Gathering Realistic Authentication Performance Data Through Field Trials

Adam Beautement
University College London
Department of Computer Science
Malet Place, London. WC1E 6BT
+44 20 7679 7214
a.beautement@cs.ucl.ac.uk

M. Angela Sasse
University College London
Department of Computer Science
Malet Place, London. WC1E 6BT
+44 20 7679 7214
a.sasse@cs.ucl.ac.uk

ABSTRACT
Most evaluations of novel authentication mechanisms have been conducted under laboratory conditions. We argue that the results of short-term usage under laboratory conditions do not predict user performance "in the wild", because there is insufficient time between enrolment and testing, the number of authentications is low, and authentication is presented as a primary task, rather then the secondary task as it is "in the wild". User generated reports of performance on the other hand provide subjective data, so reports on frequency of use, time intervals, and success or failure of authentication are subject to the vagaries of users' memories. Studies on authentication that provide objective performance data under real-world conditions are rare. In this paper, we present our experiences with a study method that tries to control frequency and timing of authentication, and collects reliable performance data, while maintaining ecological validity of the authentication context at the same time. We describe the development of an authentication server called APET, which allows us to prompt users enrolled in trial cohorts to authenticate at controlled intervals, and report our initial experiences with trials. We conclude by discussing remaining challenges in obtaining reliable performance data through a field trial method such as this one.

Categories and Subject Descriptors
Computing, Security

General Terms
Security, Usability

Keywords
Authentication, Evaluation, Passwords

1. INTRODUCTION
Over the past 10 years, research into usable security mechanisms has increased significantly, and much of this research has focused on authentication mechanisms. Adams & Sasse [1] collected user reports on the impact of the number and complexity of passwords through a web survey and interviews. Zvrian & Haga [21] investigated the user performance with cognitive passwords in the lab. Since then, there has been a steady increase in studies investigating the performance of password schemes and alternative authentication mechanisms by themselves or in a comparative way.

User performance with novel graphical authentication mechanisms has been a particular focus of interest:

- Dhamija & Perrig [8] investigated user performance with passwords and *Déjà Vu*, where users pick their choice of computer-generated images. They found that authentication performance with Déjà Vu was better than passwords. Renaud [17] extended this work with a study that found that hand drawn 'doodles' are superior to computer selected images and personal photos are unsuitable for use in a security setting.
- Perfomance of *Passfaces* – a system where users pick "their" face from 4 panel of 9 faces each - has been the subject of many studies, including Brostoff & Sasse [3], Monrose & Reiter [16] and Everitt et al [11]. [3] found that user reports of Passfaces were positive, but login frequency was significantly lower than with passwords, because participants felt that the longer login time meant it was not worth to login in for brief sessions . [16] found systematic biases in the selection of images, which means the Passfaces selected were vulnerable to guessing attacks. [11] found that the introduction of a second Passfaces login (comparable to having a second password) caused memory interference, leading to a significantly increased number of failed logins.
- The Passpoints system – where users click on a series of points within an image - was developed and tested by Wiedenbeck at al. [20]. They reported that – even though authentication performance was better than passwords, the time required to enroll and authenticate took longer. Chiasson et al. [4] also examined click-based graphical password systems finding that field performance was worse than lab trials had suggested (although they concluded it was still adequate in terms of usability. They also found that image selection

affects performance and graphical passwords suffer from interference affects.

- De Angeli et al. [7] compared user performance of two graphical authentication mechanisms to replace PINs. Neither authentication time nor recall performance of the graphical methods tested were better than PINs, but the authors attribute this to poor design of the user interaction, and suggested that better design graphical authentication would have potential.
- Dunphy & Yan [10] developed and tested BDAS, an image-based version of of the drawmetric system Draw-a-secret. They found that BDAS produced significantly more secure drawmetric passwords, with no decline in authentication performance.
- Tao & Adams [19] adapted the game of Go to create Pass-Go. In an extensive study they found that by using intersections rather than cells the password space was increased without compromising usability.
- Chiasson et al. [5] have conducted a study on Persuasive Cued Click Points (PCCP), an image-based authentication system where users click on points in a series of images. They found that recall performance was better for PCCP than for passwords, and the login speed with PCCP not significantly longer - and this was the first time that a graphical authentication mechanism achieved this.
- Davis et al. [6] examined graphical passwords based on the Passface system and found that user choice leads to predictable passwords with such low entropy that that system is in their view insecure.

Whilst these studies did produce valuable insights into user interaction on with authentication mechanisms, performance results are anything but conclusive. Having examined the way in which the studies were conducted, we argue that the results are not a valid predictor of user performance with such mechanisms "in the wild", as Dourish & Grinter [9] put it.

With the exception of [3], [4], [6], [7], [17] and [19], all of the studies above were conducted exclusively under laboratory conditions. The specific reasons why we feel that performance results have to be treated with caution are as follows:

1. User performance is usually tested only once after successful enrolment; often quite often, within the same session as enrolment. This means the time-span between enrolment and authentication is rather short.
2. In [10], participants were re-tested after one week, and in [8] and [18], after two weeks. Testing authentication performance after a longer interval provides a better insight into memorability for infrequently used mechanisms, but does not provide much insight into authentication performance in regular use. The criteria for what is usable authentication are different for frequently and infrequently used credentials. With frequent authentication (once a day or more), recall of the credential becomes automatic for most users, and fast execution becomes a priority. Execution times largely depend on the number of interaction steps, the response time of the system, and whether correct execution of the recalled credential is difficult (e.g. typing errors; note that execution errors are generally lower with frequent use, but some credentials present execution problems for some users even with frequent use .With infrequent authentication (once a week or less), most users have difficulties recalling the credential correctly, so fast execution is less important for performance than the ability to recall the credential correctly.
3. Only in [11], a subset of users performed more frequent authentication - again over 2 weeks. Arguably, this is a sufficient period to test performance with frequent authentication; but at the same time, it is not long enough to provide meaningful predictions of performance with infrequent usage.
4. In all of the above studies, authentication was the primary task; in real-world interactions, authentication is a secondary task performed in the context of a production task [18]. The production goal is the focus of user behavior, and they are interrupted on their path to the goal by the - secondary - authentication task. For a valid test of performance (and, for that matter, user satisfaction), authentication be tested in the context of a primary task..
5. A final point is that the numbers of participants in these trials is generally low – typically 40 or less – and in many cases, they are students. Small samples sizes and over-reliance on students as participants means that the results cannot be generalized to performance of other users.

[4] Also concluded that relying on solely on lab studies can be problematic after comparing results from lab and field studies. In [3] authentication frequency dropped significantly in the Passfaces cohort because authentication took too long for the primary task – answering multiple-choice questions for course credits, which students typically did in 5 minute sessions, [3] produced a large set of objective performance data under conditions that replicated usage "in the wild" very faithfully, and [13] logged all authentications to online services over 3 months. Tao & Adams [19] also gathered data over a similar time period to analyse their 'Pass-Go' authentication system.

The data gained in this way are valuable sets of objective performance dates, but obtaining ethical clearance for such a study requires significant effort and safeguards.

Information about longer-term performance and user experience can, of course, be obtained from user reports, e.g. the password diary studies used in [13]. However, the data obtained in this way are subjective and affected by memory effects; without some objective data for validation, they cannot be taken as a reliable basis for predicting user performance.

Our aim was to devise a data gathering approach that allows recruit large groups of participants for meaningful field trials of authentication performance. The studies must yield objective data on authentication performance, as well as subjective data on user experience with the authentication mechanism. It must support trials in which a direct comparison of user performance with different authentication mechanisms can be made, and let researchers to control the frequency of authentication. Finally, authentication should be performed as a secondary task, in the context of a meaningful production task. This was the starting point for the development of the Authentication Performance Evaluation Tool (APET).

2. THE APET SYSTEM

APET is a web-based system that allows authentication experiments to be set up and managed remotely. Participants can enroll and take part in authentication trials over any Internet connection. The tool is split into two major components, the core system and the authentication plugins.

2.1 The Core

At the heart of APET is an experiment management system tied to a database of participants. A researcher can log into the management side of the software and set up an experiment. At this stage they can specify its duration, the authentication mechanism involved in the experiment (from the current set of plugins), the attributes they wish to log (from a selection specified by the authentication plugin, see below) and the participant groups they wish to be associated with the experiment. New participants can be added at this stage, by submitting their email addresses through a CSV file. Alternatively, the database of previous participants can be searched and filtered by a variety of criteria - such as age and gender - or by more experiment-specific criteria - such as which studies of authentication mechanisms they have previously participated in. Multiple groups can be assigned to each experiment. This allows different conditions to be applied to the same mechanism, without having to create large numbers of experiments.

Once an experiment is active, the researcher can email the participants of their experiments. Drop-down menus allow the experiment and participant group to be selected and an (editable) message is then sent to the specified group containing a link to the APET system. It is via this link that the participants themselves interact with APET. Using any web browser from any location, they can follow the link to a page that will ask them for their email address. Once they have confirmed their address (and thus that they are taking part in the correct experiment) they will be shown a screen containing whichever authentication mechanism is being tested. Their performance is then logged to a data file that the experimenter can download at the end of the experiment. The participant's interaction with APET is largely controlled by the authentication plugin.

2.2 Authentication Plugin

Each authentication mechanism used in APET needs to have its own plugin. This allows the system to be continually expanded as new technologies are developed without necessarily rewriting the core code. The plugin controls what the participants see when the follow the link emailed to them. For example the password plugin can be configured to allow or disallow password resets, reminders or any other feature the experimenter wishes. Additionally the plugin specifies which logging options the experimenter has. The logging options can be added or removed during the set up phase. The function of the authentication mechanism cannot be modified at this time and so to change the functionality of the authentication procedure a new plugin would need to be created. Typical logging options are such attributes as the number of attempts the participant took to successfully authenticate, whether they requested a reset or reminder, and the time and date of the authentication attempt. Any attribute can be logged assuming it has been coded into the plugin.

2.3 Primary Task Scenarios

The final component of APET is the ability to host a number of different web services that provide the primary task for which participants log into the system. APET can function as an authentication service for live web services, or direct participants to pages that support other experimental scenarios. This allows us to create different primary task and authentication scenarios. The first primary task scenario we have implemented is *Barterworld* – an online marketplace where members provide services to other members for credits, which can be used to buy other members services. Participants receive emails from another member with a confirmation code to claim services they have delivered – e.g. "Member 151 has confirmed you have completed 2 hours of gardening – please log in and enter claimcode 70933 within the next 12 hours to have the credit added to your page." Participants are paid according to the numbers of hours they have logged on the website by the end of the trial. If participants fail to authenticate within 3 attempts, they receive a reminder of their credential, but have 25% of an hour credit subtracted as a "fee" for the re-set.

3. CURRENT EXPERIMENTS

APET has primarily been used to conduct collect data on password performance over time. This is one of the main strengths of the system. Without the need to bring participants into the laboratory (or for experimenters to travel to meet the participants) extended trials over time can be undertaken with relative ease. In this case the experiment ran over two periods of two weeks. The participants were asked to enroll with a memorable password that confirmed to a simple policy. They were instructed not to write their password down or use any other memory aids and to not register a password they used for any other system. This was partly for their own security but

largely to attempt to avoid the influence of prior experience with the password on the experiment.

For the first two week period the participants were sent an email once a day every working day (Monday to Friday) asking them to login to APET. The timing of the email varied through the two weeks. Email distribution is under the manual control of the experimenter (as opposed to being an automated part of the APET system) so the precise timing of the emails can be managed as needed. The second two week block took place 6 weeks after the first and this time the participants were sent emails asking them to login three times a week (Monday, Wednesday and Friday). In each case the number of attempts taken to successfully log in and the number of password resets requested were logged to form the data set for the experiment.
The entire experiment including recruitment and enrolment was run remotely over the internet. At no time were the experimenters face to face with the participants and the participants were able to take part from their own home or workplace as suited them. This meant that the experimental tasks feel within their normal working practices and would have taken on a similar priority and level of intrusion as that of any other login procedure during their day. Certainly the intrusion and disruption level was substantially lower and more realistic than that a trip to a laboratory would entail.

4. FUTURE EXPERIMENTS

As well as including the *Barterworld* primary task future experiments will focus on frequency of authentication and its impact on performance. Subject cohorts will be asked to login with varying frequencies varying from several times a week through to once every 6-8 weeks. The aim of these experiments is to search for the point at which password resets become the norm rather than recall and entry; this also being the usage frequency at which passwords become inappropriate as an authentication method. Additionally we are planning to use the APET system to investigate the effect of interference (both inter- and intra- authentication mechanism) on performance. As well as the password plugin we will be using plugins for a PIN system and the Gridsure [2].

5. PROBLEMS AND LIMITATIONS

The APET system has proved a useful vehicle for collecting objective data on user performance with different authentication mechanisms “in the wild” but there are some aspects that we cannot control. The more realistic and rewarding the primary task services or scenarios are, the more likely it is that participants will want to make sure that they do not fail. Despite being very clear with our experimental instructions that participants should not write down passwords or other authentication credentials, we cannot guarantee that participants will not do this, or undertake other behaviours that would affect their performance in the trial. Additionally, the reliance on plugins reduces the flexibility of experimental design without expending time recoding them. However when weighed against the chance to test the performance of authentication systems in a far more natural environment we believe that the benefits overwhelmingly outweigh the costs.

6. REFERENCES

[1] A. Adams & M. A. Sasse (1999): Users Are Not The Enemy: Why users compromise security mechanisms and how to take remedial measures. *Communications of the ACM*, 42 (12), pp. 40-46 December 1999.

[2] S. Brostoff, P. Inglesant & M. A. Sasse (2010) (forthcoming) Evaluating the usability and security of a graphical one-time PIN system. To appear in *People and Computers XXIV. Proceedings of HCI 2010*.

[3] S. Brostoff, S. & M. A. Sasse, A. (2000). Are Passfaces more usable than passwords? A field trial investigation. In McDonald S. et al (Eds) ‘People and Computers XIV - Usability or Else’, *Proceedings of HCI 2000*, Sunderland, UK, pp 405-424, Springer.

[4] S. Chiasson, R. Biddle, and P. C. van Oorschot (2007): A second look at the usability of click-based graphical passwords.*3rd ACM Symposium on Usable Privacy and Security (SOUPS)*, July 2007.

[5] S. Chiasson, A. Forget, R. Biddle & P. C. Van Oorschot (2008): Influencing Users Towards Better Passwords: Persuasive Cued Click-Points. *Proceedings of HCI 2008*.

[6] D. Davis, F. Monrose, and M. Reiter, "On user choice in graphical password schemes," in 13th USENIX Security Symposium, 2004.

[7] A. De Angeli, L. Coventry, C. Johnson & K. Reynaud (2005): Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, 63 (2005) 128–152.

[8] R. Dhamija & A. Perrig, Déjà Vu: a user study using images for authentication, *Proceedings of the 9th conference on USENIX Security Symposium*, p.4-4, August 14-17, 2000, Denver, Colorado

[9] P. Dourish & R. E. Grinter (2004): Security in the Wild. *Personal Ubiquitous Computing 2004*(8) 391-401.

[10] P. Dunphy & J. Yan (2007): Do Background Images Improve “Draw a Secret” Graphical Passwords? *Proceedings of CCS 2007*.

[11] K. M. Everitt, T. Bragin, J. Fogarty, & T. Kohno (2009): A Comprehensive Study of Frequency, Interference, and Training of Multiple Graphical Passwords. *Proceedings of CHI 2009*. pp. 889-898.

[12] D. Florencio & C. Herley (2007): A Large-Scale Study of Web Password Habits. *Proceedings of WWW 2007.*

[13] P. Inglesant & M. A. Sasse: The true cost of unusable password policies: password use in the wild. *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010)*, Atlanta, GA, USA, April 2010

[14] I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A.D. Rubin. The Design and Analysis of Graphical Passwords, *Proceedings of the USENIX Security Symposium, 1999.*

[15] M. Just & D. Aspinall (2009). Personal choice and challenge questions: a security and usability assessment. *Proceedings of SOUPS 2009*.

[16] F. Monrose & M. Reiter (2005): Graphical Passwords. In L. Cranor & S. Garfinkel [Eds.]: Usability and Security. O'Reilly 2005.

[17] K. Renaud (2009): On user involvement in production of images used in visual authentication. *Journal of Visual Languages and Computing*, vol. 20, no. 1, pp. 1-15, February 2009.

[18] M. A. Sasse, S.Brostoff, & D. Weirich (2001): Transforming the "weakest link": a human-computer interaction approach to usable and effective security. *BT Technology Journal*, Vol 19 (3) July 2001, pp. 122-131.

[19] H. Tao and C. Adams (2008): Pass-Go: A proposal to improve the usability of graphical passwords. *International Journal of Network Security*,vol. 7, no. 2, pp. 273-292, 2008.

[20] S. Wiedenbeck, J. Waters, J.C. Birget, A. Brodskiy & N. Memon (2005): PassPoints: Design and longitudinal evaluation of a graphical password system", International J. of Human-Computer Studies (Special Issue on HCI Research in Privacy and Security), 63 (2005) 102-127.

[21] M. Zviran & W. J. Haga (1999), "Password Security: An Empirical Study", *Journal of Management Information Systems* 15 (4): 161–185.