned multiple graphical passwords over time) had the lowest failure rate (0%). These relatively low failure rates demonstrate the potential of recognition-based schemes to be useful in a multi-graphical password context. Note that an important difference between this study and that of Moncur & Leplâatre [88] is that this system operated in a multi-grid, one-key-per-screen configuration, whereas Moncur & Leplâtre trialed a mono-grid configuration.

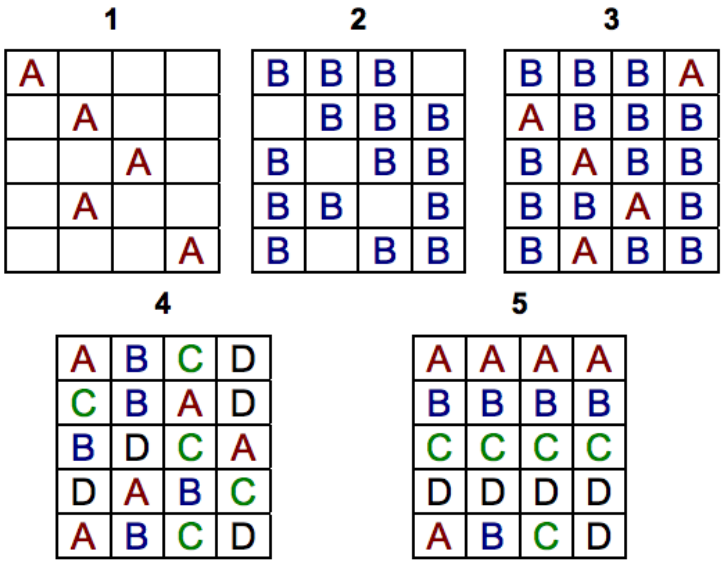


Figure 9: The five conditions from the study carried out by Everitt et al. [40] exploring the usability of remembering multiple Passfaces graphical passwords. Rows represent the 5 study weeks (Tuesday, Wednesday, Thursday, Friday) and each distinct character represents a distinct graphical password. The recurrence of those characters depicts how frequently users had to login using that graphical password.

2.4.2.3 Security

The key security benefit of using images in recognition-based authentication is that users are more likely to have the cognitive resource to remember randomly assigned credentials. However, there are a number of subtle, but significant attacks that must be considered:

- Replay Attack: an attacker can reuse captured credentials.

- Brute Force Attack: due to the small password space (usually < 20 bits) a brute force attack is trivial. Therefore this genre is only appropriate for local authentication where automated attacks are less accessible.

- Intersection Attack: where variation exists in the images presented at each login, patterns in frequency (i.e. some images appear more than others) could provide clues as to whether an image is a key image [31].

- Observation Attack: due to the greater memorability of images, key images are likely to be more memorable to casual observers.

- Educated Guess Attack: where credentials are not assigned randomly, attackers can either use knowledge of the user to guess predict user choice [133, 97], or exploit knowledge of known user preferences in image selection.

Davis et al. [27] conducted a security focused study of user choice in the context of a variant of Passfaces, called Faces, and a novel mono-grid scheme called Story. In the story system, the image stimulus comprises photographs curated to represent diverse content ranging from faces to everyday objects. The design rationale is that such semantically different images could assist the users to develop a narrative around chosen images to aid memorability. Users were allowed to choose key images, and in a user study across a four-month period they observed interesting trends of user choice. They noted that users of the Faces system were much more likely to select female images as opposed to images of males; over 40% of choices by white males were of images of white female models; over 60% of male choices were for female models. They also observed a race affiliation effect whereby users were more likely to choose images of faces from within their own racial group (Figure 2 illustrates the exact effect). For the Story scheme, user choice patterns were not so pronounced. In the analysis of usability, users were accurate in the authentication protocol as they exhibited a daily mean success rate of over 90%; most login errors originated from users in the Story condition who could not remember the correct order to enter images. Indeed, despite advising users to create a story around the chosen images as a usability scaffold, users were still inclined to make errors.

Pop.	Asian	Black	White
Asian Female	52.1%	16.7%	31.3%
Asian Male	34.4%	21.9%	43.8%
Black Male	8.3%	91.7%	0.0%
White Female	18.8%	31.3%	50.0%
White Male	17.6%	20.4%	62.0%

Table 2: The race affiliation effect noted by Davis et al. [27], white females chose faces to comprise their graphical password within their own race 50% of the time.

While Davis et al. [27] demonstrated a particular set of bias effects for facial images, whether systematic selection biases exist for other types of image stimulus has not been extensively explored. However, in a context where personal images are used as authentication credentials, research has explored the possibility of an *educated guess attack*. The threat arises if an attacker has personal knowledge of the life of a user and may be able to distinguish between images that may and may not belong in their personal collection (e.g. using knowledge of friends or holidays). Pering et al. [97] proposed a system where users must remember one-time authentication images from their own image collections to authenticate at public terminals. The authentication task for a user is to identify personal images amongst decoy images provided by other users of the same system. In user study exploring both the usability and security of this configuration they found that 2/12 attackers given a small number of personal images from a given user were able to extrapolate themes in the images to break-in to their account. Tullis and Tedesco [133] also conducted a study of an educated guess attack where attackers were allowed to target a known person to guess their password images. In the worst-case configuration (i.e. best case for the attacker) guessing accuracy was 38% on a per-image basis (where 100% corresponds to an attacker successfully guessing all 4 images used per login). This would indicate that where the authentication task is for a user to distinguish personal images amongst non-personal images, a diligent attacker would have much to gain from observation attacks and very simple social engineering.

In addition to user choice, the very composition of the graphical password login challenge can constitute a vulnerability. Dhamija and Perrig [31] identify *intersection attack* as a particular concern. Observation attack is an obvious threat to this genre of authentication mechanism, hence introducing visual diversity into each login challenge is desirable, to complicate the attacker capturing and reusing credentials. However, where this diversity is anything but random, observing the frequency that an image appears at login relative to others could be used to reveal its role as a key or decoy image. Current wisdom to defend against this threat is that all login challenges should be kept identical. Tari et al. [128] carried out a empirical study to determine the vulnerability to observation attack of Passfaces [96] graphical passwords and alphanumeric passwords. Users were asked to pose as attackers and observe live input. They discovered that alphanumeric passwords were more vulnerable to observation attack than Passfaces where a mouse was used. However, perhaps a consequence of the usability of Passfaces, users perceived Passfaces entry with a mouse to be the configuration most vulnerable to this form of attack.

2.4.2.4 Scaffolding Secure Behaviour for Recognition-based Graphical Passwords

Komanduri and Hutchings [75] compared the usability of a recognition-based system where images were comprised of simple icons, to that of alphanumeric passwords. The credentials assigned to users in both conditions were manipulated to ensure that the password entropy was 50 bits. This level of entropy in the graphical password condition was primarily achieved by enforcing that users remembered a specific ordering of the key images. A recall test was conducted one week after credentials had been assigned, and both groups performed relatively poorly. However, relaxing the constraint that users had to remember the ordering of individual credential components (i.e. particular images or characters) meant that the success rate for graphical password users increased to 100%, compared to 67% for alphanumeric passwords. This mirrors a trend observed in other studies [28, 27, 85] that increasing password entropy by enforcing recall of a the ordering of a stimuli is not a usable approach to maximising entropy and that a multi-grid system is the only practical way to scaffold choice and recall of stronger recognition-based graphical passwords.

2.4.2.5 Summary

Recognition-based graphical passwords are considered to have significant limits in terms of the entropy of the credentials that user can remember, as the act of authentication becomes increasingly difficult and slow as password entropy is increased, that is, users would be required to remember more images and navigate more grids. However, within certain bounds of entropy it appears they show promise as a usable and secure authentication mechanism. Some significant results to note are the following:

- Recognition-based graphical passwords have good memorability characteristics, even over extended periods of non-use [134].
- Recall of image order negatively impacts memorability [28, 75].
- User involvement positively impacts memorability [101, 133].
- User choice may introduce systematic biases that give rise to vulnerability to guessing attacks [27].

Finally, it is interesting to note that although we have described seven well known memorability studies reporting very favourable performance results, Passfaces is the only known deployment of a recognition-based graphical password.

2.4.3 Cued recall-based Graphical Passwords

Cued recall-based graphical password systems have been proposed to provide a middle ground between recall and recognition-based systems, in terms of both usability and

security. To assist the user to authenticate, a memory cue is provided that assists the user to reconstruct the correct credentials. There are few cued recall-based systems, however the most widely studied is Passpoints[4] [143]. Passpoints requires the user to choose memorable locations (usually four or five) within an image to comprise their authentication credentials. Visual features of the image itself assist users to choose and remember the chosen locations; to authenticate, the user must be able to select the same specific locations within the image bounded by a small pixel tolerance. Figure 10 illustrates example user interactions with the system.

Figure 10: Example visualisation of 5 chosen points in an image as required by the Passpoints system, taken from [143].

2.4.3.1 Usability Studies

Wiedenbeck et al. [143] conducted a user study to compare the usability of Passpoints and alphanumeric passwords across six weeks. They recruited 40 users to take part in a laboratory-based study where recall tests were conducted one week, and six weeks after enrolment. Tey discovered that Passpoints users made fewer mistakes during authentication, yet in a questionnaire perceived their authentication task to be more difficult than using alphanumeric passwords.

Chiasson et al. [14] conducted a laboratory-based study and a field study to explore the impact on user performance of varying the size of the pixel tolerance allowed around click points, and the impact of image choice. The lab study confirmed the presumption by Wiedenbeck et al.[142] that success rates could vary significantly depending on the image stimulus. A key finding was that the traditional measure of the efficacy of a password scheme: success rate, was significantly higher in the lab study than in the field study. In the lab study, calculating a success rate for the authentication phase of the study across all 17 images was 94%; in the field study login success rates ranged from 78%-83% This raises questions as to the most appropriate

[4]The origins of Passpoints can be found in Blonder's graphical password patent [7] in which he proposed the concept of a graphical password that authenticates users through selecting areas of an image.

Figure 11: The cued-recall graphical password scheme proposed by Weinshall [139]. Users begin in the top left corner, and traverse the grid by moving down if they see a key image, or right otherwise. The user must then enter the number they arrive to at the edge of the grid.

way to evaluate a user authentication system and what can be learned from studies conducted in each environment. In a later study, Chiasson et al. [18] found that that those users asked to remember six Passpoints graphical passwords, had a significantly higher success rate than users asked to remember six alphanumeric passwords.

2.4.3.2 Security

Cued recall-based graphical passwords in general face a number of threats:

- Replay Attack: an attacker can capture the credentials and reuse them to obtain unauthorised access to an account.
- Brute Force Attack: a 5-click graphical password with a tolerance of 19x19 pixels on a 451x331 pixel image, leaves a theoretical password space of 43 bits [132]. However, just five or six click points one can make more passwords than an eight character UNIX password based upon an alphabet of 64 characters [143].
- Observation Attack: as with all graphical passwords, observations of user input can be replayed to the system to gain unauthorised access.
- Image Processing-based Guessing Attack: a dictionary attack is seeded using automated detection of image features likely to be incorporated into click point sequences [132, 33].
- Human Seeded Guessing Attack: a dictionary attack is seeded by gathering a corpus of user-chosen click point sequences to serve as predictors for a different, larger sample [132].

Dirik et al. [33] present an analysis of Passpoints user choice for two particular images; they used image processing techniques to identify image features that users may be likely to choose as click points, and used them to compile an attack dictionary (see Figure 12). By using this dictionary to attack a corpus of user-chosen graphical passwords, they were able to guess 80% of user click positions (although this did not consider a full sequence of click points). Thorpe and van Oorschot [132] formulated a dictionary attack against full Passpoints passwords using two methods: a human-seeded attack and a fully automated attack. In the human seeded attack they collected a corpus of click points from one set of users, and used these to assemble an attack dictionary to use against a different set of collected click points. Their results suggest that the success of the method is highly dependent upon qualities of the particular image; for one image, using permutations of raw collected click points yielded a success rate for guesses of 34% (using a 36.7 bit dictionary), while for another image the success rate was 52% (using a 37.1 bit dictionary). For the same two images the automated attack was less successful, yielding a success rate of 9.1% and 0.9% respectively (using a 35 bit dictionary). In fact, they had varying degrees of success across different types of images and their approach was more successful at guessing click points collected during a lab study than guessing those collected in a field study. One explanation for this could be that the fairly intense nature of the lab study could have enticed participants to behave more predictably than normal. In later work Salehi-Abari et al. [107] developed more sophisticated image processing techniques and assumptions of user choice by focusing upon particular user selection strategies, using the same images analysed by Thorpe and van Oorschot [132] and a 34.7 bit dictionary they increased the accuracy of automated guess attacks to 48% and 54%.

2.4.3.3 Scaffolding Secure Behaviour in Cued Recall-based Graphical Passwords

Empirical studies of the Passpoints system created a number of questions regarding the extent to which biases in user choice of click points could be attacked, and in response, which methods were suitable to scaffold more secure user choice of Passpoints credentials. Wiedenbeck et al. [142] explored the most appropriate tolerance square for Passpoints, and the most appropriate image type. The smaller the tolerance square is, the greater the password space will be, but the more precise user recall must be to correctly remember the target point in the image. Through a process of experimentation they discovered that for their particular input device, a 10x10 pixel tolerance around the click point produced more login errors than a 14x14 pixel square. They also found that some images were more appropriate to be used with Passpoints than others, with the swimming pool image in their study giving rise to the highest error rates, and a map image the lowest number of errors.

Figure 12: Dirik et al. [33] used image processing techniques such as corner detection to predict likely sequences of Passpoints graphical passwords. (left) an overlay of click points collected from a corpus of users; (right) overlay of click points predicted using their image processing methods.

Chiasson et al. [19] proposed a method to resist predictable user choice in Passpoints. In *Cued Click Points* (CCP) each click point in the sequence is made upon a distinct image. This reduces the number of click points a user must choose on a single image, and spreads the guessing load for an attack across a larger number of images. In addition, this approach provides implicit feedback to the user, as the presentation of subsequent images is contingent upon the location of the click point selected; if an unexpected image appears onscreen, it is likely the user has made an error. In a short-term recall test, 24 participants had a 96% success rate overall when using this system. In a later study, Chiasson et al. [16] proposed *Persuasive Cued Click Points* (PCCP) which augments the enrolment phase of CCP by encouraging users to select click points within a randomly chosen region of the image. The reported success rate for their short term recall test of PCCP was 91%. Chiasson et al. [17] went on to analyse the effects of scaffolding secure user choice by comparing user choice between PCCP, CCP and Passpoints; a number of patterns emerged including that users of Passpoints were more likely to choose click points positioned closer together, or positioned adjacent across the horizontal axis of the image. User choice of click points for PCCP and CCP appeared more randomly distributed in the image which suggests a security benefit as these credentials appear less guessable than those chosen by users of the original Passpoints.

A small number of cued recall-based graphical password systems differ from Passpoints, yet have the scaffolding of secure user behaviour as a key design motivation. Wiedenbeck et al. [144] proposed the Convex Hull Click (CHC) scheme which requires users to click within the convex hull formed by the on-screen position of key images, rather than clicking the images themselves. In a one-week recall test with 15 users, 14 were able to repeat their assigned graphical password. However, the complexity of the login task led to a mean login duration of 71.66 seconds. Weinshall [139] proposed

a cued-recall based protocol designed to resist observation attack which requires the user to remember an assigned shared secret comprising a set of images, and then to use these to perform pre-determined moves to navigate across a grid of images containing decoy images (see Figure 11). The daily success rates of users recorded in a longitudinal user study were consistently over 90% for both low and high entropy instantiations of their proposed system. Although, Golle and Wagner [54] showed that the indirection introduced in Weinshall's protocol could be reverse engineered more efficiently than was suggested in the original proposal.

2.4.3.4 Summary

Cued-recall based graphical passwords, primarily in the form of the Passpoints system (and extensions), have been extensively studied. Scaffolding secure user choice has in particular received attention. Interference has also been observed to be significant where multiple click passwords are selected on the same image [14]. In addition, it has been highlighted that lab studies and field studies have resulted in very different usability and user choice results. The PhD thesis of Chiasson explores a number of issues related to the usability and security of Passpoints graphical passwords [13].

2.5 Thesis Overview

The study of user authentication has expanded in recent years from a purely theoretical focus, to encompass research approaches that capture more holistic insights into user behaviour, technology, and context; this shift reflects the contemporary challenges faced in securing computer systems for unauthorised access. This approach has generated insight into the ways authentication mechanisms can be appropriated in ways not envisioned by designers; however, such insight has been generated from the study of existing deployments of alphanumeric passwords. The effective absence of real-world deployments of representative graphical password systems receiving detailed study means that similar insight cannot be generated, which creates inertia for potential adopters who would be reluctant to introduce systems into an environment that could negatively impact an already delicate balance between usability and security. In response, there is a need for either: real-world longitudinal deployments of graphical password systems; or empirical methods that enable a more lightweight approach to begin to map the space of the future challenges that could be brought about by future deployments; this thesis focuses upon the latter.

The literature review of graphical passwords has highlighted three promising exemplars of graphical password system: DAS [68], mechanisms based upon image recognition, and Passpoints [143]. Passpoints has received much study [13], however the remaining two classes of system appear to provide an interesting lens through which to study user authentication phenomenon in this graphical password context.

As there are a myriad of issues in the design and evaluation of usable and secure user authentication, our case study approach must inevitably capture only a subset of interesting issues that the deployment of usable and secure graphical passwords might generate. Figure 13 provides an overview of the primary issues captured by each chapter of the thesis; each issue has been highlighted as a complication to the understanding of the usability and security of graphical passwords. Our research approach did not seek to champion one particular authentication system, however, made appropriate selection of a system that would enable us to generate the most interesting insight into a particular phenomenon. Our chosen contexts were *user choice*: a principle concern in knowledge-based authentication; *observation attack*: a threat thought to be particularly accessible and potent to graphical passwords; *password sharing*: one technique adopted by users to cope with the excessive memory load imposed by alphanumeric passwords. The final chapter focuses upon the problem of image filtering: the process of understanding the usability and security properties of images, and using that to identify images appropriate for use by a particular system; such a process likely must take different forms for different schemes. This is a recurring issue experienced during the projects described in the thesis and in other work (as indicated in Tables 3,5), yet is often placed out of the scope of research despite its important role in underpinning the usability and security properties of a system.

We started the research with a focus upon DAS due to its proposal as a method to generate crypto-level entropy [5], and the entropy of authentication credentials traditionally being the principal recipient of theoretical research in knowledge-based user authentication. Through this study we became particularly aware of the importance of the interplay between usability, security and deployability, and the potential for context-specific scaffolding of user behaviours that could impact all three aspects.

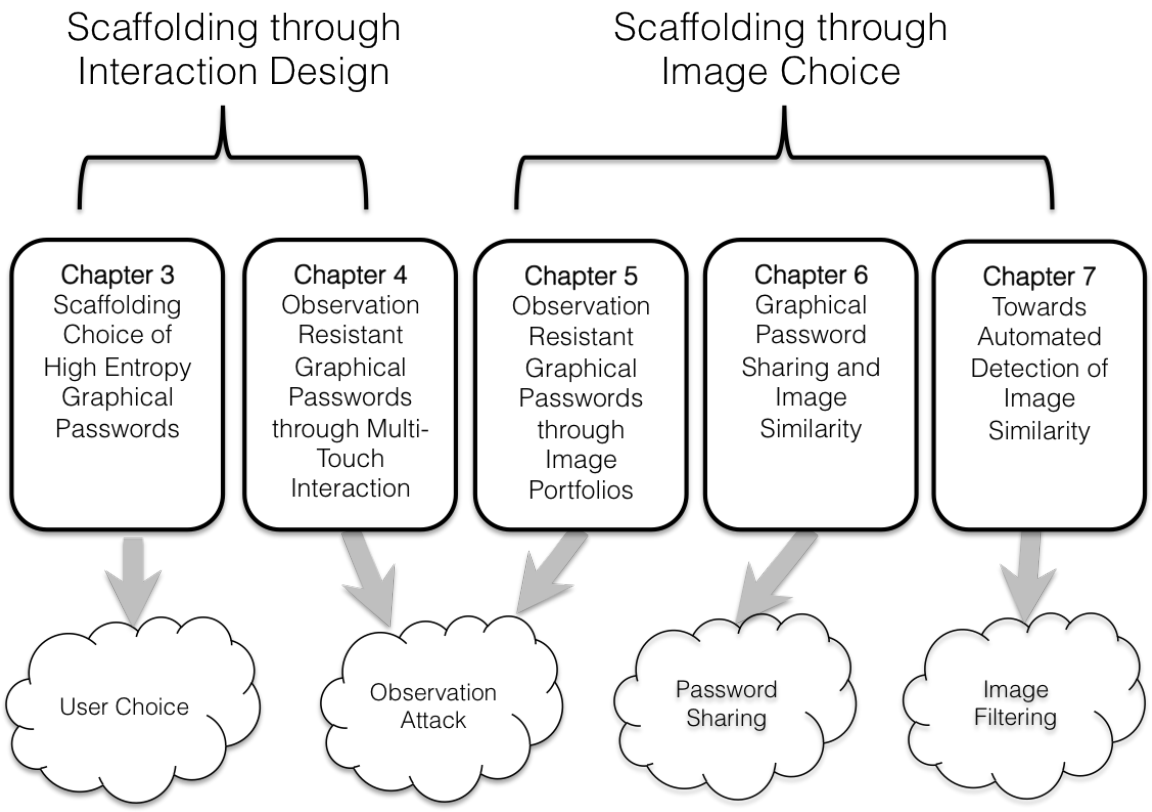


Figure 13: Illustration of the sub-domains of graphical passwords encountered in each chapter of the thesis.

System	Scaffolded Behaviour	Deployment Assumption	Evaluation Length	Evaluation Platform	Socio-Technical Threat Evaluation
Deja Vu [31]	Observation resistance	Image filtering	1 Week	Desktop Computer	-
Awase-E [125]	-	Image filtering	1 Week	Mobile device	-
VIP 1 [28]	-	Image filtering	1 Week	Desktop Computer	-
VIP 2 [28]	Observation resistance	Image filtering	1 Week	Desktop computer	-
VIP 3 [28]	Observation resistance	Image filtering	1 Week	Desktop computer	-
Passfaces [11]	-	Image filtering	12 weeks	Desktop computer	-
Pering et al. [97]	Observation resistance	Large image collection Image filtering	1 day	Desktop computer	Educated guess
Story [27]	-	Image filtering	≈11 weeks	Desktop computer	User choice
Faces [27]	-	Image filtering	≈11 weeks	Desktop computer	User choice
Dynahand [102]	-	Handwriting capture	4 weeks	Desktop computer	-
Tullis & Tedesco [133]	-	Image filtering	8 weeks	Desktop computer	Educated guess

Table 3: A collection of evaluations of recognition-based graphical passwords, with characteristics highlighted that appear relevant to the discussion of the future deployability of graphical passwords.

System	Scaffolded Behaviour	Deployment Assumption	Evaluation Length	Evaluation Platform	Socio-Technical Threat Evaluation
Nali & Thorpe [88]	-	Pen input	1 day	Pen & paper	User choice
Goldberg et al. [53]	-	Pen input	1 week	Pen & paper	-
Pass-go [127]	Secure user choice	-	13 weeks	Desktop computer	-
YAGP [52]	-	-	2 weeks	Desktop computer	-
Zakaria et al. [151]	Observation resistance	Pen input	1 day	PDA	Observation attack

Table 4: A collection of evaluations of recall-based graphical passwords, with characteristics highlighted that appear relevant to the discussion of the future deployability of graphical passwords.

System	Scaffolded Behaviour	Deployment Assumption	Evaluation Length	Evaluation Platform	Socio-Technical Threat Evaluation
Passpoints [143]	-	Image filtering	6 weeks	Desktop computer	-
Passpoints [142]	Secure user choice	-	1 week	Desktop computer	-
Passpoints (field study) [14]	-	Image filtering	8 weeks	Desktop computer	-
Cued Click Points [19]	Secure user choice	Image filtering	1 day	Desktop computer	-
Persuasive Cued Click Points [16]	Secure user choice	Image filtering	1 day	Desktop computer	User choice
Convex Hull Click [144]	Observation resistance	Image filtering	1 week	Desktop computer	-
Weinshall (High complexity) [139]	Secure user choice Observation resistance	Image filtering	≈28 weeks	Desktop computer	-
Weinshall (Low complexity) [139]	Secure user choice Observation resistance	Image filtering	≈1 year	Desktop computer	-

Table 5: A collection of evaluations of cued recall-based graphical passwords, with characteristics highlighted that appear relevant to the discussion of the future deployability of graphical passwords.

Chapter 3

Scaffolding Choice of High Entropy Graphical Passwords

The literature review of alphanumeric passwords highlights that forcibly scaffolding secure user choice can be held partly responsible for the insecure user practices that result. Indeed, it has been demonstrated that there is only a weak relationship between strictness of password composition policies and the choosing of strong passwords [76, 99]. The greater interface assumptions of graphical passwords (when compared to alphanumeric passwords) potentially allows for a larger design space of interventions that support the creation and memorization of credentials with desirable levels of password entropy. However, while various attempts have to made to improve the security of user choice in cued recall-based schemes such as Passpoints [16, 19], recall-based schemes such as Draw a Secret (DAS) [68] have been largely ignored. DAS is of particular interest in this context of user choice as it has a large theoretical password space; however, there have been a number of predictions that the usable password space of DAS could be much smaller in practice [130, 131] which could undermine the significance of these theoretical estimates. These predicted user behaviours had been made in the absence of formal empirical study; as such, there remain open questions regarding the type of DAS drawings users would be likely to create, and indeed how user interaction could be designed to support users to avoid these predicted biases.

We propose that the user choice biases predicted for everyday usage of DAS may actually be exacerbated by the use of the bare drawing grid (see Figure 3) which has few distinctive features that might be used to aid memorization of a drawing. Consequently, we created the Background Draw a Secret (BDAS) system, a variant of DAS that we hypothesized would passively support secure user creation and memorization of more complex drawings than DAS. This intervention to the user interaction with the DAS scheme could be one promising way to enhance the usability and security of DAS.

3.1 Threat Model

The principal threat to the security of DAS we consider is that an attacker can perform an offline brute force attack in a reasonable amount of time. If an attacker obtained the encoding of a DAS drawing which has been passed through a one-way function [32], then it is possible to make an unrestricted number of guesses to recover the plain-text DAS drawing. Thorpe and van Oorschot [131] report that drawings with a small number of strokes can reduce a password space of 58 bits to 40 bits; and propose that the stroke count of user-chosen drawings should be at least half of the drawing length $S > \frac{L}{2}$ where L is the length. Today, 40 bit searches are feasible in a reasonable amount of time [147] and so any valuable authentication credential should aim to exceed this level of security; if the entropy of a drawing is less than 40 bits, it is likely an attacker can recover the graphical password with few optimisations to the search.

3.2 Background Draw a Secret (BDAS)

BDAS is a system that comprises an enhancement to the user interface of DAS through the placement of an image underneath the drawing grid (see Figure 14), essentially turning BDAS into a cued-recall graphical password system. The addition of new visual cues to the grid could allow users to further leverage the picture superiority effect [117] to remember the positioning and composition of their drawing. Procedures of enrolment and login remain the same as in the original DAS proposal, however, the added visual cues in the grid intuitively could aid users to choose drawings that contain increased complexity, and also remember those more complex drawings.

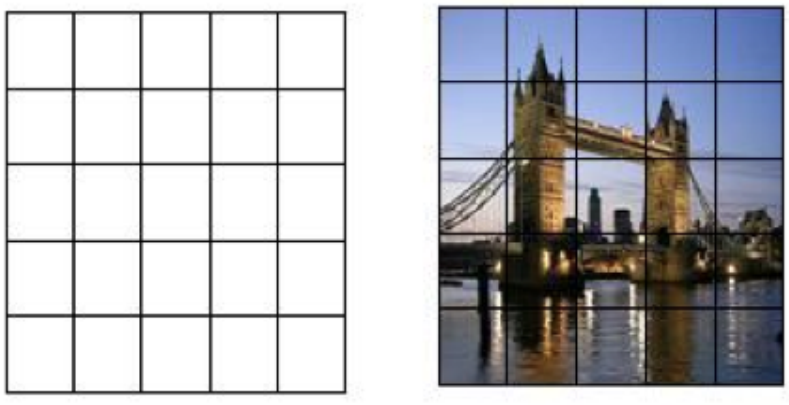


Figure 14: Example of a DAS drawing grid (left); example of a BDAS grid (right).

Of course, no formal or empirically verified criteria exist to pre-select images for the purpose to support DAS drawings; intuitively, desirable images are likely to contain diverse image content to ensure that user attention across the image is distributed more randomly about an image; this would reduce the opportunity for an attacker


Figure 15: An image with few hotspots (left) and a large number of hotspots (right).

to predict how a particular background image might influence user choice of drawing content. Desirable BDAS images are likely to include a large number of *hotspots*: memorable discrete areas within an image where objects or the local context is particularly salient. Such areas of an image are components of images that are likely to be memorable to users in the challenge to recall their drawing. Figure 15 illustrates two example images that illustrate extremes in the presence of hotspots that are useful to consider from both a usability and security perspective. Discussion on the security impact of the background images is provided in Section 3.4.1.

3.3 User Study

We conducted an empirical study[1] to explore whether users of BDAS would be more likely to choose stronger graphical passwords than users of DAS; also, to determine whether DAS and BDAS drawings would be memorable if in widespread usage. Such results would help us to explore whether theoretical predictions made by Thorpe and van Oorschot [130, 131], would be born out in tests with users, and whether these biases could be corrected for by BDAS.

3.3.1 Method

We carried out a between-subjects study where the independent variable was the system to which participants were randomly assigned (DAS or BDAS), and the dependant variable was the memorability of the user chosen credentials. We recruited 46 participants, 32 male and 14 female. All participants were undergraduate students at the time of the experiment, and none had previously heard of DAS or BDAS. The typical age range of subjects was 18-25 with one participant in the group 50+. Most participants had technical backgrounds (20) e.g. majoring in computer science or engineering, and the remaining 26 subjects non-technical (majoring in modern languages, business etc).

We chose a paper-based experiment as paper prototyping is a common technique in human-computer interaction [104] to test new ideas, but also because previous

[1]A pilot study is reported in addition to this study in a publication at the ACM Computer and Communications Security conference in 2007 [39].

short-term empirical studies of DAS were also conducted on paper [88, 53]. A few days in advance of the experiment, each participant was sent an information sheet (see Appendix A) that provided information on both the experiment and the DAS and BDAS passwords (depending on their assigned group).To begin, each participant in the BDAS group is presented with five images, and instructed to choose one to use as a background image to the grid. We allowed BDAS users to choose their background image, as we wished to gauge the kind of images that might prove to be popular. In an initial training phase the participant was permitted to practice with the assigned scheme (either DAS or BDAS). The experiment commenced in the enrolment phase in which a participant was asked to draw upon the enrolment form to set their graphical password. The path of a drawing was captured on video, but also noted by an experiment facilitator (for later evaluation of the correctness of the drawing created).

The participants would be tested after two time intervals to test their retention of their chosen graphical password: after a five-minute delay, and one week later; at each recall test participants were given three attempts to correctly authenticate, where correctness was assessed by the experiment facilitator.

3.3.2 Study Materials

The experiment used a number of custom paper forms which we used to capture participant interactions: (i) the enrolment form has two rows of three drawing grids, one row for enrolment and one row for the short-term recall test (the first row of grids were covered during the recall test); (ii) the login form has three drawing grids, one for each login attempt (see Appendix B); and finally (iii) the practice form, a nine drawing grid form on which participants could practice creating valid drawings. The size of all drawing grids was chosen to be 3.7 (corner to corner) replicating the screen size of a popular Personal Digital Assistant (PDA) at the time. The resolution of the drawing grid was set to 5x5 based on previous research [130] that suggests that this configuration provides an appropriate trade-off between usability and security. The DAS and BDAS forms were printed on transparency, this ensured both groups used the same materials, plus allowed BDAS users to overlay the grid on a printed background image (which themselves were printed on high quality A4 paper). Another form was developed so that the facilitator could make notes in a structured way whilst observing each participant. This too included drawing grids so that they could trace the route of the drawing being created and record interesting traits exhibited by the participant. Coloured marker pens were made available to BDAS participants to take account of the possibility that the colour of a background image clashed with a standard black marker.

The background images made available to the participants in the BDAS condition are shown in Figure 16. Due to our lack of intuition regarding the type of image

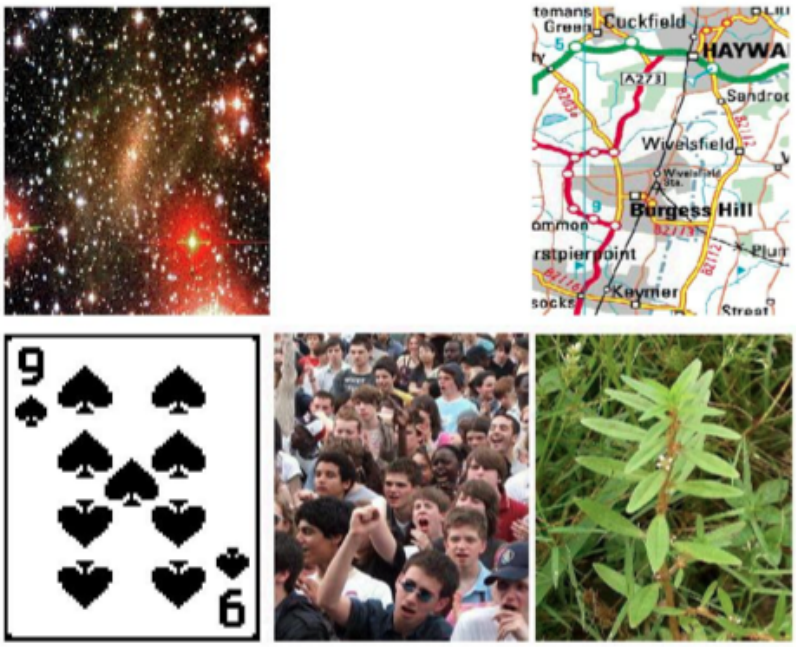


Figure 16: The background images selected for the user study, BDAS participants were able to choose one upon which to draw their BDAS graphical password.

that would be most suitable for BDAS, we provided a small yet diverse set from which participants could choose. These images were selected on the basis that each exhibited one or more desirable qualities for a background image (i.e. in relation to detail, hotspots and the distribution of each within the image). The stars image (Figure 16, top-left) provides users with the opportunity to join the dots (i.e. stars) in their drawings. The crowd image (Figure 16, bottom-middle) and the map image (Figure 16, top-right) both provide a large number of evenly distributed hotspots, and also share characteristics with the crowd and stars images in showing high levels of detail. The plant image (Figure 16, bottom-right) and the playing card image (Figure 16 bottom-left) were chosen as exemplars of images with a small number of hotspots containing low-levels of detail.

3.3.3 Results

We collected 23 drawings in the DAS group and 23 in the BDAS group. Analysis of the memorability results, and the drawings that participants created is presented in the following sections.

3.3.3.1 User-chosen Drawings

The most common drawings that users created were apparently random constructions of lines and shapes which accounted for 30% of all drawings; everyday objects accounted for 20% of user drawings. The most intricate drawings were that of a basketball, and the name of one participant written in Persian script (see Figure 17); both were successfully repeated in the recall tests and were of high complexity. While

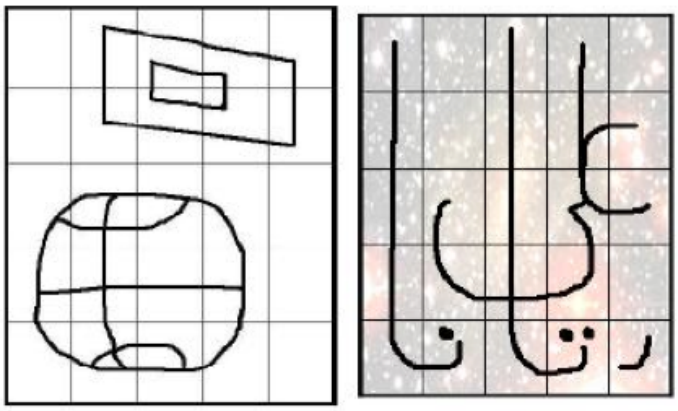


Figure 17: Examples of complex drawings that participants were able to remember: (left) Basketball and backboard created using DAS (7 strokes, length 39) (right) Persian name written with BDAS (9 strokes, length 27).

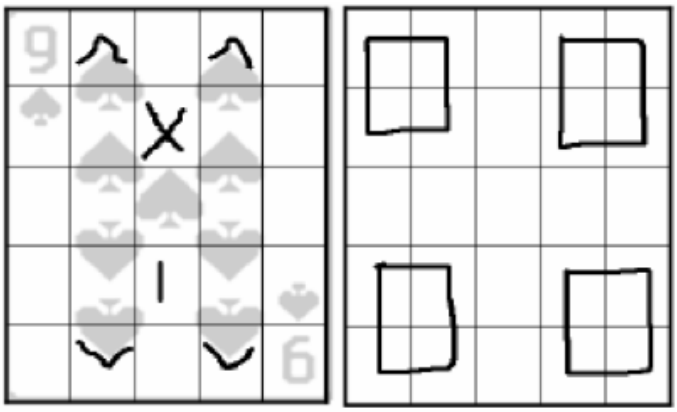


Figure 18: Examples of simple drawings created in the study that had a low stroke count, (left) from a BDAS participant (7 strokes, length 7), (right) from a DAS participant (4 strokes, length 4).

we noted meaningful drawings of high complexity, we also noted drawings of abstract shapes of low complexity. Figure 18 shows examples of drawings that exhibited little complexity. Both drawings involved the creation of shapes that followed a simple pattern. While such drawings proved memorable, they appear to be in the weak class of drawings outline by Thorpe and van Oorschot [131]. In the figures that follow of participant drawings, drawings created with DAS are presented on the blank drawing grid, and those created with BDAS are presented upon the background image, with the colour of the background image slightly faded so as to not occlude the drawing.

The playing card was the most popular background image chosen in the BDAS group (8), closely followed by the plant image (7); stars (4), crowd (3) and map (1) were less popular choices.

3.3.3.2 Drawing Complexity

The complexity of the drawings chosen in each of the two experimental groups is presented in Table 6. The mean stroke count of drawings created using BDAS was

Group	Strokes				Length			
	μ	σ	Max	Min	μ	σ	Max	Min
BDAS	7.22	2.21	12	4	21.43	7.76	37	6
DAS	5.30	2.44	10	1	18.26	9.19	42	6

Table 6: Complexity of drawings created by participants in both experimental groups.

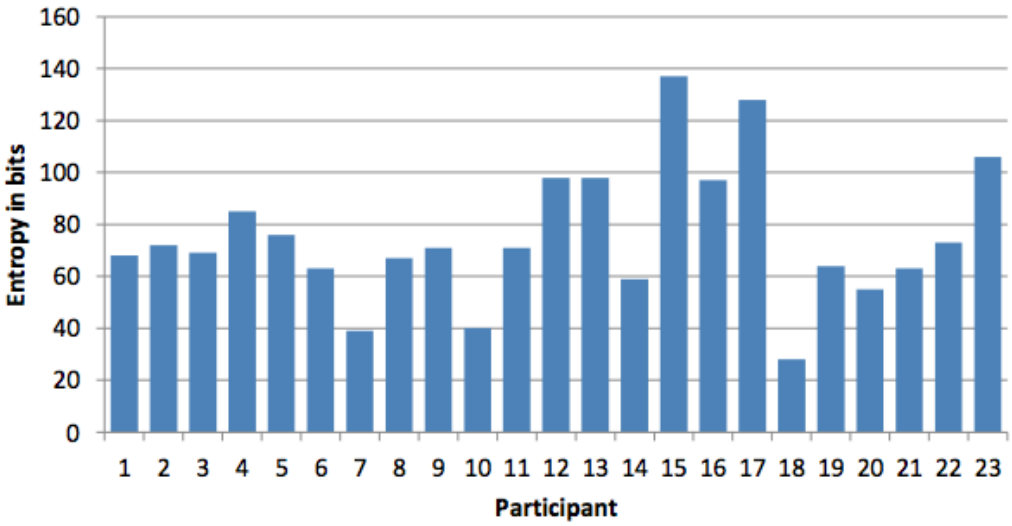


Figure 19: Entropy of the drawings (in bits) created by participants in the BDAS group. Calculations are made from the tables provided by Thorpe and van Oorschot [131]; off chart calculations are estimated.

7.22, compared to 5.30 with DAS ($t(44) = 2.78$,$p < 0.01$), thus the use of background images led to drawings with a higher stroke count. The mean length of BDAS drawings was greater than for DAS (21.43 vs.18.26) although this difference was not statistically significant. The entropy of a graphical password with the mean characteristics shown in Table 6 is 70.2 bits for BDAS and less than 60 bits for DAS; both results show promise for the creation of strong authentication credentials. We noted instances of very high entropy credentials created by participants in both groups, however the entropy of created drawings was significantly greater in the BDAS group (Mann-Whitney U=157.5, $p < 0.05$). The median entropy of a BDAS drawing was 71 bits, while the median entropy of a DAS drawing was 49 bits; the overall distribution of the security of the drawings created by BDAS participants is illustrated in Figure 19, and in Figure 20 for DAS users.

3.3.3.3 Symmetry & Centring

The creation of drawings that incorporate some level of symmetry in its content, and are centred in the grid are thought to be predictable techniques that users could adopt to aid the memorisation of drawings [130]. To explore the prevalence of such features in our own dataset, we attempted to use the qualitative criteria used by Nali and Thorpe [88] in their small study of DAS. We found that 10 drawings (43%) created

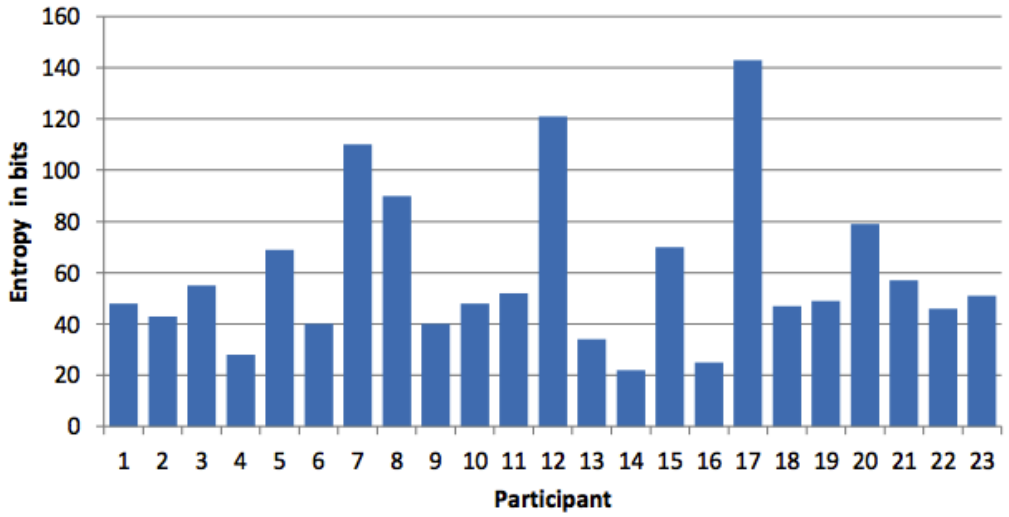


Figure 20: Entropy of the drawings (in bits) created by participants in the DAS group. Calculations are made from the tables provided by Thorpe and van Oorschot [131]; off chart calculations are estimated.

Group	Responses	Correct	Success Rate
BDAS	23	22	96%
DAS	23	23	100 %

Table 7: Recall results for both experimental groups at the five minute recall test.

with BDAS exhibited global symmetry, compared to 13 (57%) in the DAS group; also that 10 (43%) BDAS drawings exhibited centring, while 20 (87%) DAS drawings appeared to be centred.

3.3.3.4 Short Term Recall Test - 5 Minutes Later

Table 7 summarises the results of the recall test conducted five minutes after the enrolment phase; success rates are calculated using the participant success rate introduced in Equation 2 in Chapter 2. Only one person in the BDAS group was unable to repeat their drawing (see Figure 21). The problem encountered by the participant was not related to remembering the visual appearance of the drawing, but the point at which to start drawing the circle that formed the head of the drawn person. Table 8 compares the complexity of drawings that were successfully recalled in both groups. The average stroke count of a BDAS drawing was 7.45 compared to 5.3 for DAS drawings $t(44)=2.9$, $p < 0.01$ indicating that the stroke count of memorable BDAS drawings was significantly higher than that of those created using DAS. The mean length of recalled BDAS passwords (21.7) was not significantly longer than for DAS (18.26). Only 9 (41%) of recalled drawings in the BDAS group had global symmetry, compared to 13 (57%) in the DAS group; 9 (41%) of recalled drawings in the BDAS group were centred, compared to 20 (87%) in the DAS group.

The difference in magnitude of the entropy of mean memorable drawings produced

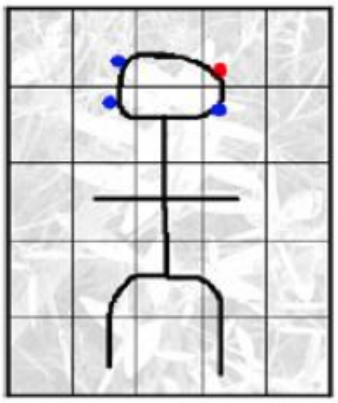


Figure 21: The BDAS drawing that one participant could not recall in the five minute recall test. The blue dots indicate erroneous attempts to remember the starting point of the circle that forms the head of the person (4 strokes, length 18).

Group	Strokes				Length			
	μ	σ	Max	Min	μ	σ	Max	Min
BDAS	7.45	2.26	12	4	21.7	8.31	37	6
DAS	5.30	2.44	10	1	18.26	9.19	42	6

Table 8: Complexity of successfully recalled drawings in the five minute recall test.

in each condition is greater than 10 bits. We also examined the number of attempts users needed to recall a drawing successfully, for which the performance of DAS users was marginally better than BDAS: 18 DAS users succeeded in recalling their drawing first time, compared to 16 with BDAS; 5 users in both schemes needed 2 attempts, and 1 BDAS user required 3 attempts; no DAS user needed more than 2 attempts. This would indicate the usability of BDAS is broadly similar to that of DAS.

3.3.3.5 Long Term Memorability - One Week Later

One week after the initial enrolment and recall test participants returned to conduct a longer term recall test. Two participants from each group failed to return at this stage and so are excluded from further discussion. Of the remaining 42 participants, two participants (one from each group) were unable to repeat their drawing within 3 attempts (see Table 9). Figure 22 illustrates the drawing that a participant could not repeat successfully in the BDAS group, and Figure 23 illustrates the drawing the participant in the DAS group could not repeat successfully.

Group	Responses	Correct	Participant Success Rate
BDAS	21	20	95%
DAS	21	20	95%

Table 9: Recall results for each experimental group at the 1 week recall test.

Table 10 compares the complexity of drawings that were successfully recalled in the

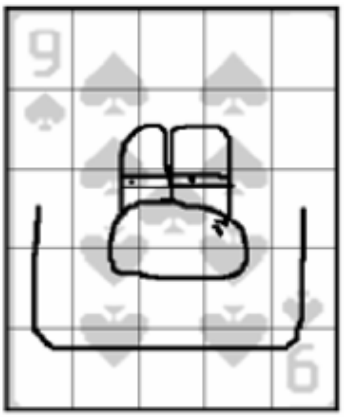


Figure 22: The BDAS drawing that a participant could not recreate correctly in the one week recall test (12 strokes, length 34).

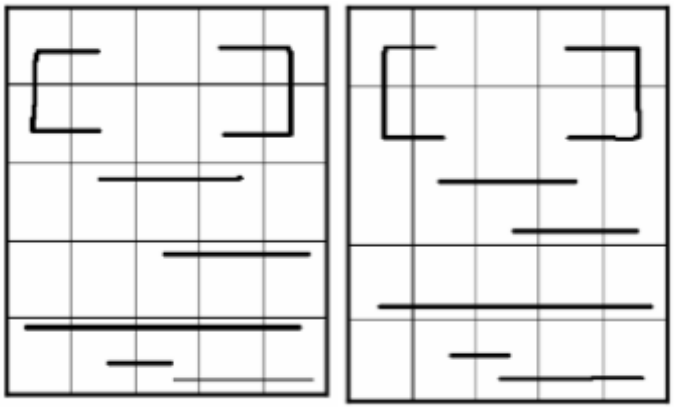


Figure 23: The drawing incorrectly repeated by the DAS participant in the one week recall test: (left) the original drawing; (right) the drawing as created by the participant one week later (7 strokes, length 24).

two conditions at the one week recall test. The complexity of drawings recalled in the BDAS condition was significantly greater than those in the DAS group ($t(40)=2.96$, $p < 0.01$). The length of memorable BDAS passwords was not significantly longer than for DAS passwords (20.9 vs. 17.45). Furthermore, only 10 (46%) of the memorable drawings in the BDAS condition exhibited global symmetry, compared to 13 (59%) in the DAS condition; 10 (46%) memorable drawings in the BDAS condition were centred, compared to 19 (86%) in the DAS condition. The entropy of a graphical password with the mean characteristics observed in Table 10 for BDAS was greater than 70.2 bits, but less than 60 bits for a mean DAS-generated password.

We again also collected the number of attempts each user needed to recall their drawing. DAS users performed marginally better, with first-time 16 recalls, compared with 13 using BDAS. Five BDAS users needed 2 recall attempts, compared with 4 using DAS, and only 1 person (a BDAS user) required 3 attempts. Again we can conclude that although BDAS drawings were inherently more complex, this complexity did not significantly hinder users ability to recall them, even after a period of one week.

Group	Strokes				Length			
	μ	σ	Max	Min	μ	σ	Max	Min
BDAS	7.1	2.16	12	4	20.9	7.71	37	6
DAS	5	2.44	10	1	17.45	7.63	37	6

Table 10: Complexity of successfully recalled drawings in the one week recall test.

3.4 Discussion

The user study has confirmed that BDAS did increase graphical password entropy by significantly increasing the stroke count of memorable graphical passwords. This is encouraging as it has been said that stroke count has a more important role in the entropy of a DAS graphical password than the overall length [131]. In a pilot study we also noted that BDAS significantly affected the length of the graphical passwords created [39]. Qualitatively, it also appears to make drawings less predictable by reducing global symmetry and centring. Furthermore, it appears the DAS graphical passwords were just as memorable as the more complex BDAS graphical passwords created. This outcome is important as any added gains in terms of complexity are not useful without a sufficient gain in memorability. It is possible that if we had conducted a longer study that interesting differences in the ability for participants to recall their drawings would emerge, this would introduce more stress into the memory task and provide a stronger test of memorability.

Background image choices in the study show that the playing card was the most popular image, followed by the plant image. Informal feedback we received suggests that the apparent simplicity of the image gave the impression it would be more usable for purposes of creating a drawing. The feedback from participants choosing the plant image included that the plant divided the picture conveniently and provided a good focal point around which to construct a drawing without being too distracting. The word distracting was a recurring one; our own guess at what makes a BDAS image distracting includes a number of contrasting colours and few elements of symmetry. The remaining three images show much less structure but contain the most diverse colours; this is possibly why they were chosen so few times. The participants selecting the stars and crowd images commented that they simply liked the image as a whole.

Due to the lack of constraints placed upon the drawings users were asked to create, it is not surprising that there were drawings of low entropy created in both experimental groups. For this reason it seems while we were able to significantly improve the complexity of the drawings created by users, it is unlikely we have removed the reliance upon mechanisms such as proactive checking to detect low-quality passwords in both the DAS and BDAS schemes.

In the majority of cases where participants incorrectly recalled their drawing, the visual appearance of the drawings was usually remembered. The nature of a number of recall failures was due to either mixing up the order of the strokes in a drawing, or

forgetting the starting point of a symmetrical shape such as a circle or square. This seems to echo observations by Goldberg et al. [53] that recall errors would result due to the need to remember the composition of the drawing.

3.4.1 Security Implications of BDAS

Throughout this chapter we have calculated the entropy of BDAS passwords using a method suited to unbiased selection of DAS drawings. As discussed earlier, background images in principle could introduce bias to drawings reducing their theoretical entropy. Would this imply that the observed increase in BDAS password complexity may not in fact indicate increased password security? Our estimate of BDAS password entropy did ignore possible negative biases that could be introduced by background images. However, the following are open problems in this domain:

- What are the biases caused by background images?
- How these biases would aid attackers?
- Can security reduction caused by image biases can be compensated by reduced symmetry and centring?

Our study did not collect sufficient data per image to meaningfully explore these questions. However, recent studies [132, 33] demonstrate that a relationship between image hotspots and user choice could impact graphical password entropy in the Passpoints [143] system. The weakness highlighted by these studies is a concern, however at present we have little knowledge on the implications of these results in the different BDAS environment. Firstly, there is a difference in the user-facing protocol of each scheme as BDAS users must create a free-form drawing upon an image, whereas Passpoints credentials are a sequence of explicitly chosen salient positions in the image. This means there are more ways a BDAS user could use a hotspot:

1. The hotspot content is used to aid memorability of the position or form of strokes.
2. The hotspot position is used to aid recall of the positioning of strokes.
3. A combination of 1 and 2.

Due to the way the BDAS image is visually partitioned into relatively large cells by an overlaid grid, it is possible a relatively small number of well distributed hotspots can be sufficient to align the size of the password space targeted at hotspots, with the full DAS theoretical space. For instance, if every cell contained a hotspot in a 5x5 grid, 25 hotspots suitably positioned, could give the attacker no obvious guessing advantage. Due to the finer granularity of selections in Passpoints, many more hotspots would be required to achieve a similar effect. Hotspots are likely to beneficial to attackers when

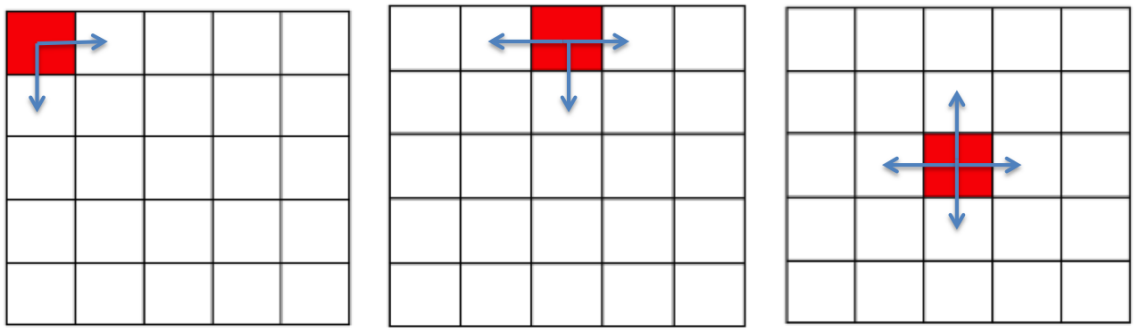


Figure 24: Based upon a password space of at most 40 bits, The resulting impact of the location of a hotspot upon the password space using the assumption that a user will visit a hotspot in every stroke: (left) the hotspot is a corner cell (middle) the hotspot is a border cell (right) the hotspot is a central cell.

there are few of them. An interesting example of the background image appearing to exert influence over users is apparent in the left image in Figure 18. Here the participant has chosen to augment the structure of the background image with simple shapes.

However, If it were empirically observed that hotspots create bias over the drawings users were likely to create in BDAS, then the location of the hotspot could likely contribute to a reduction in entropy. Using a modified version of the calculation provided by Thorpe and van Oorschot [131], Figure 24 illustrates the types of hotspot according to its position within the drawing grid; this suggests that if the user visited a corner cell in every stroke of their drawing, as much as 18 bits of entropy can be removed. These differences in reduction can be rationalised due to the differing number of cells that can be visited from each type of cell. (visualised by the arrows in Figure 24) Further empirical research is required to verify that such an effect would exist in real usage of BDAS.

3.5 Study Limitations

The study was carried out under controlled laboratory conditions which can provide results that illustrate user performance when the participant has no distractions from the task at hand. However, security is often described as a secondary task for users [150], complementary results could be obtained by moving the experiment out of the laboratory to an unsupervised environment that mirrors usage of a system of real value. A knock-on effect of this is that those taking part in our studies had no personal incentive to perform as they were not choosing or using a graphical password to access anything of real-life value to them, which could have impacted motivation in terms of choosing and remembering a complex drawing. Also we did not consider

the impact of user choice bias introduced by background images upon calculations of password entropy, future work could gather a larger amount of data per background image to explore this further. Finally, we did not note significant differences in the memorability of drawings created with either system, however a recall test across a longer period of time could provide a better opportunity for such effects to emerge.

3.6 Conclusion

This chapter has described an empirical study that has shown the potential to passively improve the entropy of DAS drawings through an intervention to the interaction design of DAS; this was achieved without onerous password composition policies and through design of the user interaction with the system. This provides an example of how to harness the greater interface assumptions of graphical passwords to potentially create a new space of interactions to support the user selection of high entropy credentials. In our empirical study, users of our BDAS system chose stronger graphical passwords than their counterparts using DAS, and included fewer instances of weak graphical password characteristics such as global symmetry and centring within the drawing grid. The median entropy of a BDAS drawing was 71 bits, while the median entropy of a DAS drawing was 49 bits. BDAS also appeared to facilitate the memorability of these more complex drawings, as both experimental groups performed similarly in terms of recall success in tests conducted 5 minutes and 1 week after enrolment. This suggests that BDAS could simultaneously support enhanced usability and security.

BDAS did not completely solve the problem of low entropy password selection, as we recorded low entropy drawings from participants using both systems. The need for users to choose and remember drawings that form high entropy credentials could be considered as a compliance defect [26]. This could indicate that DAS and BDAS must in the immediate term retain a reliance upon a graphical password equivalent of proactive password checking [74]. We believe that proactive checking would be more obstructive when applied to DAS due to the likelihood that participants would choose comparatively weaker passwords than BDAS participants. Imposing proactive password checking in this graphical password context could entice users to add complexity to drawings in predictable ways, as has been noted with alphanumeric passwords [76, 99]. One benefit of BDAS in a proactive checking context is that users may be less likely to choose sufficiently weak passwords that fail proactive password checks, and thus avoid the need to spontaneously add new complexity.

The need to entice users to choose high-entropy credentials clearly does not constitute the only way to ensure the security of knowledge-based authentication credentials. The increased difficulty of sophisticated guessing attacks will likely make less sophisticated attack vectors such as phishing [66] or observation attack much more attractive. Observation attack appears to be a pressing problem for graphical pass-

words, however BDAS and DAS to some extent have some protection as replication of the drawing visually is not sufficient to obtain a successful login. Recent research has indeed explored the possibility to redesign the user interaction of DAS to provide observation resistance [151, 79]. Recognition-based graphical passwords are thought to have a particular vulnerability to observation attack due to the simplicity of the user interaction and the memorable nature of the stimulus presented at login. This is a pressing concern, particularly with the increasing prevalence of ubiquitous computing devices in widespread usage.

Chapter 4

Observation Resistant Graphical Passwords through Multi-Touch Interaction

In Chapter 3 we found that an intervention to the user interaction design of Draw a Secret (DAS) [68] supported users to choose and remember significantly more secure graphical passwords, than those using DAS in its archetypal form. That approach was hardware neutral, however Sasse et al. [109] remind us that designing security around the affordances of the specific technology upon which mechanisms are deployed can result in security mechanisms that better fit the context. Such an approach appears particularly appropriate when considering the threat of observation attack due to the increasing diversity of ubiquitous computing [140] technologies that require a user authentication protocol to be performed in public e.g. (Automatic Teller Machine) ATM, mobile devices. This attack involves the use of techniques to surreptitiously view the sensitive user interactions with the authentication system. While this is not an attack that is scalable digitally, the impact of observation attack in everyday life is routinely encountered in an ATM context by the UK Financial Ombudsman [45]; indeed, as more technical expertise is required to carry out sophisticated guessing attacks against large password spaces, attacks would likely be drawn towards more accessible methods to capture login credentials.

Recognition-based graphical passwords are assumed to be particularly vulnerable to this attack, due to exemplar systems being simple to understand, the high memorability of visual stimulus [95, 117], and the user interaction directly exposing the secret credentials in their entirety. The difficulty of designing user authentication resistant to observation attack is embedded in the need to design observation resistance as a core part of the user interaction, in a way that is compatible with the social context. Poor compatibility with the social context can lead to either: (i) sloppy adherence to secure protocols on the part of the user; or (ii) users not performing security behaviours at all. An accepted tenet in security research is the ease with which people

can be persuaded into insecure behaviours, because of the imperative to comply with normative social protocols [84].

In this chapter we populate the design space for techniques for observation resistant input of user authentication credentials; we operationalise our analysis in the context of multi-touch tabletops, a public computing technology that allows particularly expressive forms of user interaction, and is proposed to become commonplace as commercial products such as Microsoft Pixelsense (formerly known as Microsoft Surface) [82] begin to penetrate the marketplace. Such interactive tabletop systems are designed to afford co-located collaboration between groups of users, i.e. the tabletop becomes a communal work-space shared by a small group of friends or colleagues; indeed, future applications are envisioned that include financial transactions, and other security sensitive interactions that require differentiation between collaborators with different levels of security clearance. Where user authentication is required on this platform, it will likely be focused upon something you know authentication as it is cheap and easy to understand. In an empirical study we explore the usability and security of one particularly promising and novel mechanism: *Pressure Grid*, which exploits multi-touch interaction to provide observation resistance to PINs and recognition-based graphical passwords with no adaptation to the underlying authentication scheme.

4.1 Threat Model

Tabletop interfaces and other public displays potentially pose new challenges for privacy and security-related interactions. Indeed, they create a worst case scenario for observation attacks because the displays are are: large, high resolution and designed to accommodate multiple collocated users. These systems will likely be placed in environments such as hotels or restaurants which suggests areas with high public traffic in the surrounding area; people in the surrounding area could be strangers, individuals accompanying the user, either of which could take the role of an attacker. The challenge is made still more pressing by the social context of tabletop usage – close colleagues will not wish to signal mistrust in their fellow users and are therefore less likely to adhere to proper security compliant behaviours (such as shielding PINs). Our threat model consists of resisting at least one observation attack from an attacker who can take any position they please around the tabletop; resisting an attack means that one observed authentication session should not yield sufficient information to perform a successful replay attack. The importance of resisting a single attack is due to the assumption that requiring an attacker to observe an additional authentication session from the same user could be a sufficient deterrent to this kind of threat, as circumstances may not allow this for a period of time. We assume camera-based attacks are possible and can at best be hindered by our approach to design interaction techniques; such a threat is alleviated by the assumption that social norms will prevent the direct

video recording of login sessions by attackers.

4.2 A Framework for Observation Resistant Interactions

To assemble a framework for observation resistant interactions we focus upon software methods that do not rely on additional hardware devices. Where a designer must develop a system to defeat or at least hinder observation attack, which strategies can be called upon to provide authentication that is intuitive for the user,but confusing for the observer? Observation attacks can be complicated by interfering with one or more steps in the observers process of sense making and knowledge acquisition. These strategies are summarised in Table 11.

The strategy to *reduce visibility* involves designing interactions that themselves mask the sensitive areas of the screen, or the usage of computer graphics techniques to actively conceal the secret parts of the user credentials. This approach results in minimal addition to the cognitive load of the user as intervention is confined to the interaction technique itself; one example of this in the context of alphanumeric passwords include the masking of the credentials onscreen with dots, or asking users to shield the keypad.

Subdivide action involves removing the one-to-one mapping between a single user action and a single component of the authentication credential. This increases the number of interactions a user must perform, however increases the number of interactions an attacker must observe. One example of this is provided by Roth et al. [105], who describe a protocol to permit observation resistant entry of PINs in a *cognitive trapdoor game*. In this procedure, the user must perform a number of rounds of a simple protocol per PIN digit to convince the system that each correct PIN digit is known, however the correct PIN digit is never explicitly revealed. However, in an empirical study they found that their system increased login durations by a factor of ten over standard PIN entry.

Dissipate attention involves the presentation of extraneous information onscreen to overwhelm the memory capacity of the attacker; this carries the disadvantage that the legitimate user must also navigate this extra information. This is a common method to create uncertainty for an attacker and a number of schemes harness this technique at least to support other methods of indirection. Tan et al. [126] developed an on-screen keyboard for public displays to protect against observation of alphanumeric passwords. This method also incurred a heavy time penalty for legitimate users, with mean login times when using their novel keyboard 50 seconds greater than when users used a conventional keyboard.

Transform knowledge involves the user being asked to enter credentials in a form that makes it difficult to decipher the original credentials. For example, instead of

Principle	Summary
Reduce visibility	Reduce the saliency of areas on a display where sensitive actions are taking place. This can be achieved through computer graphics techniques e.g. to reduce visual quality or exploit orientation or can rely upon interactions that occlude the sensitive area.
Subdivide action	Remove the one-to-one mapping between one user action and one part of the authentication credential. For instance by dividing a single action temporally or spatially and performing the resulting sub-actions sequentially or concurrently. This increases the work required of an attacker to capture the credentials.
Dissipate attention	Display redundant information to hinder the observer identifying information on the interface that is useful to memorize.
Transform knowledge	Enter the credentials in a form that is difficult, in isolation, to be used to reconstruct the correct credentials. The user could be asked to perform some function over the credentials in memory and enter the result.

Table 11: Principles for designing observation resistant user authentication.

entering the credentials x, the user enters f(x). This is a widely considered technique to defend against observation attack, however it is likely that this approach introduces additional cognitive load to the user. The *Convex Hull Click* scheme [144] is a recognition-based graphical password scheme designed specifically to complicate observation attack. The user is assigned a number of icons to comprise key images that they must locate amongst a large number of decoy images (also icons). At each login the user must locate three key image icons and click within the convex hull formed by their on-screen positions, this can be repeated multiple times to reduce the likelihood of a random guess attack. In their empirical study, the mean successful login duration was 72 seconds, although users were accurate in recognising the images. The patent of Baker [3] describes a simple input mechanism where the user is presented with a 6x6 grid (populated with randomly positioned alphanumeric characters) and must identify a row or column in which each particular character of a memorised password resides. A drawback of this approach as a whole is that while the user does not explicitly reveal their credentials, the interaction still leaks useful information over time.

Each of these identified methods illustrates a trade-off between the usability of the scheme and the desired level of resistance to observation attack.

4.3 Observation Resistant User Authentication on Multi-Touch Surfaces

Multi-touch interaction contributes to the collaborative context as it affords the possibility to exploit a number of interactions not possible in traditional desktop settings. Firstly, visually complex bi-manual manipulations are relatively easy to perform but difficult to reproduce based on observation alone. Secondly, the physicality and directness of multi-touch interaction means that interface elements can be directly touched and direct physical metaphors can be exploited this could improve usability and comprehension of group activities. We attempted to populate the design space of multi-touch interaction methods that could provide observation resistance. The particular affordances of multi-touch interaction we exploit include physical gestures, concurrent actions, and detection of changes in finger pressure. Each approach can potentially be used or adjusted to be used on a number of specific multi-touch platforms such as mobile devices, tablets etc; in each case we discuss an exemplar system and discuss its implications for usability and security [1]. Each example is not intended to provide a perfect solution, but serve to exercise the design space we have introduced.

4.3.1 Exploiting Intuitive Physical Gestures

ShieldPIN (Figure 25) is an interaction technique that forces users to perform a shielding gesture around a keypad before the entry of authentication credentials is permitted; this forms an interlock mechanism that ensures a physical barrier to visibility is in position before the credentials can be disclosed, as the keypad appears and disappears in response to the detection of the shielding gesture. This ensures that the shielding gesture is no longer a voluntary action that could be interpreted as an indicator of mistrust. The entry procedure of PINs and recognition-based graphical passwords would be unchanged which has significant usability and comprehensibility benefits. The defence is based upon security through obscurity, with observation attack likely to be difficult due to the small screen real estate occupied on the interface and the comparative size of the users shielding gesture. In the configuration illustrated in Figure 25, an attacker is most likely to be successful from a vantage point behind the user; careful design of the size and shape of the input area can minimise this possibility.

4.3.2 Exploiting Multiple Concurrent Touches

Colour Rings (see Figure 26) is an interaction technique that augments a typical recognition-based graphical password scheme, and requires the user to select images through the placement of coloured rings within the image grid; the rings must be

[1] Added discussion of this project is provided in our publication at the ACM SIGCHI conference in 2010 [71]

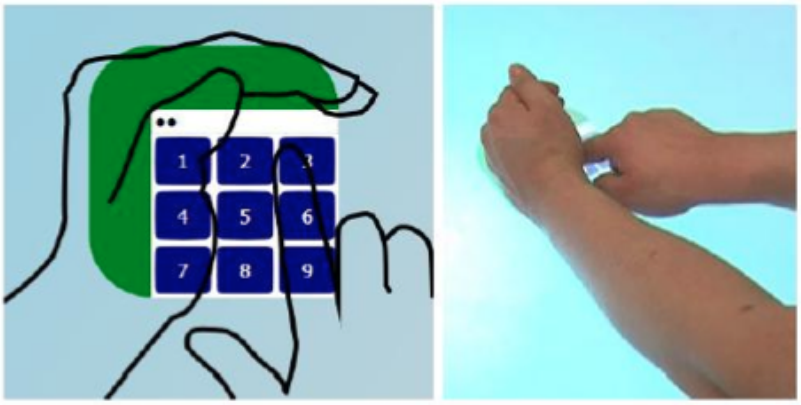


Figure 25: *Exploiting Intuitive Physical Gestures* - ShieldPIN screenshot with added example interaction (left), in situ (right): the PIN keypad only appears once the shielding gesture is detected in the green zone which serves to force the user to enter the PIN and visually cover the keypad.

moved concurrently, and can capture more than a single image. The capture of multiple images at the same time creates uncertainty for an attacker regarding which identified images comprise the graphical password. Also, this overcomes one particularly problematic feature of this genre of graphical password with regard to observation attack which is the highly observable point and touch interaction. In the illustrated instantiation, the user is assigned k key images that are collectively assigned one single colour-ring (from red, green, blue or pink). At login the user is presented with k grids of icons where 72 icons are displayed per grid and one key icon is presented in each. To begin, the user must place 4 fingers down on the display (ideally index finger and thumb from each hand) around which four rings of different colours then appear. Using the four rings the user must drag them concurrently and drop them in the grid, three of the rings should capture images that comprise dummy selections, while the fourth ring (that is their assigned colour) should capture at least the key image.

At each login the position of the icons is randomised and distinct icons may be displayed per grid (although the images are the same between logins). As users can select more than one icon per-ring, there is a need to consider the security impact on guessability. Key determinants of security are the number of rings n, and the number of distinct icons in a grid i and capacity of the rings c. The probability of a random guess $k((\frac{1}{n})(\frac{c}{i}))$ is significantly smaller than the probability of a random guess attack against PIN where n = 4, c = 5,i = 72, k = 4. Clearly, knowing the colour of the correct ring further reduces the number of possibilities an attacker must try after a single observation. Over time this scheme leaks information, however it is likely the task of deciphering the key images on- screen across a number of logins would be difficult where the adversary is restricted to the use of short-term memory. After recording a single successful login the attacker has narrowed down the password space to $log_2(k(nc))$, which still provides greater resistance than provided by a PIN which

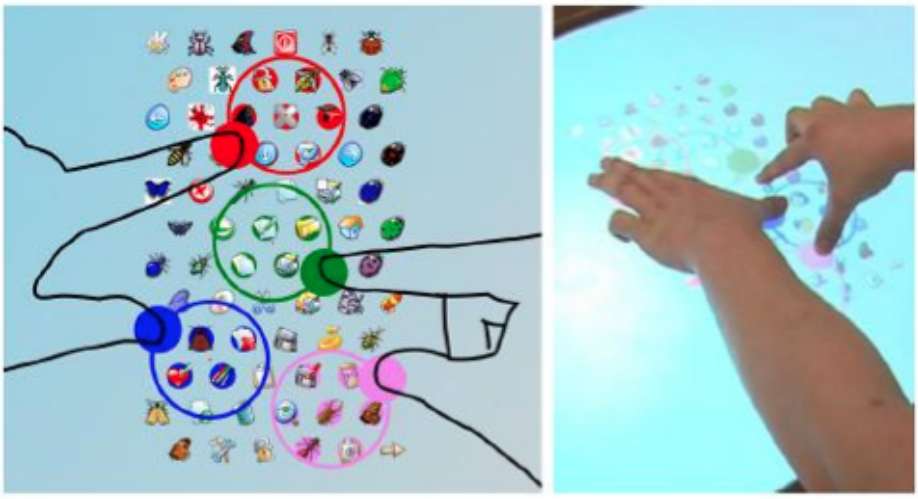


Figure 26: *Exploiting Multiple Concurrent Touches* - Colour Rings screenshot with added example interaction (left), in situ (right): The user drags coloured rings to select key icons amongst decoys, one ring lassos a key image of the user, the rest represent decoy selections to create confusion for an attacker.

is revealed in a single login. Limitations of this approach include that it introduces additional cognitive load to the user as a result of the added need to remember the assigned secret colour. In terms of both usability and accessibility the scheme requires good hand dexterity, and it is possible interaction biases may occur in practice such as the consistent placement of a ring around a key image before any other image.

4.3.3 Exploiting the Invisibility of Pressure Change

Pressure Grid (see Figure 28) is another interaction technique that aims to avoid the point and touch interaction associated with recognition-based graphical passwords and even numeric PINs. The Pressure Grid is a dynamically positioned and sized area of the screen that enables the user through discrete changes in finger pressure to communicate an (x,y) coordinate to identify an object in a grid. This approach exploits the difficulty to identify changes in pressure applied to individual fingertips already resting upon a surface and is particularly suited to multi-touch technology that can detect changes in finger pressure, for example Frustrated Total Internal Reflection (FTIR). Figure 27 illustrates the extent to which differences in finger pressure can be detected from hands resting upon the tabletop exhibiting similar postures.

The user begins by placing three fingers of each hand in calibration areas on the interface; we chose to design for three fingers per hand due to intuition that the 1st, 2nd, and 3rd fingers of each hand appeared to be the most comfortable to use for this purpose. The system uses the location of these touch points around which to dynamically position a grid of objects, and assign pressure zones to each finger, the dimensions of which are dynamically customised by the size of the hands and the spacing between fingers. Once the grid appears, the user is presented with an NxN grid of objects where N should also correspond to the number of fingers per hand used

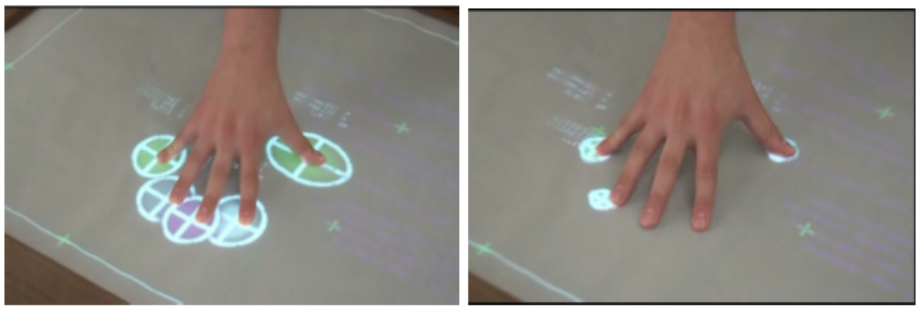


Figure 27: *Exploiting the Invisibility of Pressure Change* - Two images of a hand resting upon a multi-touch surface (based upon Frustrated Total Internal Reflection (FTIR) [57]) where different levels of pressure upon each finger are illustrated by different sized circles upon the interface.

in the interaction. Each cell is referenced by a (x, y) coordinate where x increases from left-to-right and y from bottom-to-top. Each finger on the left hand is assigned a value of y and those on the right hand values of x. For example on the right hand the 3rd finger is assigned x = 3, the 2nd finger x = 2 and the 1st finger x = 1. To select a particular cell, the user must apply additional pressure on one finger per hand. The system can attribute this additional pressure to particular pressure zones, and thus derive an (x, y) coordinate, which can be interpreted as selection of object (x, y). This can be repeated until an entire sequence of objects is selected. If fingers are completely removed from the surface during the input, the login is cancelled as the user may be at risk of exposing components of their credentials.

The key element that underpins the security of this technique is that attackers will have difficulty attending simultaneously to sources of pressure from both hands and the object to which the pressure maps. One possible limitation of this approach is in terms of accessibility as it requires good dexterity of the hands. A camera attack seems difficult, although one useful approach could exploit technology described by Marshall et al. [80], where cameras are used to detect the change in the colour of the flesh beneath the finger nail, caused by pressure of the finger upon a surface.

In the implementation we chose a static pressure threshold used across both hands to distinguish resting fingers and those exerting additional pressure. However, in future the pressure values recorded in the calibration step could be used to assign each finger an individual threshold as the strength and size of a finger likely impacts the pressure it can apply.

4.3.4 Reflection

We conceptually categorise the interaction techniques proposed, along with related work in terms of the four approaches to resisting observation attacks in Table 12. Each of our proposed interaction techniques attempts to reduce visibility, ShieldPIN in the form of the design of the physical gesture to block the view of the keypad, while Colour Rings and Pressure Grid do this due to the reduced quality of the

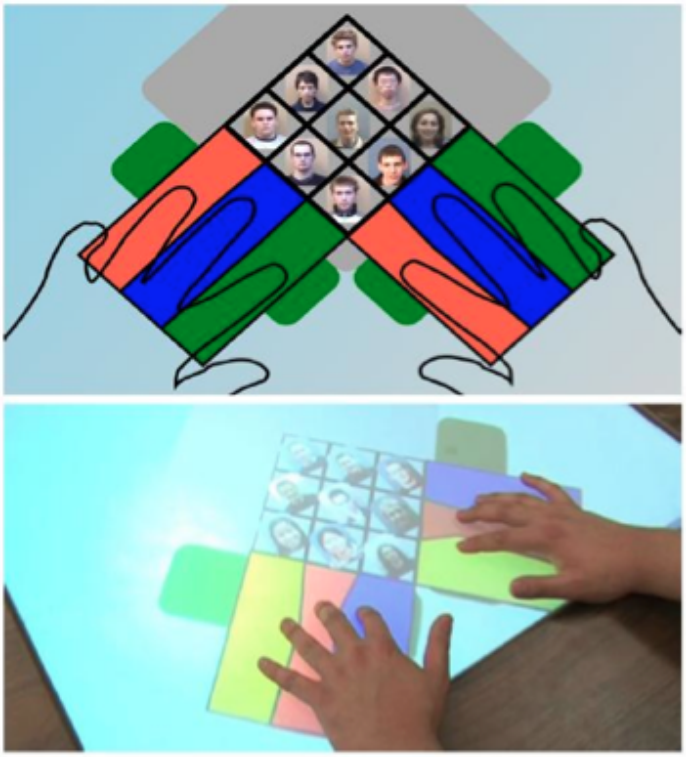


Figure 28: PressureFaces screenshot with added example interaction (top), photo (bottom). The user increases pressure on one finger per hand in the coloured pressure zones to communicate an (x, y) coordinate and select an object.

images onscreen. The Pressure Grid appears promising as it does not increase the cognitive load for the user, as opposed to the Colour Rings technique which relies upon transforming knowledge and requires users to additionally remember a colour. Indeed, related work appears to rely quite heavily on the need to dissipate attention and transform knowledge, two techniques that carry a usability penalty for the user, who must sacrifice usability to gain security [126, 105, 144].

The ShieldPIN system appears to be a very attractive mechanism due to its simplicity, however could still permit observation attack from particular vantage points behind the user, which appears to be a likely attack position. In early user-based pilot work, the Pressure Grid was well regarded and appeared to allow a fast interaction. We chose to evaluate the usability of this method further to explore whether this approach was usable in an empirical study, but also whether perceived benefits of observation resistance would be born out in practice.

4.4 Observation Attack User Study

We carried out a user study to discover the efficiency of the Pressure Grid in terms of both usability and its resistance to observation attack. To reason over the resistance to observation we developed a method similar to Tari et al. [128] where participants would be asked to perform an observation attack upon live input. We

System	Reduce Visibility	Subdivide Action	Dissipate Attention	Transform Knowledge
ShieldPIN	*			
Colour Rings	+		*	+
Pressure Grid	*	+	+	
Cognitive Trapdoor [105]		*	+	+
Spy Resistant Keyboard [126]		+	*	+
Convex Hull Click [144]			+	*
Vibrapass [30]	+		*	+

Table 12: Observation attack resistance techniques used in methods proposed in this chapter and in related work (* = primary; + = supporting).

believed that the most likely real-world manifestations of the Pressure Grid based on current research trends included the PIN, and recognition-based graphical password due to them sharing a reliance upon a point and touch interaction. To provide a graphical password system to evaluate we implemented a system where the image stimulus comprised face images from a stock database; this was designed to resemble Passfaces [96], a system which is a commercial exemplar of this genre of graphical password system. One key operational difference between PINs and Passfaces is that traditional PINs are entered on keypads with fixed digit positions, whereas the Passfaces system randomises the locations of face images at each login. This difference was included when implementing both Faces and PressureFaces. This meant we evaluated four systems: basic PIN, basic Faces, PressurePIN and PressureFaces. For the PIN system the credentials would be a four digit PIN comprised of the digits 1-9, for Faces the user would see four 3x3 grids of faces, in each grid one image would be a key image. Differences in entropy between the two systems were present, but these configurations represent common instantiations of the respective systems.

4.4.1 Method

We chose a within-subject user study where the independent variable was the system, and the dependant variables were the observability of the system and the login time. We recruited 21 participants (14 male, 7 female) to take part in the study. Twelve participants were in the age group 18-30, and nine in the group 31- 50. Participants were either undergraduates or postgraduates in the university. In advance of the study, each mechanism was randomly assigned a correct authentication sequence, which remained consistent for each system throughout the study. Groups of 3 participants were invited to a one hour session, the protocol of the study was explained, and all participants were given time to familiarise themselves with each of the 4 systems. One participant was then randomly assigned the role of inputter for the entire session, while the remaining two were assigned as observers (attackers). Then, a system was chosen at random (from PIN, Faces, PressurePIN, PressureFaces), and the inputter

Figure 29: The custom-built FTIR table used for the evaluation (49x95x105cm) and the user study context.

given time to master the entry of the correct credentials for that system. This was judged by successful input three times consecutively. The observers would then return to the interface and position themselves anywhere around the tabletop interface they chose, and the inputter asked to achieve three consecutive successful logins in their presence. Mistakes by the inputter were ignored. After viewing a login the observers performed a 30 second distractor task (reading a short text). The use of a distractor task is common in memory studies, often to simulate memory stress; its use here was motivated by our assumption that an attacker cannot immediately make use of observed information, and may be required to retain the information over an extended time period or perform other tasks before they can commence an attack. After the distractor task the observers were invited back to the tabletop individually to attempt to re-create the credentials they had observed. Each observer had three attempts to input the correct credentials. If successful in less than three attempts they were not required to login again using that system; The procedure was repeated for each of the four systems.

4.4.2 Study Materials

Figure 29 illustrates the custom built FTIR [57] tabletop system, the dimensions were 49x95x105cm. The figure also illustrates a typical user study context including one instance of the free choice of positioning given to inputter and observers. Each system was instrumented to record the duration of a login (from the first touch to the last touch), and the accuracy of the input.

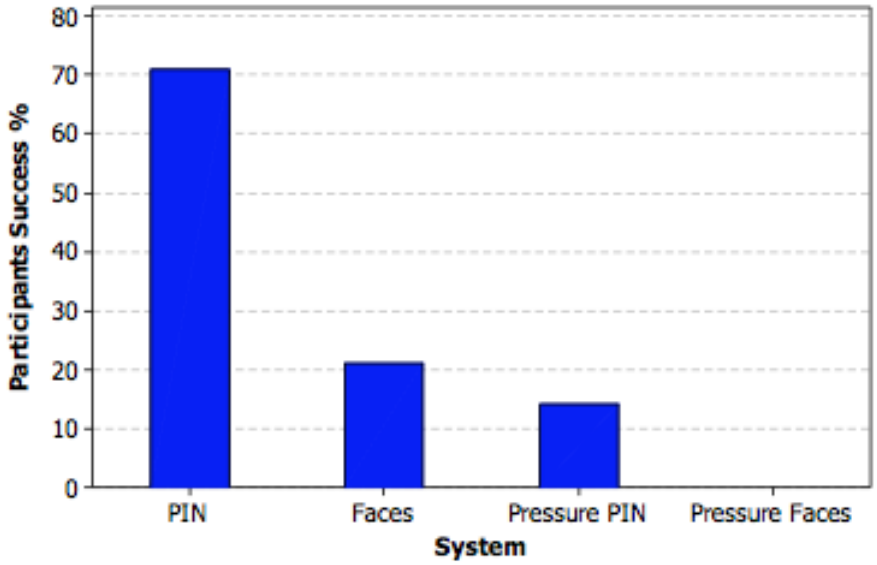


Figure 30: Percentage of observers able to replicate the inputters credentials (by authentication method).

4.4.3 Results

4.4.3.1 Observation Resistance

The key results regarding observation resistance are summarised in Figure 30. Out of the 14 observers, 10 of the 14 (71%) were able to login using an observed PIN. The PIN was considerably more vulnerable to observation than the remaining three systems, confirming our earlier assumption that this mechanism in its traditional form is not appropriate for authentication in public contexts. Faces was considerably more resistant to observation attack with only 3 observers (21%) able to perform a successful login. This could be due to the difficulty of forming fast and effective memory associations with faces, combined with the face locations being shuffled at each attempt (though our methodology does not illustrate which aspect is the most significant). PressurePIN was successfully compromised by 2 observers (14%), which is a significant improvement over a PIN in its traditional form. PressureFaces was not successfully compromised by any observer. This led us to analyse the extent to which components of authentication sequences were recalled (i.e. how many of the 4 faces, or 4 digits, each observer correctly identified). Table 13 shows the accuracy of participants on a per-system basis. Although observers were able to select one correct component of a PressureFaces sequence in 40% of attempts, this success rate appears similar to the probability of making random guesses on every grid and successfully guessing one correct component overall ($\frac{4}{9} = 44.4\%$), particularly given that all observers claimed to have no knowledge of face components when questioned after the study.

System		Components Guessed				
	Logins	0	1	2	3	All
PIN	22	14%	18%	14%	9%	45%
Faces	36	25%	19%	36%	11%	8%
Pressure PIN	38	42%	32%	18%	3%	5%
Pressure Faces	42	57%	40%	2%	0%	0%

Table 13: Rounded percentage of logins where participants guessed a particular number of authentication components (138 attempts collected across all systems).

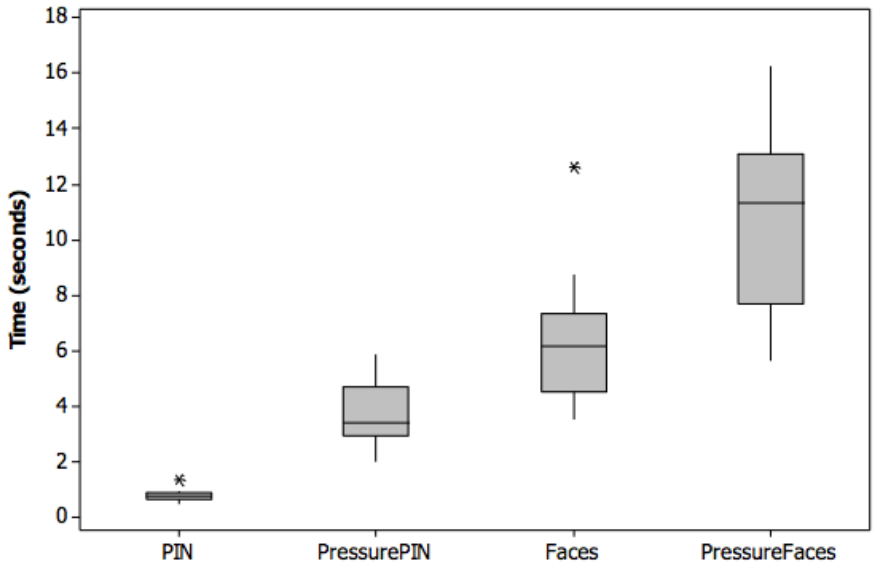


Figure 31: The distribution of successful login durations recorded for each system.

4.4.3.2 Login Durations

In addition to observer success rates, we recorded the login durations for the designated inputters, this would help us determine the impact of the Pressure Grid upon usability. We did not analyse timings for observers as we specified that their concern was to observe and repeat the captured credentials. The mean login time for PIN was 0.79 seconds, and 3.71 seconds with PressurePIN. The mean login duration of a Faces login was 6.3 seconds compared to 10.8 seconds using the PressureFaces. In both cases the impact of the Pressure Grid upon was approximately 3 seconds. These login times were subject to a 2 (PIN vs. Faces) x 2 (pressure grid vs. no pressure grid) analysis of variance using SPSS that demonstrated significant main effects on both factors, with PIN logins proving faster than faces ($F(1, 20) = 61.89, p < 0.001$), and pressure systems proving slower than no-pressure systems ($F(1, 20) = 234.51, p < 0.001$). There was no significant interaction between conditions. The distribution of login times for each of the four conditions is illustrated in Figure 31.

Question	Mean σ
How confident are you *PIN* is safe from Observation Attack?	1.9 (1.2)
How confident are you *Faces* a is safe from Observation Attack?	3.5 (1.3)
How confident are you *PressurePIN* is safe from Observation Attack?	3.1 (1.2)
How confident are you *PressureFaces* is safe from Observation Attack?	4.7 (0.6)

Table 14: Perceptions of PIN, Faces and Pressure Grid on a 1-5 likert scale where 1 is not very confident and 5 is very confident.

4.4.3.3 Questionnaire

After the experiment we asked participants to complete a short questionnaire to elicit opinions on each of the systems and the problem domain. Overall participants were experienced with multi-touch interfaces with 66% having previously used one; 72% were concerned about the ease of observing passwords and PINs entry in everyday life, and 50% of participants reported no confidence in the privacy of their PIN when entered in public environments. Participants were also asked a number of questions to rate on a 5 point Likert Scale regarding their confidence in the observation resistance provided by each mechanism, where 1 indicates no confidence and 5 is very confident. Table 14 illustrates the results. A Mann-Whitney U test reveals that participants believed the Faces system to be significantly more resistant to observation attacks than PIN ($U = 62, p < 0.01$). In addition participants felt that PressureFaces was more resistant to observation than PressurePIN ($U = 30, p < 0.01$). When asked about any fatigue induced to the hands or fingers by Pressure Grid (1 is no fatigue, 5 is high fatigue) participants on average rated this at 1.88 ($\sigma = 0.9$); when asked to rate the usability of the Pressure Grid (1 is not usable, 5 is very usable) the mean participant response was 3.8 ($\sigma = 0.9$).

4.5 Discussion

The user study results confirmed our hypothesis that the Pressure Grid would be a significant defence against observation attack for PIN and graphical password systems on multi-touch interfaces. We were surprised that four participants were unable to login using a captured PIN; reasons for this are a mix of the difficulty to control the manner in which the PIN was entered by the inputter (instructions were not to shield the input), and over-confidence from observers that the PIN had been captured on the first attempt. Another surprising occurrence was that observers were able to compromise PressurePIN. These observers commented that their strategy was to use knowledge of the system workings to focus attention on one hand per observation, and use the third observation to validate the information obtained.

The use of the Pressure Grid added approximately three seconds to the average login duration of both PIN and Faces (see Figure 32). In our study, we noted that our Faces graphical password system was significantly more resistant to observation

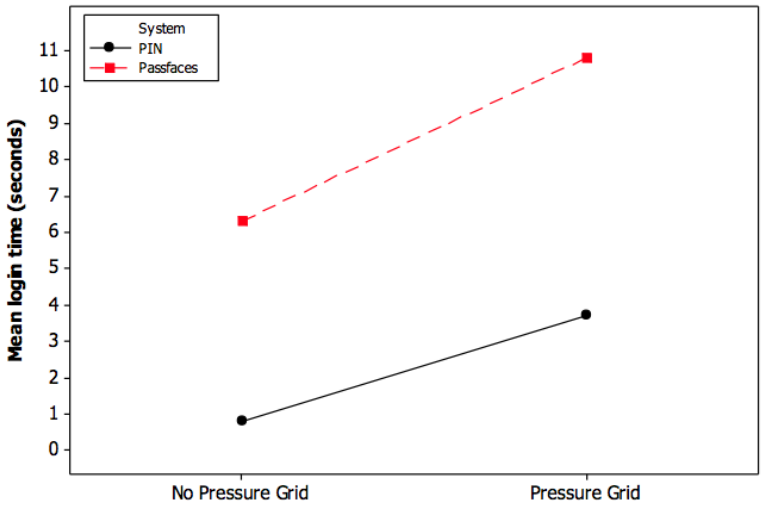


Figure 32: The impact of the Pressure Grid on the login durations of participants using *Faces* and PIN.

attack than PIN, as without the Pressure Grid 50% more participants were able to successfully observe and re-enter a PIN over the Faces system. This is also despite the lower entropy of Faces vs. PIN. A number of explanations are possible, firstly the change in positioning of faces at each login could have made the task of observation more difficult than it was for PIN. Also human face recognition has the interesting property that it is heavily orientation dependent [119] which could have complicated the task for an observer not in an optimal position behind the inputter. Tari et al. [128] also discovered that alphanumeric passwords (not comprising meaningful words) were more vulnerable to observation than a sequence of 5 Passfaces selected with mouse input although the difference was not large. More research with greater numbers of participants is required to firstly prove or disprove this effect, and also determine whether any observation attack resistance is unique to faces, or extends to other images too.

4.6 Study Limitations

The focus of the user study was the observability of the entered credentials, and the usability of the Pressure Grid. The participants were aware of the artificial scenario of being observed while authenticating to our systems, and it is a risk that the ecological validity could be questioned due to the fact that etiquette and typical user behaviour around tabletop interfaces is not yet widely established. Our goal was to test the memory retention of images by attackers given similar conditions to our user study participants. It is possible that more accurate results could be obtained if users did not know they were being observed, although this is likely to be a difficult user study to design.

4.7 Conclusion

In this chapter we explored the approach of securing recognition-based graphical passwords and PIN for observation attack by harnessing the interaction affordances of multi-touch technology. We firstly enumerated the methods available to designers who may seek to defeat observation attack. Designers can restrict visibility, subdivide action, dissipate attention or transform knowledge.

In an empirical study we evaluated the Pressure Grid, a novel interaction technique that relies upon the particular interaction affordances of FTIR [57] multi-touch technology. We discovered that the Pressure Grid significantly improved the resistance of the authentication credentials to observation attack, and only increased the interaction time for PIN and our variant of a Passfaces graphical password by approximately 3 seconds. The results showed that participants observing the Pressure Grid in combination with graphical passwords could not compromise a single login, indeed, the closest any participant came to achieving this across all groups was guessing two of the four components of the credential; conversation with participants after each group suggested any correct entries were guesses. This resistance was likely due to both the image stimulus and the fact that objects in the grid did not occupy a static position between logins.

A number of broad design guidelines can be suggested to assist in the defence against observation attacks against recognition-based graphical passwords: *do not rely upon an active user*: defence should not depend upon a user volunteering to perform an action pertaining to security, such sentiments are echoed by De Luca [29]; *exploit the interaction affordances of specific technologies*: security should exploit the technology upon which it is deployed, this can enable security to better fit the context in which it is deployed; *consider the social context* interactions should not require that a user breaks social conventions to behave securely, as users will work around any interactions that do not comply or will take steps to avoid the system completely.

The approach taken in Chapter 3 and Chapter 4 has been to scaffold desirable security behaviours based upon the design of new interactions built on top of existing graphical password schemes. This has produced promising results, however based upon previous experience with alphanumeric passwords is more likely that in practice less invasive system configuration changes involving the strategic selection of images could also be applied to achieve similar results. In the following Chapter we again consider the threat of observation attack but consider how defences could be instead be implemented based upon strategic presentation of the login images, and not through such a design approach that makes specific assumptions about the deployment environment or technology.

Chapter 5

Observation Resistant Graphical Passwords through Image Portfolios

In Chapters 3 and 4 we demonstrated that redesign of the user interface and interaction with existing graphical password schemes could provide scaffolding for secure user behaviours. This approach considered interventions at the interface level and exploiting the interaction affordances of multi-touch technology. One limitation of such an approach is that it is quite invasive from a system perspective, as flexibility may not exist in the environment to constantly adapt to novel technologies and redesign the interface. For these very cases it is important to explore how users can be shaped in a less invasive, and more spontaneous manner by parameterising aspects of the system. In the context of recognition-based graphical passwords, one underexplored system feature that can likely be manipulated to shape user behaviours in this way is the manipulation of the frequency of presentation or visual content of the images presented to users; it is possible that techniques in this domain could provide scaffolding to both positive and negative behaviours which makes it important to explore the space of possible interventions and their effects.

To provide resistance to observation attack within the aforementioned constraints of preserving the simplicity of the user-facing features of the system and ensuring minimal impact on surrounding infrastructure, De Angeli et al. [28] proposed the key image *portfolio*, which means that a user would be assigned a set of key images but would only be challenged to identify (and thereby expose) a random subset of that portfolio at each login. This potentially presents benefits for observation attack as every login challenge is likely to be visually diverse, however, this approach is undermined by a vulnerability to intersection attack [31] because in its proposed form introduces specific patterns in the frequency that key images appear in a login challenge; these patterns can be monitored over time to enable an attacker to discover the key images.

In this chapter we secure the portfolio key image mode for intersection attack, the resulting system serves to complicate observation attack without the need for users to take part in specifically designed protocols, by creating visual variation in each encountered login in terms of the key and decoy images presented. We evaluate the impact of the resulting system upon usability, grounded in the context of mobile devices to provide a further case study to explore the impact of ubiquitous computing technologies [140] upon the design of security; recent estimates have suggested that in 2012 there will be more data enabled mobile devices than people on the planet [21] which provides critique for the keyboard and mouse assumptions for user authentication. Indeed, the increasing significance of devices for everyday email and banking has made them an indispensable technology for many. We evaluate resistance to observation attack using methodology similar to that used in Section 4.4.1, and carry out a field study where participants used our proposed mechanism on their own personal devices.

5.1 Threat Model

The observation attack threat on mobile devices consists of attacks from an observers collocated with the legitimate user during the authentication challenge; attackers could be individuals known to the user, such as friends, colleagues, and even children. This is due to the user being less likely to suspect malicious (malicious could mean even gaining playful unauthorised access to the device) intentions from those individuals in close proximity; we call this a *friend attack*. Social pressures can prevent users from practicing security conscious behaviour in the presence of known individuals; indeed, the security of an authentication system rests not only upon reliable technology and robust security protocols, but also upon acceptability of the required security behaviours in the social context. Insecure behaviour has been noted as a response to such social pressures [109]. Impersonal attackers could also attempt an observation attack to precede a device theft, although this may alert the user due to the need to be relatively close to see the small display; official UK Home Office statistics report that 700,000 devices were stolen throughout the UK in 2001 [58].

Different from the tabletop context, the mobile device is assumed to always be with the user who must regulalrly authenticate to their device in public to access the device features, giving our proposed attackers many more opportunities to view an authentication session; therefore in an image-based authentication context on mobile devices it should be assumed that observation attack is a core attack. We assume attackers must retain any sensitive information in short-term memory, although the complication of camera-based attacks is a desirable feature; an attacker who records every authentication session has the best chance to comprise a system and this is difficult to defeat without one-time credential entry. A *lunchtime attack* is where an attacker tries to compromise the device protection over time while the legitimate user

is briefly parted from their device; each time the attacker has the intention not to leave a trace to ensure that future surreptitious access to the device will not be hindered, by for example never triggering PIN lockout. There are two classes of lunchtime attacker: firstly a *naive attacker* who can only make random guesses (most likely to be an impersonal attacker), and secondly a *knowledgeable attacker* who by means of observation attack, or another eavesdropping technique, has gained some knowledge that will assist in the login procedure; we propose that the latter is most likely to be a friend attack.

5.2 Intersection Attack vs. Observation Attack

Currently with recognition-based graphical passwords there exists a trade-off between defence against observation attack, and the possibility of intersection attack. Intersection attack can be first seen as a threat in the VIP3 [28] system; the user is assigned a set of key images (called a key image portfolio), and is challenged to recognise a random subset of those images at each login. The desired effect is that the variation in the key images between logins can hinder a replay attack after an observation attack. Intersection attacks can occur because in the VIP3 system the decoy images presented are fixed between logins, but each key image has probability of $\frac{1}{2}$ to appear. The attacker can know that the images that change between authentication sessions are key images. Current wisdom to approaching this trade-off is to sacrifice the goal of observation resistance and ensure every login challenge is identical [31, 59].

We propose a simple yet novel approach that allows the presentation of different images per login and is resistant to intersection attack. In addition to a key image portfolio we introduce a *decoy image portfolio*. If there will be variation in the key images presented across login challenges, there should be exactly the same variation in the presentation of decoy images so that patterns in frequency do not emerge. To achieve this, key and decoy images should both be randomly selected from larger, fixed sized portfolios where the same ratio exists between images randomly selected to appear at login and the size of the portfolio. In other words, if k_l and d_l refer to the number of keys and decoys displayed at any given login challenge respectively, and k and d refer to the total number of key and decoy images being assigned to a user, then the ratios should be chosen as in Equation 3:

$$\frac{k_l}{k} = \frac{d_l}{d} \tag{3}$$

In practice a system administrator would initially choose values for k; k_l; d_l, then calculate $d = d_l \frac{k}{k_l}$. Calculation of the minimum number of images required to bootstrap a system for the given parameters is given by d + k.

5.3 Simulated Observation Attacks

The observation attack resistance offered by our proposed approach does not render systems immune to attack, however increase the difficulty and time required to capture sufficient information to gain an unauthorised login. In order to quantify the benefit of image portfolios approach we created a simulation of an attack that incorporated appropriate behaviour of a user, an attacker and the image portfolios, where $k_l = 4$; $d_l = 32$. The value of k was an independent variable and d was adjusted accordingly to preserve the correct ratios. The intention was to explore the capabilities of an attacker having differing memory accuracies, against increasing size of the image portfolios, where the attacker is attempting to observe and reuse key images. The model encompassed the following behaviour:

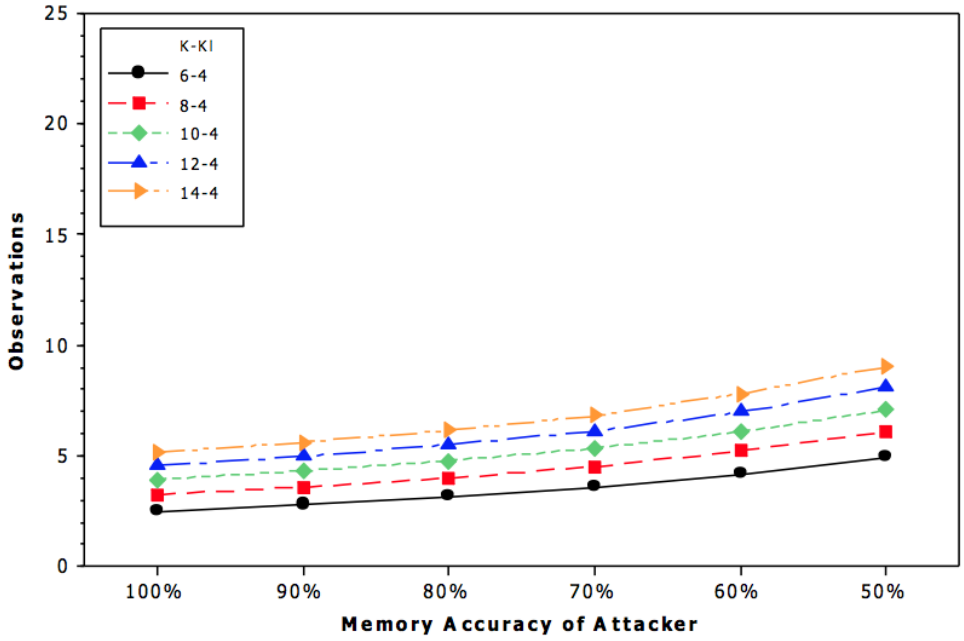
1. A login challenge is presented, the attacker observes user entry of each key image and given a fixed probability p (a parameter to our model) remembers it.

2. The attacker is presented with a new login challenge and has one attempt to authenticate. If the attacker can identify 4 images previously observed being selected, the attack is successful, otherwise unsuccessful.

3. Stages 1 & 2 repeat, with the knowledge of the attacker increasing each time until an attack is successful.

Figure 33 illustrates the mean number of observations required for an attacker to obtain: i) a single login ii) all key images which would guarantee unimpeded future access to the system. Intuitively the number of required observed logins increases with the size of the key image portfolio, and the decreasing accuracy of observer memory. Considering the most likely case where the goal of the attacker is to achieve a single login, a camera equipped attacker on average requires less than five observations (100% memory accuracy), and needs more than 10 only when the size of the key image portfolio is increased to 14. In practice, increasing the size of the key image portfolio to this level is likely to be detrimental to usability; even where the attacker has only a 50% chance to remember observed images the key image portfolio would need to be of size 10 to force more than five observations.

Another case may arise where an attacker would like to obtain the entire key image portfolio; the effect of inaccurate memory has a greater impact in this scenario, particularly for the larger key image portfolios. This is intuitive as if an attacker misses one image, the random nature of the images selected to appear at login means that the missed image may not reappear for some time. The previous model assumes that an attacker is only observing the key images, however it is clear that in viewing a successful login, an attacker with the means for perfect recall (e.g. camera equipped) can also learn from the images the user does not select. In this new scenario, the functionality of the previous model is preserved but the attacker records all images



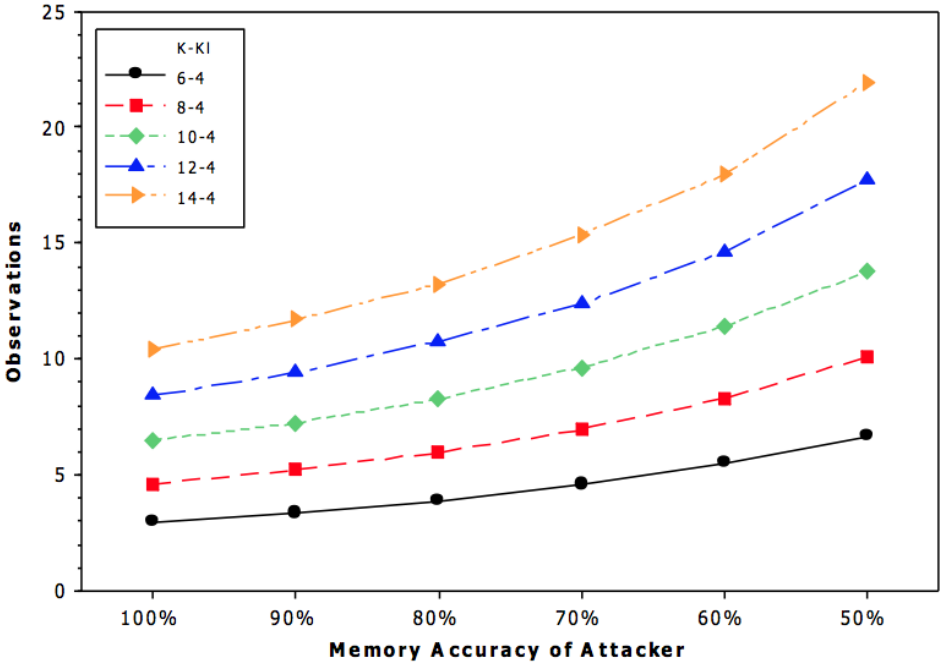


Figure 33: Top: mean number of observations required by an attacker to learn a sufficient number of key images to gain a single successful login; (bottom): mean number of observations required by an attacker to learn the entire key image portfolio, useful for unrestricted future access. Both represent averages of 10,000 simulated sessions.

and selections (where $k_l = 4$; $k = 6$; d = 48; $d_l = 32$). The goal of the attacker is to obtain sufficient knowledge to perform a single successful login. There are three desirable scenarios for the attacker: 1) the attacker can identify all key images in the login challenge; 2) the attacker can identify all decoy images in the login challenge; 3) The attacker knows the key/decoy role of every image in the login challenge. Figure 34 is a Pareto chart that illustrates the means by which an attacker with means for perfect recall can obtain a successful login. On average 84% of attacks are successful due to the attacker concentrating purely on key images presented in the challenge set. In 12%of cases the attacker knows the role of all images in the challenge set, and in only 4% of cases the attacker can identify all decoys in the challenge set (and so derive the keys). It is interesting to note that scenario three is more likely than scenario two. Since the likelihood of scenario three is intertwined with the likelihood of scenario one (which is high), scenario two is least likely since this involves knowing the decoy images and not the key images.

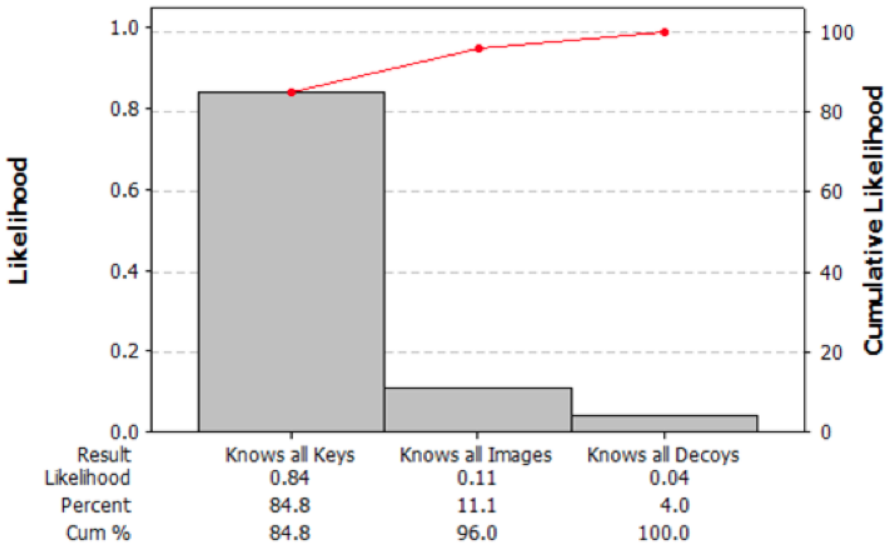


Figure 34: Pareto chart that illustrates the outcomes of 10,000 simulated observation attacks and the likelihood of various contexts that provide an attacker with a successful login.

5.4 Implementation

We developed two implementations to operationalise our analysis, with the ethos that what works in theoretical analysis and in practice can often be disappointingly different. We developed two systems, one with *high entropy* and one of *low entropy* (see Figure 35); this was to reflect that recognition-based mechanisms within certain bounds of complexity are flexible as to the entropy they can be configured to usably provide; both make use of the image portfolios discussed previously, and have the

capability to be bootstrapped using images taken directly using the camera functionality of device. A high level overview of the user experience is as follows: firstly the user provides images to a client application that resizes the images and initiates the image filtering process; the resulting set of images is then transferred to the device, from which image portfolios are populated randomly. We decided to ask users of both systems to 4 random key images (k_l) from the key portfolio of 6 (k). This number was informed by intuitive usability concerns of asking the user to retain more than this number. At each login the selected images from each portfolio are shuffled together and displayed across the chosen number of screens; details of each system are described in the following sections.

5.4.1 High Entropy Graphical Passwords

The high entropy system is the most resistant recognition-based mechanism to brute force attack we have seen on mobile devices; and provides six-times more entropy than a randomly generated PIN, as the probability of a random guess is $\frac{1}{\binom{36}{4}}$. The user interface is visually similar to those seen in previous research [125, 59], where images are presented in a 3x3 grid; research has favoured this design due to an intuitive keypad mapping available between the onscreen images and the numeric keys on a device keypad. The disadvantage of this configuration is that images are displayed small in size which can cause problems for users with imperfect vision. In this system 36 images are displayed across 4 screens in the 3x3 grid; the enforced ratios with respect to the image portfolios are k_l:k = 6:4 and d_l:d = 48:32 indicating a user must provide at least 54 images to bootstrap the system.

5.4.2 Low Entropy Graphical Passwords

The low entropy implementation adopts a new user interface convention for this genre of system on mobile devices; images appear larger, and 24 are spread across 6 screens in a 2x2 grid. The intention was this system would be designed to offer entropy comparable to a randomly generated PIN; the probability of a random guess is $\frac{1}{\binom{24}{4}}$. It could be argued that low screen resolutions seen in many current devices is a transient problem, however designs to accommodate this scenario in the immediate term have implications for usability and accessibility, as users are better able to identify images on-screen. In this configuration the chosen image portfolio ratios are k_l:k = 6:4 and d_l:d = 30:20 indicating in total a user requires 36 images to bootstrap the mechanism.

5.5 User Study

Previous work published in the usable security community has suggested laboratory studies can offer misleading results of password recall when compared to field studies

Figure 35: Screenshots of the *high entropy* and *low entropy* systems. In the high entropy system, images were displayed in a 3x3 grid, and in the low entropy version displayed in a 2x2 grid.

[14]. In the years preceding this work in the wider HCI literature, arguments over the validity of results obtained in these two configurations have long existed; this debate is particularly strong in the field of Mobile HCI due to the recognition that mobile devices are typically used in more dynamic contexts than desktop computers, and so should not be evaluated in the same environment. Nielsen et al. [91] argues that field studies are most effective in uncovering issues of cognitive load and interaction style. Rogers et al. [103] comment that field studies are good at demonstrating how people appropriate technologies in their intended setting, but are expensive and difficult to conduct.

Of course there are arguments that dispute this added value, Kjeldskov et al. [73] comment that field studies are not worth the added value and a good lab study uncovers just as many usability issues. To date, controlled laboratory studies have yielded high success rates in all instances for recognition-based systems; due to only subtle design differences with these systems, we had no reason to believe our mechanisms would perform differently in such a context. This motivated us to shift our attention away from a controlled lab study, to a more pressing issue of how the high entropy and low entropy systems might perform in everyday use on the personal device of each participant. Our research questions concerned the level of user performance that could be expected from such systems in everyday life; how human performance in an observation attack would compare to our modelling of an attack; and finally how users would appropriate the mechanisms into daily life.

5.5.1 Method

We chose a between-subjects study design and recruited 17 participants within the Nokia Research Centre through internal mailing lists, with the incentive of free cinema tickets; in the early phases of the study this was reduced to 16 as the device of one participant failed and participation could not continue. We split participants randomly between the two systems, this meant 8 using the low entropy system and 8 using the high entropy version. All participants were smartphone users for regularly accessing work email and other web-based services, which we believe placed them within a key target group given our perceptions of who might need enhanced user authentication on mobile devices.

Of the 16 participants, 10 were male and six female, 12 in the range 18-28 and four in the 29-39 range and nobody older than 39. Education levels were high with six to BSc level, nine to MSc and one to PhD level. The mobile devices owned by participants were all Symbian S60 devices, with 11 owning the Nokia N95, others included the Nokia E61 and E65; screen resolutions of devices were 320x240 (E61) and 240x320 (N95 and E65). To initiate the study, participants visited the research lab with a portfolio of approximately 80 images either already on their personal device or on removable storage. The assigned mechanism was installed on the personal device of each participant and the resized, filtered images [1] are imported automatically. For the enrolment period the system assigned key images to the user and asked them to achieve 3 consecutive correct logins in the presence of the moderator. The mechanism was not actively securing the device, but was an application that allowed the user to test retention of key images throughout the study, away from study facilitators. Performance data such as success/failure of the login, time/date of the login and login duration were logged automatically, and upon entry of a secret key combination could be output to a file.

The study design was similar to that of Everitt et al. [42] where participants would be sent emails when it was desired they should perform a login; this configuration had the potential to be effective since all participants read emails on their device, the same location on which the mechanism would be installed. For the first week, participants would be asked to login twice per working day; for the second week, participants were asked to login twice per day, every two days. We hoped to gauge the effect of reduced usage on success rates after an intensive first week of usage.

On days when a login was required, we sent participants an email reminder at 10am and 3pm; upon receiving the email the participant would be required to open the application and attempt to authenticate successfully. Upon success, the program disappeared into the background, but if a participant could not login after three attempts they would be *locked out*, and offered a logged reminder of their key images so they could continue with the study. At the end of the study, the participants

[1] More details are given of our image similarity filtering procedure in our publication at SOUPS 2010 [37]

visited the lab so that the log file could be extracted.

5.5.2 Study Materials

Each authentication mechanism was developed as a Java Midlet and contained logging functionality that would record the time of day, the content of logins and the success or failure of each attempt. An example entry from an extracted log file (with added formatting) can be seen in Figure 36. From these entries we were able to analyse statistics of user performance across the period of the study.

**********Login**********
Tue Sep 01 10:42:40 GMT+01:00 2009 Success:false
6.322seconds Keys: 8528634 5027079 7641793 8493317
Decoys: 5328514 836515 2090387 2921383 4222148
288468 7874519 6318636 2298412 4005625 7220047
7546626 9930905 6975851 5766799 3251317 7881604
52555 2854716 8836977
**********End Login**********

Figure 36: Example log entry for a single authentication attempt. The sequence of numbers refers uniquely to images presented at login.

An example log file from the two week study and observation attack study can be found in Appendix C.

5.5.3 Results

5.5.3.1 Success Rates

We define the success rate as $(\frac{successful\ logins}{number\ of\ logins})$ and calculated this across all logins for each system. We collected 319 logins across two working weeks: 178 from users of the low entropy system and 141 from users of the high entropy version. Of the 319 logins, 30 occurred at the weekend outside of the requested time period, however we included these in our analysis and added data to week one. As the number of logins was not strictly controlled, on average participants in the high entropy group logged in 17.6 times ($\sigma = 7.9$), compared to 22.3 ($\sigma = 5.9$) in the low entropy group. Participants were accurate in authentication trials, as only two lockouts were experienced, both from the high entropy group both from the same person in the week of reduced usage. Tables 15 and Table 16 break down the performance into each week of the study, and Figure 37 illustrates the daily fluctuations in success rates across the study.

Across both systems, performance remained stable between week one and week two, with 77% in week 1 and 78% in week two. Analysing success rates per system, this was 77% for both systems. From week one to week two the success rate of low entropy users increased from 70% to 89% and was statistically significant (Mann-Whitney U = 3, $p < 0.05$) however the decreased performance in the high entropy

System	Attempts	Success Rate	Lockouts
High Entropy	85	84%	0
Low Entropy	123	70%	0

Table 15: Success rates and attempts recorded for each mechanism during week one where participants were requested to login twice per working day.

System	Attempts	Success Rate	Lockouts
High Entropy	56	67%	2
Low Entropy	55	89%	0

Table 16: Success rates and attempts recorded for each mechanism during week two, where participants were requested to login twice, every two days.

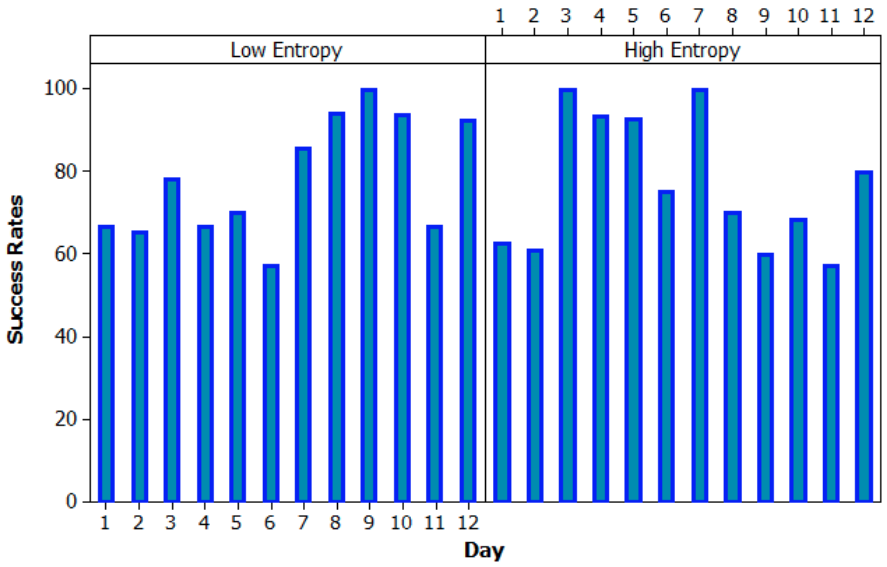


Figure 37: Success rates per day per system; participants were reminded to use the system less frequently during the second week and this affected success rates in the high entropy group.

system, from 84% to 67%, was not significant. The performance difference between the two systems in week two was significant (Mann-Whitney U=3, $p < 0.05$). To understand the success rates in more detail we classified logins using a convention similar to that used by Renaud and Olsen [102] (high entropy- low entropy):

- A single successful attempt not preceded by an erroneous attempt (90- 107).
- 1-2 failed attempts followed by a successful attempt (18-29).
- A failed attempt not followed by a new attempt within 30mins (6-3).

We were also able to consider how success might be affected by the time of day participants carried out their authentication attempts. Across both systems, 145 logins were recorded in the morning (AM), and 174 in the afternoon (PM). This analysis contained no significant differences, nevertheless considering AM logins the

success rate was 78%, and for PM logins this was 76%. Users of the high entropy system experienced degradation in performance from AM to PM. In the mornings the success rate was 83% and in the afternoons this fell to 69%. Users of the low entropy system showed a more consistent performance with an mean AM success rate of 73%, and this increased to 75% for PM.

5.5.4 Login Durations

As well as the accuracy of user recall performance, login durations were recorded (see Figure 38). Login durations were recorded from the user first seeing the login challenge until the final key press; the following discussion refers to successful login durations. The mean login duration across both groups was 19.8 seconds ($\sigma = 3.8$). Considering the high entropy group alone this was 19 seconds ($\sigma = 4.7$) compared to a mean of 21 seconds ($\sigma = 4.9$) using low entropy; in a two-sample t-test this difference was not significant ($p = 0.366$).

Considering the change in login duration for each week of the study however was most interesting; in the case of both systems, login durations became faster. In the first week the mean high entropy login lasted 22 seconds ($\sigma = 4.9$), in the second week this fell to 15 seconds ($\sigma = 3.6$). This was significant in a two-sample t-test ($t(30) = 4.09$, $p < 0.01$). A similar effect was noted for users of the low entropy system, in the first week the mean login duration was 23 seconds ($\sigma = 4.7$) and in the second week this fell to 17 seconds ($\sigma = 2.7$). Again this was significant in a two-sample t-test ($t(30)=2.84$, $p < 0.05$). While in the second week there was less data, both changes were significant. Users of both systems experienced a similar level of improvement in terms of login durations. However comparing systems on a week by week basis did not produce significant results. The fastest instance of a correct login was 9 seconds, with the slowest being 76 seconds; it is likely that the user generating the latter was multitasking at the same time as performing the login.

5.6 Observation Attack User Study

This study was designed to complement the formal analysis of observation attack conducted earlier in Section 5.3 to explore the number of observations required for a human attacker to compromise each implementation in the context of a friend attack. This phase took place one week into the data collection study, where participants attended with others who had been using the same system. A benefit of conducting the study at this point was that participants had already gained one week of experience with their assigned system and so its functionality had become habitual; we believe this equipped participants sufficiently to launch their own observation attacks. Participants were randomly paired and randomly assigned roles as either an attacker or a victim as seen in a similar study [128]. The scenario offered to participants was

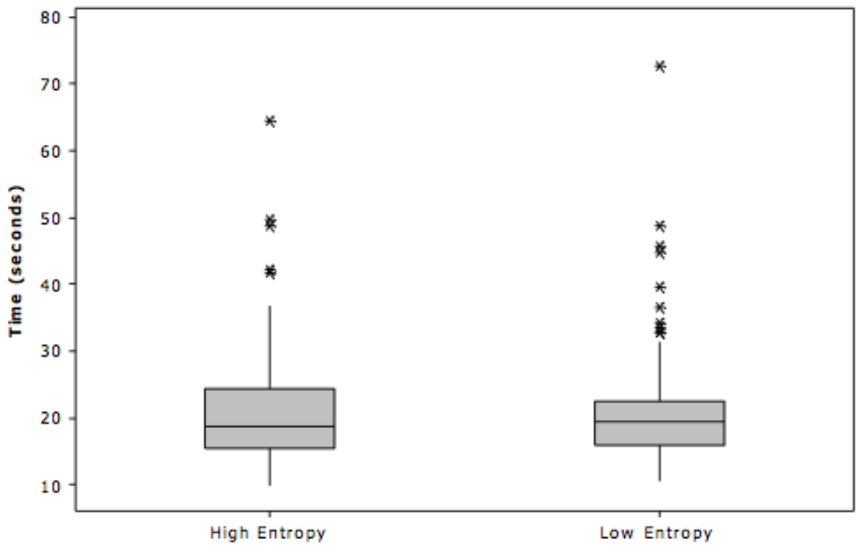


Figure 38: The distributions of login durations recorded across both conditions across the user study.

the following:

> *You and your assigned partner are friends, the victim has just called over the attacker over to show a new application on their device. However while both of you are looking at the screen, the device asks the victim to login to continue. The victim does not know their friend is untrustworthy, and is actively trying to learn their key images. So the victim continues to login...*

The victim was asked to login to their own personal device, holding it in a way that was not sharing the screen with the attacker unrealistically (see Figure 39 for the context). After viewing a login the attacker had a decision to make: the attacker has learned sufficient information to attempt to login, and was given a maximum of 3 attempts to reflect a "three strikes" policy or the attacker asks the victim to perform another login. This occurred a maximum of 10 times. After this phase the participants switched roles and repeated the procedure where the new victim used their own personal device to repeat the process.

5.6.1 Results

The mean number of observations required for observers to login was 7.5 ($\sigma = 1.8$) against the high entropy system compared to 4.5 ($\sigma = 0.6$) against the low entropy version. A Mann-Whitney test shows this to be a significant difference (U = 0, $p < 0.05$). Referring back to the simulations presented in Figure 33, high entropy participants performed as well as attackers with 30% memory accuracy (not detailed on the graph), whilst low entropy participants were approximately 50% accurate. This difference is reasonable since the shoulder surfing task for high entropy participants

Figure 39: The context of the observation attack component of our user study, participants could sit or stand.

was more difficult, more images were displayed and of a lower quality. There were 4 instances where participants were unable to login as an attacker, three using high entropy and one using low entropy. The mean login durations of legitimate users was 16 seconds ($\sigma = 9$) and 23 seconds ($\sigma = 24$) for an attacker ($U = 469$, $p < 0.01$). We calculated this using knowledge of the time and date of this study in the system log. Figure 40 illustrates the difference in the distribution of login durations between legitimate users and observers. There is a clear difference between the login durations of legitimate users and observers. During the study we did observe that the key image portfolio did provide some temporary resistance to an impostor login. In the context of a lunchtime attack this temporary delay could resist attack for a significant period of time depending on the access gained to the device by an attacker.

5.6.2 Questionnaire

After the study we distributed simple questionnaires and had informal discussions with all participants to elicit opinions of the mechanisms and some security practices in general. The small sample size prevented us from gaining meaningful statistical consensus however we still interested to capture user experiences across the study. We hoped that after using the mechanisms on their own devices intensively for two weeks they would have stronger and more interesting comments than if we had performed a short lab study. A selection of questions and comments are presented in Table 17.

When participants were asked if they normally used PINs on their mobile devices, 57% responded yes and 43% no. When asking those who responded negatively if there were items on their device they would consider to be private, everybody responded that there were. This suggests that those users attribute security to their assumption that they will not be separated from their device. Ratings of the time required

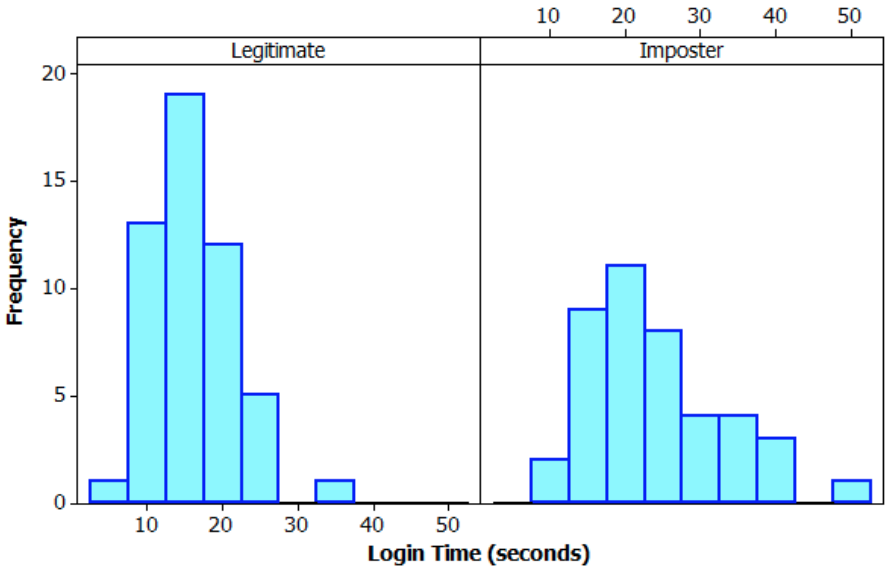


Figure 40: Histograms of the login durations collected from legitimate users and those posing as observers. There is a clear difference between the login durations of legitimate users and imposter users. This could be used to inform design of a login timeout.

Question	Yes	No
Do you normally use a PIN on your device?	57%	43%
Were the login durations of the system acceptable	64%	36%
Did using the mechanism feel secure?	21%	79%

Table 17: Questionnaire results at the end of the two week study.

to login are important as the most crucial driver to user acceptance is often the convenience of use. A disadvantage of graphical password schemes more generally is that it typically takes much longer to authenticate than PIN, due to the visual search required. 64% of users said the time required to login was acceptable, while the remaining 36% thought the time cost was unacceptable. Providing the users with a feeling of security is something that the mechanisms both lacked. 79% of Participants indicated they felt more secure using PINs but could not come up with concrete reasons why. This is possibly because of the transparency of the mechanisms and their game-like nature. Some interesting and recurring comments received in a free-text section of the questionnaire were the following:

"I would prefer to choose my own images"
"When I was walking around I had to concentrate much more than when using a normal PIN, for that I dont need to see the keypad"
"During one login a particularly funny image appeared so I showed it to my colleague"
"Its much harder to crack numbers than images isnt it?"

The majority of users expressed a desire to choose their own key image portfo-

lio. Although, It is common knowledge that users are likely to choose authentication credentials in predictable ways [74]; this trend has been noted in graphical password studies too [27, 33, 132]. Where the distribution of credential is anything but uniform, an attacker can prioritise a guessing attack by any known biases. One feasible concession to this rule could be the system choosing a random subset of images slightly larger than the number of required key images and allow user selection from within this set.

A number of users lamented that the mechanisms demanded their visual attention for use; with PINs they made it clear they were able to enter numeric digits without viewing the screen (due to the tactile nature of the keypad) and multi-task more effectively. A number of users commented how they treated the logins as an enjoyable means to view the images on their device, one in particular commented how they would show particularly amusing images that appeared to work colleagues; this hints that our hypothesis of a friend attack is potentially realistic. The final comment could provide an insight into why many users felt the mechanisms were less secure than PIN. The user has an incorrect mental map of what makes credentials crackable and perceives the game-like nature of the mechanism to be a reflection of its seriousness regarding security. In addition, all their previous experience with mobile device authentication had been with PINs, which likely also greatly informs their preference. A similar effect was noted in a user evaluation of device pairing methods [135] where users associate more difficult with more secure.

5.7 Discussion

During the field study we collected more login attempts than we anticipated; with hindsight this this was inevitable given the novelty of the mechanisms and our decision not to limit the number of logins collected per day. Although this occurrence resulted in an interesting dataset where participants would sporadically use the authentication system we provided, and the overall shape of the data was as we hoped, one week of intensive image followed by a week of reduced usage. The maximum number of logins recorded from a particular person, on a particular day was six. Accuracy was similar across both mechanisms, in addition to the success rates, 60% of logins from the low entropy group could be categorised as correct first time compared to 63% of logins on the high entropy system. All the user accuracy data indicates that one intuitive hypothesis that participants of the high entropy system would have a reduced performance over those performing the low entropy task does not appear to be valid. After considering that this could be an anomaly attributed to the amount of data collected, another possibility is that in a visual search task, practice can flatten out the difficulty of the task. This could suggest that once participants have gained enough practice with the mechanism, performance is not predicted by system entropy.

The results of the observation attack study show both mechanisms to be vulnerable

to observation attack to some extent. Impostors observed on average 7.5 logins using the high entropy version before being able to login, whereas using the low entropy version this was an average of 4.5 logins. While the scenario did not take into account the likely time gap between a lunchtime attacker observing a challenge and having the opportunity to login, the results at least suggest a lower bound to attackers purely using human memory to record images.

In designing and carrying out the field study a number of methodological insights were gained that are relevant to others considering similar studies. Firstly participants expressed apprehension towards full deployment on personal devices; our participants were active smartphone users and busy members of an organisation, and due to the increasing importance of mobile devices they were concerned that unexpected software problems could block them from working. Secondly recruitment is based on specic criteria, participants should own devices on the targeted platform (or devices should be provided). This can reduce the size of the participant pool considerably. Using multi-platform programming languages such as Java can help, although ability to make low-level system calls is reduced. Finally, the devices on a particular platform can be diverse; one platform can contain different devices that can provide different user experiences e.g. screen resolutions and keypads. This hints at the difficulty of studying security mechanisms on emerging ubiquitous computing platforms.

5.8 Study Limitations

The two week duration of the study was relatively short, however was chosen as we hoped to capture the particularly stressful time of learning and using new credentials; this proved to be a sufficient timescale to provide a glimpse of how users would appropriate the systems into their daily routines, and enable us to provoke some strong opinions. Also the mechanism was not actively securing the device and users had to remember to open our application in order to authenticate which is slightly different to a classic user authentication scenario. However , this enabled us to obtain a glimpse of usability away from a laboratory environment. The success rates we report are likely to be under-estimates to what may be obtained in a laboratory study as using the mechanism was likely not the primary task for participants. Also with hindsight we would have limited the number of login attempts that a participant could make on the device per day, although the data we collected provided and interesting dataset where different participants had different habits for using their assigned authentication mechanism.

The observation attack study attempted to recreate an observation attack scenario, and we must consider the ecological validity of this method. This is a difficult phenomenon to recreate in an artificial setting, as typically a victim is unaware of an attack taking place. However, our setup potentially fits well to our friend attack threat model as the victim would know they were under observation. An alternative

approach to recreating such a scenario could involve observers viewing a video.

5.9 Conclusion

In this chapter we have identified an intuitive recognition-based graphical password protocol based upon the strategic presentation of images that provides a layer of observation resistance, is not vulnerable to intersection attack, and does not force excessive levels of indirection into the user interaction. To evaluate the usability of this approach in an everyday context we installed mechanisms on the personal mobile devices of participants, bootstrapped it using personal photographs, and asked them to use it daily. User performance was good irrespective of the mechanism as the success rate was 77% for each system. The success rates reported are similar to those reported in a field study of Passpoints [14] where success rates ranged from 78%-83%; other field studies that provide means of comparison include Passfaces (95%) and Dynahand (97.4%) [102]. While this may seem like a large difference, the low entropy system exhibited higher entropy than each of the aforementioned systems (excluding Passpoints), and it is likely that a longer study would have allowed any performance extremes to stabilise.

In a study that explored the observation attack capabilities of user study participants, we intuitively discovered that users of the high entropy system where images were smaller and of reduced quality needed a mean of 7.5 observations to obtain a successful login compared to 4.5 logins where participants attacked the low entropy system where images were displayed larger and in higher quality. In the worst case 88% of participants attacking the low entropy system were able to successfully login by capturing graphical password credentials. The analysis of the observation attack threat suggests new defence methods: analysis of *login durations*: if an authentication system requires a visual search, or cognitive processing of the login challenge legitimate users are likely to be more skilled and thus faster than attackers; also it appears important to *treat aborted logins as failed logins*: simply viewing the images present in an authentication challenge can enable an observer to train to recognise particular images during an attack.

The strategic use of the image portfolios to provide observation attack resistance can be readily used by system administrators to provide some deterrent to observation attack. However, these image portfolios do not appear to be a scalable method to achieve observation resistant graphical passwords as while increasing their size increases difficulty for an attacker, difficulty is perhaps increased most significantly for the user. Also, the randomness at which key images are presented could mean that some images rarely appear and such rare appearances would likely cause login errors. This could indicate the need to develop functionality for *secure reminders*, or that a focus upon observation resistant interaction techniques is the most fruitful approach to minimise errors and login durations. The design of an effective enrolment pro-

cess such as that seen in the Passfaces [96] online system could serve to provide a more longitudinal encoding of the images in memory where memorability limitations remain.

Chapter 6

Graphical Password Sharing and Image Similarity

In Chapter 5 we explored how regulating the frequency at which particular images were presented to users could provide means upon which to design observation resistant graphical passwords. This approach took no account of the visual characteristics of the images themselves; however, designing security around judicious selection of images based upon their visual characteristics represents an as yet unexplored design space for graphical passwords. In this chapter we explore how strategic selection of images based upon considerations of image similarity could provide security benefits in terms of impacting the ability of users to share a graphical password. Currently, the usability of current methods of knowledge-based authentication (KBA) methods can be attributed to a fine balance between memorability, recordability, and password sharing. The sharing and recording of credentials has become an indispensable (yet often unspoken) coping technique in remembering KBA credentials; Adams and Sasse [1] report that at least 50% of participants in a survey of 139 wrote down passwords as a result of the difficulty of remembering multiple passwords and compliance with password expiration policies. In addition, the practice has been legitimately encouraged where computer systems use group passwords [1], and can serve a social function for purposes such as delegation of access to computer systems.

Recently, the debate regarding the legitimacy of sharing alphanumeric passwords and writing them down has been moderated by an awareness of the need to take into account the realities of different contexts (e.g. that different contexts pose different levels of risk). In an appropriate setting, and when carried out responsibly, writing down and even sharing passwords is considered by some as one usable and secure solution to the password management problem [23].

It appears to be a reasonable assumption that if graphical passwords were to become widely used, users would be likely to attempt to adopt the same coping strategies as observed with other KBA systems; that is, record the password externally or communicate it to a trusted friend or colleague. The role of graphical passwords

in this debate appears interesting, as they are assumed to be particularly resistant to being written down and verbally communicated [31]. However, that users will find it difficult to share graphical password credentials is an unchallenged assumption. Uncertainty regarding the efficacy of users to carry out this practice likely creates further reluctance for wider deployment for graphical passwords due to the significant role password recording already plays in everyday password management. Of course, no methodology exists to measure the extent to which users can share graphical passwords. In this chapter[1] we pursue an empirical method that enables us to reason over the ability of users to share Passfaces [96] graphical passwords and explore the extent to which assembly of image grids based upon qualities of image similarity can facilitate or complicate this practice.

6.1 Threat Model

We define password description as any non-digital attempt to record or communicate a password, using either an external representation or verbal/non-verbal means. This encompasses sketches, written and spoken descriptions and instructions and even accompanying physical gestures. Technologies to facilitate the sharing of graphical passwords clearly exist; given the ubiquity of mobile devices with built-in cameras; these provide an obvious and familiar way to record images in recognition-based schemes. However, we assume that sharing such representations may be unattractive due to their permanence, and likelihood the representations can be identified as graphical password images if discovered by an adversary. It is likely that a more transient and spontaneous way to share graphical passwords will be based upon description. Users may wish to record a description in order to remember the credentials associated with a particular account, or to be able to distribute the credentials in order to grant others access to their system privileges. An attacker may actively seek out such descriptions if they are recorded externally in order to have a route to gain unauthorised access to a system.

6.2 Graphical Password *Description*

Unlike other forms of KBA, most graphical password schemes can neither be precisely written down (excluding Draw a Secret [68]) using a static media (e.g. pen and paper) nor verbally communicated. Users therefore have to revert to the production of a written or spoken *description* of their credentials. In relation to graphical passwords, the nature of users' descriptions of graphical password credentials, and the vulnerability of password schemes to description, have not been previously explored. Description itself is a phenomenon of some interest given the widely held assumptions

[1]The content of this chapter has been published at the ACM Symposium on Usable Privacy and Security (SOUPS) 2008 [36].

related to Passfaces and other graphical schemes. One explicit claimed advantage of the Passfaces scheme over conventional alphanumeric passwords is that Passfaces: *can't be written down or copied" and can't be given to another person"* ([96], pg. 3). A number of subtle configurations have been adopted to mitigate against the risk of easy description, such as that the grids of faces in Passfaces are grouped by gender and are selected to be equally distinctive so that Passfaces cannot be described by gender or obvious characteristics: *"None of the faces stand out from the others"* ([96], pg. 5); and that *"Passfaces can be used in grayscale on all platforms in order to make it even harder for a user to describe their Passfaces to someone else"* ([96], pg.4). As previously mentioned, no methods exist to enable us to quantify the difficulty of sharing graphical passwords. Our objectives were to study four aspects of Passfaces in relation to description: to analyse approaches to description; to quantify Passfaces' vulnerability to description; to evaluate approaches to reduce its vulnerability to description through the choice of decoy images; to explore whether there are any significant gender differences in relation to creating and interpreting descriptions of Passfaces in our sample.

Cognitive gender differences are the subject of active research and debate [56]. However, there are three widely accepted differences: linguistic ability, visuo-spatial ability, and perceptual-motor ability.The term linguistic ability encompasses but is not limited to: verbal fluency, grammar, writing, and vocabulary. One of the key ingredients of effective communication (and thereby description) is a shared and subtle understanding of any vocabulary used, yet potentially males and females are subject to quite different cultural influences. In relation to password description, any differences in terms of linguistic ability are particularly relevant; these appear developmentally though it is not clear if these differences are maintained into old age. A study by Huttenlocher et al. [64] revealed that on average there is a 16 word difference in vocabulary between females and males at 13 months, increasing to 51 words at 20 months and 115 at 24 months. Horgan [34] discovered that females between 2 and 4 years old used longer utterances at a younger age than males, also showing more linguistic maturity by demonstrating ability with the passive voice, participles and adjectives.

Our approach to affect the vulnerability of Passfaces to description centred upon the procedure by which decoy images would be associated to a given key image. To explore these objectives we performed two user studies, the first of which supported the second. The first study involved description collection where we gathered and analysed a corpus of descriptions for faces in a collection gathered from the Passfaces website. This was followed by the second user study, where in a web-based experiment we asked participants to authenticate to a mocked up Passfaces system using only descriptions of the key images in each grid.

6.3 User Study 1 - Description Collection

The first user study was intended to collect a number of face descriptions to seed the second user study. We collected descriptions for a set of 45 (27 female and 18 male) face images taken from the Passfaces online demo.

6.3.1 Method

We recruited 18 participants (9 male and 9 female). The mean age of participants was 26 ($\sigma = 8$). Participants were recruited opportunistically and comprised university undergraduate and postgraduate students or graduate level employees. Each participant was classified according the subject of their highest level qualification, and although computer scientists dominated, there was an approximately even split between computing science and mathematics on one hand, and arts and humanities on the other (Figure 41). The recording sessions took place in a recording studio at Newcastle University, a quiet environment free of distractions and background noise. Participants were provided with an A3 sized sheet of paper containing the 45 images, and asked to examine each face and describe it in isolation (without reference to other faces on the paper) in their own time; that is, using as much time as they required both in preparation for and during the recording. No advice was given on how to approach the descriptions, other than being given the scenario that they were describing faces to a friend. Participants were left alone in the studio and afterwards the recordings were assessed in terms of the sound quality. In two cases it was deemed that the quality of the recordings was not high enough to be used in the subsequent study (due to participant interaction with the microphone e.g. not holding the microphone close enough to the mouth) and two new participants were recruited. The following discussion refers only to participants whose descriptions were usable. We allowed elements of hesitation in the recording, although no examples of this were found to be detrimental to the quality of the description. Each participant recorded verbal descriptions of 15 random faces from the set of 45, providing us with six descriptions per face, 3 male and 3 female.

6.3.2 Study Materials

The key materials required for this study were faces collected from the Passfaces online demo. These were printed out at size 3cmx4cm which is a similar size to how they appear online, and were fixed to A3 paper in a three rows of 15 configuration with each assigned an ID number (see Figure 42). In addition, handheld voice recording equipment was obtained to capture the verbal descriptions from participants.

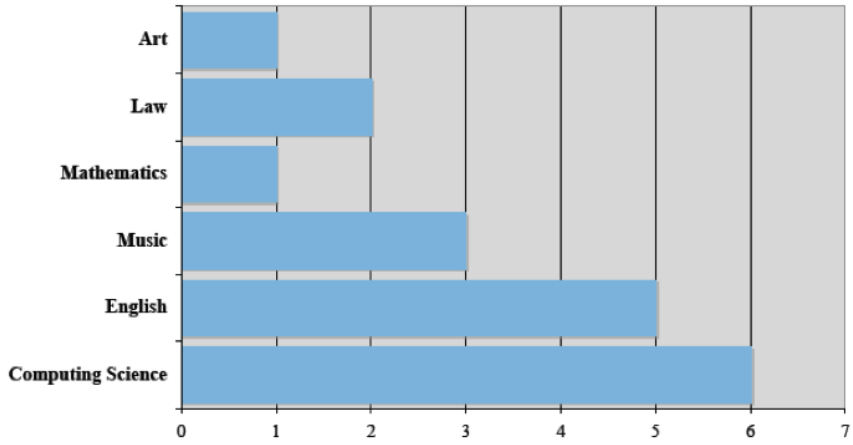


Figure 41: The classification of educational backgrounds of participants recruited to the description collection study.

Figure 42: The sequence of faces given to participants in the Description Collection study.

6.3.3 Results

Our analysis of the results of this phase encompassed simple qualitative and quantitive analysis of the content of the collected descriptions. The most readily apparent strategy used by participants was to first describe a distinguishing feature, either relating to the person (e.g. race) or their face (e.g. hair colour or length, nose or ear size and shape). Where participants did not detect features that were sufficiently distinct, they would systematically describe the face in a feature by feature manner. Where there was a sense that participants became frustrated by their own ability to formulate satisfactory descriptions, they often used the slight presence of clothing in addition to their description of the face itself. Anecdotally, female participants appeared much more likely to adopt a more holistic approach to description. By contrast male participants often resorted to systematic list-like feature-by-feature descriptions of faces. Females often attempted to present a richer sense of the person behind the face and included judgements as to whether the person was happy, likely personality traits, stereotypes, and even social class:

> **Female Describer (feature by feature description):** *"A white caucasian female. Uhm, mousy brown/blonde highlights. Quite big ears. Big mouth. Uhm, wide jaw, big eyes."*

> **Female Describer (Holistic description incorporating. facial features, personality traits and social class):** *"This girl looks young and slightly embarrassed, again kind of, uhm, highlighted blonde hair. She's kind of, uhm, I don't know, uhm, her hair is again kind of half falling out, it's kind of up in a kind of tied back way so, uhm, she looks friendly enough again, looks like she's probably well bred and,uhm, does lots of sport, particularly with horses."*

The mean length of all recorded descriptions was 23 seconds ($\sigma = 11$). Interestingly, the mean recording length for females was 27 seconds, which was greater than that of males at 20 seconds; this difference was statistically significant ($t(269) = 5.37$, $p < 0.01$). Furthermore, the male participants used 567 distinct words (from a total of 4329), and the female participants 654 distinct words (from a total of 5560).

Further examples of faces and collected descriptions are provided in Appendix D.

6.4 User Study 2 - Passfaces Authentication using Description

In the second user study we modelled the scenario of the participant receiving a description of a sequence of Passfaces to login to a resource on behalf of a colleague.

6.4.1 Method

We chose a within-subjects study design and recruited 56 participants for controlled trials that took place during undergraduate computing science practical teaching seminars. Of the 56 participants, 31 were male and 25 female; none had taken part as participants in the previous study. The mean age of participants was 22 ($\sigma = 7$). Rather than mirroring the typical enrolment-login procedure as seen in most user studies of KBA, the study commenced at the login phase. At the presentation of each grid of face images an audio description would be played relating to one of the face images displayed onscreen. In response, participants were required to click the face they believed was being described. To login successfully a participant had to match each of the 5 spoken descriptions with the corresponding face in the grid. For each grid, one face was chosen at random as the target face, with the decoys generated depending on the experimental condition. A particular face could not be the target more than once in the same condition. We also ensured that each description was delivered by a randomly chosen speaker, reducing dependence on a particular speaker. In order to study gender differences, we imposed the additional constraint that participants heard descriptions from speakers that were either all male, or all female (decided at random upon starting the study). Descriptions could be replayed as many times as required using on-screen controls. Participants were asked to perform all 3 conditions in an order determined randomly by our software. To provide a small incentive we offered 5 to participants who were able to match all descriptions in any condition. Though the exact procedure by which decoy selection is undertaken in the commercial Passfaces system is not known to us, we explored three different methods of doing so, each corresponding to a different decoy selection procedure. In this way we could evaluate the potential for image selection to provide resistance against description:

- Random groups: decoys for a target face were selected randomly for each grid (control case).
- Visual groups: decoys are selected on the basis of visual similarities between the target face and the decoys. Visual grouping seeks to maximise the number of decoys that might match based on visual similarity judgments alone.
- Verbal groups: decoys are selected on the basis of the similarity of the verbal descriptions of the faces of the target face and the decoys. Verbal grouping seeks to maximise the ambiguity relating to verbal descriptions.

In advance of the study we assembled image grids according to the indicated criteria for each grouping condition. Where image grids would be assembled according to visual or verbal groups, methods were required to gain consensus regarding which images could be considered visually or verbally similar. In the following sections we describe these methods. Examples of descriptions generated by the participants for

the target face chosen to illustrate the three study conditions in the following sections are as follows:

> *"Long, red brown hair, parting in the middle. Uhm < break > happy girl."*
>
> *"Okay, female, smiling. Uhm, Caucasian I would guess. Slightly longer than shoulder length reddy-brown hair. Dark reddy brown hair. Dark eyes.Uhm, not think not fat, average sort of, uhm,chubbiness. Uhm, slightly pointed chin. Slightly odd looking smile. Uhm, quite big hair, straight again, or slightly curly."*
>
> *"Sort of late 20's, uhm, white girl with long brown hair. Looks a bit like Kate Winslet."*

Though these are all descriptions of the target face, aspects of each do create ambiguity when descriptions are applied to other collocated faces.

6.4.1.1 Random Groups

The base condition was the random grouping of key images and decoy images. When a face was randomly selected to be the target face, the system selected 8 random faces of the same gender to be decoys. This was repeated for each of the 5 grids and within a trial a particular face was only used once as either a target or decoy. Figure 43 shows an example of a randomly generated grid where the target face is highlighted in red.

Figure 43: Randomly assembled grid: decoys are selected at random within the set of faces of the same gender (target face highlighted in red).

6.4.1.2 Visual Groups

Decoy images in the visual groups condition were selected based on their visual similarity to the target face. To gain a consensus on which faces were visually similar, we recruited a second group of participants (similarity judges), whose similarity rankings were used in the selection of eight visually similar decoys for each target face. These volunteers were recruited in computer labs around Newcastle University campus and each was assigned five faces out of our bank of 45. Of these 5 faces they were asked to select as many other faces as possible that were lookalikes. Typically participants selected six or seven, though some were assigned more faces depending on their perceived difficulty of the task. Participant responses were recorded by an experiment moderator. To determine consensus we manually looked through the results for any agreement between participant choices of lookalikes so as to create a set of eight decoys for each face. Where it was not possible to select eight decoys for a particular face based on the similarity judgements alone (i.e. an insufficient number of faces had been identified as similar) we manually selected the short-fall based on our own judgement of similarity. At most only 2 decoys were selected at any one time using this ad-hoc technique. Figure 44 shows the visually similar grid assembled for the same target face. As expected the visual similarity of the members of the visual group is readily apparent.

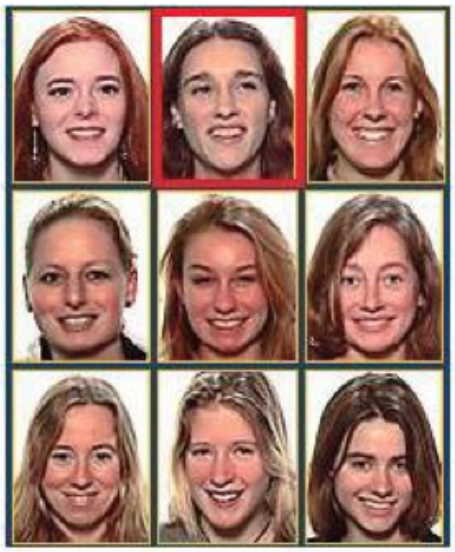


Figure 44: Visual groups: a grid of faces grouped by visual similarity (the target face is highlighted).

6.4.1.3 Verbal Groups

The decoys in verbal groups were selected based on the similarity of the verbal descriptions to the target face. Two descriptions were deemed similar if the same features were emphasised and similar values used to describe those features. We developed a systematic way to construct these groupings based on the set of 250 descriptions. The

first step was to transcribe each collected description, and use the TextSTAT concordance software to collate key terms used in accordance with each face. By looking at the common descriptive terms arising in the descriptions we identified the following features the participants were most likely to include in their descriptions (and the set of values used):

- Hair: used 338 times in descriptions including the values: blond, black, dark/brown, long, tied-back, short, red, curly, fringe.
- Face shape: used 162 times in descriptions including the values: round, oval, small, long, pointed.
- Eyebrows: used 91 times in descriptions including the values: heavy/bushy, shaped, dark, groomed, thin.
- Nose: used 68 times in descriptions including the values: big, crooked, button/small, long, wide, thin/pointy.

Face Shape	
Value	Faces
Round	10,11,12,13
Long	2,7,8,9
Short	34,37,39
Pointed	14,15

Nose	
Value	Faces
Long	2,3,4,5,6,7,8
Flat	22,23,24,25
Pointed	2,23,24,27
Wide	14,15,16

Hair Colour	
Value	Faces
Blonde	2,3,4,5,6,7,8
Brown	14,16,17
Black	33,36,39,42
Red	41,44

Figure 45: Illustrative verbal grouping procedure: Example tabulation of raw data for a cross-section of possible facial features. For face 2, all face numbers occurring on the same row as face 2, for any feature, are candidate decoys.

Figure 45 illustrates the verbal description grouping process. In selecting the verbal group decoys for face 2, we would first make a note of all faces where the adjective long was used. This is repeated for hair, eyebrows and nose descriptors, and scores are incremented for each matching feature value (and decremented for explicit conflicts). The eight faces with the greatest tally are used as decoys for the target face. As already described, in a small number of instances we had to use our own judgement to select decoys where this verbal similarity ranking did not yield 8 decoys. Figure 46 shows an image grid assembled to be verbally similar; the verbal groups exhibit more variation in appearance than the visual groups.

6.4.2 Study Materials

The central component of the user study was a web-based login screen that closely simulated a typical Passfaces login, where users would be presented with 5 image grids

Figure 46: Verbal groups: a grid of faces grouped by verbal similarity (the target face is highlighted).

sequentially with each displayed in a 3x3 formation. The 45 faces presented were the same as used in the first study, which enabled us to construct three grids of female faces and two grids of male faces. The sexes were grouped to mirror standard practice in the Passfaces system. The study was implemented as a website with a PHP and MYSQL backend and logged the content of the authentication challenge and also the user response. An example screen is seen in Figure 47.

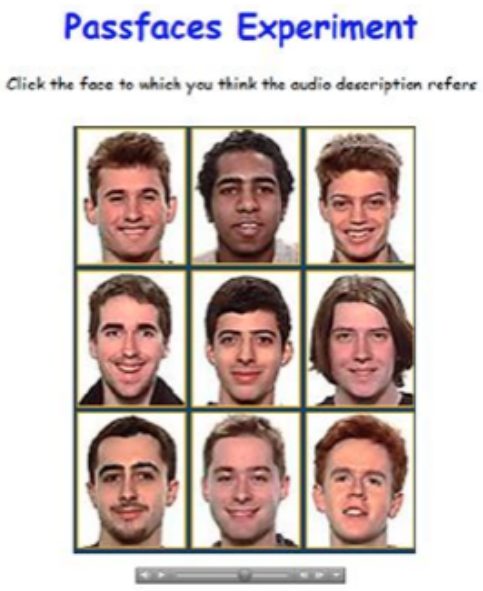


Figure 47: Example Passfaces grid in the description study. Participants were required to select the face to which an audio description refers. The interface widget below the image grid is the audio control panel.

6.4.3 Results

Across the 56 users, we collected 158 login attempts collectively, 54 in the random groups condition, 55 in the visual groups condition, and 49 in the verbal groups condition. Of those 158 logins, 13 (8%) of those logins were successful; that is, where the participant correctly associated all 5 verbal descriptions with the correct target face.

Condition	Attempts	Male/Female	Success
Random Groups	54	29/25	8
Visual Groups	56	31/25	4
Verbal Groups	49	26/23	1

Table 18: Number of successful logins in the different experimental conditions.

Table 18 shows the number of attempts made in each condition, the male/female participant split in that group, along with the number of successful login attempts. Login success was greatest for the random groups and lowest for the verbal groups. The mean score (out of 5) in the random condition was 3.57 ($\sigma = 0.91$) compared to 2.87 ($\sigma = 1.07$) in the visual groups condition t(107)=3.63 $p < 0.01$. The mean score in the verbal groups condition was 2.81 ($\sigma = 1.14$). The difference between the verbal groups condition and the random condition was also statistically significant (t(101)=3.64 $p < 0.01$). Scores in the random condition are concentrated about 3 and 4 out of 5 with 40 of the 54 participants scoring in this range. This theme of scoring highly was also reflected by 8 participants achieving 5/5, higher than both the visual and verbal conditions combined.

The shape of the distribution of the visual groups scores is bell shaped with scores concentrated mainly around 2/5 and 3/5. It is surprising that 5 more participants scored 4/5 in the verbal condition than the visual condition, while so few in the verbal condition scored 5/5. A possible contributing factor is that, as the number of participants in each condition was not overly large, 8 fewer participants in the ?verbal condition did not allow extremes to even out.

Although participants were not placed under time pressure, we measured the time taken to complete each authentication attempt. Our assumption was that participants would take the least time to complete the random groups condition due to the audio descriptions appearing less ambiguous. In fact, the mean timings differed little between the three conditions, the mean was 155 seconds ($\sigma = 172$) for the random groups, 152 seconds ($\sigma = 151$) for the visual groups, and 149 seconds ($\sigma = 155$) for the verbal groups; there was no statistical significance in the difference.

6.4.3.1 Descriptions

Overall participants associated audio descriptions with the correct face 62% of the time (482/780). Using male descriptions participants were correct 60% of the time

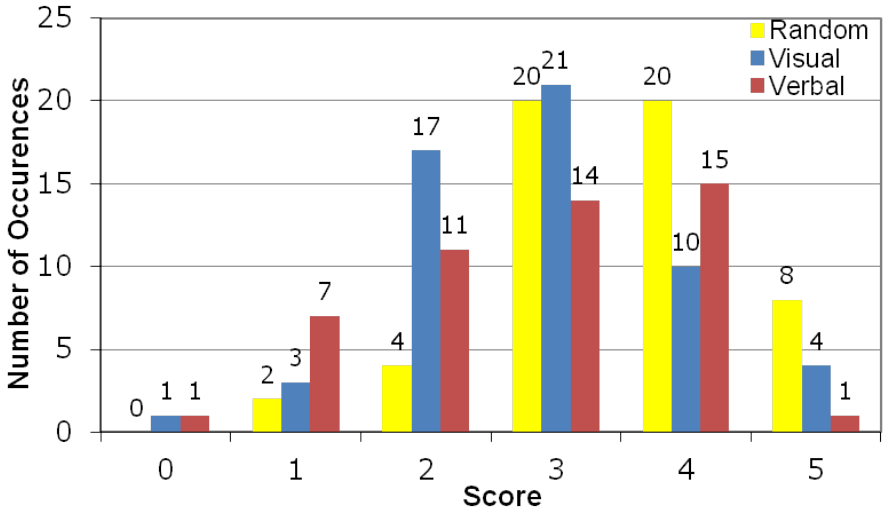


Figure 48: Random vs Visual vs Verbal groups: A breakdown of the scores achieved in each condition.

(238/395) and using female descriptions this figure was 63% (244/385). Some descriptions were clearly more effective than others; the following discussion refers to the descriptions listed in Table 19. For example, we correctly expected the image in Figure 49(a) to be easily distinguished as only two faces out of the 18 male faces had red hair, indeed, though relatively short, these descriptions were highly effective in supporting the identification of the correct face. Description 1 was used correctly 8 times out of the 8 it was played to participants; description 2 and description 3 were both used correctly 3 times out of the 3 they were heard. Unlike face (a), face (b) does not have such distinctive features and as a result the task of constructing an effective verbal description is made significantly more difficult. Description 4 was used twice with no correct responses while description 5 was used three times and also received no correct responses. Describers of faces (c) and (d) experienced similar difficulties and the worst performing descriptions in the experiment were associated with these faces. Descriptions of each of these faces were heard 4 times and resulted in no correct responses:

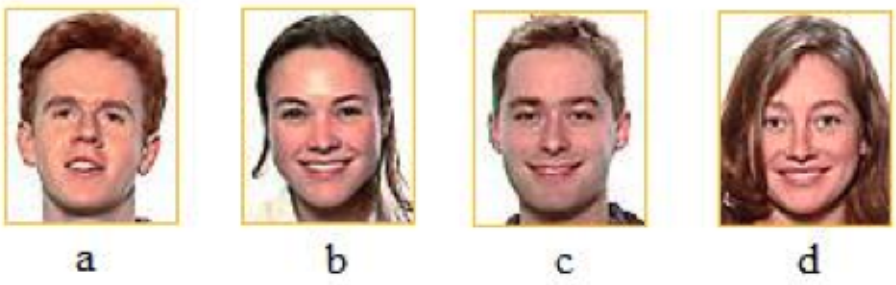


Figure 49: Example faces: participants showed diverse performance depending on the image.

#	Describes	Description
1	Figure 49(a)	*"Big red hair, large forehead."*
2	Figure 49(a)	*"Pale skin, ginger hair, almost smirking rather than smiling."*
3	Figure 49(a)	*"Male, ginger hair. Ginger eye-brows. Uhm, slightly curved eyebrows. Uhm, broadish nose, head slightly tilted back. Not smiling so much. Ears showing around the hair-cut, quite curly hair. Can't see any clothing, uhm, the way the shot's taken"*
4	Figure 49(b)	*"Female with, straggly < break > dark < break > brown with slight lighter blondish< break > highlights or areas within her hair. Tied back presumably. < break > Looking up in the corner, slightly chubby face"*
5	Figure 49(b)	*"A white female, with her hair up but some of it coming across her face like two bits down the side of her face. < break > Uhm< break > uhm < break > like a brown jumper on with a bit of a white collar."*
6	Figure 49(c)	*"Uh, he's got blonde short hair, and his eyebrows are quite prominent, quite thick eyebrows, uhm, but also quite big eyes. Uhm he's got quite a long face."*
7	Figure 49(d)	*"Uhm, large eyes. Uhm, long blondy hair. Happy looking."*

Table 19: Examples of descriptions collected that provided interesting authentication behaviours. These descriptions refer to images in Figure 49.

6.4.4 Gender Differences

On average females outperformed males with a mean score of 3.21 vs. 2.99 (see Table 20 for a breakdown for each condition). However, female performance was more consistent than that of males as displayed (see Figure 50). The upper tail ratio for female participants either achieving 4 or 5 is 1.14. This means that for every 100 males achieving 4 or 5 you would expect 114 females to do the same. We also noticed trends that females scored better using female descriptions. Aside from females scoring the most successful logins using female descriptions, they also performed better overall using female descriptions rather than male descriptions (mean 3.39 vs. 2.97). Interestingly males performed better using male descriptions rather than female descriptions (mean 3.04 vs.2.92) although this was not statistically significant. Table 21 shows a breakdown of successful logins in terms of the gender of the participant and the gender of the speakers in the audio descriptions. Females using female audio descriptions were the most successful, with male listeners using female descriptions a close second. Females marginally outperformed males in every condition.

Condition	Female	Male
Random Groups	3.72	3.42
Visual Groups	2.92	2.83
Verbal Groups	2.96	2.68

Table 20: Female vs male performance in each experimental condition.

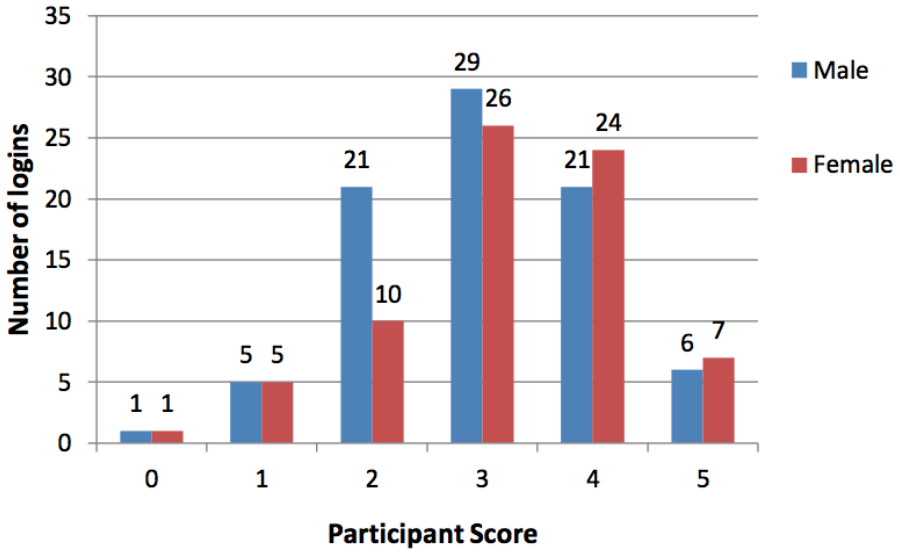


Figure 50: Comparison of male and female performance in the login task, 5/5 indicates a successful login.

Condition	M:M	M:F	F:M	F:F
Random Groups	1	0	1	5
Visual Groups	1	3	0	1
Verbal Groups	0	1	0	0

Table 21: Gender combinations of describer and listener and the number of successful logins that resulted with that combination (e.g. M:M = male listener using male description).

6.5 Discussion

The results highlighted that graphical passwords are not prohibitively difficult to share by description, indeed, although 8% of participants could authenticate successfully with a description overall. Participants in the random groups condition achieved an overall mean score of 3.57, compared with 2.87 in the visual groups condition and 2.81 in the verbal groups condition. As these figures indicate, participants performed best in the random condition, where it was most likely faces were distinctive enough for descriptions to have maximum chance of distinguishing the correct image. Interestingly there was no significant difference between performance in the visual group and the verbal group which suggests each has a similar detrimental effect on the sharing of Passfaces graphical passwords.

Real-world implications of future research confirming and extending these results could involve the similarity of Passfaces decoys being adjusted depending on the security context of system deployment. In corporate environments where password sharing is not desirable, the similarity of decoys to the target face according to some metric could be increased, whereas for other contexts this could be relaxed. While our results did not reveal any significant performance differences between male and female participants, it was interesting to note that females did perform better on

average across all conditions over males. In addition, descriptions created by females tended to include more detail and be significantly longer in duration than those of male participants. As such, females performed significantly better when using female descriptions over those of males. One simple technique that could be adopted to make description of recognition-based credentials difficult is for the system to vary the order of the challenge images presented to the user. In this situation the best a describer could do is to describe all their faces in the hope the listener could identify with one of the descriptions. We speculate this technique would be highly prone to error on the part of the listener.

6.6 Study Limitations

The ability of our participants to authenticate is likely to be a lower bound rather than an upper bound to likely user performance. When collecting descriptions there was such variation in participants' ability that a participant could have been hindered in scoring authenticating successfully in our second study by being randomly assigned a description which to them was not so meaningful. In this situation a participant is almost reduced to a random guess. In a set-up incorporating a two-way dialogue between describer and listener we feel the results would have been more resounding. In addition the experiment could have been performed using a larger bank of images and have been a more socially embedded affair. Such an experimental setting would give a better indication of the likely manifestation of graphical password sharing in the real world.

6.7 Conclusion

This chapter has contributed a methodology and empirical results that demonstrate the degree to which Passfaces can be verbally shared between users, and also how judicious choice of decoys based upon their visual content can reduce the vulnerability to description. Our empirical study has highlighted the reality that, contrary to common wisdom, users can share Passfaces graphical passwords. 15% of logins in the random groups condition were successful, with this being reduced to 7% in the visual groups condition and 2% in the verbal groups condition.The results suggest that the topic of description warrants levels of investigation on a par with other accepted issues facing graphical passwords such as observation attack. We also anticipate that the vulnerability of graphical password schemes to description could have impact on issues of both memorability and observation attack which we did not explore in this work.

Recognition-based schemes such as Story [27] and VIP [28] use images representing very different themes and are likely to be easily described based on the content alone. Schemes such as Deja Vu [31] in part aim to explicitly address description, yet the

computer generated random art yields unique images. Thus Deja Vu may admit description via the very distinctive features that are intended to provide its immunity.

In an alphanumeric password setting users are typically forbidden to write down and share passwords. However, conventional passwords are so well suited to this type of distribution that users are able to share them with anyone they please. If system administrators do not wish users to write down and share passwords they might use a scheme that by its very nature mitigates against the sharing and external recording of authentication secrets. Indeed, as we better understand the nature of password description, the ability to write down and describe credentials might in future be incorporated as an explicit selection criterion for authentication schemes. The significant differences observed highlight the importance that image choice can have upon the usability and security of graphical passwords beyond enticing predictable image choice as observed by Davis et al. [27].

Chapter 7

Towards Automated Detection of Image Similarity

Chapters 4,5 and 6 presented user studies that explored the scaffolding of user behaviours in the context of recognition-based graphical passwords. One issue consistently encountered during these works was the uncertainty over suitable criteria to assemble a usable set of images. While the usability of this genre of graphical password is increasingly understood, it appears likely that the usability benefits are contingent upon subtle attributes of the image sets that are presented to users. It has been noted that visual and semantic similarity exhibited between images has the potential to disrupt the picture superiority effect by causing errors in visual search [35] and rote learning [121]. For security and simplicity, it would be desirable that image sets are assembled randomly; however, the constraints that exist when creating a usable login challenge imply that some degree of skill is required to do this effectively in a manner that preserves security. In Chapter 6 we discovered that the presence of image similarity in the login challenge could hinder the ability of the user to share Passfaces graphical passwords; combined with our proposal that it could also hinder or help usability, it seems likely that image similarity can either positively or negatively scaffold user behaviours during authentication. Without developing a good understanding of the conditions under which such scaffolds are created and manifest, there is likely to be reluctance to deploy recognition-based graphical passwords. Figure 51 illustrates different extremes of assembling decoy images for a particular key image.

Zurko and Simon [153] remind us that user-centred security should be proportioned between end-users, developers and system administrators. Unfortunately such a holistic consideration is lacking in this graphical password context, as there is currently no systematic or empirically verified convention to reason over the similarity in an image set, and as a result this process must likely be performed by hand on the basis of commonsense judgments of image semantics [28]. Users themselves could be asked to tag for similarity, but this can present security threats if users attempt to circumvent the process to obtain an overly simplistic login task. This absence of a

systematic means to evaluate image similarity, combined with the potential impact of inappropriate levels of similarity on authentication error rates, constitutes a significant barrier to the real world deployment of graphical passwords. Indeed, the need for such a spontaneous approach to the filtering of images becomes more pressing when considering other deployment level phenomena such as password resets, where new image sets would need to be generated in response to user demand.

One as yet unexplored approach to solve this problem is to harness image processing research from the field of content-based image retrieval (CBIR) [55]. One fundamental challenge is to determine whether two images are similar. In this field the underlying assumption is that images with similar visual characteristics are more likely to be semantically similar [138]. In this chapter[1] we explore the efficacy of a systematic method to identify instances of visual similarity between images, and explore the impact of its careful manipulation upon usability and security of a recognition-based graphical password login, a context thought to be particularly suitable for graphical passwords yet complicated for similarity judgments [101].

Figure 51: Extremes of decoy image selection for the same key image: (left) decoys are semantically different; (centre) semantic and visual similarity to key image; (right) decoys are semantically similar yet different from the key image.

7.1 Usable and Secure Graphical Passwords and Image Filtering

We firstly define some terms: the *image set* comprises all the images available to the authentication system; the *login challenge* is a subset of the image set, which is comprised of both key images and decoy images and is presented to the user at login. There are a number of conventions regarding the presentation of the login challenge to users, however for simplicity we constrain our discussion to the mode where the

[1]The content of this chapter was published at the Annual Computer Security Applications Conference (ACSAC) 2012 [38].

login challenge is displayed across a sequence of grids, and where one key image is certain to appear in each grid. *Image filtering* is the process of reducing an image set into a login challenge through a process of evaluating the usability and security features of particular images.

The absence of an accepted automated process to perform image filtering has likely, in part, motivated recent research pursuing the identification of an optimal image type for recognition-based graphical passwords [62]. This optimal image type is intuitively one that minimises the burden placed upon a person to undergo the process of image filtering by hand, and one that allows users to perform favourably in recall tests with the login challenges assembled using that process. The drive to satisfy both constraints has led to a focus upon particularly contrived image types that by design permit a narrow range of possible interpretations as to their content (e.g. clipart). The lack of explicit attention given to image filtering even in these contexts appears to assume that it is a one-off procedure, and that the resulting login challenge can be reused for each user of the system. However, likely realities of deployment might make the use of a finite image set unrealistic. For instance, inevitable password resets would mean that previously seen images must be discarded from the image set for a particular user. In addition, if the image set or login challenge is static between users, then attackers can build up knowledge regarding user behaviour with those images e.g. user choice, can permit phishing, and spontaneous distribution of credentials e.g. password sharing, observation attack (due to the images providing a common frame of reference shared between users). This approach also takes little account of results that have suggested users have better memory retention for images they have created [101], nor context-specific defenses that result from strategic selection of image content (as we proposed in Chapter 6).

There is a general lack of knowledge regarding the impact of strategies of image filtering upon usability and security. The assumption so far has been that images should all be semantically and visually different for purposes of usability. A different strategy is illustrated by Passfaces [96], a commercial system where the image stimuli are drawn from a database of normalized face photographs. A brief description offered regarding their image filtering procedure is the following: *"The grids of faces in Passfaces are grouped by sex and are selected to be equally distinctive so that Passfaces cannot be described by gender or obvious characteristics"* [96] pg. 5. This illustrates sensitivity to risks of large semantic differences in the login challenge and the ability for users to share the graphical passwords. Usability concerns must result, however one study of human memory involving 2500 images presented in pairs showed that participants could be accurate at remembering precise image details. Even where images were visually and semantically identical and exhibited only small differences in detail, e.g. orientation, user recognition rates were only marginally worse than when images exhibited semantic differences [10]. The assembly of a login challenge based upon distinct semantics is perceived to improve usability, but those assembled

to incorporate similar semantics could be harnessed to improve security.

In either case, while the curation of a usable and secure login challenge remains a skill residing with those with the greatest experience of doing it, the propagation of such systems more generally is limited. The spontaneous use of everyday uncurated image collections (e.g. photographs) in this context is perceived to be particularly challenging, however, this image type is in some ways attractive, as sets of uncurated images are ubiquitous in personal collections and online. It is possible that if methods of automated image filtering based upon judicious analysis of image content are identified, this could reduce the imperative to identify an optimal image type.

7.1.1 Similarity in the Recognition-based Graphical Password Login

There is currently little convention to follow regarding where to apply systematic analysis of image similarity. Figure 52 outlines points in a typical recognition-based graphical password login challenge that could comprise the image filtering procedure. Analysis can occur on a per-grid and a per-login basis. On a per-grid basis, intra grid key-decoy similarity refers to the similarity between a key image and collocated decoy images. The most usable visual search is one where decoy images appear distinct from the key image [35]. High similarity in this dimension suggests that users might confuse the key image with a collocated decoy image, whereas low similarity suggests the key image would appear to be easier to identify amongst the decoy images. Intra grid decoy similarity refers to the difference between collocated decoy images. In isolation such consideration provides few usability issues, however high intra grid decoy-decoy similarity and high intra grid key-decoy similarity indicates that a grid may overall appear visually similar, which could complicate usability, observation attack and description [36].

The per-grid consideration should be complemented with per-login analysis. Inter grid key similarity refers to the similarity between key images. If there is high similarity in this dimension there is a threat that an attacker might infer that pattern (e.g. all key images are of specific objects or contain particular colours). If this difference is too great, it is likely that images will be more difficult to remember for the user who must remember visually and semantically disconnected images. Inter grid key-decoy similarity refers to the similarity between a decoy image and non-collocated key images. This is important from a usability perspective, as a decoy image may appear to be similar to a non-collocated key image and entice users to select it erroneously. Inter grid decoy similarity considers the similarity of decoy images across a whole login challenge. High similarity in this regard, along with high intra grid decoy similarity, indicates that decoys across the whole login could appear visually similar, whereas high inter grid decoy similarity and low intra grid decoy similarity indicates that there exists similarity within the decoys in each grid, however each grid appears

visually different.

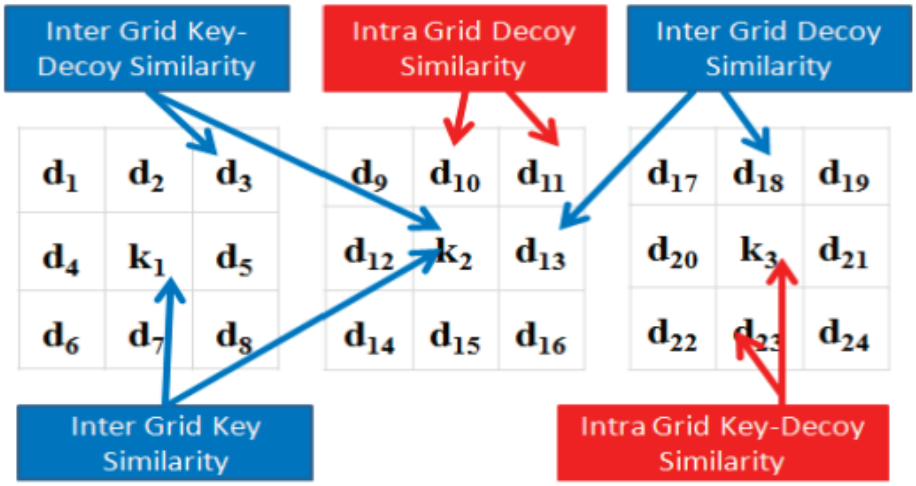


Figure 52: Points at which to consider image similarity across an example login. Red indicates a per grid requirement, and blue indicates a per login consideration.

7.2 User Study 1 - Human Consensus of Image Similarity

Perceptions of image similarity are subjective. However, in order to measure the success of a proposed image processing intervention, it is necessary to first obtain some ground truth notion of pairwise similarity that exists within a particular image set. To do this we carried out a user study to capture a human consensus of similarity within an image set to provide the basis for further study.

7.2.1 Method

We assembled a set of 101 digital photographs and recruited 20 participants (14 male, 6 female with ages $\mu = 27$, $\sigma = 6$) who were staff and students in the research lab. Each participant was asked to organise the printed set of 101 images into piles on a tabletop according to a similarity ranking method proposed elsewhere [123]. This involved the participant being asked to organise the set of images into piles, with the only criteria being that those perceived to be similar should be placed in the same pile. No further advice is offered. The raw data per-participant were the image numbers present within each pile. Across all participants this was aggregated into a score n for each image pair (x, y), where $(x, y) = n$ means that image x and y appeared on the same pile n times, where $n \leq 20$, and high values of n indicate high agreement of similarity. The set of 101 images was intended to be representative of a typical photograph collection. The size of the image set was chosen to provide a manageable sorting task for participants. The images were printed onto high quality paper (100mm x 80mm) and the reverse of each was numbered. For descriptive purposes only we

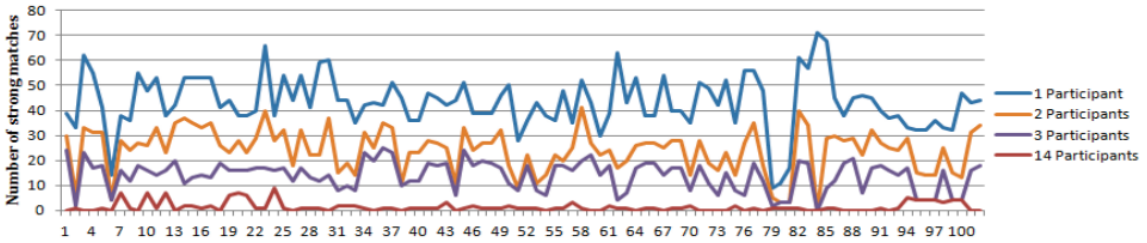


Figure 53: Visualisation of the number of strong matches identified per image in the first study. A strong match is determined by a threshold upon the number of participants that must have classed an image pair as being similar. Overall: 1 participant=4424 strong matches; 2=2462 strong matches; 3=1425 strong matches, 14=148 strong matches.

labeled the images according to the following informal categories: People (9): focus is a person or group of individuals; Scene (30): the focus is purely a landscape scene; Object (14): the focus is purely an object; People/Scene (47): the focus is upon people and scenery; People/Object (1): the focus is upon both people and an object. The image collection contained images taken to a wide range of photographic quality, and was sourced by aggregating a number of personal collections.

7.2.2 Results

Figure 53 gives an overview of the raw output for this study, which highlights the subjective nature of image similarity judgments even across a relatively small image set. For each image, the graph illustrates the number of other images considered to be a strong match for similarity. For a pairing to be considered a strong match, we applied a threshold to n that represented the minimum number of times images should have been placed on the same pile. As we increase the threshold, fewer images are classed as a strong match. The median number of piles participants sorted the images into was 21.5 (IQR = 12.25) with a minimum of 6 piles and a maximum of 32 piles. Images in the people and scene categories generally had the highest number of image matches (Median=45, IQR=13) and those in the Object category had the least (Median=33; IQR=23). No systematic investigation of the strategies used to group images was conducted, but in general it was apparent that these ranged from matching particular objects in the image, to matching the overall context, contrast level or principal colours. Although we only use 101 images in the results, the graph shows 102 (the original number) since one image was misplaced in the course of the study (#84).

7.2.3 Choosing an Automated Similarity Metric

A final phase of this study was to test a number of image processing methods to identify one that was most appropriate to detect the most severe instances of similarity as identified in the sorting task. The threshold of $n \geq 14$ (that is: in our first study,

fourteen or more participants judged two images as similar) provided a basis for us to identify the most severe cases of similarity that any automated mechanism should detect. The field of Content-based Image Retrieval is fast moving; our approach was to test a number of candidate image signatures that would not require extensive expertise in image processing to understand and implement. We reused the images from the first study and performed analysis with those images in the CIE(L*a*b*) color space [50] which is more perceptually linear than RGB or HSV. We implemented each of the following in OpenCV:

- Statistical Moments: treat each channel of a digital image as a probability distribution and calculate the first three statistical moments. To compare two image signatures we calculated the Euclidean distance between the statistical moments of each colour channel and threshold the result.

- Colour Histogram: the histogram bins contain the frequencies of particular pixel values. Firstly we initialise a histogram with 16x16x16x16 bins, which divides each 8 bit color channel into 16 bins. In the normalisation phase, each bin is set to a value between zero and one representing its relative frequency with regard to the other bins. Then we remove any bins with less than 1% of the volume as this can be attributed to noise. To compare histograms we calculate the (EMD) [106] which treats the histograms as piles and provides the minimum cost of turning one pile into the other. The threshold for similarity was 0.9.

- PerceptualDiff [149]: this is not an image signature but is a suite of algorithms that contains a model of the human visual system. Its canonical task is to optimise the computer graphics task of global illumination, by determining whether two scenes are perceptually similar. We were interested to see if a more sophisticated approach held promise.

For each method we made a single pass of the digital images from the sorting study where each was resized to 384x286. We took each image in turn, calculated the corresponding image signatures and compared to every other image in the set, noting the images that were judged to be similar in each case. To calculate the success of these routines, we employed widely used metrics for information retrieval: recall and precision:

$$Recall \;=\; \frac{|relevant\ images| \cap |retrieved\ images|}{|relevant\ images|} \tag{4}$$

$$Precision \;=\; \frac{|relevant\ images| \cap |retrieved\ images|}{|retrieved\ images|} \tag{5}$$

We had knowledge of *relevant* images from the first user study in the form of the strong matches identified by participants for each image. *Retrieved* images are the set of images that the particular automated method judged to be similar. The

metric of recall provides a measure of the fraction of relevant images that a particular method returned. The precision provides the fraction of the returned images that were relevant and is sensitive to false positives. Where thresholds had to be chosen to make a decision of similarity for a particular image processing intervention, they were selected to balance precision and recall. To incorporate spatial information into the calculations we also augmented the statistical moment and colour histogram methods with a vertical or horizontal *region of interest* (ROI). This involved partitioning images with a vertical or horizontal line, calculating the image signature for both halves and using the mean of the two as the result for that image. Table 22 summarises the filtering results obtained for each method. In addition to precision and recall we calculated the F_1 score, which is used to aggregate both precision and recall and represents a weighted average of the two.

Overall, the colour histogram image signature applied to whole images provided the best recall at .58. The addition of spatial information to the image signature through the vertical ROI gave higher recall than the horizontal ROI but also introduced more false positives. The use of statistical moments was less effective than the colour histogram in all configurations, as recall was .34, and this in fact dropped with the introduction of ROI, although ROI eliminated false positives. PerceptualDiff produced a lower recall than both colour histogram and statistical moments. The use of PerceptualDiff was most effective at returning very strict matches where visually the objects and colours in the scene appeared similar. As might have been anticipated, the recall using the ROI was consistently lower than signatures based upon whole images, but ROI augmentation also yielded fewer false positives. This was reflected in a high score for precision.

Method	Recall	Precision	F_1
Colour Histogram	.58	.95	.4
Colour Histogram & Vertical ROI	.48	.8	.3
Colour Histogram & Horizontal ROI	.41	1	.3
Statistical Moments	.34	1	.3
Statistical Moments & Vertical ROI	.2	1	.2
Statistical Moments & Horizontal ROI	.07	1	.1
PerceptualDiff [149]	.24	.9	.2

Table 22: Results from filtering procedure on an image set with 800 photographs, resized to 384x286.

Since the color histogram approach provided the best recall and the highest F_1 score, we chose to use this in our second study. The efficacy of the colour histogram is likely because that representation captured the diversity of colour without being restrictive spatially. This method did not provide a perfect recall score; however, we believed this score was difficult to better given the set of images in use.

7.3 User Study 2 - Recall Test Using Automatically Selected Image

The first user study suggested that the optimal image signature we tested was the colour histogram, as it provided the closest predictor of the human similarity judgments we collected. Our remaining research question concerned whether systematic manipulation of thresholds chosen for this image signature could have a predictable impact upon the short-term recall of users in a typical graphical password login.

7.3.1 Method

We chose a between-subjects study design where the independent variable was the similarity between a key image and its decoy images, and the dependant variables were user performance in terms of recall and login time. We developed a web-based system that would challenge the user to identify four key images across four grids of nine images in a 3x3 layout, with one key image certain to appear in each grid, providing theoretical entropy of 12.7 bits. We chose three experimental conditions where similarity between the key image and its decoy images was controlled by a threshold upon the EMD distance d between the colour histograms of the images:

- Similar: where $1 > d \geq 0$
- Middle: where $4 > d \geq 3$
- Dissimilar: where $6 > d \geq 5$

Studies in psychology [35] have observed how the difficulty of the visual search should decrease with decreasing similarity between target and non-targets. We were hoping to recreate a similar trend. To generalise our results more effectively we firstly discarded the image set used in the first study and obtained a set of 1000 images used in other image processing research [138], and removed any portrait oriented images for display consistency (reducing to 800). The database is highly categorical, which provides a worst case scenario for this study. In advance, we also chose 8 key images that represented exemplars of particular categories in the image set (see Figure 54). These key images were persistent across conditions. For each key image and experimental condition we automatically chose eight decoy images according to the condition-specific similarity criteria. We also enforced distances between other images in the login to respect the image filtering concerns discussed in Section 7.1.1. Within a particular condition, once an image was selected as a decoy to be associated with a particular key image, it could not be selected to appear as a decoy image for a different key image within that condition. Also, a key image could not reappear as a decoy image. Figure 55 provides an example of decoy image selection for one particular key image across all three conditions.

Figure 54: The key images chosen for the study, these were the same in all study conditions.

We recruited participants from the crowdsourcing platform Amazon Mechanical Turk. Kittur et al. [72] provide hindsight from conducting user studies on this platform, in particular that the most suitable tasks are those that have a verifiable answer. Clearly those carrying out studies on crowdsourcing platforms must design robust experiments that do not rely on literacy, and due to its remote nature take measures to detect behaviour that may undermine the integrity of the study. The user sample on Mechanical Turk was suitable as they are likely to be technology savvy adults. There were a number of study phases: registration: participants were requested to give information such as worker ID and demographic information; enrolment: where the participant would be given 30 seconds to view four key images randomly selected from our set of eight (see Appendix E for an example); wait: A JavaScript enforced stoppage of 30 minutes where participants could not progress to the next phase, but were free to carry out other tasks on Mechanical Turk. If participants attempted to progress beyond the wait period too quickly this could be detected via use of server side timestamps. The last phase was recall: the participant attempts to recognise the images assigned to them and has a single attempt to do so.

Due to the remote nature of the study we designed the following study defenses in order to have increased confidence in our results: *anti-image caching*: the images presented at login were drawn from a different location on the server than those at enrolment. This removed the threat that the key images would load faster due to caching; *anti-print screen*: participation was restricted to Internet Explorer and via a JavaScript we cleared the clipboard of the participant every 100 milliseconds. If consent to do this was not granted the experiment would not continue; *dynamic key images*: key image sequences were not static across participants and there were $\binom{8}{4}$ different possibilities for the sequence of key images that could be presented. This meant that if images were recorded they may not be immediately reusable by another participant; *HTTP GET parameter protection* ensured that we could detect where parameters were maliciously altered in the browser or the back button was pressed.

Figure 55: Example image grids assembled for key image #4 using similar, middle, and dissimilar decoy image criteria.

7.3.2 Results

We received 364 completed logins across a 6 day period. We treated as outliers those who completed the login procedure identifying no key images in less than 5 seconds. This reduced the numbers down to 343 with 117 in the similar condition, 112 in the dissimilar condition and 114 in the middle condition. In terms of demographics, 72% of participants were from India, with the United States the next prominent location at 7%. Most of the participants were male (73%). In terms of age, 269 were in the age group 18-30, 67 in the group 31-40, 15 in the age group 41-50, and 13 were 51+ years of age.

7.3.2.1 Accuracy

We firstly calculated a login success rate on a per-participant basis i.e. to compare the participants who correctly identified all 4 images. This was calculated by (successful logins/total logins). The raw data comprised success/fail value to represent a login. There was a significant difference between the performance of participants in the dissimilar group (70%) and the similar group (29%) $\chi^2(1, N = 229)$=37.716, $p < 0.01$. In addition there was a significant difference between the success rate in the middle (59%) and similar condition $\chi^2(1, N = 231)$=20.716, $p < 0.01$. The difference between the dissimilar and the middle condition was not statistically significant. Table 23 presents a more detailed illustration of participant performance. We also calculated a per-click success rate that represented $\frac{correct\ clicks}{total\ clicks}$ for each condition. The benefit of this calculation is that it can give insight into accuracy in a manner less sensitive to a single mistake by a participant. For example, a single problematic image grid could reduce login success rates considerably, whereas in reality this would reduce the per click success rate less severely. There was a significant difference between the success rates in the dissimilar condition (90%) and the similar condition (67%) $\chi^2(1, N = 916)$=57.679, $p < 0.01$. There was also a significant difference between the success in the middle (80%) and the similar condition, $\chi^2(1, N = 924)$=19.758, $p < 0.01$. The difference between the dissimilar and the middle condition using this metric was not significant.

Condition	Score Distribution				
	0	1	2	3	Success
Similar ($n = 117$)	6 (5%)	14 (12%)	26 (22%)	37 (32%)	34 (29%)
Middle ($n = 114$)	5 (4%)	8 (7%)	14 (12%)	20 (18%)	67 (59%)
Dissimilar ($n = 112$)	0%	6 (5%)	6 (5%)	8 (7%)	74 (70%)

Table 23: The number of key images correctly identified (out of four) in study two. Success is 4/4.

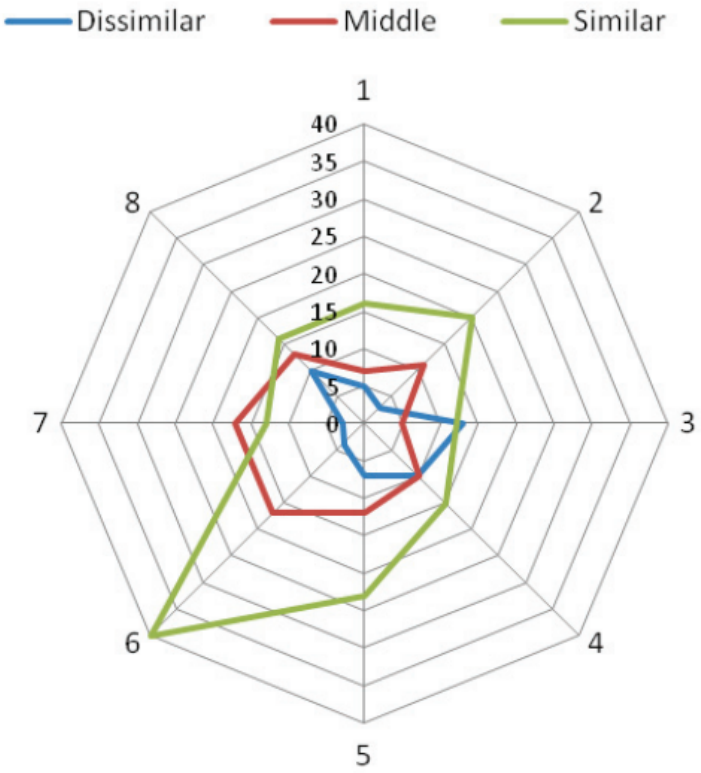


Figure 56: The number of login errors made per key image and per experimental condition.

Figure 56 illustrates the number of errors made per condition per image. Overall the errors do follow an intuitive pattern, they increase as the decoy images become more similar to the key image. However there are two exceptions, image 3 and image 7. Looking closely at the grids for these images, it is likely these errors can be explained by inter grid key-decoy similarity: a decoy image was visually similar to a non- collocated key image. This likely created confusion as to which image the user should select. This could indicate that the threshold we imposed on this instance of similarity was not sufficiently high. The graph also illustrates the interesting case of image 6 in the similar condition: there was a large number of user errors recorded when they were asked to identify this image. The decoy images for this image appeared visually and semantically similar. Analysis of errors made on a per-image basis across conditions highlighted a number of significant results too (see Table 24).

Image	Success %	Success %	χ^2
	Similar	Dissimilar	
1	71%	91%	$\chi^2(1,120)=5.729$, $p < 0.05$
2	66%	95%	$\chi^2(1,119)=14.540$, $p < 0.01$
5	55%	87%	$\chi^2(1,105)=14.207$, $p < 0.01$
6	32%	93%	$\chi^2(1,117)=46.230$,$p < 0.01$
7	77%	95%	$\chi^2(1,81)=31.777$, $p < 0.01$
	Similar	Middle	
5	55%	83%	$\chi^2(1,116)=10.449$, $p < 0.01$
6	32%	70%	$\chi^2(1,116)=16.726$, $p < 0.01$
	Middle	Dissimilar	
2	74%	95%	$\chi^2(1,101)=8.614$, $p < 0.01$
6	70%	92%	$\chi^2(1,115)=10.125$, $p < 0.01$
7	73%	95%	$\chi^2(1,123)=10.908$, $p < 0.01$

Table 24: Significant differences noted in user performance across experiment conditions on a per-image basis.

7.3.2.2 Login Durations

We also recorded time required for users to login in each condition. This was recorded from the first grid appearing onscreen, until the final click. We treated the data as non-parametric due to the existence of a number of particularly long login durations distorting the mean. The median login duration was 57 seconds in the similar group, in the middle group 40 seconds, and in the dissimilar group 36 seconds. The difference between the similar and dissimilar conditions was significant in a Mann-Whitney U test (Z=-4.730, $p < 0.01$). Using a Wilcoxon 1-sample sign test we estimated the 95% confidence interval for the medians. This estimates that participants in the dissimilar condition would take between 33-40 seconds, in the middle condition 38-51 seconds, and in the similar condition 48-68 seconds. This suggests that the choice of decoy selection method could also have a significant impact upon the login durations. Figure 57 shows the distribution of login durations recorded for successful logins for each condition.

7.4 Discussion

The vision of this research is that a system can take an arbitrary set of images and perform a filtering operation to generate a usable and secure login challenge, or else conclude that an image set does not contain suitable images for this purpose. Such a spontaneous approach to the generation of a login challenge becomes more useful when considering deployment level phenomena such as password resets, where ineffective recycling of images could cause confusion between new key images and old. A perfect automated semantic separation of images appears to be a difficult goal; however, we have shown that taking a coarse grained approach to the problem can affect usability.

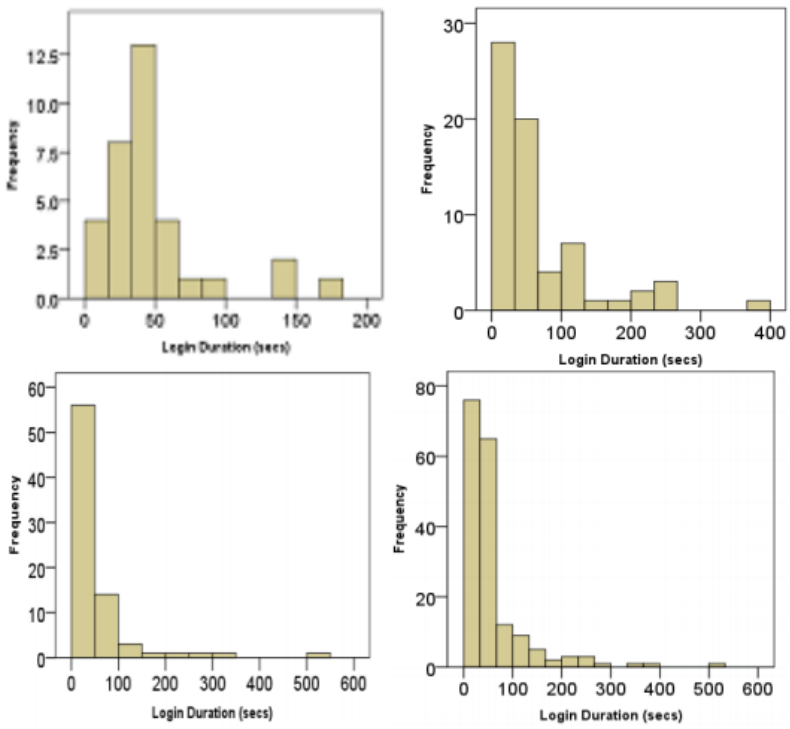


Figure 57: Length of successful logins in each condition: top left) similar; top right) middle; bottom left) dissimilar; bottom right) overall.

As a result it seems possible that ensuring a specific visual difference between images using pixel-level image signatures could assist in automated image selection strategies for recognition-based graphical passwords.

The recall test results suggest that comparison of pixel-level image signatures can affect the usability of recognition-based graphical passwords in terms of both user accuracy and the time required to login. We observed significant accuracy results between the similar and dissimilar condition, and the similar and middle condition. We did not observe significant differences between the middle and dissimilar condition. The login durations were also significantly impacted between the similar and dissimilar group with a significant difference of 21 seconds in the medians. The most damaging type of similarity we noted was inter grid key-decoy similarity, which is the similarity between a decoy image and a non-collocated key image. In this case the user erroneously selects a decoy image that appears similar to a non-collocated key image. Particularly conservative thresholds should be employed when considering this type of similarity. The remote scenario places success rates in a realistic zone, as participants were not within the sphere of influence of experimenters. Such results have important implications for deployment of recognition-based graphical passwords, as they serve to highlight the impact that seemingly subtle image choices can have upon the usability of the system.

The results have implications for graphical password systems of all genres. This work has focused upon recognition-based graphical passwords where there are multiple grids and one key image on each screen. However, our taxonomy of image filtering

is relevant to systems where images are static for all users, or for configurations where key images are randomly distributed across the grids [31, 59] and our proposed mechanism in Chapter 5. The results also have implications for other systems such as Passpoints [143], where future research could focus upon some notion of similarity between images to predict whether similar user choices could be expected between a number of images.

7.4.1 Security Implications

The introduction of deliberate and measurable differences in visual similarity between images creates the potential for traces of this process to be left behind and exploited by attackers, who could infer key images and gain unauthorised access to systems. The particular threat is *guessability*, that patterns in the composition of the login challenge could be reverse engineered to allow an attacker to make better than random guesses. The goal for an attacker is to obtain a successful login without any interaction with the user through activities such as coercion or observation attack. We should assume that an attacker knows the method used for image filtering and any thresholds employed, and is able to capture the login challenge specific to a particular user. If the system provides direct authentication [120] we assume the attacker has compromised the username of the legitimate user and can capture the login images for offline analysis. If the infrastructure is local authentication then an attacker may be able to take a high quality photograph of the images, although this could be considered less likely. Other threats to be considered when introducing visual differences in the image filtering procedure include observation attack and description.

7.4.1.1 Key Relative Filtering

Key relative filtering [102] has been proposed in previous work as a method that is suitable for small image sets, as it imposes few constraints upon the login challenge composition. The procedure is to firstly identify the key image, reject all images within a similarity distance d of the key image and select decoy images randomly from the remaining images. However, an attacker could identify candidate key images even without knowledge of the thresholds being used, by recalculating pairwise similarities to search for patterns. One way to do this is to create a similarity matrix (see Figure 58) which is a simple visualisation that captures pairwise EMD distances between images in a single nxn grid. In Figure 58 each large square represents the location of a single image, and the smaller squares within contain the EMD distance between that image and every other image in the 3x3 grid. The figure represents a particularly vulnerable case where the key relative similarity threshold is $d < 3$. In this case the attacker could conclude that the centermost image has a good chance of being the key image, since it is the only image that exhibits such a careful pattern ($d > 3$) in pairwise similarity values. Even without knowledge of the threshold the attacker

-	1.8	3.1	1.8	-	2.3	3.1	2.3	-
4.0	3.2	0.8	4.0	3.1	5.0	2.0	3.3	1.7
1.3	3.0	6.0	2.9	4.9	3.8	4.0	5.0	5.0
4.0	4.0	2.0	3.2	3.1	3.3	0.8	5.0	1.7
-	3.5	3.0	3.5	-	3.0	3.0	3.0	-
0.6	4.1	1.4	3.2	3.4	4.0	3.0	1.9	5.0
1.3	2.9	4.0	3.0	4.9	5.0	6.0	3.8	5.0
0.6	3.2	3.0	4.1	3.4	1.9	1.4	4.0	5.0
-	4.0	5.0	4.0	-	4.0	5.0	4.0	-

Figure 58: *Similarity matrix* that illustrates the pairwise EMD distance d between images in a single 3x3 image grid. This grid has been assembled with key relative filtering [102].

could make a good guess at its value based upon the minimal d observed for each image in the matrix.

7.4.1.2 Exhaustive filtering

The analysis of key relative filtering has shown that for purposes of security a more holistic approach to image filtering should be taken, in order to hide traces of the filtering procedure. One approach is based upon ensuring a *minimum distance* exists between all images in the grid. One limitation of this approach is that while it enforces a minimal distance between all images, there is no upper bound, which could leave the login vulnerable to observation attack, as images exhibiting large visual differences could remain in the login challenge. An alternative approach that could eliminate this threat is based upon *similarity intervals*, where additionally an upper bound of similarity is also enforced. However, this approach would likely be difficult to implement in small image collections, as a greater number of images are likely to be rejected due to the increased number of similarity constraints upon a permissible image. There is a trade-off between the volume of images rejected in the filtering procedure and the number of constraints that are enforced. As a compromise, a minimum distance approach is likely to be suitable in smaller image sets and where observation attack or description is less likely to be a threat. An example of the visual differences that may result is illustrated in Figure 59. In order to minimise the number of images that must be rejected, a useful strategy in general involves:

Figure 59: (Left) grids assembled using minimum distance approach where $d = 2$; (right) similarity intervals where $4 > d > 0$.

1. Choosing a strategy for decoy selection i.e. similarity or dissimilarity.
2. Choosing a candidate key image, and calculating the distribution of pairwise similarity between it and the rest of the image set.
3. Sorting the images in ascending order of EMD, then, if choosing for dissimilarity, choosing from the back of the list, and if choosing for similarity, choosing from the front of the list.
4. Repeating 2-4 for each key image.

7.5 Study Limitations

In this study we did not consider the longitudinal memory impact of the recall task; we chose to model a short-term memory task as password enrolment is a particularly traumatic period for committing new credentials to memory. However, we believe our scenario introduced sufficient stress into the enrolment procedure to enable us to have confidence in the results. The success rates are constrained by the fact that users were not working with their own images and only had one attempt to identify the images. In addition, the participants were not logging into a real system and so not authenticating to access anything of value. Finally, the image processing intervention we chose only operated at the pixel-level. Study of more sophisticated techniques that incorporate object segmentation may prove to be a fruitful future research direction.

7.6 Conclusion

In this chapter we considered the usability, security and deployment issues inherent in the process of choosing images to comprise a graphical password login. This process would likely be one which varies greatly between instantiations of graphical password systems. In particular we explored the extent to which automated image processing

techniques applied at the pixel-level can be used to systematically affect the amount of visual similarity, and so usability, in a typical login. We found that using a colour histogram in the LAB colour space as an image signature, and comparison of signatures using the Earth Movers Distance [106] was the most fruitful approach. In a recall test with more than 300 people recruited via Amazon Mechanical Turk we found that selecting decoy images to differing levels of similarity could impact the number of errors that occurred during an authentication session. We found significantly fewer errors made by users viewing grids with dissimilar decoys compared to those viewing the most similar decoys. In the most significant case we found that our automated decoy selection method could affect login success rates by 40%. While this result may appear intuitive, the contribution is that we noted the potential for recall to be affected systematically. We also highlighted that there were secure and insecure ways to choose images for login using automated methods.

These results are significant as they show the importance that an apparently subtle deployment decision such as the choice of images can have upon the usability and security of graphical passwords. A system administrator can without intent create a particularly difficult recognition task. Future work in this domain could consider more sophisticated methods of searching for image similarity. Such a direction of research would likely have implications for graphical passwords of all genres. For instance in the Passpoints [143] system some notion of similarity between images would be useful to help predict whether similar user choices could be expected between a number of images.

Chapter 8

Conclusion

As Herley and van Oorschot [60] note, despite the widespread desire to revamp the current state of knowledge-based authentication (KBA), not only have we failed to make significant changes, but our reliance upon conventional methods of KBA continues to grow. The management of alphanumeric passwords and personal identification numbers (PINs), have shaped security practices that are firmly embedded into our everyday lives, and their deployment is supported by a near ubiquitous technical infrastructure. Alternative approaches to KBA face the significant challenge of demonstrating real-world value in terms of being usable, secure, cheap, easy to deploy and maintain, provide credentials that are both easy to replace and cannot be shared; note, alphanumeric passwords and PINs also fail to meet these criteria, which adds weight to the reluctantly held assumption that a universally usable and secure authentication mechanism is an improbable ambition for the future of security and privacy. Instead, the choice of authentication mechanism must be made as a trade-off that incorporates the relevant model of likely adversaries, as well as an understanding of the characteristics of the user population [100], the context and the technology.

In this thesis we have contended that the design and evaluation of graphical password systems should take account of usability and security, but also *deployability*; currently, graphical password research has a limited understanding of the interplay and interdependence of these three factors. Our own conceptualisation of *deployability* was centred upon the problem of providing scaffolding for desirable authentication behaviours to users, and understanding the impact of socio-technical threats that could emerge in the context of graphical passwords. In the absence of large scale deployments of candidate schemes and accompanying ethnographic studies, empirical research has a responsibility to develop lightweight methods to gain insight into the likely appropriation of novel systems, lest systems introduce *revenge effects* [129] that impact security in unforeseen ways. Our hope is that this inquiry has shed light on new considerations pertaining to KBA that will help both researchers and practitioners better understand the difficult trade-off between usability, security and deployability. In this chapter we revisit the research questions posed in Section 1.3, summarise our contributions, and propose potentially fruitful avenues for future research.

8.1 Research Questions

8.1.1 How can interaction design scaffold secure behaviour in graphical password schemes?

The diversity of potential deployment environments means that it is unlikely that graphical password systems will be deployed in an archetypal form. Consequently, it is important that research explores the impact upon usability that results from the fine-tuning of systems to satisfy context-specific security constraints. In particular, it is important to explore how this fine-tuning can be performed to limit the potential for revenge effects [129], negative side effects that emerge when usability is overexerted for purposes of security. In this thesis[1] we considered two classes of interaction design interventions that aim to support users to avert threats that result either as a result of low levels of compliance with standard security advice or as a result of the unusually public nature of a shared collaborative interface. In Chapter 3 we focused upon improving the strength of Draw a Secret (DAS) [68] graphical passwords, and in Chapter 4 the design of multi-touch interaction techniques to resist observation attack in recognition-based graphical passwords.

In Chapter 3 we proposed an interface modification to DAS and evaluated its potential to increase the usability of strong graphical passwords (DAS is a system specifically designed to provide a large password space). Models of cognition suggest that users in practice are unlikely to create graphical passwords that exploit the security benefits of DAS's potentially large password space [130, 131]; if born out in practice this would render DAS considerably less secure than the theoretical analysis of its password space might suggest. In Section 3.2 we proposed that these weak trends may arise due to the visual appearance of the drawing grid; in response, we proposed the addition of a background image to introduce additional cues to support the creation and recall of more complex drawings. In Section 3.3 we described an evaluation of the resulting system: Background Draw a Secret (BDAS) and showed that users of BDAS created (and successfully recalled) more complex graphical passwords than users of DAS. However, while the background images did significantly improve the complexity of graphical passwords chosen by our participants in general, it did not give rise to strong graphical passwords for *all* participants. As with DAS, BDAS and indeed alphanumeric passwords, the most convenient user behaviour is still to choose simple credentials that are easy to remember. Indeed, future work should to consider how background images could introduce a new set of biases as to the drawings users would create (Section 3.4.1 speculated on the impact such biases could have upon security).

In Chapter 4 we considered the threat of observation attack to recognition-based

[1] in parallel to this thesis we considered the use of eye trackers as an entry method for observation resistant graphical passwords [36]

graphical passwords. The typical defence against such attacks has interesting social implications as systems tend to assume a user who will proactively shield input from bystanders. However, in the case of collocated users (such as at a shared multitouch interface) the act of shielding interactions is an implicit signal of mistrust and social pressure will influence users not to engage in such behaviour. Our response was to consider how the interaction affordances of emerging multi-touch technology could be harnessed to permit the design of observation resistant interactions within the authentication procedure; without impacting upon the authentication scheme itself. In Section 4.3 we explored the design space for observation resistant interactions on multi-touch surfaces. In an empirical lab-based study we found that the Pressure Grid (Section 4.4.3.1) proved to be an effective defence against observation attacks and only had a small impact upon the time taken to perform a login with either Personal Identification Numbers (PINs) or a recognition-based graphical password. The design of discreet user interactions that exploit affordances of the deployment platform appears to be a more effective means of defending against observation attacks, than the image portfolio approach described in Section 5.2 (which had a more significant impact upon successful login durations in the user study).

The solutions described in Chapters 3, and 4 stand as examples of interaction design that aim to align user behaviour with a secure path of behaviour. The problem at large appears to be that graphical passwords (and KBA more generally) maintain a reliance on an active user who makes decisions that prioritise security ahead of usability. Reconfiguring interaction in widely studied graphical password systems constitutes a promising approach to the promotion of more security compliant behaviour, without redesigning the underlying authentication system. Notable recent work that considers such issues includes the use of principles of persuasive technology [49] to improve user choice of alphanumeric passwords [51] and the Passpoints graphical passwords system [16].

8.1.2 How can strategic selection of images provide scaffolding in graphical password schemes?

While user interaction with existing authentication systems can be re-envisioned to support users to behave desirably with authentication credentials, such an approach assumes that systems and infrastructure are sufficiently flexible to admit such adaptation. However, recognition-based graphical passwords are intrinsically configurable through the choice of images provided to the user; in Chapters 5, 6, and 7 we explored strategies for image delivery, and how careful consideration and manipulation of image content can provide both usability and security benefits.

In Chapter 5 we explored the challenge of defence against observation attack through manipulation of the frequency that secret images were presented in the login challenge. By incorporating randomness in the exposure of the secret images, we

aimed to increase the difficulty for an attacker to observe and reuse the secret images to gain a successful login. Most importantly, this defence could be realised without assuming that a user would actively defend their input. In Section 5.2 we described how image portfolios could increase the complexity of an observation attack, and resist intersection attack; we explored how much protection this might bring in practice in Section 5.3 by developing statistical models of observation attacks.

In Chapter 6 and Chapter 7 we focused upon the relationship between image content and usability and security. In Chapter 6 we explored the impact of judicious choice of images (based upon visual and verbal image similarity) on the verbal sharing of Passfaces [96] graphical passwords. In our empirical study, described in Section 6.4, we found that the assembly of a login challenge that contains instances of visual and verbal similarity could impact the ability of users to login to a Passfaces system using recorded verbal descriptions. Indeed, we found that images selected to be visually and verbally similar had a similarly detrimental impact on user; this suggests that the manipulation of image similarity could reduce the impact of password sharing by inhibiting description.

In Chapter 7 we conducted two user studies to assess the effectiveness of automated image selection schemes. This considers both the choice of image content and the image selection strategy. In our first user study, described in Section 7.2 we found that pixel-level image signatures based upon colour histograms were a low-computation cost, but effective, approach to automatically estimate visual similarity. In the second study, described in in Section 7.3, we applied this result and used our histogram-based estimation algorithm to systematically manipulate the pairwise similarity between elements of a login challenge. We discovered that by manipulating the similarity threshold we could directly influence recall performance for a recognition-based graphical password scheme based on everyday images.

The impact of image choice on the usability and security of recognition-based graphical passwords is large, yet in the majority of studies conducted in graphical password research, very little emphasis has been placed on the processes or strategies for assembling a login challenge. In recent work [90] we made initial investigations how automated judgments of image similarity could impact observation attack and description in the context of a novel recognition-based graphical password system. The challenge for alternative authentication solutions is to demonstrate evidence of new types of value that impact the trade-off between security, usability, deployability, and convenience. Future work exploring automated image choice could help graphical passwords exploit such results in an automated manner, and provide such added value. This thesis has explored how to, and highlights the need to, develop a more systematic understanding of image choice, so that those who are seeking to deploy graphical password systems can harness the security benefits of appropriate image choice whilst being aware of the potential impact of such choices on usability.

8.1.3 What is the impact of graphical passwords upon socio-technical threats?

The usability and security of KBA are both sustained and threatened by the practices of the user. On the one hand users develop memory aides to remember credentials, for example by sharing them with others and writing them down, and on the other hand they can gain unauthorised access to accounts based upon non-technical attacks such as observation attack. One requirement placed on future KBA is to not make worse the current state of play; this causes us to ask how we might determine if graphical passwords are subject to the same insecure behaviours (such as password sharing or writing down), or whether they even promote such behaviour more than PINs and alphanumeric passwords. The most common research method to uncover the extent of these workarounds is the self-report survey [114, 113], which inevitably can only uncover the range of practices that subjects are prepared to reveal.

To this end we have made a number of methodological contributions. In Chapters 4 and 5 we explored the vulnerability of recognition-based graphical password systems to observation attack. Our focus on observation attack highlighted the impact of emerging ubiquitous computing technologies and their environment upon threats to a KBA. Firstly we explored threats to collaborative multi-touch tabletop systems, and secondly, threats to users of mobile devices. In both studies we cast participants as attackers attempting to discover the credentials being entered by other participants. in Chapter 6 we proposed an approach to modelling the ability of users to verbally share Passfaces graphical passwords.

In Chapter 4 we explored the design space for mechanisms that harnessed multi-touch interaction to defend against observation attack, and evaluated the ability of the Pressure Grid against two collocated observers (see Figure 29 for the study context). This enabled us to discover that our variant of Passfaces appeared less vulnerable to observation attack than PINs, but also that design of user interaction around that scheme could make that scheme significantly more difficult to compromise.

In Chapter 5 we considered a different approach to defence against observation attack based upon delayed exposure of images in the login challenge. In Section 5.2.1 we proposed that a friend attack could be particularly potent against mobile devices, due to the social imperative not to signal mistrust to collocated people by shielding the entry of sensitive information. To explore the efficacy of the observation resistant graphical password configuration we proposed, in Section 5.3 we assembled a statistical model of an attack and illustrated the protection that it provided given an adversary observing multiple logins over time; this was complemented with an empirical study of observation attack (see Figure 39). We discovered that a low entropy version of our graphical password system that had higher quality images was more vulnerable to observation attack than the high entropy version.

Chapter 6 considered the problem of users being able to verbally share graphical

passwords and the generally accepted assumption that the use of images in KBA would make this difficult. To gain an empirical understanding of users ability to do this in practice, we collected a corpus of verbal descriptions (see Section 6.3) of faces and explored the extent to which users could login to our mock-up of a Passfaces system using those descriptions (see Section 6.4). We found that the careful selection of decoy images, to be either visually or verbally similar to key images, had the potential to reduce users ability to effectively verbally share Passfaces-style graphical passwords.

The design of our empirical studies, in which subjects attempt attacks and carry out insecure workarounds has allowed us to explore the efficacy of novel forms of defence, and also fine-tune the usability of those defences. Such an approach to future KBA systems and contexts can provide insight into the severity of phenomena, before deployment in a real context with real users.

8.2 Summary of Contributions

The contributions of this thesis can be summarised as follows:

1. We highlighted the importance of the role of deployability when considering the usability and security of graphical passwords. Our conceptualisation of deployability comprised the explicit recognition of the problem of providing context-specific scaffolding to desired user authentication behaviours, and the impact of interventions upon usability and security. Our empirical studies explored how interventions could positively and negatively shape user behaviours.

2. We proposed a framework for supporting designers to scaffold user behaviours to defend against observation attack. By applying this framework to the contemporary problem of authentication on shared public multi-touch interfaces, we were able to articulate a design space of observation resistant augmentation to a standard graphical password.

3. We identified *description* as a threat to recognition-based graphical passwords and performed an empirical study to explore the ability of users to verbally share Passfaces [96] graphical passwords. We discovered that through the manual manipulation of similarity of decoy images and key images we could directly impact on users ability to identify key images from audio descriptions.

4. We proposed novel empirical methods to facilitate the study of socio-technical threats in laboratory-based studies, and the usability of systems outside of a laboratory environment. We measured user performance in matching audio descriptions to face images as an estimator for the prevalence of *description*. Also we refined a method proposed elsewhere to explore the observation resistance of graphical passwords on shared displays and mobile devices, asking users to

perform as attackers and observe live authentication sessions. Finally we performed the first field study of graphical passwords on mobile devices, where our system was installed on the personal devices of participants.

5. We performed the first attempt to harness image processing techniques to identify instances of image similarity that could be detrimental to usability if included in a recognition-based graphical password grid. The need for a human to perform this task is a severe limitation in terms of the deployability and scalability of this genre of graphical password. We conducted a usability evaluation of the best method we tested with over 300 participants recruited using Amazon Mechanical Turk where login challenges were assembled to include differing levels of similarity. We found that manipulation of the similarity levels allowed us to provide both positive and negative scaffolding of user authentication behaviours behaviours, highlighting the importance of image choice in the usability and security of graphical passwords.

8.3 Future Work

The results and discussion in this thesis suggest a number of interesting potential future research directions.

8.3.1 Scaffolding User Choice of High Entropy Graphical Passwords

In Chapter 3 we observed that users could be supported to create and recall high entropy graphical passwords through interventions to the interaction design of DAS [68]. Future work should consider the impact of image choice upon the drawings users would likely create. As entropy is still a core concern for user authentication credentials, there is still a need to consider approaches that support users to choose strong graphical passwords. Of course, innovation in the support of users to choose stronger authentication credentials is of no value without further innovation to support users to remember those credentials. DAS, Passpoints [143], and Pass-Go [127] are schemes that by their design provide an environment for users to choose high entropy graphical passwords (should they wish); researchers should further consider methods to align ideal notions of secure user choice with the choices users are most likely to make in practice. Promising work in this regard has been carried out by Chiasson et al. [16] who trialled the use of persuasive [49] techniques to encourage users to choose less predictable click points in Passpoints. In addition their intervention increased the difficulty for users to choose the more predictable click points.

Although visual stimuli have been shown to be more memorable than words or numbers [117], it is still inevitable that at some point users may forget their graphical

password, particularly if forced to manage a large number of high entropy credentials across many systems. This could suggest that greater consideration should be given to *designing for forgetting*. It is likely that usable and secure schemes could be developed to enable users to record and recover their graphical passwords; FacePIN [40] was designed as a method to enable users to remember PINs, that avoided explicit recording of the PIN, and potentially does not need to be associated to any system infrastructure in particular. Finally, Chiasson [13] proposes that designing for giving implicit feedback to the user during the authentication session, this appears to be a promising direction to design with the assumption that users, even with the best intentions, will forget their credentials one day.

8.3.2 Mass Longitudinal Comparisons of Representative Graphical Password Systems

Despite the field of graphical passwords having existed for over 12 years [5] there are still uncertainties over the most appropriate evaluation method to uncover insights into usability and security, at a time when computing platforms are evolving rapidly. Traditional studies of usability have taken place in the laboratory to ensure the participant is not distracted from the task that has been carefully designed by experimenters; however, the turn to ubiquitous computing [140] technologies, and in particular the widespread adoption of mobile devices, means that laboratory-based experiments have significantly less ecological validity [104] compared to field-based approaches, that is, their results appear less generalisable to in the wild use. Indeed, Chiasson et al. [14] report large discrepancies in user performance recorded between a usability study conducted in a laboratory and one conducted in a field study. This raises interesting questions about the combination of studies that are required to form a holistic understanding of how appropriate mechanisms are for their usage context. Now that the design space of graphical passwords appears increasingly explored, it could be that a turn to methodology is required, and future user studies could seek to place a greater emphasis on consistent platforms and research tools. Systems to support such research exist, for instance the MVP web authentication framework [15].

8.3.3 Experience-centred Security

In this thesis the consideration of the human aspects of security has focused exclusively upon usability. This domain of *usable security* is inherited from the roots of Human Computer Interaction (HCI) in its second wave [9] form of fine-tuning systems to make workers more efficient in their usage of computer systems in the workplace. This approach conceptualises the user based upon their cognitive ability, and theoretical models of action developed to describe human decision making. This makes the assumption that by making systems easier to use, users will find it easier to make

the correct security decisions. This is true, usability is important, however this focus alone does not appear sufficient to produce a holistic understanding of human-centred security. Technology, and computer security are no longer encountered solely in the workplace by young adults; security-facing systems now play an active role in everyday life, with systems used by diverse groups of users such as older people, children and those with disabilities. This implies new requirements aside from ease of use for technology and its associated security to complement our everyday lives.

Security is simultaneously a feeling and a reality, and they are not the same [112]. Third-wave HCI research considers that user interaction with technology is simultaneously emotional, aesthetic, sensual and intellectual [81]. Research in the HCI community has considered how to probe user experiences, and also how to design for them. Research attention has considered, for example, to design for fun experiences [8], and uncomfortable experiences [4]. A number of disparate works in the security domain hint towards the need for this holistic account of user experiences: Singh et al. [118] discuss the password sharing behaviour in contexts ranging from married couples to Australian indigenous islanders who delegated finances to each other. Vines et al. [136] explore the drivers of trust around financial management for people over eighty years of age. In addition, Vines et al. [137] discuss how trust practices emerged around financial transactions conducted using paper cheques.

In computer security we have much to learn about our intersection with the literature on user experience. Implicitly we have designed computer systems to create unpleasant interactions in the form of providing unreasonable rules and creating guilt when users circumvent those rules. Security researchers should not be exempt from taking account of the context, and developing rich understandings of the user; the sustaining of personal relationships often proves to be a key driver for the breaking of security rules, this appears to be a natural entry point in this domain. Of course, such an approach requires innovation in methods of engaging with users, reflection upon how this affects the process of *design* [43], and how to produce security as a result of such an understanding. By taking steps in this direction we can edge closer to designing *user-centred security* in its broadest sense.

Bibliography

[1] Adams, A., and Sasse, M. A. Users are not the enemy. *Commun. ACM 42*, 12 (Dec. 1999), 40–46.

[2] Anderson, R. Why cryptosystems fail. In *Proceedings of the 1st ACM conference on Computer and communications security* (New York, NY, USA, 1993), CCS '93, ACM, pp. 215–227.

[3] Baker, D. G. Non-disclosing password entry method, 2008.

[4] Benford, S., Greenhalgh, C., Giannachi, G., Walker, B., Marshall, J., and Rodden, T. Uncomfortable interactions. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 2005–2014.

[5] Biddle, R., Chiasson, S., and van Oorschot, P. Graphical Passwords: Learning from the first twelve years. *ACM Computing Surveys 44* (2011).

[6] Bishop, M. Psychological Acceptability Revisited. In *Security and Usability: Designing Secure Systems That People Can Use*, L. Cranor and S. Garfinkel, Eds. O Reilly, 2005, pp. 1–12.

[7] Blonder, G. Graphical Passwords, US Patent 5559961, 1995.

[8] Blythe, M., Overbeeke, A., Monk, A., and Wright, P. *Funology: from usability to enjoyment*. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[9] Bø dker, S. When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (New York, NY, USA, 2006), NordiCHI '06, ACM, pp. 1–8.

[10] Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* (2008).

[11] Brostoff, S., and Sasse, M. A. Are Passfaces more usable than passwords? A field trial investigation. *Computer pages* (2000), 405–424.

[12] Burr, W. E., Dodson, D. F., and Polk, W. T. Electronic Authentication Guideline. Tech. rep., NIST Special Publication 800-63, 2006.

[13] Chiasson, S. *Usable Authentication and Click-Based Graphical Passwords.* PhD thesis, Carleton University, 2008.

[14] Chiasson, S., Biddle, R., and van Oorschot, P. C. A second look at the usability of click-based graphical passwords. In *Proceedings of the 3rd symposium on Usable privacy and security* (New York, NY, USA, 2007), SOUPS '07, ACM, pp. 1–12.

[15] Chiasson, S., Deschamps, C., Stobert, E., Hlywa, M., Freitas Machado, B., Forget, A., Wright, N., Chan, G., and Biddle, R. The MVP Web-based Authentication Framework. In *Financial Cryptography* (2012).

[16] Chiasson, S., Forget, A., Biddle, R., and van Oorschot, P. C. Influencing users towards better passwords: persuasive cued click-points. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1* (Swinton, UK, UK, 2008), BCS-HCI '08, British Computer Society, pp. 121–130.

[17] Chiasson, S., Forget, A., Biddle, R., and van Oorschot, P. C. User interface design affects security: patterns in click-based graphical passwords. *Int. J. Inf. Secur. 8*, 6 (Oct. 2009), 387–398.

[18] Chiasson, S., Forget, A., Stobert, E., van Oorschot, P. C., and Biddle, R. Multiple password interference in text passwords and click-based graphical passwords. In *Proceedings of the 16th ACM conference on Computer and communications security* (New York, NY, USA, 2009), CCS '09, ACM, pp. 500–511.

[19] Chiasson, S., van Oorschot, P., and Biddle, R. Graphical Password Authentication Using Cued Click Points. In *12th European Symposium On Research In Computer Security (ESORICS)*, J. Biskup and J. Lpez, Eds., vol. 4734 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2007, pp. 359–374.

[20] Chiasson, S., van Oorschot, P. C., and Biddle, R. A usability study and critique of two password managers. In *Proceedings of the 15th conference on USENIX Security Symposium - Volume 15* (Berkeley, CA, USA, 2006), USENIX Association.

[21] Cisco. Visual Networking Index: Global Mobile Data Traffic Forecast Update 2011-2016. Tech. rep., 2012.

[22] CNet. Gates Predicts the Death of the Password. http://news.cnet.com/2100-1029-5164733.html, 2004.

[23] CNet. Microsoft security guru: Jot down your passwords. http://news.cnet.com/Microsoft-security-guru-Jot-down-your-passwords/2100-7355_3-5716590.html, 2005.

[24] Coventry, L. Usable Biometrics. In *Usability and Security: Designing Secure Systems That People Can Use*, L. Cranor and S. Garfinkel, Eds. O'Reilly, 2005, pp. 175–197.

[25] Cranor, L., and Garfinkel, S., Eds. *Security and Usability: Designing Secure Systems That People Can Use*. O' Reilly, 2005.

[26] Davis, D. Compliance defects in public-key cryptography. In *Proceedings of the 6th conference on USENIX Security Symposium, Focusing on Applications of Cryptography - Volume 6* (Berkeley, CA, USA, 1996), USENIX Association, p. 17.

[27] Davis, D., Monrose, F., and Reiter, M. K. On user choice in graphical password schemes. In *Proceedings of the 13th conference on USENIX Security Symposium - Volume 13* (Berkeley, CA, USA, 2004), SSYM'04, USENIX Association, p. 11.

[28] De Angeli, A., Coutts, M., Coventry, L., Johnson, G. I., Cameron, D., and Fischer, M. H. VIP: a visual approach to user authentication. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2002), AVI '02, ACM, pp. 316–323.

[29] De Luca, A. *Designing Usable and Secure Authentication Mechanisms for Public Spaces*. PhD thesis, University of Munich, 2011.

[30] De Luca, A., von Zezschwitz, E., and Hu\ssmann, H. Vibrapass: secure authentication based on shared lies. In *Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), CHI '09, ACM, pp. 913–916.

[31] Dhamija, R., and Perrig, A. Deja Vu: a user study using images for authentication. In *Proceedings of the 9th conference on USENIX Security Symposium - Volume 9* (Berkeley, CA, USA, 2000), USENIX Association, p. 4.

[32] Diffie, W., and Hellman, M. New directions in cryptography. *Information Theory, IEEE Transactions on 22*, 6 (Nov. 1976), 644–654.

[33] Dirik, A. E., Memon, N., and Birget, J.-C. Modeling user choice in the PassPoints graphical password scheme. In *Proceedings of the 3rd symposium on*

Usable privacy and security (New York, NY, USA, 2007), SOUPS '07, ACM, pp. 20–28.

[34] D.M.Horgan. *Language development*. PhD thesis, University of Michigan, 1975.

[35] Duncan, J., and Humphreys, G. W. Visual search and stimulus similarity. *Psychological review 96*, 3 (July 1989), 433–458.

[36] Dunphy, P., Fitch, A., and Olivier, P. Gaze-contingent Passwords at the ATM. In *4th COGAIN Annual Conference on Communication by Gaze Interaction (COGAIN 2008)* (2008).

[37] Dunphy, P., Heiner, A. P., and Asokan, N. A closer look at recognition-based graphical passwords on mobile devices. In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (New York, NY, USA, 2010), SOUPS '10, ACM, pp. 3:1—-3:12.

[38] Dunphy, P., and Olivier, P. On automated image choice for secure and usable graphical passwords. In *Proceedings of the 28th Annual Computer Security Applications Conference* (New York, NY, USA, 2012), ACSAC '12, ACM, pp. 99–108.

[39] Dunphy, P., and Yan, J. Do background images improve "draw a secret" graphical passwords? In *Proceedings of the 14th ACM conference on Computer and communications security* (New York, NY, USA, 2007), CCS '07, ACM, pp. 36–47.

[40] Dunphy, P., and Yan, J. Is FacePIN secure and usable? In *Proceedings of the 3rd symposium on Usable privacy and security* (New York, NY, USA, 2007), SOUPS '07, ACM, pp. 165–166.

[41] Eisenberg, T., Gries, D., Hartmanis, J., Holcomb, D., Lynn, M. S., and Santoro, T. The Cornell commission: on Morris and the worm. *Commun. ACM 32*, 6 (1989), 706–709.

[42] Everitt, K. M., Bragin, T., Fogarty, J., and Kohno, T. A comprehensive study of frequency, interference, and training of multiple graphical passwords. In *Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), CHI '09, ACM, pp. 889–898.

[43] Faily, S., Coles-Kemp, L., Dunphy, P., Just, M., Akama, Y., and De Luca, A. Designing interactive secure system: chi 2013 special interest group. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2013), CHI EA '13, ACM, pp. 2469–2472.

[44] Federal Information Processing Standards. Password Usage. Tech. rep., 1985.

[45] Financial Ombudsman. Ombudsman News 67. Tech. rep., 2008.

[46] Florencio, D., and Herley, C. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web* (New York, NY, USA, 2007), WWW '07, ACM, pp. 657–666.

[47] Florêncio, D., and Herley, C. Where do security policies come from? In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (New York, NY, USA, 2010), SOUPS '10, ACM, pp. 10:1—-10:14.

[48] Florêncio, D., Herley, C., and Coskun, B. Do strong web passwords accomplish anything? In *Proceedings of the 2nd USENIX workshop on Hot topics in security* (Berkeley, CA, USA, 2007), USENIX Association, pp. 10:1—-10:6.

[49] Fogg, B. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, 2003.

[50] Ford, A., and Roberts, A. Colour Space Conversions. *University of Westminster Technical Report* (1998).

[51] Forget, A., Chiasson, S., van Oorschot, P. C., and Biddle, R. Improving text passwords through persuasion. In *Proceedings of the 4th symposium on Usable privacy and security* (New York, NY, USA, 2008), SOUPS '08, ACM, pp. 1–12.

[52] Gao, H., Guo, X., Chen, X., Wang, L., and Liu, X. YAGP: Yet Another Graphical Password Strategy. In *Computer Security Applications Conference, 2008. ACSAC 2008. Annual* (2008), pp. 121–129.

[53] Goldberg, J., Hagman, J., and Sazawal, V. Doodling our way to better authentication. In *CHI '02 extended abstracts on Human factors in computing systems* (New York, NY, USA, 2002), CHI EA '02, ACM, pp. 868–869.

[54] Golle, P., and Wagner, D. Cryptanalysis of a Cognitive Authentication Scheme (Extended Abstract). In *Proceedings of the 2007 IEEE Symposium on Security and Privacy* (Washington, DC, USA, 2007), SP '07, IEEE Computer Society, pp. 66–70.

[55] Gudivada, V. N., and Raghavan, V. V. Content based image retrieval systems. *Computer 28*, 9 (Sept. 1995), 18–22.

[56] Halpern, D. F. *Sex Differences in Cognitive Abilities*, 3rd editio ed. Psychology Press;, 2000.

[57] Han, J. Y. Low-cost multi-touch sensing through frustrated total internal reflection. In *Proceedings of the 18th annual ACM symposium on User interface software and technology* (New York, NY, USA, 2005), UIST '05, ACM, pp. 115–118.

[58] Harrington, V., and Mayhew, P. Mobile phone theft. Tech. rep., Home Office Research Study 235, 2001.

[59] Hayashi, E., Dhamija, R., Christin, N., and Perrig, A. Use Your Illusion: secure authentication usable anywhere. In *Proceedings of the 4th symposium on Usable privacy and security* (New York, NY, USA, 2008), SOUPS '08, ACM, pp. 35–45.

[60] Herley, C., and Oorschot, P. V. A Research Agenda Acknowledging the Persistence of Passwords. *IEEE Security and Privacy 99*, PrePrints (2011).

[61] Herschberg, I. S. The hackers' comfort. *Comput. Secur. 6*, 2 (May 1987), 133–138.

[62] Hlywa, M., Biddle, R., and Patrick, A. S. Facing the facts about image type in recognition-based graphical passwords. In *Proceedings of the 27th Annual Computer Security Applications Conference* (New York, NY, USA, 2011), ACSAC '11, ACM, pp. 149–158.

[63] Hulme, C., Maughan, S., and Brown, G. D. A. Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language 30*, 6 (1991), 685–701.

[64] Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., and Lyons, T. Early vocabulary growth: relation to language input and gender. *Developmental Psychology* (1991), 236–248.

[65] Inglesant, P. G., and Sasse, M. A. The true cost of unusable password policies: password use in the wild. In *Proceedings of the 28th international conference on Human factors in computing systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 383–392.

[66] Jakobsson, M., and Myers, S. *Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft.* Wiley-Blackwell, 2007.

[67] Jakobsson, M., Stolterman, E., Wetzel, S., and Yang, L. Love and authentication. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 197–200.

[68] Jermyn, I., Mayer, A., Monrose, F., Reiter, M. K., and Rubin, A. D. The design and analysis of graphical passwords. In *Proceedings of the 8th conference on USENIX Security Symposium - Volume 8* (Berkeley, CA, USA, 1999), USENIX Association, p. 1.

[69] Kaye, J. J. Self-reported password sharing strategies. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 2619–2622.

[70] Keith, M., Shao, B., and Steinbart, P. J. The usability of passphrases for authentication: An empirical field study. *Int. J. Hum.-Comput. Stud. 65*, 1 (Jan. 2007), 17–28.

[71] Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J., Nicholson, J., and Olivier, P. Multi-touch authentication on tabletops. *Proceedings of the 28th international conference on Human factors in computing systems CHI 10* (2010), 1093.

[72] Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 453–456.

[73] Kjeldskov, J., Skov, M. B., Als, B. S., and Høegh, R. Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Mobile Human-Computer Interaction (MobileHCI)*, S. Brewster and M. Dunlop, Eds., vol. 3160 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2004, pp. 529–535.

[74] Klein, D. V. "Foiling the Cracker" – A Survey of, and Improvements to, Password Security. In *Proceedings of the second USENIX Workshop on Security* (1990), pp. 5–14.

[75] Komanduri, S., and Hutchings, D. R. Order and entropy in picture passwords. In *Proceedings of graphics interface 2008* (Toronto, Ont., Canada, Canada, 2008), GI '08, Canadian Information Processing Society, pp. 115–122.

[76] Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F., and Egelman, S. Of passwords and people: measuring the effect of password-composition policies. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 2595–2604.

[77] Kuo, C., Romanosky, S., and Cranor, L. F. Human selection of mnemonic phrase-based passwords. In *Proceedings of the second symposium on*

Usable privacy and security (New York, NY, USA, 2006), SOUPS '06, ACM, pp. 67–78.

[78] Kurzban, S. A. Easily remembered passphrases: a better approach. *SIGSAC Rev. 3*, 2-4 (Sept. 1985), 10–21.

[79] Lin, D., Dunphy, P., Olivier, P., and Yan, J. Graphical passwords & qualitative spatial relations. In *Proceedings of the 3rd symposium on Usable privacy and security* (New York, NY, USA, 2007), SOUPS '07, ACM, pp. 161–162.

[80] Marshall, J., Pridmore, T., Pound, M., Benford, S., and Koleva, B. Pressing the Flesh: Sensing Multiple Touch and Finger Pressure on Arbitrary Surfaces. In *Proceedings of the 6th International Conference on Pervasive Computing* (Berlin, Heidelberg, 2008), Pervasive '08, Springer-Verlag, pp. 38–55.

[81] McCarthy, J., and Wright, P. *Technology as Experience*. The MIT Press, 2007.

[82] Microsoft. Microsoft Pixelsense. `http://www.microsoft.com/en-us/pixelsense/`.

[83] Miller, G. A. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review 63* (1956), 81–97.

[84] Mitnick, K. *The Art of Deception*. John Wiley & Sons, 2003.

[85] Moncur, W., and Leplâtre, G. Pictures at the ATM: exploring the usability of multiple graphical passwords. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2007), CHI '07, ACM, pp. 887–894.

[86] Monrose, F., and Reiter, M. K. Graphical Passwords. In *Security and Usability: Designing Secure Systems That People Can Use* (2005), L. Cranor and S. Garfinkel, Eds., O'Reilly & Associates, pp. 157–174.

[87] Morris, R., and Thompson, K. Password security: a case history. *Commun. ACM 22*, 11 (Nov. 1979), 594–597.

[88] Nali, D., and Thorpe, J. Analyzing User Choice in Graphical Passwords. *Technical Report TR-04-01, School of Computer Science, Carleton University.* (2004).

[89] Neuman, B. C., and Ts'o, T. Kerberos: an authentication service for computer networks. *Communications Magazine, IEEE 32*, 9 (Sept. 1994), 33–38.

[90] Nicholson, J., Dunphy, P., Coventry, L., Briggs, P., and Olivier, P. A security assessment of tiles: a new portfolio-based graphical authentication system. In *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts* (New York, NY, USA, 2012), CHI EA '12, ACM, pp. 1967–1972.

[91] Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J., and Stenild, S. It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (New York, NY, USA, 2006), NordiCHI '06, ACM, pp. 272–280.

[92] Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people* (New York, NY, USA, 1990), CHI '90, ACM, pp. 249–256.

[93] Norman, D. *The Design of Everyday Things*. Basic Books, 2002.

[94] Openwall. John the Ripper. `http://www.openwall.com/john/`.

[95] Paivio, A. *Imagery and Verbal Processes*. Psychology Press, 1978.

[96] Passfaces Corporation. The Science Behind Passfaces. `http://www.passfaces.com/published/The%20Science%20Behind%20Passfaces.pdf`.

[97] Pering, T., Sundar, M., Light, J., and Want, R. Photographic Authentication through Untrusted Terminals. *IEEE Pervasive Computing 2*, 1 (Jan. 2003), 30–36.

[98] Piazzalunga, U., Salvaneschi, P., and Coffetti, P. The Usability of Security Devices. In *Security and Usability: Designing Secure Systems That People Can Use*, L. Cranor and S. Garfinkel, Eds. O Reilly, 2005, pp. 221–242.

[99] Proctor, R., Lien, M.-C., Vu, K.-P., Schultz, E., and Salvendy, G. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods 34*, 2 (2002), 163–169.

[100] Renaud, K. Evaluating Authentication Mechanisms. In *Security and Usability: Designing Secure Systems That People Can Use* (2005), L. Cranor and S. Garfinkel, Eds., O'Reilly & Associates, pp. 103–128.

[101] Renaud, K. On user involvement in production of images used in visual authentication. *J. Vis. Lang. Comput. 20*, 1 (Feb. 2009), 1–15.

[102] Renaud, K., and Olsen, E. S. DynaHand: Observation-resistant recognition-based web authentication. *Technology and Society Magazine, IEEE 26*, 2 (2007), 22–31.

[103] Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R. E., Hursey, J., and Toscos, T. Why it's worth the hassle: the value of in-situ studies when designing Ubicomp. In *Proceedings of the 9th international conference on Ubiquitous computing* (Berlin, Heidelberg, 2007), UbiComp '07, Springer-Verlag, pp. 336–353.

[104] Rogers, Y., Sharp, H., and Preece, J. *Interaction design: beyond human-computer interaction*, 2 ed. John Wiley and Sons, 2011.

[105] Roth, V., Richter, K., and Freidinger, R. A PIN-entry method resilient against shoulder surfing. In *Proceedings of the 11th ACM conference on Computer and communications security* (New York, NY, USA, 2004), CCS '04, ACM, pp. 236–245.

[106] Rubner, Y., Tomasi, C., and Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vision 40*, 2 (2000), 99–121.

[107] Salehi-Abari, A., Thorpe, J., and van Oorschot, P. C. On Purely Automated Attacks and Click-Based Graphical Passwords. In *Proceedings of the 2008 Annual Computer Security Applications Conference* (Washington, DC, USA, 2008), ACSAC '08, IEEE Computer Society, pp. 111–120.

[108] Saltzer, J., and Schroeder, M. The Protection of Information in Computer Systems. *IEEE 63*, 9 (1975), 1278–1308.

[109] Sasse, M. A., Brostoff, S., and Weirich, D. Transforming the 'Weakest Link' – a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal 19*, 3 (July 2001), 122–131.

[110] Schechter, S., Brush, A. J. B., and Egelman, S. It's No Secret. Measuring the Security and Reliability of Authentication via "Secret” Questions. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 375–390.

[111] Schneier, B. *Secrets & Lies: Digital Security in a Networked World*. John Wiley & Sons, 2000.

[112] Schneier, B. The Psychology of Security. Tech. rep., 2008.

[113] SearchSecurity. Employees willing to share passwords with strangers. `http://searchsecurity.techtarget.com/news/895483/Study-Employees-willing-to-share-passwords-with-strangers`, 2003.

[114] SearchSecurity. Survey: Most workers must remember six passwords or more. `http://searchsecurity.techtarget.com/news/902867/Survey-Most-workers-must-remember-six-passwords-or-more`, 2003.

[115] Shannon, C. E. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev. 5*, 1 (Jan. 1948), 3–55.

[116] Shannon, C. E. Prediction and Entropy of Printed English. *Bell System Technical Journal 30*, 1 (1951), 50–64.

[117] Shepard, R. N. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior 6*, 1 (1967), 156–163.

[118] Singh, S., Cabraal, A., Demosthenous, C., Astbrink, G., and Furlong, M. Password sharing: implications for security design based on social practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2007), CHI '07, ACM, pp. 895–904.

[119] Sinha, P., Balas, B., Ostrovsky, Y., and Russell, R. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE 94*, 11 (2006), 1948–1962.

[120] Smith, R. E. *Authentication: From Passwords to Public Keys*. Addison Wesley, 2001.

[121] Smith, T. A., Jones, L. V., and Thomas, S. Effects upon verbal learning of stimulus similarity, number of stimuli per response, and concept formation. *Journal of Verbal Learning and Verbal Behavior 1*, 6 (1963), 470–476.

[122] Spafford, E. H. The Internet Worm Program: An Analysis. Tech. Rep. Purdue Technical Report CSD-TR-823, West Lafayette, IN 47907-2004, 1988.

[123] Squire, D. M. Learning a similarity-based distance measure for image database organization from human partitionings of an image set. In *Applications of Computer Vision, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on* (Oct. 1998), pp. 88–93.

[124] Suo X., Z. Y. O. G. S. Graphical Passwords: A Survey. In *Proceedings of the 21st Annual Computer Security Applications Conference* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 463–472.

[125] Takada, T., and Koike, H. Awase-E: Image-Based Authentication for Mobile Phones Using User's Favorite Images. In *Human-Computer Interaction with Mobile Devices and Services*, L. Chittaro, Ed., vol. 2795 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2003, pp. 347–351.

[126] Tan, D. S., Keyani, P., and Czerwinski, M. Spy-resistant keyboard: more secure password entry on public touch screen displays. In *Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (Narrabundah, Australia, Australia, 2005),

OZCHI '05, Computer-Human Interaction Special Interest Group (CHISIG) of Australia, pp. 1–10.

[127] Tao, H., and Adams, C. Pass-Go: A proposal to improve the usability of graphical passwords. *International Journal of Network Security 7*, 2 (2008), 273–292.

[128] Tari, F., Ozok, A. A., and Holden, S. H. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In *Proceedings of the second symposium on Usable privacy and security* (New York, NY, USA, 2006), SOUPS '06, ACM, pp. 56–66.

[129] Tenner, E. *Why Things Bite Back: Technology and the Revenge of Unintended Consequences*. Vintage Books, 2007.

[130] Thorpe, J., and van Oorschot, P. C. Graphical dictionaries and the memorable space of graphical passwords. In *Proceedings of the 13th conference on USENIX Security Symposium - Volume 13* (Berkeley, CA, USA, 2004), SSYM'04, USENIX Association, p. 10.

[131] Thorpe, J., and van Oorschot, P. C. Towards Secure Design Choices for Implementing Graphical Passwords. In *Proceedings of the 20th Annual Computer Security Applications Conference* (Washington, DC, USA, 2004), ACSAC '04, IEEE Computer Society, pp. 50–60.

[132] Thorpe, J., and van Oorschot, P. C. Human-seeded attacks and exploiting hot-spots in graphical passwords. In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium* (Berkeley, CA, USA, 2007), USENIX Association, pp. 8:1—-8:16.

[133] Tullis, T. S., and Tedesco, D. P. Using personal photos as pictorial passwords. In *CHI '05 extended abstracts on Human factors in computing systems* (New York, NY, USA, 2005), CHI EA '05, ACM, pp. 1841–1844.

[134] Tullis, T. S., Tedesco, D. P., and McCaffrey, K. E. Can users remember their pictorial passwords six years later. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (New York, NY, USA, 2011), CHI EA '11, ACM, pp. 1789–1794.

[135] Uzun, E., Karvonen, K., and Asokan, N. Usability analysis of secure pairing methods. In *Proceedings of the 11th International Conference on Financial cryptography and 1st International conference on Usable Security* (Berlin, Heidelberg, 2007), FC'07/USEC'07, Springer-Verlag, pp. 307–324.

[136] Vines, J., Blythe, M., Dunphy, P., and Monk, A. Eighty something: banking for the older old. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction* (Swinton, UK, UK, 2011), BCS-HCI '11, British Computer Society, pp. 64–73.

[137] Vines, J., Dunphy, P., Blythe, M., Lindsay, S., Monk, A., and Olivier, P. The joy of cheques: trust, paper and eighty somethings. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (New York, NY, USA, 2012), CSCW '12, ACM, pp. 147–156.

[138] Wang, J. Z., Li, J., and Wiederhold, G. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 9 (Sept. 2001), 947–963.

[139] Weinshall, D. Cognitive authentication schemes safe against spyware. In *Security and Privacy, 2006 IEEE Symposium on* (May 2006), pp. 6 pp. –300.

[140] Weiser, M. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev. 3*, 3 (July 1999), 3–11.

[141] Whitten, A., and Tygar, J. D. Why Johnny can't encrypt: a usability evaluation of PGP 5.0. In *Proceedings of the 8th conference on USENIX Security Symposium - Volume 8* (Berkeley, CA, USA, 1999), USENIX Association, p. 14.

[142] Wiedenbeck, S., Waters, J., Birget, J.-C., Brodskiy, A., and Memon, N. Authentication using graphical passwords: effects of tolerance and image choice. In *Proceedings of the 2005 symposium on Usable privacy and security* (New York, NY, USA, 2005), SOUPS '05, ACM, pp. 1–12.

[143] Wiedenbeck, S., Waters, J., Birget, J.-C., Brodskiy, A., and Memon, N. PassPoints: design and longitudinal evaluation of a graphical password system. *Int. J. Hum.-Comput. Stud. 63*, 1-2 (July 2005), 102–127.

[144] Wiedenbeck, S., Waters, J., Sobrado, L., and Birget, J.-C. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *Proceedings of the working conference on Advanced visual interfaces* (New York, NY, USA, 2006), AVI '06, ACM, pp. 177–184.

[145] Wood, C. C. Effective information system security with password controls. *Comput. Secur. 2*, 1 (1983), 5–10.

[146] Wood, D., Bruner, J. S., and Ross, G. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry 17*, 2 (1976), 89–100.

[147] Yan, J., Blackwell, A., Anderson, R., and Grant, A. The Memorability and Security of Passwords. In *Security and Usability: Designing Secure*

Systems That People Can Use, L. Cranor and S. Garfinkel, Eds. O Reilly, 2005, pp. 129–142.

[148] Yan, J. J. A note on proactive password checking. In *Proceedings of the 2001 workshop on New security paradigms* (New York, NY, USA, 2001), NSPW '01, ACM, pp. 127–135.

[149] Yee, H. PerceptualDiff. `http://pdiff.sourceforge.net/`.

[150] Yee, K.-P. Guidelines and Strategies for Secure Interaction Design. In *Security and Usability: Designing Secure Systems That People Can Use* (2005), L. Cranor and S. Garfinkel, Eds., O'Reilly & Associates, pp. 247–273.

[151] Zakaria, N. H., Griffiths, D., Brostoff, S., and Yan, J. Shoulder surfing defence for recall-based graphical passwords. In *Proceedings of the Seventh Symposium on Usable Privacy and Security* (New York, NY, USA, 2011), SOUPS '11, ACM, pp. 6:1—-6:12.

[152] Zhang, Y., Monrose, F., and Reiter, M. K. The security of modern password expiration: an algorithmic framework and empirical analysis. In *Proceedings of the 17th ACM conference on Computer and communications security* (New York, NY, USA, 2010), CCS '10, ACM, pp. 176–186.

[153] Zurko, M. E., and Simon, R. T. User-centered security. In *Proceedings of the 1996 workshop on New security paradigms NSPW 96* (1996), vol. 1, ACM Press, pp. 27–33.

[154] Zviran, M., and Haga, W. J. Cognitive passwords: the key to easy access control. *Comput. Secur. 9*, 9 (1990), 723–736.

[155] Zviran, M., and Haga, W. J. A Comparison of Password Techniques for Multilevel Authentication Mechanisms. *The Computer Journal 36*, 3 (1993), 227–237.

Appendix A

Draw a Secret (DAS) Information sheet

Information for participants

Are you a person that always forgets passwords? Or has anyone ever cracked your password? To combat both these problems a new genre of passwords is on the horizon…*Graphical Passwords*. *Draw a secret* (DAS) is a *graphical password* scheme that aims to be better than its textual counterpart in both security and usability. An example of a DAS grid is seen below.

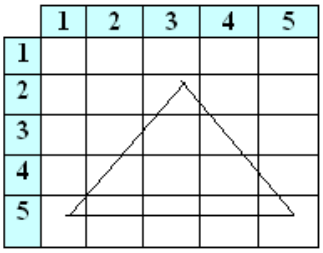


Fig 1: Example password

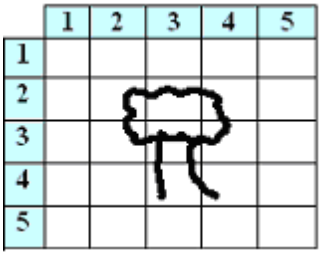


Fig 2: Example password

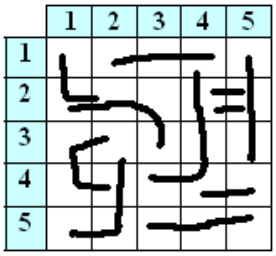


Fig 3: Example password

Fig 1 – Is the equivalent of setting your password as "password" very poor!
Fig 2 – A tree, could this be something you would draw?, and be able to remember ?
Fig 3 – Difficult to remember, impractical.

Procedure…

You will be presented with a drawing grid and asked to draw an image to represent your graphical password**.** You will notice there is a coordinate system on the grid**. To correctly repeat a drawing you must draw through the same squares in the same order, also lifting your pen in the same places.** This means your drawing can be slightly different each time you draw but still acceptable**.** Then you will be asked to return a week later to see if you remember it…

Jargon…

- Stroke – Separate Lines! a line drawn, at the end of which you lift the pen up from the paper. Drawings can contain many strokes (see fig 3)
- Length – The number of cells you cross in total.

It is essential to create a drawing which is as complex as you can remember

Illegal moves….

fig 4: tracing grid lines

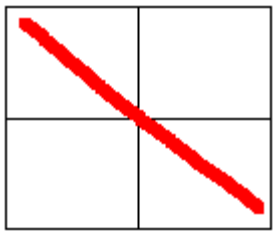


fig 5: cutting diagonals

The above diagrams show two illegal means of constructing an image, you cannot create an image using lines like the above. If you do so, you will have to start again…

Appendix B

Background Draw a Secret (BDAS) Enrolment and Login Sheet

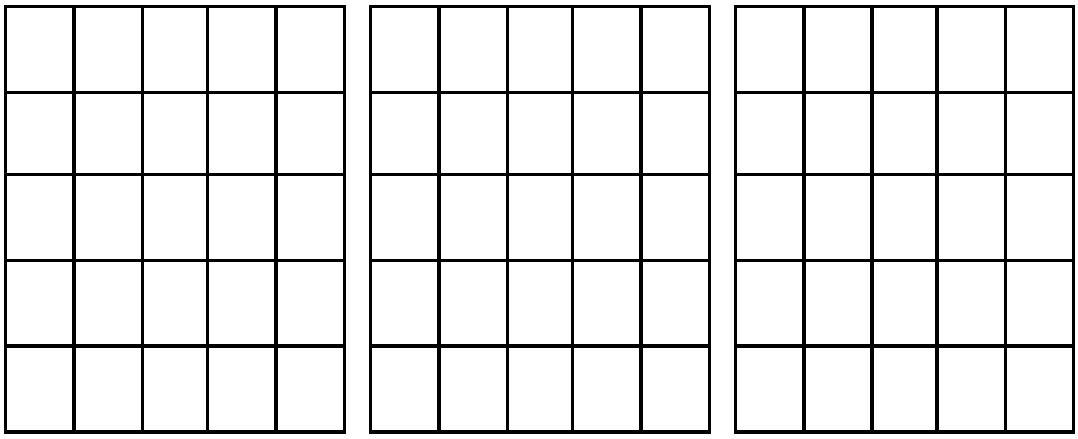
Note. Dimensions altered to fit margins.

Draw-a-Secret Experiment

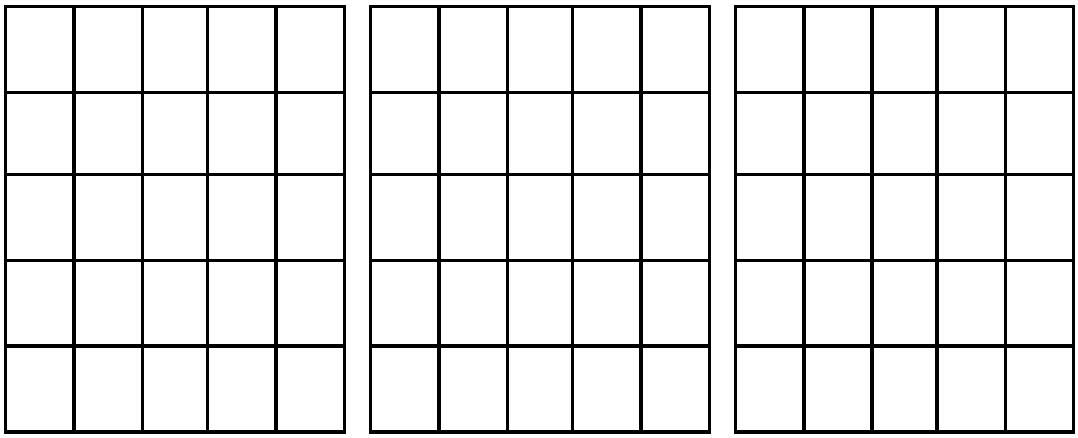
Name	

Set your password….

Set the password you will use here:



Repeat your password….



Remember

1) Lifting your pen from the paper ***DOES*** constitute part of your password
2) Passwords that contain crossings that pass very close to diagonals are void.

Appendix C

Example log file from two weeks of usage for the mobile-based graphical password system

Note. Monday 3rd contains the observation attack study.

*********Enrol********** Mon Oct 27 12:54:18 GMT+02:00 2008 8888494 8801138 7963511 4442334 8078067 4102145

***Login *** Mon Oct 27 13:13:26 GMT+02:00 2008 true 58.244 Keys: 4102145 4442334 8888494 7963511 Decoys: 8769762 7448873 6493381 9912283 5759174 6272508 8002273 306148 6077224 9242885 6191203 814571 5066761 7382931 4010575 9988092 4206270 2644439 3357672 7575297 4806405 5144585 1945808 7105789 6321093 6206394 3754854 195791 9830985 3250059 6602862 5787659

********TRY AGAIN********** Mon Oct 27 13:13:35 GMT+02:00 2008

***Login *** Mon Oct 27 13:14:14 GMT+02:00 2008 true 34.918 Keys: 8801138 7963511 4442334 4102145 Decoys: 9988092 9830985 6191203 5144585 5066761 306148 6272508 6321093 8769762 6206394 6892001 7575297 9003944 1406885 8002273 8254532 5759174 6077224 5803717 3357672 5304269 6493381 3754854 945608 6148842 9242885 7448873 2644439 7382931 7105789 5862731 1945808

********TRY AGAIN********** Mon Oct 27 13:14:19 GMT+02:00 2008

***Login *** Mon Oct 27 13:14:46 GMT+02:00 2008 true 24.998 Keys: 4442334 8078067 8801138 8888494 Decoys: 6892001 6191203 4010575 9003944 9910993 8972678 9830985 8002273 9242885 814571 6077224 6053809 195791 5759174 2644439 7382931 6321093 5066761 4210635 306148 6272508 7575297 5787659 3357672 5803717 9912283 8254532 7448873 3754854 4806405 5304269 5144585

***Login *** Mon Oct 27 15:21:37 GMT+02:00 2008 true 34.773 Keys: 8801138 8888494 4102145 4442334 Decoys: 945608 5304269 4010575 8002273 6272508 6148842 7448873 6892001 3357672 8769762 7105789 5759174 7382931 3557552 9912283 814571 4806405 4210635 6053809 6077224 9910993 3754854 195791 3250059 9242885 5144585 9003944 4206270 306148 6493381 9988092 5803717

***Login *** Tue Oct 28 07:48:14 GMT+02:00 2008 true 27.48 Keys: 4442334 7963511 4102145 8888494 Decoys: 1406885 5759174 3250059 6892001 5066761 9003944 9830985 4206270 6077224 7448873 8769762 4014623

5803717 7575297 3357672 6272508 9988092 6602862 9242885 9910993 4010575 945608 7382931 5304269 814571 5787659 8002273 8254532 4210635 306148 8972678 3754854

***Login *** Tue Oct 28 14:50:45 GMT+02:00 2008 true 26.624 Keys: 7963511 8078067 8888494 8801138 Decoys: 6206394 1406885 8972678 1945808 8254532 306148 2644439 6272508 3357672 814571 4010575 6321093 7105789 5759174 5862731 7575297 9003944 4210635 9830985 195791 4206270 9242885 7448873 945608 4806405 5803717 8002273 3557552 6602862 3754854 9910993 5787659

***Login *** Wed Oct 29 07:16:46 GMT+02:00 2008 true 27.837 Keys: 8078067 4442334 8888494 4102145 Decoys: 6493381 7575297 4210635 7105789 1033254 3250059 6053809 4806405 306148 6602862 3754854 2644439 6892001 9912283 6148842 5066761 6206394 8002273 9988092 5803717 5144585 945608 8769762 3357672 6191203 1945808 5787659 814571 1406885 4206270 6077224 8254532

***Login *** Wed Oct 29 19:40:50 GMT+02:00 2008 true 22.698 Keys: 4442334 4102145 7963511 8078067 Decoys: 5787659 5759174 2644439 1945808 4014623 3557552 8002273 306148 6148842 945608 6206394 7448873 5304269 1033254 9242885 8769762 5862731 6493381 3250059 5144585 8972678 7382931 9988092 5803717 6892001 6602862 4206270 9830985 9003944 3754854 4010575 6077224

***Login *** Thu Oct 30 16:19:32 GMT+02:00 2008 true 16.367 Keys: 8801138 4442334 7963511 8888494 Decoys: 6206394 3357672 6892001 2644439 5803717 7105789 6077224 1945808 7448873 8254532 814571 8972678 7382931 4010575 6053809 306148 6272508 5862731 945608 1033254 9242885 9912283 8002273 3557552 6493381 5066761 5787659 9910993 5759174 5144585 4014623 9988092

***Login *** Fri Oct 31 09:01:08 GMT+02:00 2008 true 19.796 Keys: 4442334 7963511 8888494 8801138 Decoys: 3357672 6493381 6148842 814571 6191203 9910993 8769762 945608 5787659 7105789 4010575 2644439 1406885 5862731 3557552 7448873 8254532 1033254 9003944 6077224

4806405 5144585 5803717 9912283 8972678 4206270 3754854 4210635 7575297 1945808 9242885 8002273

***Login *** Fri Oct 31 14:27:41 GMT+02:00 2008 true 20.592 Keys: 4442334 4102145 8801138 8888494 Decoys: 5066761 6892001 945608 8769762 4010575 9910993 6148842 7382931 3357672 6053809 6077224 1033254 814571 7448873 5304269 5787659 4014623 9830985 5144585 7105789 6191203 1406885 5803717 8972678 8254532 6602862 3250059 9988092 6493381 195791 5759174 2644439

***Login *** Fri Oct 31 17:37:00 GMT+02:00 2008 true 36.862 Keys: 4442334 4102145 7963511 8888494 Decoys: 3250059 5787659 8769762 5862731 4806405 6191203 1945808 3557552 306148 195791 7105789 8254532 9003944 6321093 4210635 1406885 9830985 9988092 9242885 5304269 7448873 6602862 8972678 3357672 6206394 5066761 9912283 6148842 5759174 6272508 6493381 5803717

***Login *** Mon Nov 03 08:53:13 GMT+02:00 2008 true 17.357 Keys: 4442334 8888494 7963511 8078067 Decoys: 945608 9910993 1945808 1033254 2644439 3557552 4806405 6191203 5144585 6206394 8002273 3250059 6493381 195791 5304269 1406885 7382931 5066761 4210635 6053809 8254532 5862731 8769762 3357672 9242885 8972678 814571 7448873 6321093 9912283 6148842 306148

***Login *** Mon Nov 03 14:20:07 GMT+02:00 2008 true 17.952 Keys: 8801138 7963511 4442334 8888494 Decoys: 7448873 1033254 9988092 1406885 8254532 9003944 7105789 814571 4210635 8769762 6206394 6602862 5803717 306148 5066761 3754854 1945808 7575297 945608 3357672 2644439 6148842 6892001 5304269 6077224 6053809 195791 3250059 6272508 4014623 6191203 9912283

***Login *** Mon Nov 03 14:21:02 GMT+02:00 2008 true 15.981 Keys: 7963511 4442334 8078067 4102145 Decoys: 9912283 5803717 8769762 8002273 5144585 8972678 5304269 1033254 6272508 6493381 1406885 6148842 814571 3357672 9830985 1945808 4806405 6191203 5066761 9003944 195791 3250059 6892001 9988092 6321093 7575297 4206270 5862731 5787659 4014623 9910993 5759174

***Login *** Mon Nov 03 14:21:31 GMT+02:00 2008 true 17.253 Keys: 8078067 8888494 8801138 4102145 Decoys: 2644439 4010575 5759174 7105789 7448873 4014623 5862731 195791 6602862 6493381 1945808 814571 3557552 3250059 8002273 7575297 8254532 8972678 3357672 4206270 5787659 9003944 6077224 6272508 6206394 945608 5304269 6321093 5803717 4806405 5066761 6892001

********TRY AGAIN********** Mon Nov 03 14:21:44 GMT+02:00 2008

***Login *** Mon Nov 03 14:22:17 GMT+02:00 2008 false 33.141 Keys: 4102145 7963511 8078067 8801138 Decoys: 306148 5759174 7448873 6272508 945608 3754854 5304269 5803717 8972678 4010575 1033254 5066761 3557552 5144585 7105789 9830985 6493381 4206270 1945808 2644439 9910993 6077224 6148842 4210635 9912283 814571 6191203 6321093 7382931 5787659 8002273 5862731

********TRY AGAIN********** Mon Nov 03 14:22:19 GMT+02:00 2008

***Login *** Mon Nov 03 14:22:50 GMT+02:00 2008 false 29.948 Keys: 4102145 7963511 8078067 8801138 Decoys: 6191203 5759174 3754854 1406885 6321093 5144585 9910993 5803717 9988092 2644439 6892001 195791 6602862 9830985 945608 3357672 5066761 7448873 6493381 7382931 8002273 3250059 1945808 5787659 6053809 4206270 6148842 814571 3557552 6272508 4014623 8254532

********TRY AGAIN********** Mon Nov 03 14:23:17 GMT+02:00 2008

***Login *** Mon Nov 03 14:23:35 GMT+02:00 2008 true 17.292 Keys: 8801138 4102145 8078067 4442334 Decoys: 8002273 7448873 4010575 306148 5304269 6321093 6191203 4806405 7575297 7105789 7382931 3754854 9988092 8972678 8254532 6602862 6148842 1033254 5759174 8769762 6206394 4014623 5144585 4206270 9830985 9912283 9003944 9910993 3557552 3250059 6077224 9242885

********TRY AGAIN********** Mon Nov 03 14:23:42 GMT+02:00 2008

***Login *** Mon Nov 03 14:24:21 GMT+02:00 2008 false 38.578 Keys: 8078067 7963511 4102145 8888494 Decoys: 9988092 6053809 8972678 6191203 4210635 5066761 3557552 4010575 3754854 5803717 5759174 6892001

306148 4014623 1945808 7382931 6272508 5862731 3357672 9003944 6206394 9910993 814571 8002273 5304269 5787659 6602862 9912283 5144585 8769762 6148842 6077224

LOCKOUT! Mon Nov 03 14:24:21 GMT+02:00 2008

********TRY AGAIN********** Mon Nov 03 14:24:23 GMT+02:00 2008

***Login *** Mon Nov 03 14:24:55 GMT+02:00 2008 false 30.603 Keys: 7963511 8078067 4102145 4442334 Decoys: 7575297 3557552 6602862 6206394 5787659 306148 1033254 7105789 6148842 2644439 8769762 6272508 6321093 8972678 8002273 9988092 7382931 3250059 6077224 5862731 6053809 4206270 5803717 5066761 9912283 3357672 8254532 6892001 4014623 4010575 6493381 1945808

********TRY AGAIN********** Mon Nov 03 14:24:57 GMT+02:00 2008

***Login *** Mon Nov 03 14:25:18 GMT+02:00 2008 true 13.479 Keys: 8801138 4102145 8078067 8888494 Decoys: 6191203 945608 6053809 9830985 9988092 9912283 3250059 5144585 8254532 5304269 814571 3557552 6206394 5787659 3357672 6321093 4210635 1033254 4014623 5803717 5759174 6077224 2644439 8972678 6272508 4806405 9242885 6493381 1945808 7575297 9003944 4206270

********TRY AGAIN********** Mon Nov 03 14:25:23 GMT+02:00 2008

***Login *** Mon Nov 03 14:25:50 GMT+02:00 2008 false 26.467 Keys: 7963511 8078067 8801138 4102145 Decoys: 9910993 3250059 5144585 6272508 814571 9003944 3754854 6321093 5304269 6077224 5066761 6892001 8254532 5803717 6053809 1033254 4210635 5862731 195791 5787659 7575297 1406885 6191203 8972678 306148 1945808 6206394 4014623 5759174 6493381 7382931 3357672

********TRY AGAIN********** Mon Nov 03 14:25:51 GMT+02:00 2008

***Login *** Mon Nov 03 14:26:36 GMT+02:00 2008 false 44.15 Keys: 8801138 7963511 4442334 8078067 Decoys: 4014623 9242885 195791 5803717 5144585 3557552 6148842 1033254 945608 8254532 1945808 3754854 9003944 6892001 9988092 9910993 4010575 8769762 1406885 5759174

7575297 2644439 7448873 5787659 814571 306148 6053809 6321093 9912283 6077224 7382931 6191203

********TRY AGAIN********** Mon Nov 03 14:26:43 GMT+02:00 2008

m***Login *** Mon Nov 03 14:26:59 GMT+02:00 2008 true 15.437 Keys: 8078067 4442334 7963511 8801138 Decoys: 6148842 3250059 5803717 4206270 6892001 6272508 4210635 6077224 9910993 6206394 3754854 6321093 6053809 5862731 9003944 306148 9988092 7105789 1406885 5304269 8002273 7382931 814571 2644439 9912283 6191203 4806405 8769762 945608 5787659 8972678 5759174

********TRY AGAIN********** Mon Nov 03 14:27:05 GMT+02:00 2008

***Login *** Mon Nov 03 14:27:45 GMT+02:00 2008 false 38.65 Keys: 8888494 4442334 8801138 4102145 Decoys: 9910993 6077224 4010575 2644439 9242885 306148 3250059 195791 5759174 6206394 5803717 6892001 7448873 4014623 7382931 814571 4206270 7575297 9003944 6602862 6493381 6321093 5862731 9830985 1406885 6148842 4210635 5787659 3557552 1945808 8254532 7105789

LOCKOUT! Mon Nov 03 14:27:45 GMT+02:00 2008

********TRY AGAIN********** Mon Nov 03 14:27:46 GMT+02:00 2008

***Login *** Mon Nov 03 14:28:06 GMT+02:00 2008 false 19.034 Keys: 7963511 8801138 8888494 4102145 Decoys: 6892001 5144585 9910993 3754854 945608 6493381 5862731 6206394 8769762 4014623 8972678 9242885 2644439 5803717 814571 7382931 5304269 6077224 3250059 7575297 6272508 4806405 9830985 8002273 5787659 4206270 6321093 6602862 1033254 6148842 7448873 5066761

********TRY AGAIN********** Mon Nov 03 14:28:11 GMT+02:00 2008

***Login *** Mon Nov 03 14:28:28 GMT+02:00 2008 true 15.925 Keys: 8078067 8801138 7963511 4102145 Decoys: 5144585 6602862 306148 9910993 6191203 5803717 3357672 3557552 9003944 6272508 4806405 3250059 1945808 4014623 6206394 5759174 8002273 945608 8769762 4010575

5304269 6053809 9912283 3754854 2644439 6148842 7382931 9242885 6321093 5787659 9988092 195791

********TRY AGAIN********** Mon Nov 03 14:28:32 GMT+02:00 2008

***Login *** Mon Nov 03 14:28:58 GMT+02:00 2008 false 25.409 Keys: 8801138 8078067 8888494 7963511 Decoys: 9910993 3250059 3754854 195791 814571 5862731 3357672 1033254 7105789 8972678 5304269 9830985 6493381 5759174 6191203 7448873 1945808 4806405 4210635 945608 9912283 4206270 306148 5803717 8002273 3557552 4010575 6272508 6148842 6077224 9003944 6321093

********TRY AGAIN********** Mon Nov 03 14:29:11 GMT+02:00 2008

***Login *** Mon Nov 03 14:29:46 GMT+02:00 2008 false 14.116 Keys: 4442334 8078067 8888494 4102145 Decoys: 945608 4206270 6148842 6321093 9988092 5304269 8769762 1945808 9912283 5144585 4210635 4014623 2644439 9910993 5803717 6602862 3557552 7105789 5066761 1406885 8972678 9242885 7382931 6191203 6272508 195791 7448873 3250059 8002273 8254532 9003944 1033254

********TRY AGAIN********** Mon Nov 03 14:29:48 GMT+02:00 2008

***Login *** Mon Nov 03 14:30:03 GMT+02:00 2008 false 14.662 Keys: 7963511 8078067 4102145 8888494 Decoys: 9912283 6053809 9242885 5862731 4010575 6206394 8972678 5144585 3557552 6892001 4806405 8769762 6077224 7575297 7448873 4014623 5066761 9003944 1406885 5787659 6602862 4206270 9910993 1033254 2644439 6321093 8002273 5304269 8254532 5803717 195791 6272508

l***Login *** Wed Nov 05 15:15:37 GMT+02:00 2008 true 17.254 Keys: 8078067 4442334 8801138 8888494 Decoys: 3357672 3557552 945608 4210635 1033254 2644439 1406885 195791 9830985 5144585 306148 9003944 7382931 8254532 9912283 7575297 4010575 6191203 8769762 1945808 5759174 6077224 5066761 9910993 5304269 6148842 6053809 7448873 6892001 3250059 814571 6272508

***Login *** Wed Nov 05 17:02:45 GMT+02:00 2008 true 18.682 Keys: 8888494 7963511 4102145 4442334 Decoys: 5066761 6191203 6892001 1033254

4014623 6272508 195791 7105789 5787659 7448873 5862731 3557552 8769762 6206394 8002273 7382931 306148 945608 3754854 9988092 4210635 814571 5803717 9003944 6077224 7575297 3357672 5304269 6602862 8972678 4206270 6493381

***Login *** Fri Nov 07 09:17:04 GMT+02:00 2008 true 13.578 Keys: 8801138 4442334 7963511 8078067 Decoys: 7105789 3357672 3250059 6892001 6191203 9910993 4806405 5759174 7448873 3754854 6053809 6321093 9830985 9912283 5144585 5862731 9003944 4014623 4206270 5304269 8002273 2644439 5066761 8254532 9242885 1406885 8769762 195791 5803717 6148842 6493381 814571

Appendix D

Examples of collected descriptions

Female describer: *Uhm, <break> a blonde young woman, with a quite hesitant look about her. She's got her hair tied back and there's <break> little bits com-ing down the sides. <break> Scarf tied around her neck.*

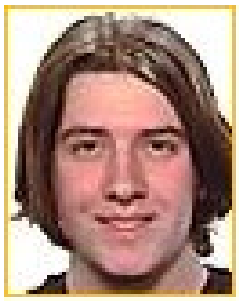
Female describer: *Okay, she's a girl and she's got kind oflongish blonde hair, looks as though it's been high-lighted, so. Uhm, quite straggly wearing a scarf. Uhm, looks quite friendly. Uhm, got her hair tied back. That's about it really.*

Female describer: *She's got blonde hair. Uhm, quite a wide open face, quite pretty. <break> Uhm*

Female describer: *Okay, female, uhm, sort of longish, uhm, darkish blonde hair with a some sort of yellow top and some sort of white-ish collar. Erm. <break> I don't know, slightly long chin <laughs>.*

Male describer: *Female, blonde, dark eyes, thick red lips.<break> Round face, good skin*

Male describer: *Er, blonde woman, quite sticky out chin. Big lips, quite attractive*



Female describer: *He's got, uhm, hair that come down to his jaw line, uhm, it's quite thick hair almost like a girl would wear in a bob. Uhm, he's got, he's not smiling a great deal but even he's got quite a small mouth as well, he's got dark eyes which are quite prominent, uhm, and quite, uhm, a reddy-pink complexion as well.*

Female describer: *Okay, the next one's a man. Uhm <break>he's got quite <break> an oval face, quite a pointy chin <break> he's got brown hair cut down to his, uhm,*

chin. It's quite, a bit wavy a bit unkept.<break> Uhm, he's got a small mouth <break>.He's got brown eyes and quite bushy eyebrows. Uhm <break> he's got quite a wide face, and a bit of a vacant look about him in this picture.

Female describer: *White male. Uhm, chin length brown hair that's quite wavy. Uhm <break> he's got a brown jumper on. Uhm.*

Male describer: *A boy with long brown hair. A small smile on his face. Wide coloured face.*

Male describer: *Male with <break> chin length, slightly unkept, dark brown hair, long dark brown hair. Ovalish square type face.*

Male describer: *Uhm, First thing to notice is the hair. Uhm <break> quite big open eyes, uhm, <break> again face isn't symmetrical, uhm ,<break> first thing to notice is the hair. The mouth, not much of a defined jaw line, cheek bones or anything like that*

Female describer: *Chinese or Taiwanese girl. Smiling, slightly. Erm, looking full face sort of onto to me with her dark hair tied back, middle, uhm, side-parting, sorry. <break> Just looking down slightly but still looking right at me.*

Female describer: *She looks quite young. She's got short, dark hair. Nice eyes, nice eye shadow. Uhm, she looks young. I think she's, sort of Asian. Round face, and hair is short and dark, with a severe parting.*

Female describer: *Okay, this girl has either got, uhm, quite short hair or it is tied back. Oriental with, uhm, brown eyes and very dark hair. Smiling, doesn't look desperately friendly, but I think it's just the camera in front of her.*

Male describer: *Uhm, female, uhm, slightly Oriental looking, dark hair, comb parting on the left hand side, uhm, wearing some sort of white top.*

Male describer: *Uhm, female dark eyes, dark hair, Asian origin, uhm <break>, round face, round chin.*

Male describer: *Er, female, Asian origin, dark hair*

***Female Describer**: "This girl looks quite young, she's got uhm dark hair that looks as though she's had high-lights in but it's on the whole very dark. Uhm, she's a uhm, got dark eyes and sort of quite heavy eyebrows again. Sort of pronounced chin this girl. She looks quite friendly, her forehead's quite big and her cheeks quite pink."*

***Female Describer:** "uhm <break> smiling a little bit shyly. Uhm, she's got dark brown hair hanging loosely probably in a long bob along her face, centre part-ing again, no fringe. Probably quite slim. <break> Winona Ryder type if I had to describe her like anybody."*

***Male Describer:** "White female, sort of light brown to red-dish hair. Long. Uhm, wideish mouth, quite tanned, dark eyes <break> square head. Prettyish."*

***Male Describer: "**Fairly attractive woman with long brown hair."*

***Female Describer:** "He's got very neat, short dark, very dark hair. Uhm, quite a low, wears his hair, style, uhm, which is quite low over his forehead. Uhm, per-haps quite a low hairline. And he's got very prominent dark eyebrows that do very long, is actually one side of his face to the other, really. Uhm, he's got, uhm, very slight mustache. Uhm, and when he smiles he doesn't, his teeth don't show. Uhm, and he looks quite young as well like he doesn't, uhm, the fact that he, uhm, he has quite a youthful complexion also she's got quite slightly tanned skin."*

***Female Describer: "**Okay, this one's, uhm, a man. He's got short, very dark brown hair that looks like its got some curl to it. Huge bushy eyebrows. Uhm, dark brown eyes and a bit of stubble where a moustache could grow but not on the rest of his chin. He's got quite full lips. Uhm <break> and he looks he's got an European kind of*

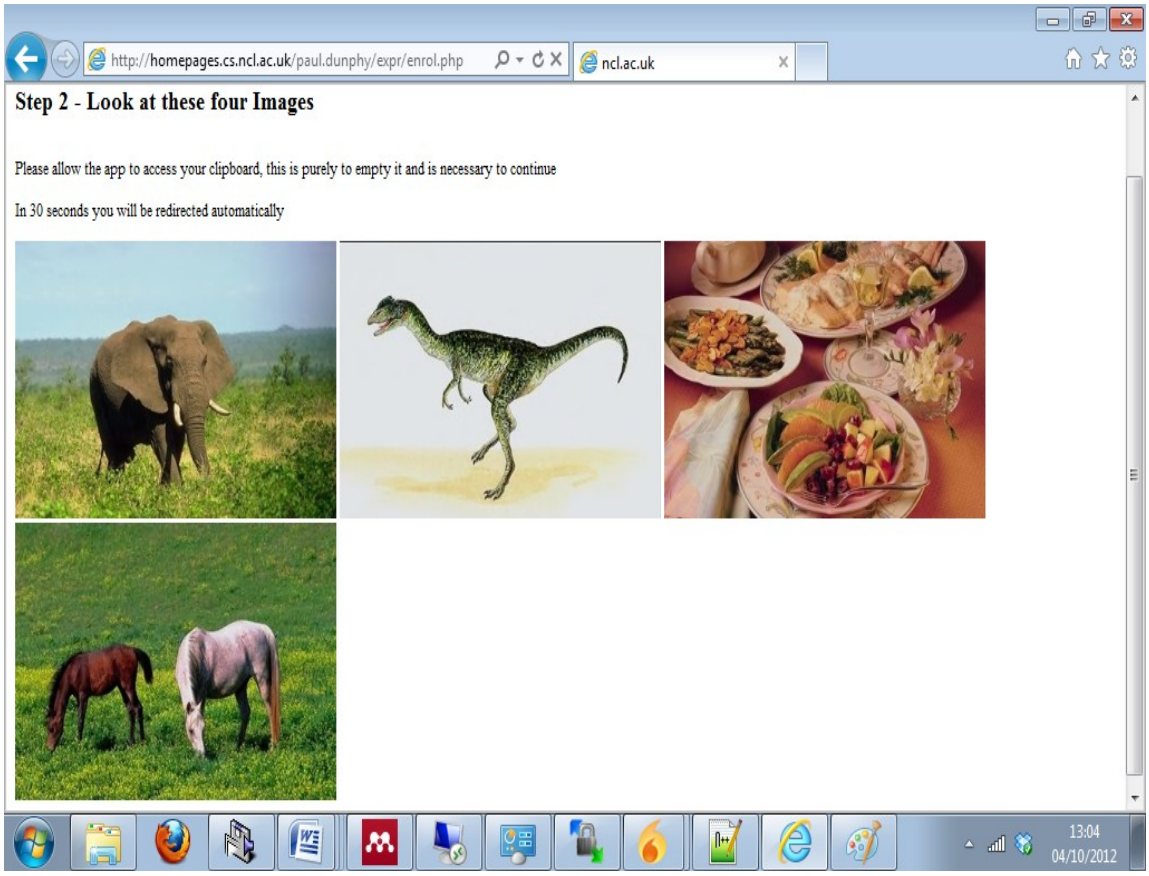
Italian or Spanish colouring to his skin, or a sun tan. Uhm, he's got quite a wide nose at the bottom of his nose but the bridge of his nose is quite narrow. Uhm."

Male Describer: "*Boy. Laughing without showing his teeth. Looking straight ahead, thick eyebrows."*

Male Describer: *"Male with, very thick dark black hair. Very strong hairline. Has dark bushy eyebrows and slight moustache stubble."*

Appendix E

Example enrolment screen for *Mechanical Turk* user study


http://homepages.cs.ncl.ac.uk/paul.dunphy/expr/enrol.php
ncl.ac.uk
Step 2 - Look at these four Images
Please allow the app to access your clipboard, this is purely to empty it and is necessary to continue
In 30 seconds you will be redirected automatically
13:04
04/10/2012