Predicting Graphical Passwords

Matthieu Devlin

St Anne's College
University of Oxford

Supervised by Dr. Duncan Hodges, Dr. Jason Nurse and
Prof. Michael Goldsmith

September 4, 2014

Abstract

PassPoints is a graphical password scheme that has been proposed as an alternative authentication mechanism to passwords. This thesis examines whether a PassPoints password can be predicted from information known about the creator.

A web survey was developed and deployed to obtain demographic, learning style and risk appetite information about participants together with PassPoints passwords created specifically for the study.

The analysis found no evidence that a PassPoints password can be predicted using this information but did confirm previously known security issues such as hotspots. There was some evidence which suggested that the image used can be responsible for different click-order patterns of a password. The study also introduced a rudimentary password strength meter for PassPoints.

Acknowledgements

I would like to thank my supervisors Dr. Duncan Hodges and Dr. Jason Nurse for their help and support over the duration of this project.

Contents

Chapter 1

Introduction

1.1 Motivation

The ability to authenticate and login securely to a remote service has become vitally important in recent times. There are nearly 3 billion Internet users in the world today[1] and most use it multiple times a day to interact with friends on social media sites, to do Internet banking, and to purchase items through e-commerce. Each of these tasks requires a level of security: often information stored within these services contains confidential details so authorization and authentication, particularly through the use of passwords, has become integral. These passwords must be strong enough to be resilient to attack but memorable enough for the user to remember.

In theory, the alphanumeric passwords that are widely used today can provide the necessary security, but in practice they are often insufficient due to our ineptitude to memorise 'strong' passwords, typically deemed as a mix of upper-case and lower-case letters, numbers and punctuation. Many Internet users are guilty of using the same, usually 'weak', password across multiple accounts or a small variation thereof [10]. This clearly poses a risk to the security of the system and is where graphical passwords i.e. passwords based on an image, try to help.

The widespread adoption of touchscreen and handheld devices, has led graphical passwords to gain more popularity over the past decade with their promise of providing greater usability and memorability [2]. Graphical passwords have the potential to be more memorable than alphanumeric passwords but must still provide the required level of security. Many different graphical password schemes have been proposed, some of which claim to have the same high theoretical level of security as alphanumeric passwords and similar levels of memorability even when the user has no prior experience of the scheme [32]. This is encouraging since memorability is likely to increase as users get accustomed to graphical passwords. However, there

[1] `www.internetlivestats.com`

is evidence that graphical passwords suffer from the same pitfalls of human choice as alphanumeric passwords and this is the subject of this project.

1.2 Problem statement

This project examines the predictability of passwords from the cued-recall graphical password scheme PassPoints. The password is created by clicking on an image in five locations. The locations and the order of the clicks form the password.

The relationship between a participant's password and their characteristics such as age, gender and learning style will be explored as well as comparing the passwords of participants who share characteristics to identify those that might be responsible for the similarities. The participant's choice of image will also be examined and if these images encourage different patterns in where the participant clicks.

Any correlations between user characteristics and PassPoints password choice that are found are potential weaknesses to the security of the scheme and could be exploited by an attacker, for example through the use of a dictionary attack[2].

A rudimentary graphical password strength meter will be developed and tested on the data collected during this project, and could form the basis of a feedback loop for user's creating graphical passwords.

1.3 Structure

This dissertation is split into 5 sections including this introduction chapter.

Chapter 2 explains the problem at hand and provides background information on the topic of graphical passwords. The specific graphical password scheme that is explored in this project, PassPoints, is described in detail as well as some security concerns that have already been uncovered through previous research. The context in which this project is defined is also explained, followed by a discussion of how this work fits with the research that has already been conducted on this topic.

Chapter 3 presents the methodology adopted by the project, describing how the survey website was created. Design and implementation details are discussed along with how participants for the survey were recruited.

Chapter 4 reports the results of the statistical analyses of the survey data together with interpretation and discussion of these results.

Chapter 5 concludes the dissertation, giving a brief discussion of the results, reflections on the project as a whole and future work that could be carried out as a result of this project.

[2]A dictionary attack attempts passwords from a precomputed list of likely passwords (a dictionary) which can be much smaller than the password space.

Chapter 2

Background

2.1 Graphical Passwords

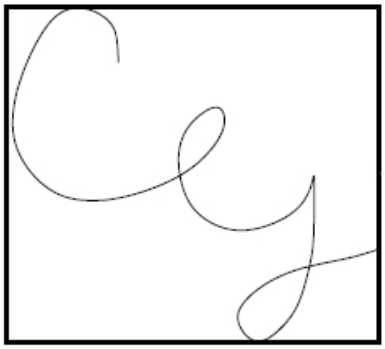
Many graphical passwords schemes have been proposed such as PassPoints, Passfaces and Passdoodle. They boast enhanced security, memorability and usability. Unsurprisingly, mobile operating systems have adopted graphical password schemes. Touchscreen devices demand a more usable interface and a level of security beyond that of the standard PIN. Android uses Pattern-Lock and more recently Windows 8 has introduced Picture Passwords.

There are three types of graphical password schemes: recall, cued-recall and recognition based [25, 27]. Recall-based schemes ask the user to input information without any help from the system, cued-recall schemes require input with some kind of cue or help from the system and recognition based schemes provide the information to the user and ask them to confirm that it is correct. Recognition based tasks are accepted as being easier than recall tasks. However, using a cue, as in the case of cued-recall schemes, increases memorability [28].

Many recall-based graphical password schemes such as Passdoodle require the user to draw on a blank canvas. A problem with Passdoodle is that stored passwords cannot be hashed; the original password is needed to compare against login attempts [2]. This is bad practice for password storage and leaves accounts at risk if the server is compromised.

Recognition based schemes such as Passfaces are generally easiest for humans [27]. They are usually based on requiring a user to pick a single image out of a grid of images [2]. The drawback of these schemes is the small password space; an extremely large grid of images must be used and/or many sets of images (picking one each time), vastly decreasing the usability of the scheme. They can also be susceptible to predictability (see Section 2.5) [11].

Cued-recall based schemes such as PassPoints usually require a user to pick out parts of an image or draw on top of one. Not only does this ease the task of recalling the password due to the cue, it also reduces the amount



(a) An example of a Passdoodle [30]

(b) An example of the Passfaces system [11]

of control the user has when they create the password; they must choose areas on an (often predefined) image. This however creates a weakness of the scheme as the population of users tend to use only a small section of the entire image (see Section 2.3). The next section describes the PassPoints scheme in more detail.

2.2 PassPoints

The graphical password scheme examined in this project is a cued-recall based scheme called PassPoints which was developed by Wiedenbeck et al. [32]. A user creates a password by clicking (or touching on a touch-screen) an on-screen canvas a prescribed number of times (usually 5). The order and position of these clicks are recorded and establish the user's password. To aid memorisation, an image is used as the canvas. To authenticate, the user clicks the same points (within a tolerance) in the same order. An advantage of the PassPoints scheme is that the password space can be larger, making guessing a password harder. An alphanumeric password over a 64-character alphabet with length 8 has a password space of $64^8 = 2.8 \times 10^{14}$ whereas a PassPoints password with an image size of 1024 px $\times$ 752 px with a tolerance of 20 px $\times$ 20 px and 5 clicks points has a password space of $\left(\frac{1024}{20} * \frac{752}{20}\right)^5 = 2.6 \times 10^{16}$.

It has also been shown that although it takes longer to create and authenticate using a PassPoints password, retention was similar to that of alphanumeric passwords over a 5 week period [32]. Although this may not seem like an increase in memorability, it was the first time that many of the participants had used graphical passwords and the first time they had

used PassPoints since it was created for this study. With more practice and familiarity, the memorability of PassPoints is expected to be greater than that of alphanumeric passwords.

Although PassPoints has a large password space size, there are concerns about the security of the scheme.

2.3 Security Concerns with PassPoints

Much like alphanumeric passwords, the security of PassPoints passwords is weakened by the predictable behaviour of users. It is not uncommon for a user to create a alphanumeric password using only lowercase letters. This reduces the password space for a password of length 8 from $64^8 = 2.8 \times 10^{14}$ with a 64-character alphabet to $26^8 = 2.1 \times 10^{11}$, a factor of 1000.

A similar user error for a PassPoints password is choosing the salient (noticeable or important) areas of an image as the points of the password. Golofit [18] showed that over 50% of user clicks were in areas encompassing only 3% of the total image area. This demonstrates a huge reduction in the size of the password space. Large areas of the image are unused because they do not have any stand-out objects, for example a blue sky, or they do have stand-out objects but they do not differ in a memorable way, for example leaves on grass. One of the supposed advantages of the PassPoints scheme is that it has a large password space.

Dictionary attacks where pre-computed lists of likely passwords are used to guess the password in question, have also been proven dangerously effective against PassPoints-style graphical passwords [13, 17, 26, 29]. Van Oorschott and Thorpe [29] presented two dictionary-based attacks. The first was a 'human-seeded' attack, which uses experimental data to predict hot-spots of the images (areas with a large number of clicks). Creating a dictionary a factor of 1000 times smaller than the password space, they were able to predict 26-46% of passwords. Using a dictionary 10^6 smaller than the password space, they were able to predict 10-17% of passwords and 6% within 5 guesses. Their second attack combined these dictionaries with click-order patterns (e.g. a user creates their password by clicking in a left to right fashion) to create a first-order Markov model-based dictionary. This dictionary found on average 7-10% of passwords within 3 guesses and led the authors to comment that "it is difficult to recommend the use of PassPoints-style graphical passwords".

Dirik et al. [13], demonstrated the importance of image choice on the vulnerability of a PassPoints password to a dictionary attack. Using image processing techniques, they created a dictionary of the most likely click locations and showed an attack on two images. They also provided an estimate of the click point entropy for each image which estimates the number of different locations for each click point in the password. In theory, an image

with a higher click point entropy is less vulnerable to dictionary attacks. This was confirmed by the results of their study. 61% of passwords on the first image, which had a click point entropy of 5.3 bits (∼39 locations per click point), were guessed with a dictionary of size 31^5 (the 31 most probable points). The second image, which had a click point entropy of 7.2 bits (∼147 locations per click point), needed a dictionary of size 80^5 to guess only 8.5% of the passwords. This result demonstrates the huge impact image choice has on the security of the password.

With these security concerns in mind, it is important that a user of the system knows what constitutes a 'strong' password.

2.4 Password Strength

There is much literature on calculating the password strength of alphanumeric passwords. Early password strength meters (PSM) used a set of rules to exclude weak passwords [3, 22]. More recent password checkers combine the earlier rules-based PSMs with machine learning and computational linguistic techniques [7, 12]. However, there is not much literature on PSMs for graphical passwords. Andriotis et al. [1] devised a PSM for the Android Pattern-Lock system that used a set of rules based on common password characteristics discovered during a pilot study, for example passwords with a length of less than six nodes are considered weaker.

2.5 Context

The majority of past research on PassPoints has focussed on the image used as the background and has neglected to study the role of the user in depth. This project explores the relationship between a user's PassPoints password and information known about them.

Davis et al. [11] performed a study using an alternative graphical password scheme based on Passfaces in which a user is shown a grid of faces. The user chooses one of the faces, constituting the first point of their password. This process is repeated multiple times (9 times in the study) and the user's password is a set of faces. The authors found that passwords "are highly correlated with the race or gender of the user. For one scheme, this effect is so dramatic so as to render the scheme insecure".

In the PassPoints scheme, there are far more points that a user can choose as part of their password but it is already known that there are particular areas of the images that will be used more often. This project will be focussing on whether users with the same demographics, backgrounds or characteristics tend to choose similar points and therefore similar passwords. Correlations between a user's age, gender, culture, risk appetite and learning styles and their choice of password will be explored. Any correlations that

do exist could be exploited by an attacker perhaps by reducing the size of a dictionary needed to guess the password or through a different attack altogether.

Chapter 3

Study

3.1 Website Overview

The survey website allowed a participant to create a graphical password following the PassPoints scheme and to answer three questionnaires; taking about 20-25 minutes to complete in total. An online survey was chosen for this study as it provided the largest potential sample size. Also, participant's needed to recall their passwords on two separate occasions, days after starting the survey, therefore it would have been impractical to perform the study in a specific location. The survey was designed to be self-contained so all the information that the participant needed was available at each stage of the survey.

Firstly, the participant was asked to read the study and ethics information as outlined in the University-approved proposal as well as providing their email address. The email addresses were collected at this stage to demonstrate a separation between the participant's email address and confidential data collected during the rest of the survey. They were used to send a reminder and link for the recall tasks and to contact the winner of the prize draw (see Section 3.7).

The participant was then given a brief introduction to graphical passwords and the PassPoints scheme before choosing an image to form the background of their PassPoints password (see Section 3.3). The images were shown in a random order to each participant to avoid bias associated with the position of the image on the page. Having chosen the image, the participant creates their password. The tolerance (see Section 3.4) area around each chosen location was shown to demonstrate the margin for error allowed when re-entering the password later in the survey (see Figure 3.2 for example).

After creating a password, the participant was asked to complete three questionnaires on demographic information, learning styles and risk appetite, respectively (see Section 3.5).

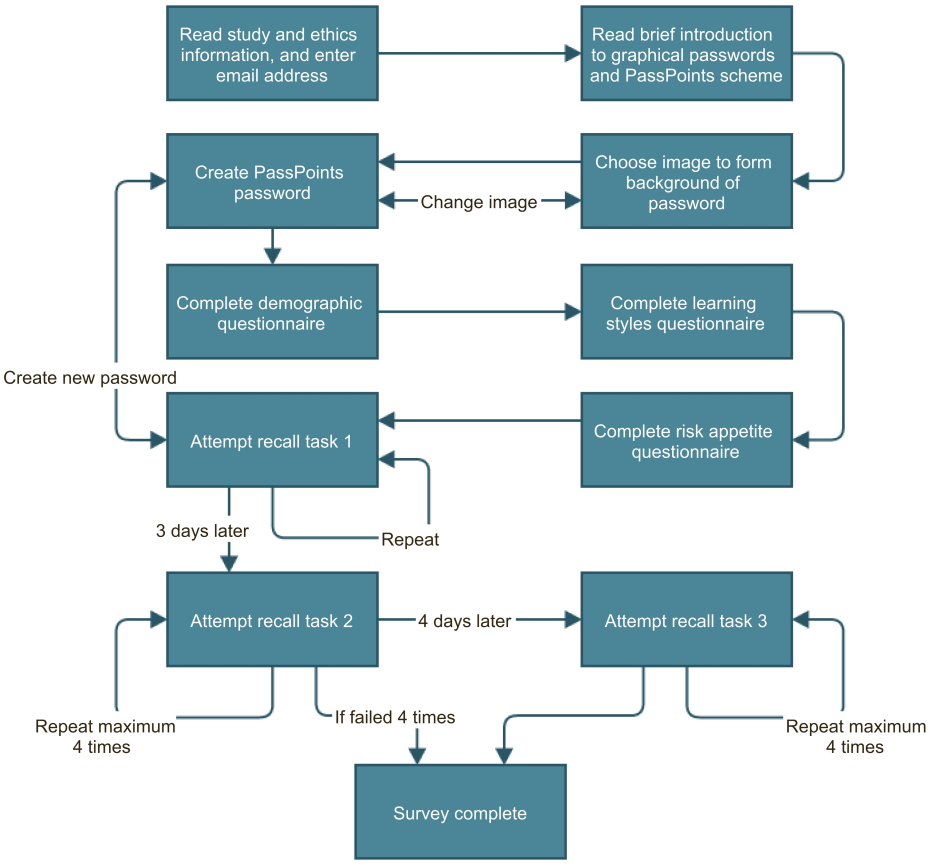


Figure 3.1: The participant's flow through the survey website

The participant was also asked to re-enter their password on three occasions; immediately after completing the questionnaires, 3 days after completing the questionnaires and then again 4 days later (i.e. 7 days after completing the questionnaires), taking less than 5 minutes each time. On each recall attempt, the participant was given 4 attempts to re-enter their password. If they failed on the first recall session (immediately after completing the questionnaires), they were given the opportunity to create a new password or continue to try and enter their original password. Users that changed their passwords at this point were not asked to complete the questionnaires again but instead were taken straight to the first recall page where they could attempt the first recall task again. If a participant failed to re-enter their password correctly on the second recall task (after 3 days), they were deemed to have completed the entire survey. It was unnecessary to test these participants again in the third recall task as it is highly unlikely that they would be able to remember their password after 7 days if they could

Figure 3.2: PassPoints password showing tolerance area

not after 3 days. Participants that could recall their password after 3 days in less than 5 attempts were asked to attempt to recall after 7 days. Like in the previous two recall tasks, the participants were allowed 4 attempts to re-enter their password and were deemed to have completed the whole survey whether they were able to recall it or not.

3.2 Development

The website was written in PHP[1] and used a MySQL[2] database. Javascript and the jQuery library[3] were used to create the PassPoints password creation and authentication functions as well as to add interactivity to the website. The jQuery validation plugin[4] was used to ensure that all required answers were completed by participants using the Safari and Internet Explorer 9 or earlier web browsers which do not support the HTML 'required' attribute.

The data collected during the survey was stored in two database tables; one storing the participant's email and the other storing the remainder of the data collected about them. This separated the participant's identifiable information (email address) from the confidential information collected during the surveys.

The Mail package from the phpPear framework[5] was used to send the participants email reminders to recall their passwords and Cron was used to

[1] `www.php.net`
[2] `www.mysql.com`
[3] `www.jquery.com`
[4] `jqueryvalidation.org`
[5] `pear.php.net`

schedule this task to be run automatically every hour.

3.3 Images

An important consideration during the design of the survey was the choice of images to be used as backgrounds for the passwords. In real-world applications of PassPoints, the user may be allowed to choose their own image. However, in this study, the participants had to choose from three images so that the choice of image by the participant could be analysed.

The images that were selected are shown in Figures 3.3, 3.4 and 3.5 and were chosen because they have different themes that might appeal to different participants, and they have multiple salient points. This was important to the study since images with few salient points are likely to result in similar passwords, making the images the limiting factor in creating similarities between passwords as opposed to the similar characteristics of the participants. The images were available to use under variations of the Creative Commons Licence. The Landscape[6] and People[7] images were found via Flickr[8] and the Animal[9] image via Pixabay[10]. All three of the images were 620 px wide by 413 px high.

Figure 3.3: People image

[6] https://flic.kr/p/8BumLU
[7] https://flic.kr/p/4ycWeu
[8] www.flickr.com
[9] http://pixabay.com/en/chameleon-lizard-multi-coloured-318649/
[10] www.pixabay.com

Figure 3.4: Landscape image

Figure 3.5: Animal image

3.4 PassPoints

The PassPoints functionality was written in Javascript using the jQuery library and adapted from the code at CSS-Tricks.com [9]. The passwords were stored in a JSON format in the database to allow for easy parsing when authenticating a participant's password. In a commercial product implementing the PassPoints scheme, the passwords would be stored as hashed

values for security reasons. This was not done for this study as analysing them would not be possible and security of the passwords themselves was not a concern as the participants were asked not to use any of their personal passwords.

A tolerance of 9 pixels around each point was used giving a 19 px × 19 px tolerance region in line with other studies on PassPoints, e.g. van Oorschot and Thorpe [29]. This tolerance region was shown to the participant during the password creation phase by a yellow box enclosing the chosen point (see Figure 3.2) but not during authentication.

3.5 Questionnaires

Participants were asked to complete three questionnaires. The first contained basic demographic questions such as age, gender and nationality.

The second questionnaire, created by Bixler [4], determined the participants learning style; auditory, tactile, visual or a mixture. Fleming [16] defines learning style as "an individual's characteristics and preferred ways of gathering, organizing, and thinking about information". For example, an auditory learner would prefer attending lectures to reading a textbook. Learning styles were calculated using a points system where 1 point was added to the learning style corresponding to the question if the participant answered 'seldom', 3 for 'sometimes' and 5 for 'often'. The points totals for each learning style as well as the overall style were stored in the database for each participant. A participant's learning style may influence the ease with which they can use graphical password schemes, for example a visual learner may be more successful than a auditory or tactile learner.

The final questionnaire assessed the participant's attitude towards risk and was adapted from Weber et al. [31]. It contained questions on 6 different domains; ethical, investment, gambling, health/safety, recreational and social items, allowing each domain to be analysed separately as well as the participant's overall attitude towards risk. It is likely that a participant's risk appetite will affect their choice of password; a risk-taker, with respect to security, may be more likely to choose a 'weaker' password than a risk-avoider. Therefore, participants with similar attitudes to risk may choose similar passwords.

See Appendix B for the list of questions in the questionnaires.

3.6 Security

It was important to ensure that the website was secure since participants were providing personal information. This was done by eliminating possible cross-site scripting attacks, using MySQL user accounts to limit user privileges, and sanitising all user inputs to help prevent SQL injection attacks.

An independent code review was also performed.

3.7 Recruitment

In accordance with ethical research practices, participation was restricted to those aged 18 or over. The survey was distributed using social media, in particular on Facebook[11], Reddit[12] and The Student Room[13], and users that participated were encouraged to share it with family and friends. Posters were also placed on noticeboards in multiple buildings within the University of Oxford. To encourage participation a £50 Amazon.co.uk voucher was randomly awarded to one participant who completed the survey.

[11] `www.facebook.com`
[12] `www.reddit.com`
[13] `www.thestudentroom.co.uk`

Chapter 4

Analysis

4.1 Population Sample

Of the total of 236 unique emails that were used to start the survey, 180 participants created a password and completed the three questionnaires, 156 completed the recall task after 3 days and 150 completed the whole survey as defined in Section 3.1. Only the data on the 150 people that completed the whole survey was used for the analysis.

Since the survey was online, social media was a particularly useful mode for distribution as it created a geographically further reaching recruitment net. As a result, the participants that completed the whole survey came from 24 different nations (although Oxford has a very multi-national population). Social media distribution also helped to reduce the bias towards participants having higher levels of education that is common in most university studies as well as raising the sample size.

4.2 Which characteristics do people who chose the same image have in common?

4.2.1 Planned Analysis

This section questioned which characteristics of the participant could be used to predict which of the three images they will use to create their password on. As the participant was allowed to change their choice of image if they could not remember their password during the first recall task, only the final chosen image was used in the analysis.

Firstly, a Fisher's exact test was used on each attribute to test whether there were any attributes that led participants to choose a certain image. The attributes, such as `gender` or `education`, are the recorded and computed data from the questionnaires that define a participant's overall characteristics. This has been done since a single characteristic can encompass

multiple attributes, for example there are 10 different attributes for the learning styles characteristic. See Appendix C for a complete list.

Secondly, a multiclass classifier was fit to the data. Since the input was a vector of attributes and the output was one of 3 categories, corresponding to the 3 images, a multiclass classifier was an appropriate technique. Stepwise regression was used to automatically choose which variables were most relevant to predicting image choice.

4.2.2 Results

Fisher's exact test was performed on each attribute that was extracted from the participant's responses to the survey questions against each of the images. Fisher's exact test is used to determine whether there are associations between two categorical variables (see Appendix A.1 for more details). For each attribute, a contingency table is created with the number of categories in that attribute as the number of columns and three rows; one for each image. If the p-value calculated from the Fisher's exact test is less than or equal to α, the null hypothesis that the row and column variables are independent is rejected. α is usually set to 0.05, the 5% level, and was done so in this study too. The alternative hypothesis that the row and column variables are dependent in some way is then favoured (not described in any way by the test). Alternatively, if the p-value is greater than α it cannot be said that the row and column variables are independent of each other, instead it is interpreted as there exists no evidence that they are not independent, hence they may or may not be independent.

For the Fisher's exact test, all continuous variables such as age or mean of responses to risk appetite questionnaire, had to be discretised into categories as the test uses nominal data. This arguably decreases the accuracy of the test but is negligible in most cases. Some of the variables were discretised multiple times using different boundaries, for example, the age characteristic was split into categories with 10 year gaps, e.g. 20-30, 30-40; into categories with 20 year gaps and into categories with 40 year gaps (this reduced to above and below 50 years old).

No significant results were obtained when applying the Fisher's exact test to each of the attributes on the three images and therefore we can conclude that the choice of image by the user is not dependent of any single attribute.

Next, a multiclass classifier was fit to the data (attributes of all the participants) which could predict which image the participant would choose given their attributes. A multiclass logistic regression model was chosen to classify the data (see Appendix A.3 for more details). The goal was to fit a classifier that would produce the most accurate results given the data from the study.

Stepwise regression was used to perform this model selection task, which

involves automatically finding which attributes can be used to accurately predict the chosen image. Stepwise regression uses some metric to compare models and the Akaike information criterion (AIC), which is most often used, was used in this study [8].

Backwards stepwise regression was used and starts with a model containing all of the independent variables. At each 'step', the independent variable which has the least impact on how the model fits the data is removed. This is found by calculating the AIC of the model at this 'step' with each remaining independent variable removed in turn. The independent variable that results in the lowest AIC value once removed, has the lowest impact on model fit. It is therefore removed from the model for the next 'step'. This process is continued until removal of any of the remaining independent variables does not decrease the AIC i.e. the AIC has been minimised. The remaining independent variables are those that have the most impact on fitting the model to the data.

To avoid overfitting in the model, the data was randomly shuffled and the first 80% of participants in the shuffled set were used to select and train the model. This process was repeated 50 times. If there existed an accurate model to predict the image that the participant would choose, then the same set of attributes (or a very similar set) would be selected in each model. The following attributes were part of all 50 models: `auditoryNorm`, `tactileNorm`, `visualNorm`, `ethical`, `gambling`, `health`, `investment`, `recreational` and `social`. The `auditoryNorm`, `tactileNorm` and `visualNorm` attributes correspond to normalised scores for the learning styles questionnaire and `ethical`, `gambling`, `health`, `investment`, `recreational` and `social` correspond to means of groups of questions in the risk appetite questionnaire (see Appendix C for more details).

The `age`, `children`, `tactile` and `visual` variables also appeared in 25 or more of the models. These correspond to a participant's age, whether they have children, and their raw score for tactile and visual questions in the learning style questionnaire respectively (see Appendix C for more details).

Although there were multiple attributes appearing in the model for each iteration, the models always performed poorly. For each iteration, the model was tested on the remaining 20% of the data and the accuracy, precision and recall of the model were calculated.

Image	Precision	Recall	Accuracy
People	0.376	0.325	
Landscape	0.447	0.621	0.399
Animal	0.294	0.163	

Table 4.1: Performance of multiclass logistic regression on participant image choice

Table 4.1 shows the average performance of the multiclass models over the 50 iterations. It clearly shows that the models performed poorly: the accuracy, which is the overall correctness of the model (number of correct classifications divided by the total number of classifications), is less than 40% and the precision and recall are also poor, especially for the Animal image. The recall for the Landscape image looks promising however this was because the model predicted most of the participants would use that image. If the model were to predict the Landscape image for each participant, the precision for the Landscape image would be 1 but the model would be very inaccurate so it is important to look at the values for each classification.

The repeated presence of a number of attributes in the model for each iteration provides some suggestion that a multiclass classifier could potentially be used to predict which image the participant chooses. However in this study the small sample size does not allow for an accurate model to be produced.

In summary, there were no significant results when applying Fisher's exact test, therefore there is no evidence that a participant's choice of image is dependent on any single attribute. A multiclass logistic regression model was successfully fit to the data however its predictions were very poor. However, there were multiple attributes that were in the model on each of the 50 iterations, suggesting that they might have an influence on image selection. A larger sample size would be needed to explore this claim.

4.3 What characteristics do people who can recall their password have in common?

4.3.1 Planned Analysis

Logistic regression was chosen to answer this question as the desired output was binary, the participant either passed or failed, and the input was a vector of attributes. The classifier was used to predict whether the participant could recall their password after 3 days i.e. the result of the second recall task.

To visually compare the difference in where participants clicked who passed and failed, a scatter plot was created with the points of participants who passed in green and those that failed in red. Clusters and patterns of people who were able or unable to recall their password could be picked out from the scatter plots.

To quantify any differences between where participants who passed and failed clicked, a test of categorical data was performed by firstly discretising the image into 20 px by 18 px boxes and counting the number of clicks in each box separately for those who passed and those who failed. This produced a 2 by n contingency table, where n was the number of discretised

boxes in the image, 713 in this case. Under the simplifying assumption that a participant's click positions are independent of each other (in practice they are not), a test of categorical data could be used.

Since many of the boxes had a count of 0, Fisher's exact test was used which does not assume a sufficiently large sample size and all expected cell counts of greater than 5. See Appendix A.1 for more details.

Another Fisher's exact test was used to determine whether there was a significant difference in where participants clicked who could recall their password after 3 and 7 days and those who could recall after 3 but not after 7. Finally, another Fisher's exact test was used to determine whether there was a difference in difficulty in recalling a password on each of the three images. Difficulty was defined by the ratio of participants who could and could not recall their password after 3 days.

4.3.2 Results

By fitting a logistic regression model to the data, a set of attributes that can be used to predict whether a participant will pass or fail can be extracted. As in Section 4.2.2 the desired model was initially unknown and therefore backwards stepwise logistic regression was used. The model was selected 50 times, shuffling the data each time and taking the first 80% of participants to select and train the model. If the same set of attributes for the model were selected each time, there is some suggestion that this set could be used to predict whether a participant will pass or fail. The `ethical`, `investment` and `riskMeans` attributes appeared in the model on more than 10 of the 50 iterations. Although this is far from unanimous, it may suggest that these attributes do have an influence on a participant's ability to recall their password after 3 days.

Accuracy	Precision	Recall	Specificity
0.728	0.753	0.957	0.003

Table 4.2: Performance of logistic regression for predicting recall or not after 3 days

Table 4.2 shows the average performance of the classifiers produced using backward stepwise logistic regression. The performance appears somewhat promising from the accuracy, precision and recall but this is misleading. A participant being able to recall their password after 3 days is defined as a positive result and a participant failing to recall after 3 days is defined as a negative. Using these definitions, the specificity is defined as the proportion of participants that the model correctly predicts will fail to recall their password. Comparing this with the recall which gives the proportion of participants that the model correctly identifies as being able to recall their password, there is clearly a fault with the classifiers; they are very good at

predicting if a participant will recall but very poor at predicting if they will not.

Now consider a classifier that gives a positive result to every participant i.e. predicts that every participant can recall their password after 3 days. Since 112 of the 150 participants in the study could recall their password after 3 days, we can assume that, on average, 22.4 participants in the test dataset (30 participants) would be able to recall their password after 3 days. With the classifier that predicts that all participants can recall their password, the classifier would have an average accuracy of 0.747 as on average it would predict 22.4 of the 30 participants correctly (all participants that could recall their password). This classifier would then have a better average accuracy than the classifiers built using backwards stepwise logistic regression suggesting that a more accurate logistic regression model using the available attributes and data is not possible. This does not necessarily mean that a classifier cannot be built using stepwise logistic regression on these attributes, however, more data is needed to select and train the model.

To determine whether there was a difference in where people who failed to recall their password after 3 days clicked and those that could recall clicked, a Fisher's exact test was used.

	People	Landscape	Animal	Combined
p-value	0.121	0.051	0.027	0.008

Table 4.3: Results of Fisher exact test on clickpoints of participants who recalled and did not recall after 3 days

Table 4.3 shows a significant result ($p < 0.05$) for the Animal image; the participant's ability to recall their password is dependent on where they clicked. Using Fisher's method to combine the p-values on each of the images (see Appendix A.2) yields $p = 0.008$ which is also significant. Therefore the null hypothesis that the participant's ability to recall their password is independent of where they clicked is rejected. This may suggest that the users that fail do so because of where they choose to click. There are many potential reasons for this, perhaps they tend or decide to choose points that are harder to remember, or fail to pick out locations that are easy to remember.

The scatter plots in Figures 4.1, 4.2 and 4.3 show the points where participants clicked on each of the three images. The green dots represent points chosen by participants who could recall their password after 3 days and the red dots represent the participants who could not.

It is clear, particularly in Figure 4.2, that the participants who failed to recall their password after 3 days pick different points to those who can recall. Often the differences in the positions of clicks are small but this may have a large impact on whether the participant can recall the password later

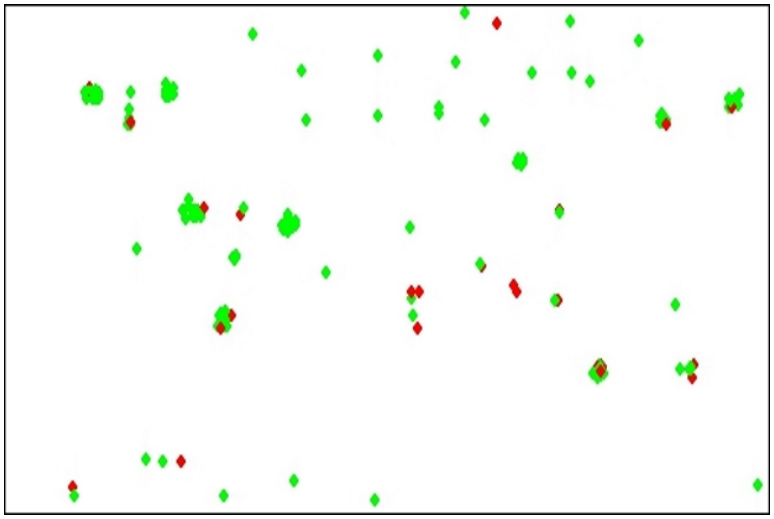


Figure 4.1: Pass/fail scatter plot for People image

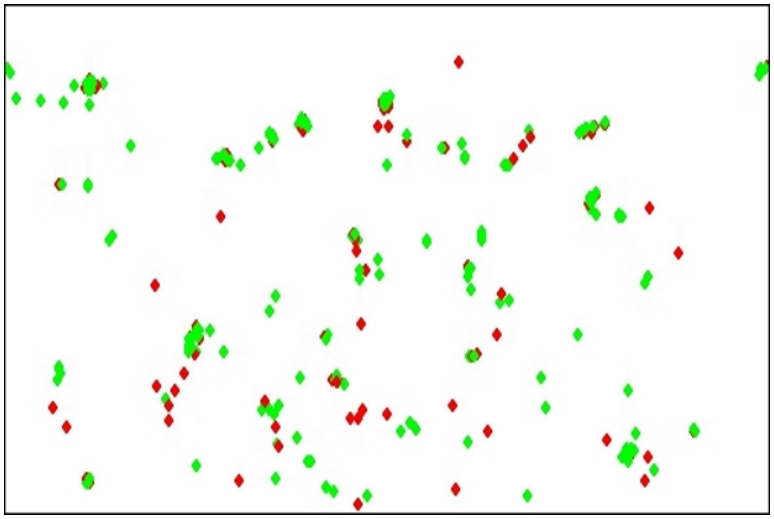


Figure 4.2: Pass/fail scatter plot for Landscape image

on.

A Fisher's exact test was used to determine whether there was a difference in where people who could recall after 3 days but not after 7, and people who could recall after 3 and 7 days clicked. When combined using Fisher's method, $p > 0.05$ so there is no evidence to reject the null hypothesis that they are independent, suggesting that there is no difference between where participants who could recall after 3 and 7 days clicked and participants who could recall after 3 days but not after 7.

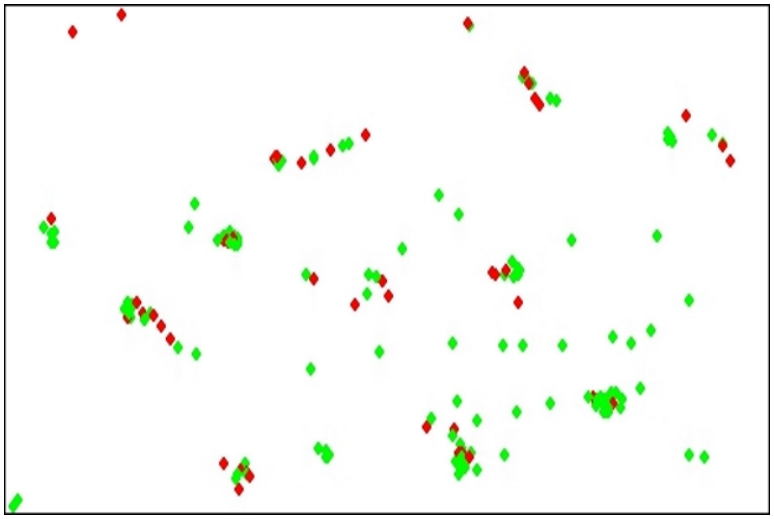


Figure 4.3: Pass/fail scatter plot for Animal image

Finally, another Fisher's exact test was used to determine if there was a difference in difficulty in recalling a password on each of the images. The resulting p-value was 0.34. Therefore there is no evidence to reject the null hypothesis that the proportion of participants that could recall their password after 3 days is independent of the image they choose. Hence there is no evidence that recalling a password on one image is any harder than recalling on one of the other images.

In summary, the logistic regression model performed no better than assigning the same outcome to all participants. However, there were three attributes that appeared in the models on more than 10 of the 50 iterations, suggesting that they may have an influence on whether or not a participant can recall their password after 3 days. A Fisher's exact test showed that there was a statistically significant difference in where participants who could recall and participants who could not recall clicked. This was also confirmed visually with the scatter plots.

There was shown to be no difference in where participants who could recall after 3 days but not after 7 days and participants who could recall after 3 and 7 days clicked. Also, there was no difference in difficulty of recalling passwords across each of the three images.

4.4 Where do people click and what characteristics do people who click in the same place share?

4.4.1 Planned Analysis

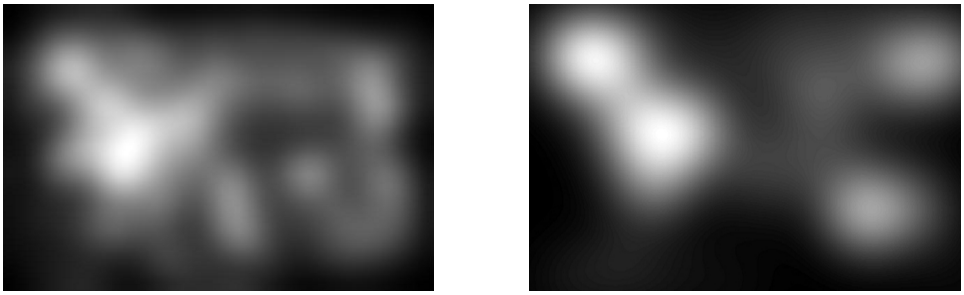
To analyse where people clicked, saliency and heat maps where created for each image. The saliency maps were produced using MATLAB [19] and the Itti-Koch-Neibur algorithm [21]. Saliency maps provide a visual representation of how much each point of an input image stands out with respect to its neighbouring points.

The heat maps were produced in R [20] and show every participant's click-points as well as areas of the images that were popular amongst participants. Previous research [18, 29] has shown that users tend to use salient points as part of their passwords; this was tested by comparing the saliency and heat maps for the data from this study.

Using the same discretisation technique as in Section 4.2.1, the attributes of participants that click in the same areas of the image could be compared. A Fisher's exact test was used on m by n contingency tables where m is the number of categories of the attribute and n is the number of boxes in the discretised image, 713 in this study. This tests whether clicking in a box is independent of the attribute category. Where a significant result was found, the boxes with a large difference where analysed visually for characteristics that could have led to the differences.

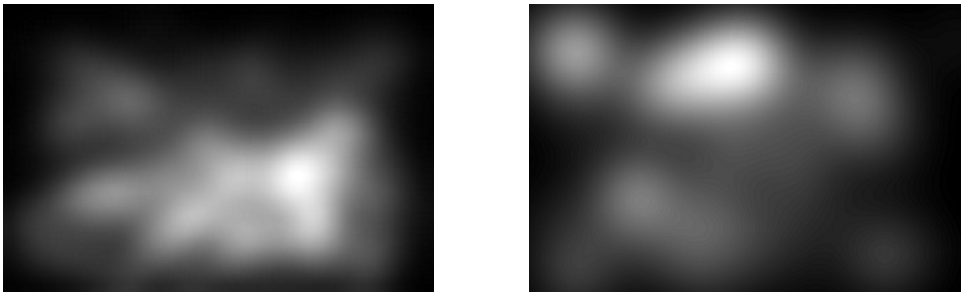
4.4.2 Results

It is already known that graphical password schemes, including PassPoints, are subject to hotspots [2, 17, 29]. Saliency maps of the three images were created using the Itti-Koch-Niebur algorithm [19, 21] and heatmaps created using kernel density estimation [20].



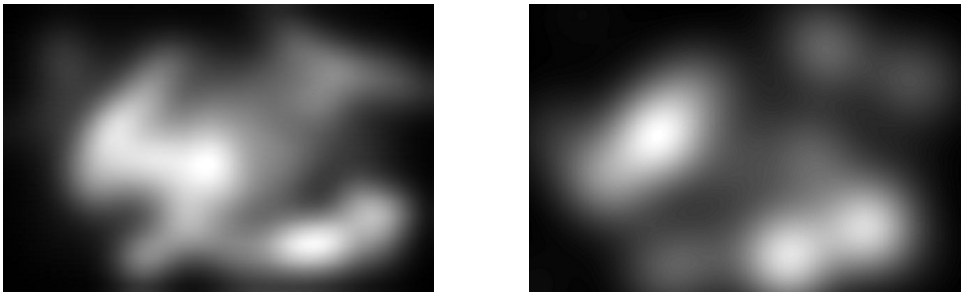
(a) Saliency map (b) Heatmap

Figure 4.4: Saliency map and heatmap for the People image



(a) Saliency map (b) Heatmap

Figure 4.5: Saliency map and heatmap for the Landscape image



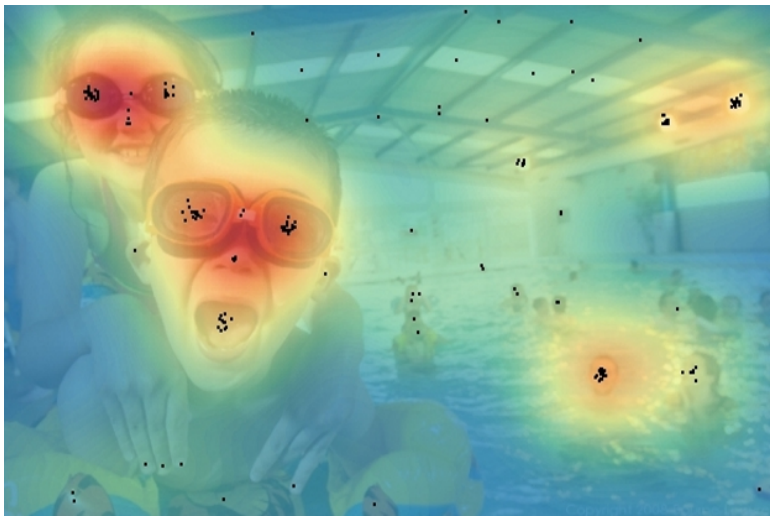
(a) Saliency Map (b) Heatmap

Figure 4.6: Saliency map and heatmap for the Animal image

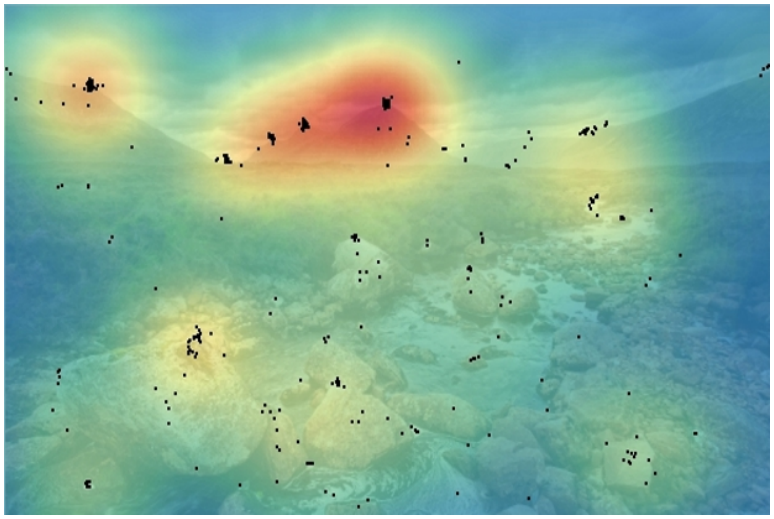
Figures 4.4, 4.5 and 4.6 show the saliency maps and heatmaps of the People, Landscape and Animal images respectively. There are some similarities between the maps particularly for the People (Figure 4.4) and Animal (Figure 4.6) maps, agreeing with the previous research that some users pick salient points as part of their passwords. On the People image, it can be seen that the children's faces in the top left corner are both salient and a popular region for participants to click, as well as the chameleon's eye, nose and feet in the Animal picture. The saliency map and heatmap for the Landscape image in Figure 4.5 are noticeably different. The saliency map focuses on the rocks and stream in the bottom right quadrant of the image, whereas the heatmap shows the most popular region as the central mountain. It is not surprising that the mountains were used by many participants as part of their password as they are memorable points; it is probably more surprising that the mountain peaks were not picked up as salient points in the saliency maps. This may be due to the Itti-Koch-Neibur algorithm that was used to implement the saliency map. An alternative algorithm may have produced a slightly different saliency map that was more consistent with expected salient regions.

Figures 4.7a, 4.7b, 4.7c show alternative heatmaps of the three images

with the heatmap superimposed on top of the image along with all the click points. The closer the colour is to red, the more clicks that are present in that region. These alternative heatmaps show that predictable points such as the children's faces, the central mountain and the nose, eye and feet of the chameleon are used by many people, demonstrating a reduction of the password space.

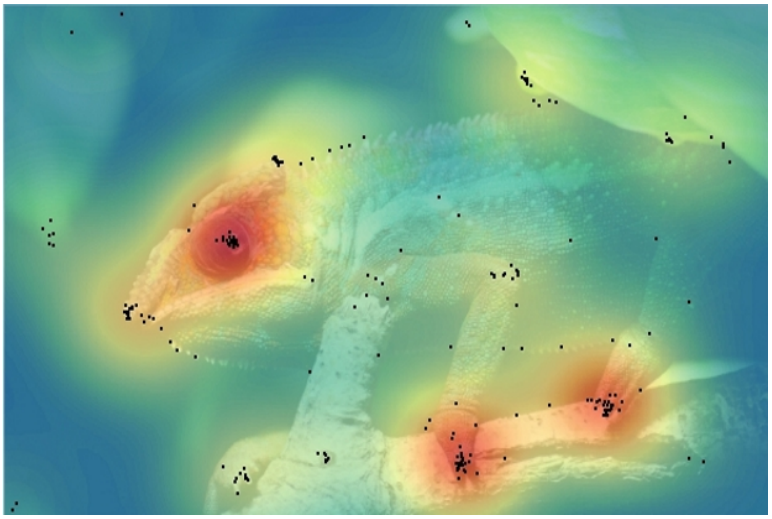


(a) Alternative heatmap for People image



(b) Alternative heatmap for Landscape image

Fisher's exact tests were performed on 43 categorical attributes of the participants against the 'boxes' in which they click on the images. Of the 43 tests, the only significant value at the $\alpha = 0.05$ level was for the `Gambling-Cats2` attribute (see Appendix C) when combining the p-values for the three



(c) Alternative heatmap for Animal image

Figure 4.7: Alternative heatmaps for each image

images using Fisher's method, $p = 0.003$. `GamblingCats2` corresponds to a participant's responses to questions in the gambling section of the risk appetite questionnaire. It is important to note that these p-values were not corrected for multiple comparisons and that of the 3 images, there is only a significant result on the People image. As the test is not significant for the attributes `GamblingCats` and `GamblingCats1` (different ways of categorising the gambling attribute) combined over the 3 images, the validity of the significant result for `GamblingCats2` is questionable. If the p-value were to be corrected for multiple comparisons using the Bonferroni correction, the result would no longer be significant. As the number of simultaneous tests performed on a dataset increases, the probability of finding a statistically significant result also increases. The Bonferroni correction is used to account for this (see Appendix A.4).

Fisher exact tests were also done on combinations of attributes, such as being married and having children, and position of clicks but all results were non-significant.

In summary, many participants choose similar points, that are often salient, but the attributes that they possess, at least the ones in this study, have no effect on the locations that they will click.

4.5 Do people share the same pattern of clicks (i.e. the order) and if so which characteristics do they share?

4.5.1 Planned Analysis

Each participant's click pattern was classified in a similar way to van Oorschot and Thorpe [29] into the following patterns:

- Left-to-right (LR), $x_i \leq x_{i+1} + \tau$
- Right-to-left (RL), $x_i \geq x_{i+1} - \tau$
- Bottom-to-top (BT), $y_i \leq y_{i+1} + \tau$
- Top-to-bottom (TB), $y_i \geq y_{i+1} - \tau$
- Bottom-left-to-top-right (LR_BT), $LR \wedge BT$
- Top-left-to-bottom-right (LR_TB), $LR \wedge TB$
- Bottom-right-to-top-left (RL_BT), $RL \wedge BT$
- Top-right-to-bottom-left (RL_TB), $RL \wedge TB$
- Clockwise (CW)
- Anticlockwise (ACW)
- None of these patterns (NONE)

where x_i and y_i are the x and y co-ordinates of i^{th} point in the participant's password and τ is the error tolerance. $\tau = 19$ px was used to match the tolerance region size used during authentication of the PassPoints passwords.

The clockwise and anticlockwise patterns where calculated using the cross products of vectors [6]: three consecutive points are taken and two vectors are created from the first two and last two points respectively. Taking the cross product of these two vectors, creates a new vector in the z-plane and is positive if the angle between them is less than 180°, an anticlockwise turn and negative if greater than 180°, a clockwise turn. If there are 5 positive cross products (corresponding to the angles at each point) in the password, it has an anticlockwise pattern, and if there are 5 negative cross products, the password has a clockwise pattern. The following formula was used to calculate the cross-products:

$$\textit{cross-product} = (x_i - x_{i-1})(y_{i+1} - y_i) - (y_i - y_{i-1})(x_{i+1} - x_i)$$

To ensure that the password follows a circular pattern, a restriction that the sum of all angles had to be less that 360°was imposed, disallowing spiral patterns.

Having assigned click-order patterns to each participant, a Fisher's exact test could be used for each pattern and attribute to test whether the pattern of a participant's password is independent of an attribute that they have. Each pattern was tested individually as the patterns are not mutually exclusive, for example a LR_BT pattern has both the LR and BT patterns. An alternative test would be the z-test for proportions but this only works if the attribute has two categories therefore Fisher's exact test was chosen for consistency across attributes and patterns.

Having determined whether participant's attributes affect their pattern, the role of the image itself was examined. A Fisher's exact test was again used to test whether click-order patterns were independent of the image the participant chose to create their password on.

4.5.2 Results

Each participant's click points were analysed and categorised into the order patterns as described in Section 4.5.1. Table 4.4 shows the distribution of click-order patterns; note that each participant can have more than one pattern for example a participant with the LR_BT pattern also has the LR and BT patterns.

Pattern	Count
LR	37
RL	3
BT	16
TB	23
LR_BT	6
RL_BT	2
LR_TB	12
RL_BT	0
CW	11
ACW	17
NONE	71

Table 4.4: Distribution of click-order patterns on the three images

From Table 4.4 it appears that the LR pattern is more common than the RL pattern. A Fisher's exact test was used to test whether there were statistical differences in the proportion of two patterns; LR versus RL, TB versus BT, CW versus ACW and any pattern versus no pattern (NONE). The results are summarised in Table 4.5.

There is a significant result for the LR versus RL patterns. This is not surprising as the vast majority of participants were from countries where reading and writing is performed from left to right so it is expected that

Pattern	p-value
LR versus RL	2.2×10^{-9}
BT versus TB	0.303
CW versus ACW	0.321
Pattern versus No Pattern	0.419

Table 4.5: Distribution of click-order patterns on the three images

this pattern would be transferred into other activities such as creating a graphical password.

To examine whether there is a difference in the pattern a participant adopts based on their attributes, a Fisher's exact test was performed on each attribute and pattern. Three combinations were significant at the $\alpha = 0.05$ level without any correction for multiple comparisons: `gender` with BT pattern ($p = 0.016$), `children` with TB pattern ($p = 0.014$) and `riskCats3` with BT pattern ($p = 0.021$). The first of these significant results shows that men and more likely than women to use a BT pattern (proportions: $men = 0.16, women = 0.03$). The second shows that participants who do not have children are more likely to use a TB pattern than those with children (proportions: $without\,children = 0.19, with\,children = 0.00$). The final result shows that participants with a medium risk average are less likely to use a BT pattern than those with a high or low average and that those with a high average are most likely to use the BT pattern (proportions: $low = 0.20, medium = 0.08, high = 0.36$). It is difficult to qualify these results; it is conceivable that there would be a difference between genders but the other two results are harder to explain. When corrected for multiple comparisons using the Bonferroni correction all three of the results are no longer significant suggesting that the significant results are just down to chance and that there is no correlation between participant attributes and the patterns they use. The Bonferroni correction is used as many attributes and patterns are being tested simultaneously.

The final analysis looks at whether the image can influence the participant to create a password with a certain pattern. A Fisher's exact test on counts of the number of participants that used the pattern in question and those who did not against the 3 images, was used. The results are summarised in Table 4.6.

There were significant results for the TB, LR_BT and CW patterns. Table 4.6 suggests that the image does have an influence on the patterns in a participant's password. The TB pattern is an example of this with 25% of participants of the People image having the pattern whereas less than 5% of participants use it on the Animal image. Grouping the CW and ACW patterns, over 27% of participants with the Animal image used a 'rotational' pattern whereas only 8% and 19% used them on the People and Landscape

Pattern	p-value	Proportion of Pattern on Image		
		People	Landscape	Animal
LR	0.911	0.222	0.243	0.273
RL	0.229	0.056	0.143	0.000
BT	0.790	0.138	0.100	0.091
TB	0.028	0.250	0.171	0.045
LR_BT	0.028	0.083	0.000	0.068
RL_BT	0.724	0.028	0.014	0.000
LR_TB	0.545	0.111	0.086	0.045
RL_BT	1.000	0.000	0.000	0.000
CW	0.020	0.000	0.057	0.159
ACW	0.896	0.083	0.129	0.114
NONE	0.587	0.528	0.429	0.500

Table 4.6: Difference in click-order patterns on the three images

images respectively. The LR pattern is prevalent in each of the images in similar proportions; 22%, 24% and 27% for the People, Landscape and Animal images respectively. This may be due to the majority of participants being nationals of countries with a left-to-right written language and this pattern creeps into other 'observational' activities.

It should be noted that if a correction for multiple comparisons using the Bonferroni correction, the significant results in Table 4.6 are no longer significant. However this seems possibly incorrect as 3 of 11 are statistically significant.

In summary, there is a statistically significant difference in the popularity of the different click-order patterns however a participant's choice of pattern is not predictable from a single attribute. There is also no statistically significant difference in the proportion of participants choosing a single pattern across the three images, however there is some evidence to the contrary.

4.6 Evaluating PassPoints password strength

4.6.1 Planned Analysis

Often when a user is creating an alphanumeric password on a website, they are provided with feedback on the strength of their password usually on a scale from 'weak' to 'strong'. It is important that the same feedback is given for graphical passwords especially since most users are unfamiliar with them and may not understand what constitutes a 'strong' password. To do this, a set of rules were created for scoring passwords based on information taken from saliency maps and click-order pattens. This allows the method to be used on any image and does not require any click-point data for that

particular image. These rules were developed after the survey was completed and used the data collected from it, hence participants were not shown the strength of their password during the study.

Using the saliency maps from Section 4.4.1, the nth percentile of saliency values for each image was calculated ($n = 90, 95, 97$). These values for n were chosen as Golofit [18] found that 50% of clicks occurred in only 3% of an image; the other two levels were included to increase the size of the used salient region without making it too large. The regions of the saliency maps for each image were then filtered such that only the highest (100-n)% of saliency values remained; the set of these regions is denoted $\phi_{n,i}$ where $n = 90, 95, 97$ and $i = 0, 1, 2$; the image used.

The password strength score was calculated as follows:

- δ_{sal} - the number of the participant's click-points in the filtered saliency map $\phi_{n,i}$.
- δ_{pat} - 1 if the participant's password exhibits a pattern described in Section 4.5.1, else 0.
- δ_{LR} - 1 if the participant's password exhibits a left-to-right pattern as described in Section 4.5.1, else 0.

The password strength score, Δ is then $\Delta = \delta_{sal}+\delta_{pat}+\delta_{LR}, (0 \leq \Delta \leq 7)$. The left-to-right pattern was chosen as an indicator of password strength as over 20% of passwords had this pattern for each of the three images (see Section 4.5.2) hence if an attacker prioritises for this common pattern, the password could possibly be obtained faster.

4.6.2 Results

The password strength score, Δ, was calculated for each participant using $n = 90, 95, 97$ however $n = 90$ was chosen as the final percentile value for the saliency maps as it produced a larger distribution of password scores. Table 4.7 shows the distribution of password scores on each image. There were no passwords with a score of 6 or 7.

Figure 4.8 shows the proportion of passwords with each password strength score for each image. The Landscape and Animal images have approximately normal distributions of the proportions and the Landscape image has a positive skew. Therefore there are more 'stronger' passwords on the Landscape image than the Animal image. This may be because the Landscape image facilitates 'stronger' passwords than the Animal image or, by chance, the participants that chose the Landscape image pick 'stronger' passwords.

The People image has an approximately uniform distribution but with a larger value for $\Delta = 1$. Therefore, there are approximately equal proportions of each password strength score for $0 \leq \Delta \leq 4$. This may suggest that the People image promotes more 'weaker' (although also more 'stronger')

Password score, Δ		Images		
		People	Landscape	Animal
Strong	0	6	9	1
	1	11	25	12
	2	5	20	16
	3	7	10	9
	4	7	5	5
	5	0	1	1
	6	0	0	0
Weak	7	0	0	0

Table 4.7: Password score distribution for each image

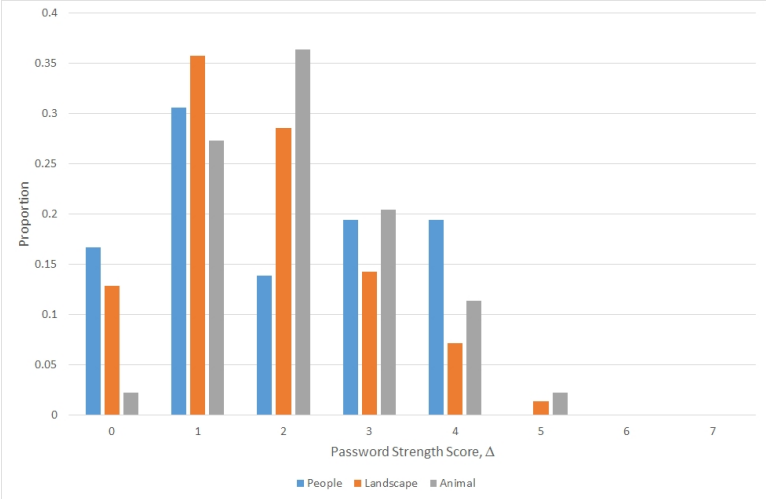


Figure 4.8: Bar chart of proportions of password strength score for each image

passwords than the other two images. This may present a weakness as a user that creates a 'weak' password on the People image, may have chosen a 'stronger' password on the Landscape or Animal image because they discourage 'weaker' passwords.

Figure 4.9 show passwords that got high ($\Delta \geq 4$) password strength scores denoting 'weaker' passwords. Figure 4.10 show passwords that got low ($\Delta = 0$) password strength scores denoting 'stronger' passwords. The lower the password strength score Δ, the 'stronger' the password.

Both passwords in Figure 4.9 could be considered 'weak' by an observer suggesting that the rules set out to calculate the strength of a graphical password can pick out a 'weak' password. However Figure 4.10b is given a password strength score $\Delta = 0$ but does not appear to be as 'strong' as the score suggests especially after comparison with Figure 4.10a, which

(a) Password Strength, $\Delta = 4$

(b) Password Strength, $\Delta = 5$

Figure 4.9: Examples of weaker passwords

also has a password strength score $\Delta = 0$. It would seem that the only non-obvious point-choice is on the chameleon's body, between its legs. The reason for this error is that calculating the password score relies heavily on the saliency maps (5 of the possible 7 points are awarded from the position of clicks), and the saliency map for the Animal image does not recognise the chameleon's nose as one of the most salient points, nor the front knee joint or either foot. However, each of these locations would likely be considered salient. Therefore to improve the accuracy of the password strength scores, the algorithm implementing the saliency maps for the images would have to

(a) Password Strength, $\Delta = 0$

(b) Password Strength, $\Delta = 0$

Figure 4.10: Examples of stronger passwords

be improved so that the most 'stand-out' features are identified.

In summary, the password strength score developed can pick out some 'strong' and 'weak' passwords but the algorithm used to implement the saliency map needs improvement to increase the accuracy of the metric.

Chapter 5

Conclusions

The goal of this study was to determine whether it is possible to predict a PassPoints password if personal information is known about the creator. The results have provided no statistical significant evidence that it is consistently possible to do so.

This study has shown that participants tend to choose similar locations for the points of their passwords thereby creating hotspots, in accordance with previous research on PassPoints. It has also confirmed that these hotspots tend to occur on regions of the image that are salient, although the saliency maps produced in this study did not always pick out what are expected to be salient points. This is possibly due to a limitation in the algorithm that implements the saliency map.

There is also some evidence that the click-point pattern that a user's password exhibits is dependent on the image that they chose as their background, exposing a potential weakness to the scheme. However, this evidence was only available after processing the password data on the image, therefore an attacker would need prior knowledge of the image and have password data on it to exploit this information.

This study introduced a password strength metric that can be used to provide feedback on the strength of a PassPoints password without the requirement of knowing the image or having password data for that image.

5.1 Future Work

It is hard, based on the results of this study, to recommend further research into the hypothesis that PassPoints passwords can be predicted from user attributes. Although this study only had a small sample size, there was no evidence to suggest that the hypothesis was true. However, the study has shown that the image may an effect on the click-order pattern that a user exhibits. Only simple patterns were explored in this study, so further work with more participants, images and patterns, could result in further

findings. It was also shown that the left-to-right pattern was common on all images and it would be interesting to determine whether this was a result of the majority of participants belonging to a nations with a language that is written and read from left to right.

The password strength metric also warrants further study. The algorithm implementing the saliency maps in this study seemed unsuitable for the task. For the Landscape and Animal images, regions that would be considered salient, such as the central mountain peak on the Landscape image and the nose and feet of the chameleon in the Animal image, were not picked out as salient regions in the saliency map. Also each point that was in the high-saliency regions was weighted the same, each point with a score of 1, however picking the most salient point as the first point in the password may, for example, create a 'weaker' password than picking it for the fifth point of the password. Similarly, a score of 1 was awarded for any password that exhibited any one of the click-order patterns explored in the study, however they may not be equally weighted in regards to the strength of the password; perhaps the patterns that are most commonly seen across all images would be worth more than those that are rarely seen. Another possible extension would be to equate a password score with how long it would take to break using a dictionary attack and/or map password scores to an alphanumeric password of equivalent strength, relating the score to a tangible attacker scenario. Overall, the password strength metric is an interesting topic for further research and has practical application if graphical passwords, and in particular PassPoints, become more popular in the future.

Appendices

Appendix A

Statistical background

This section provides some background information into the statistical methods used during this study.

A.1 Fisher's exact test

Fisher's exact test is used to determine whether there are associations between two categorical variables, although it does not give any indication of what the associations are if they do exist. It is used instead of a chi-squared test when sample sizes are small and expected cell counts are less than 5, since it exactly computes the chi-squared statistic. The sampling distribution for a chi-squared test is only approximately chi-squared and is inaccurate at low sample sizes however at large sample sizes, it is preferred as the approximation is accurate and is less computationally demanding than Fisher's exact test. Fisher's exact test is usually performed on 2×2 contingency tables however the tests used in this study often involve larger contingency tables [14].

To calculate the p-value of a Fisher's exact test, consider the following contingency table for the categorical variables X and Y which have m and n different categories respectively and $\mathbf{A}$ is a matrix of counts over these variables:

	X_1	...	X_m	Row Total
Y_1	$a_{1,1}$	...	$a_{m,1}$	R_1
...	...	...	...	...
Y_n	$a_{1,n}$	...	$a_{m,n}$	R_n
Column Total	C_1	...	C_m	N

$a_{i,j}$ represents the count of observations with category i for variable X and category j for variable Y and must be non-negative, R_i and C_j represent

the row and column sums, respectively, and the sum of all counts

$$N = \sum_{i=1}^{n} R_i = \sum_{j=1}^{m} C_j = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j}$$

Let β denote all of the possible matrices that have the same row and column sums as $\mathbf{A}$:

$$\beta = \left\{ \mathbf{B} : \ \mathbf{B} \text{ is } m \times n, \sum_{j=1}^{n} b_{ij} = R_i, \sum_{i=1}^{m} b_{ij} = C_j \right\}$$

Then the probability of observing any matrix $\mathbf{B} \in \beta$ is given by

$$P(\mathbf{B}) = \prod_{j=1}^{n} \frac{C_j!}{y_{1j}! y_{2j}! \ldots y_{mj}!} / \frac{\sum_{i=1}^{m} R_i}{R_1! R_2! \ldots R_n!}$$

The p-value for the observed count matrix $\mathbf{A}$ is the sum of all the probabilities of count matrices in β that are less than or equally as likely as $\mathbf{A}$, i.e.

$$p = \sum_{\mathbf{B} \in \phi} P(\mathbf{B})$$

where $\phi = \{\mathbf{B} : \mathbf{B} \in \beta \text{ and } P(\mathbf{B}) \leq P(\mathbf{A})\}$ [23].

A.2 Fisher's method

Fisher's method is used to combine p-values of independent tests of significance, resulting in a single chi-squared test statistic with $2k$ degrees of freedom, where k is the number of independent tests being combined. It is calculated using the formula

$$\chi^2_{2k} \sim -2 \sum_{i=1}^{k} ln(p_i),$$

where p_i is the p-value of the i^{th} test of significance. From this chi-squared test statistic, a p-value for the combined tests can be easily obtained [15].

A.3 Logistic regression

Logistic regression is a probabilistic binary classification model which can be extended to multiclass classification. It takes as input a feature vector $\mathbf{x}$ $(x_1, ..., x_n)$ of continuous and/or categorical data and returns a categorical outcome variable $\hat{y}$. The model takes the form of

$$p(y = C_1 | \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x})$$

where y is the true class corresponding to the feature vector $\mathbf{x}$, $\mathbf{w}$ $(w_0, w_1, ..., w_n)$ is the weight vector with w_0 as the constant term and w_i as the weight of the feature x_i, and

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

is the logistic sigmoid function which maps the real numbers to the range $[0, 1]$, which can be treated as probabilities. By using the decision rules

$$p(y = C_1|\mathbf{x}, \mathbf{w}) \geq 0.5 \Rightarrow \hat{y}(\mathbf{x}) = C_1$$
$$p(y = C_1|\mathbf{x}, \mathbf{w}) < 0.5 \Rightarrow \hat{y}(\mathbf{x}) = C_2$$

the model produces the categorical outcome $\hat{y}$. The multiclass extension of the logistic regression model, called multinomial logistic regression, has the form

$$p(y = c|\mathbf{x}, \mathbf{W}) = \frac{\exp(\mathbf{w}_c^T \mathbf{x})}{\sum_{c'=1}^{C} \exp(\mathbf{w}_c^T \mathbf{x})}$$

where C is the number of classes [24].

Logistic regression assumes linear relationships between any continuous input variables and the logit (the inverse of the logistic function) of the output variable, and that observations in the data set are independent, for example two observations representing the same person at two different times would violate this assumption [14].

A.4 Bonferroni correction

The Bonferroni correction is used when a large number of significance tests are performed simultaneously. Suppose that the significance level is α, then the probability of at least one significant result is

$$\begin{aligned} P(\text{at least one significant results}) &= 1 - P(\text{no significant results}) \\ &= 1 - (1 - \alpha) \end{aligned}$$

If there were κ hypotheses being tested simultaneously, then the probability of at least one significant result is

$$P(\text{at least one significant results}) = 1 - (1 - \alpha)^{\kappa}$$

Now suppose that $\alpha = 0.05$ and $\kappa = 20$, then the probability of at least one significant result is 0.64. As the number of hypothesis tests increases, the more likely you are to observe a significant result. The Bonferroni correction corrects for this by saying that a single test is significant at the α level only if the p-value

$$p < \frac{\alpha}{\kappa}$$

where κ is the number of hypothesis tests being run simultaneously [5].

Appendix B

Questionnaires

The questions asked in each of the questionnaires is listed below.

B.1 Demographic questions

1. What is your age?
2. What is your gender?
3. What is your relationship status?
4. What is your nationality?
5. Please rearrange these colours (by dragging them with your cursor) according to the ones you prefer the most (starting on the left) to those you prefer the least (ending on the right). Blue, brown, green, orange, purple, red, yellow.
6. Are you left or right handed?
7. What is the highest level of education you have completed?
8. If you have completed a degree of some kind, which faculty was it in (Select the most appropriate)?
9. Do you have past experience with graphical passwords?
10. Do you think graphical passwords are more secure than textual passwords?
11. In your opinion, are graphical passwords easier to remember?

B.2 Learning Styles questions

This questionnaire is taken from [4].

Instructions: For each question, choose either 'seldom', 'sometimes' or 'often'.

1. I can remember more about a subject through the lecture method with information, explanations and discussion.
2. I prefer information to be presented the use of visual aids.
3. I like to write things down or to take notes for visual review.
4. I prefer to make posters, physical models, or actual practice and some activities in class.
5. I require explanations of diagrams, graphs, or visual directions.
6. I enjoy working with my hands or making things.
7. I am skillful with and enjoy developing and making graphs and charts.
8. I can tell if sounds match when presented with pairs of sounds.
9. I remember best by writing things down several times.
10. I can understand and follow directions on maps.
11. I do better at academic subjects by listening to lectures and tapes as opposed to reading a textbook.
12. I play with coins or keys in pockets.
13. I learn to spell better by repeating the words out loud than by writing the word on papers.
14. I can better understand a news article by reading about it in the paper than by listening to the radio.
15. I chew gum, smoke, or snack during studies.
16. I feel the best way to remember is to picture it in your head.
17. I learn spelling by tracing the letters with my fingers.
18. I would rather listen to a good lecture or speech than read about the same material in a textbook.
19. I am good at working and solving jigsaw puzzles and mazes.
20. I play with objects in hands during learning period.

21. I remember more by listening to the news on the radio rather than reading about it in the newspaper.

22. I obtain information on an interesting subject by reading relevant materials.

23. I feel very comfortable touching others, hugging, handshaking, etc.

24. I follow oral directions better than written ones.

B.3 Risk Appetite questions

This questionnaire is taken from [31].

Instructions: For each of the following statements, please indicate the likelihood of engaging in each activity or behaviour. Select a rating from 1 to 5, using the following scale: 1 Very unlikely; 2 Unlikely; 3 Not sure; 4 Likely; 5 Very likely.

1. Admitting that your tastes are different from those of your friends.

2. Going camping in the wilderness, beyond the civilization of a campground.

3. Betting a day's income at the horse races.

4. Buying an illegal drug for your own use.

5. Cheating on an exam.

6. Chasing a tornado or hurricane by car to take dramatic photos.

7. Investing 10% of your annual income in a moderate growth mutual fund.

8. Consuming five or more servings of alcohol in a single evening.

9. Cheating by a significant amount on your income tax return.

10. Disagreeing with your parents on a major issue.

11. Betting a day's income at a high stake poker game.

12. Having an affair with a married man or woman.

13. Forging somebody's signature.

14. Passing off somebody else's work as your own.

15. Going on a vacation in a third-world country without prearranged travel and hotel accommodations.

16. Arguing with a friend about an issue on which he or she has a very different opinion.

17. Going down a ski run that is beyond your ability or closed.

18. Investing 5% of your annual income in a very speculative stock.

19. Approaching your boss to ask for a raise.

20. Illegally copying a piece of software.

21. Going whitewater rafting during rapid water flows in the spring.

22. Betting a day's income on the outcome of a sporting event (e.g. baseball, soccer, or football).

23. Telling a friend if his or her significant other has made a pass at you.

24. Investing 5% of your annual income in a conservative stock.

25. Shoplifting a small item (e.g. a lipstick or a pen).

26. Wearing provocative or unconventional clothes on occasion.

27. Engaging in unprotected sex.

28. Using someone else's personal WiFi connection to access the Internet for an extended period of time (i.e. not paying for your own). (Original question - Stealing an additional TV cable connection off the one you pay for - changed to make more applicable worldwide)

29. Not wearing a seatbelt when being a passenger in the front seat.

30. Investing 10% of your annual income in government bonds (treasury bills).

31. Periodically engaging in a dangerous sport (e.g. mountain climbing or sky diving).

32. Not wearing a helmet when riding a motorcycle.

33. Gambling a week's income at a casino.

34. Taking a job that you enjoy over one that is prestigious but less enjoyable.

35. Defending an unpopular issue that you believe in at a social occasion.

36. Exposing yourself to the sun without using sunscreen.

37. Trying out bungee jumping at least once.

38. Piloting your own small plane, if you could.
39. Walking home alone at night in a somewhat unsafe area of town.
40. Regularly eating high cholesterol foods.

Appendix C

Data variables

Below lists the variables that were used during the data analysis along with a brief description.

points PassPoints password in JSON format

imageNum Number corresponding to image [0 - People, 1 - Landscape, 2 - Animal]

attempts Number of recall attempts on first recall task

recall1Attempts Number of recall attempts on second recall task (after 3 days)

recall2Attempts Number of recall attempts on third recall task (after 7 days)

imageChanges Number of times the participant has changed their choice of image (after first recall task)

age Participant's age

ageRange Age in categories: (10, 20], (20, 30], (30, 40], (40, 50], (50, 60], (60, 70], (70, 80]

ageRangeBig Age in categories: (10, 30], (30, 50], (50, 70], (70, 90]

ageRangeBigger Age in categories: (10, 50], (50, 90]

gender Participant's gender

relationshipStatus Participants's relationship status with categories: single, in a relationship, married, divorced, separated, widowed

children Whether the participant has children

nationality Participant's nationality

colours The colours [blue, brown, green, orange, purple, red, yellow] in order of preference

favColour Participant's favourite colour from colours variable

leastFavColour Participant's least favourite colour from colours variable

hand Participant's handedness

education Participant's level of education with categories: none, gcse, alevels, foundation, bachelors, masters, doctorate, other

degree Department of participant's degree subject with categories: arts/humanities; biological and life sciences; business, management and economics; education, social sciences and law; engineering; environment; mathematics and physical sciences; medicine and health; performance, visual arts and communication; other

faculty Faculty of participant's degree subject with categories: arts, medicine, science, polysci, other, NA

experience Whether the participant has experience with graphical passwords

secure Whether the participant thinks graphical passwords are more secure than alphanumeric passwords with categories: yes, no, don't know

memorability Whether the participant thinks graphical passwords are more memorable than alphanumeric passwords with categories: yes, no, don't know

learningStyle Participant's learning style with categories: auditory, tactile, visual - can be a mixture

auditory Participant's points for questions relating to auditory style in learning style questionnaire

tactile Participant's points for questions relating to tactile style in learning style questionnaire

visual Participant's points for questions relating to visual style in learning style questionnaire

auditoryNorm Participant's auditory score divided by score for auditory, tactile and visual combined

tactileNorm Participant's tactile score divided by score for auditory, tactile and visual combined

visualNorm Participant's visual score divided by score for auditory, tactile and visual combined

auditoryNormCats Participant's auditoryNorm value in categories: (0.1, 0.2], ..., (0.9, 1.0]

tactileNormCats Participant's tactileNorm value in categories: (0.1, 0.2], ..., (0.9, 1.0]

visualNormCats Participant's visualNorm value in categories: (0.1, 0.2], ..., (0.9, 1.0]

riskMeans Participant's mean value for responses to risk appetite questions

riskCats1 Participant's riskMeans in categories: (0, 1], ..., (4, 5]

riskCats2 Participant's riskMeans in categories: (0, 0.5], (0.5, 1.0], ..., (4.5, 5]

riskCats3 Participant's riskMeans in categories: (0, 1.67], (1.67, 3.33], (3.33, 5]

riskCats4 Participant's riskMeans in categories: (0, 2.5], (2.5, 5]

ethical Participant's mean value for responses to ethical questions in risk appetite questionnaire

investment Participant's mean value for responses to investment questions in risk appetite questionnaire

gambling Participant's mean value for responses to gambling questions in risk appetite questionnaire

health Participant's mean value for responses to health/safety questions in risk appetite questionnaire

recreational Participant's mean value for responses to recreational questions in risk appetite questionnaire

social Participant's mean value for responses to social questions in risk appetite questionnaire

ethicalCats Participant's ethical value in categories: (0, 0.5], (0.5, 1.0], ..., (4.5, 5.0]

ethicalCats1 Participant's ethical value in categories: (0, 1], (1, 2], ..., (4, 5]

ethicalCats2 Participant's ethical value in categories: (0, 1.67], (1.67, 3.33], (3.33, 5.0]

investmentCats Participant's investment value in categories: (0, 0.5], (0.5, 1.0], ..., (4.5, 5.0]

investmentCats1 Participant's investment value in categories: (0, 1], (1, 2], ..., (4, 5]

investmentCats2 Participant's investment value in categories: (0, 1.67], (1.67, 3.33], (3.33, 5.0]

gamblingCats Participant's gambling value in categories: (0, 0.5], (0.5, 1.0], ..., (4.5, 5.0]

gamblingCats1 Participant's gambling value in categories: (0, 1], (1, 2], ..., (4, 5]

gamblingCats2 Participant's gambling value in categories: (0, 1.67], (1.67, 3.33], (3.33, 5.0]

healthCats Participant's health value in categories: (0, 0.5], (0.5, 1.0], ..., (4.5, 5.0]

healthCats1 Participant's health value in categories: (0, 1], (1, 2], ..., (4, 5]

healthCats2 Participant's health value in categories: (0, 1.67], (1.67, 3.33], (3.33, 5.0]

recreationalCats Participant's recreational value in categories: (0, 0.5], (0.5, 1.0], ..., (4.5, 5.0]

recreationalCats1 Participant's recreational value in categories: (0, 1], (1, 2], ..., (4, 5]

recreationalCats2 Participant's recreational value in categories: (0, 1.67], (1.67, 3.33], (3.33, 5.0]

socialCats Participant's social value in categories: (0, 0.5], (0.5, 1.0], ..., (4.5, 5.0]

socialCats1 Participant's social value in categories: (0, 1], (1, 2], ..., (4, 5]

socialCats2 Participant's social value in categories: (0, 1.67], (1.67, 3.33], (3.33, 5.0]

Bibliography

[1] P. Andriotis, T. Tryfonas, and G. Oikonomou. Complexity metrics and user strength perceptions of the pattern-lock graphical authentication method. In T. Tryfonas and I. Askoxylakis, editors, *Human Aspects of Information Security, Privacy, and Trust*, volume 8533 of *Lecture Notes in Computer Science*, pages 115–126. Springer International Publishing, 2014.

[2] R. Biddle, S. Chiasson, and P. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Comput. Surv.*, 44(4):1–41, Sept. 2012.

[3] M. Bishop and D. V. Klein. Improving system security via proactive password checking. *Computers & Security*, 14(3):233–249, 1995.

[4] B. Bixler. Learning styles inventory. `www.personal.psu.edu/bxb11/LSI/LSI.htm`, (n.d.). [Online; accessed 2-May-2014].

[5] J. M. Bland and D. G. Altman. Multiple significance tests: The Bonferroni method. *BMJ*, 310(6973):170, 1995.

[6] P. Bourke. Determining whether or not a polygon (2D) has its vertices ordered clockwise or counterclockwise. `paulbourke.net/geometry/polygonmesh/`, 1998. [Online; accessed 8-July-2014].

[7] C. Castelluccia, M. Dürmuth, and D. Perito. Adaptive password-strength meters from Markov models. In *19th Annual Network and Distributed System Security Symposium*, NDSS. The Internet Society, 2012.

[8] J. E. Cavanaugh. 171:290 model selection - lecture 2: The Akaike information criterion. `myweb.uiowa.edu/cavaaugh/ms_lec_2_ho.pdf`, Aug. 2012. [Lecture online, accessed 4-August-2014].

[9] CSS-Tricks. Get x, y coordinates of mouse within box. `css-tricks.com/snippets/jquery/get-x-y-mouse-coordinates/`, 2009. [Online; accessed 4-May-2014].

[10] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The tangled web of password reuse. In *Proceedings of NDSS*, 2014.

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

[12] M. M. A. Devillers. Analyzing password strength. `http://www.cs.ru.nl/bachelorscripties/2010/Martin\_Devillers\_\_\_0437999\_\_\_Analyzing\_password\_strength.pdf`, 2010. [Online; accessed 16-August-2014].

[13] A. E. Dirik, N. Memon, and J.-C. Birget. Modeling user choice in the passpoints graphical password scheme. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS '07, pages 20–28, New York, NY, USA, 2007. ACM.

[14] A. Field. *Discovering Statistics Using SPSS*. Sage, 3 edition.

[15] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, 4 edition, 1932.

[16] N. Fleming. *Teaching and Learning Styles: VARK Strategies*. N.D. Fleming, 2001.

[17] H. Gao, W. Jia, N. Liu, and K. Li. The hot-spots problem in Windows 8 graphical password scheme. In G. Wang, I. Ray, D. Feng, and M. Rajarajan, editors, *Cyberspace Safety and Security*, volume 8300 of *Lecture Notes in Computer Science*, pages 349–362. Springer International Publishing, 2013.

[18] K. Golofit. Click passwords under investigation. In *Proceedings of the 12th European Symposium on Research In Computer Security*, ESORICS '07, pages 343–358, Berlin, Heidelberg, 2007. Springer-Verlag.

[19] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007.

[20] M. Harrison. Heatmap of Toronto traffic signals using rgooglemaps. `www.r-bloggers.com/heatmap-of-toronto-traffic-signals-using-rgooglemaps/`, 2014. [Online; accessed 25-June-2014].

[21] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[22] D. V. Klein. Foiling the cracker: A survey of, and improvements to, password security. *Proceedings of the 2nd USENIX Security Workshop*, pages 5–14, 1990.

[23] C. R. Mehta and N. R. Patel. A network algorithm for performing Fisher's exact test in r x c contingency tables. *Journal of the American Statistical Association*, 78(382):427–434, 1983.

[24] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[25] J. G. W. Raaijmakers and R. M. Shiffrin. Models for recall and recognition. *Annual Review of Psychology*, 43(1):205–234, 1992. PMID: 1539943.

[26] A. Salehi-Abari, J. Thorpe, and P. C. v. Oorschot. On purely automated attacks and click-based graphical passwords. In *Proceedings of the 2008 Annual Computer Security Applications Conference*, ACSAC '08, pages 111–120, Washington, DC, USA, 2008. IEEE Computer Society.

[27] E. Stobert and R. Biddle. Memory retrieval and graphical passwords. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 1–14, New York, NY, USA, 2013. ACM.

[28] E. Tulving and Z. Pearlstone. Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, 5(4):381–391, 1966.

[29] P. C. van Oorschot and J. Thorpe. Exploiting predictability in click-based graphical passwords. *J. Comput. Secur.*, 19(4):669–702, Dec. 2011.

[30] C. Varenhorst and L. Rudolph. Passdoodles: A lightweight authentication method. 2004.

[31] E. U. Weber, A.-R. Blais, and N. E. Betz. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4):263–290, 2002.

[32] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. Authentication using graphical passwords: Basic results. In *Human-Computer Interaction International*, HCII 2005, 2005.