A Large-Scale Study of Web Password Habits

Dinei Florêncio and Cormac Herley
Microsoft Research
One Microsoft Way
Redmond, WA

ABSTRACT

We report the results of a large scale study of password use and password re-use habits. The study involved half a million users over a three month period. A client component on users' machines recorded a variety of password strength, usage and frequency metrics. This allows us to measure or estimate such quantities as the average number of passwords and average number of accounts each user has, how many passwords she types per day, how often passwords are shared among sites, and how often they are forgotten. We get extremely detailed data on password strength, the types and lengths of passwords chosen, and how they vary by site. The data is the first large scale study of its kind, and yields numerous other insights into the rôle the passwords play in users' online experience.

Categories and Subject Descriptors

K.6.5 [**Management Of Computing And Information Systems**]: Security and Protection—*Authentication*

General Terms

Security

Keywords

password, authentication, measurements

1. INTRODUCTION

Passwords play a large part of the typical web user's experience. The are the near universal means for gaining access to accounts of all kinds. Email, banks, portals, dating and social networking sites all require passwords. So important are they that HTML has a special form field to allow for the special treatment they require, and an important rôle of SSL is protecting the secrecy of passwords from observers of the connection.

Alternative to passwords certainly exist. Hardware authentication, *e.g.* [1], is sometimes used for access to corporate networks. However, this requires an issuing authority and seems to be limited to environments that justify the cost, such as in the employer-employee relationship. Challenge response authentication has the advantage that observing a single successful sign in does not allow an attacker to gain the secret. However, challenge response systems are generally regarded as being more time consuming than password entry and do not seem widely deployed. One time passwords have also not seen broad acceptance. The difficulty for users of remembering many passwords is obvious. Various Password Management systems offer to assist users by having a single sign-on using a master password [7, 13]. Again, the use of these systems does not appear widespread. For a majority of users, it appears that their growing herd of password accounts is maintained using a small collection of passwords. For a user with, *e.g.* 30 password accounts, the problem becomes not remembering 30 distinct passwords, but rather remembering which of 5 or 6 passwords was used. This appears to be done using a combination of memory, pieces of paper, trial and error (trying each of the passwords in turn), and password resets.

Since passwords protect accounts with valuable assets they have increasingly been subjected to harvesting attacks. Phishing attacks, where a victim is lured into submitting her password to a malicious site masquerading as a trusted institution have increased enormously in the last few years [11]. Incidences of keylogging malware, which record keystrokes on a PC have also been rising rapidly. Unlike brute force attacks on passwords, both phishing and keylogging harvest strong passwords as easily as weak ones. Thus the nature of the risk surrounding password authentication has altered greatly. The longstanding problem of users choosing passwords that are too easily brute forced [12, 6, 3] has been joined by the new problem of users unwittingly revealing their passwords in the clear.

The convenience of web access to accounts is extremely compelling, and thus the rôle they play in the average web users life seems likely to increase. However we find that firm data on users' actual password habits is hard to come by. It is conventional wisdom that users choose weak passwords, frequently re-use passwords across multiple sites, and often forget them. In this paper we report on a large scale study of web users habits where we measured and report these and other patterns for the first time. We obtained data from over half a million users over a period of three months. This is more than 100 times more participants than any previous study we are aware of.

Among our interesting findings is how large a rôle web passwords play in users lives. The average user has 6.5 passwords, each of which is shared across 3.9 different sites. Each user has about 25 accounts that require passwords, and types an average of 8 passwords per day. That users choose weak passwords has been known informally for some time;

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

we are able to measure exactly how weak. Users choose passwords with an average bitstrength 40.54 bits. The overwhelming majority of users choose passwords that contain lower case letters only (*i.e.* no uppercase, digits, or special characters) unless forced to do otherwise. We were able to measure that 0.4% of users type passwords (on an annualized basis) at verified phishing sites, and at least 0.2% of users actively maintain their own router. Finally users forget passwords a lot: we estimate that at least 1.5% of Yahoo users forget their passwords each month.

In the next section we cover details of the client and the data gathered. In Section 3 we present our results, broken into logical sections. In Section 4 we discuss related work.

2. EXPERIMENTAL METHOD

Our client software shipped as a component of one skew of Windows Live Toolbar. Not all toolbar users received the component. The component was optional, and users were presented with an opt-in agreement. The toolbar was first available for download on the Microsoft web on 7/24/2006, and a total of 544960 clients received, opted in and activated by 10/1/2006.

2.1 Client Implementation

The client consists of a module within the toolbar that monitors and records Password Re-use Events (PRE's). It contains the following main components.

HTML password locator: this component scans the document object model in search of filled-out password fields, and extracts the passwords. The first task merely involves searching the HTML for fields declared

`inputtype="password"`

and extracting the `value` field. This search is initiated every time the browser `BeforeNavigate2` event occurs. Thus we find completed password fields before they are sent to the server. Once the password is found it is hashed and added to the Protected Password List (PPL).

Protected Password List: This list contains the password hash, the full URL of the receiving server, the bitstrength of the password, the current time, and minutes since both the first and last time (if any) that password was sent to that server. All of the information in the PPL is stored using the Data Protection API (DPAPI) provided by Windows [14] (the same API that is used to protect passwords that Windows stores). Thus any passwords that the HTML Password Locator finds are stored at least as securely as passwords that a user elects to save. Passwords with bitstrength < 20 bits generate no entry in the PPL.

Realtime password locator: this component maintains a 16 character FIFO that stores the last 16 keys typed while the browser had focus. Call this string $f_0, f_1, \cdots, f_{15}$. At every keystroke, while the browser has focus, the realtime password locator first checks whether the string $f_0 f_1 \cdots f_6$ matches *any* of the already stored passwords in the PPL. It then checks whether $f_0 f_1 \cdots f_7$ matches, and so on until finally it checks whether $f_0 f_1 \cdots f_{15}$ matches any of the hashes. Thus it performs a hash check a maximum of ten times the number of entries in the PPL. When a match occurs (*i.e.* when a typed sequence in the FIFO matches a password in the PPL) it checks whether the URL of the current server is among the URLs previously associated with the password. If not a Password Re-use Event (PRE) report is sent to the server.

PRE Report: this contains:

- the current (primary) URL
- all of the URLs previously associated with the password (secondary URLs)
- time since last login at each URL previously associated with the password
- time since first login at each URL previously associated with the password
- the password strength
- number of entries in the PPL, and number of PREs filed by client
- number of unique passwords used by this client
- the age of the client.

The format of the report is

$$[U_p, \{sU_0, sU_1, \cdots, sU_{N-1}\}, \{t_0, t_1, \cdots, t_{N-1}\},$$
$$\{\tau_0, \tau_1, \cdots, \tau_{N-1}\}, \text{PwdStr, PPLSz, NPREs, NPwds, CAge}].$$

Suppose for example that a user has a password that is used at PayPal, Yahoo, eBay and YouTube. The first time the password is typed (say at eBay) it will be added to the PPL, and no report made to the server. This password can then be typed at eBay over and over and will generate no PRE reports and no additions to the PPL. The next time it is typed at a site other than eBay (say Yahoo) a PRE report will be sent listing `www.yahoo.com` as the primary URL U_p and `www.ebay.com/login` as the secondary sU_0. Now typing it at PayPal will cause a PRE report listing `www.paypal.com` as U_p and `www.yahoo.com` and `www.ebay.com/login` as sU_0, sU_1. Observe that neither the password, nor its hash are sent in the report. There is no personally identifying information in the report.

Note: The reason that we perform the realtime password check is that we wish to be sure that we catch every Password Re-use Event. If a user enters a password at URL A it will be entered in the PPL by the HTML password locator. However it is possible that the password could be typed at another site that does not use a HTML password field. We wish to capture and report any case where a previously used password is typed at another site.

2.1.1 Privacy

A number of measures were taken to protect the privacy of those who opted in. No Personally Identifying Information was gathered from the clients. Neither passwords, nor usernames, nor their hashes were sent to the server. IP addresses from which reports were received were not stored at the server. In addition the time at which PRE reports were received was timestamped at the server with granularity 10 minutes to make identifying users by login times difficult. Finally, the recorded password strengths were quantized to eliminate the possibility that a unique password strength might identify a user. A privacy audit was performed and published [10]. This confirmed, among other things that no personally identifying information was transmitted, and that URL information sent could not be used to identify a user's personal information (*e.g.* if a userid or name appeared as part of a URL). None of the aggregated statistics we present here involve data from fewer than 250 users.

2.1.2 Server

The server records each received report and stores with a per-PRE report ID and a timestamp. It does not record any

location information such as IP address that might allow for identification of the user or his/her location.

2.2 The Data

Downloads began almost as soon as the client became available on the web, and data from the clients began to flow shortly thereafter.

2.2.1 Canonicalization of URLs

The client reports the full primary and secondary URL. Some sites use a unique string per login event, so that the URL will never appear the same. In T-SQL we canonicalize by setting `URLc` to

```
substring(URL+'/',1,charindex('/',URL+'/',9))
```

As an example the URLs `www.foo.com/bar/` and `www.foo.com` would be mapped to `www.foo.com/`.

2.2.2 Cleaning the data

As with any large scale study unforeseen issues caused the data to be noisier than anticipated. The client reports a PRE whenever a password is typed at a site other than the first URL with which it is associated. For example `www.bigbank.com/foo1` and `www.bigbank.com/foo2` will count as different sites. We extracted all reports where the authority is determined to be the same.

As detailed in Section 2.1 our component's strategy for adding to the PPL was to check for a non-empty password field on the page when the `BeforeNavigate2` event fired. This certainly occurs every time a user types into a password field and submits. However, there are a significant number of sites that have a login form pre-populated with default values. The gaming site `www.iwin.com`, for example, has a login form pre-populated with the values "username" and "password" and this form is on many of the pages on the site. A single visit to the site causes a new list to be formed in the PPL. As the user visits other pages on the site these generate PRE reports against the original page. Since the user has not actually typed a password we view these entries as spurious (and indeed they tell us nothing about the user's password habits). By removing all records where the primary and secondary URLs derive from the same authority we eliminate the effect of such lists. Note that this operation merely groups reports from the same client; it does not allow us to identify the client.

2.2.3 Generating a per list identifier

The PRE reports from clients do not contain an identifier that would allow us to group reports from a single client. Suppose a client has k passwords, $p_0, p_1, \cdots, p_{k-1}$ and p_i is used at n_i sites. Thus we would receive $\sum_i (n_i - 1)$ reports from this client, $n_i - 1$ for each of the passwords. Since we do not have a identifier that uniquely identifies the client there is no exact method to group together all of the reports generated by a single client.

It is possible, however, to group the $(n_i - 1)$ reports that come from the same client using the same password at n_i different sites. Grouping $PwdStr, CAge, \{sU_0, sU_1\}, \{\tau_0, \tau_1\}$ (*i.e.* the strength, client age, first two URLs visited using this password, and the visit times) suffices to uniquely identify all $(n_i - 1)$ of the PRE reports for password p_i most of the time. In only a tiny minority of cases do two distinct users instal the toolbar at the same time, have a same strength password, and visit the same first two URLs at about the same time.

2.3 Limitations, Caveats and Sources of Bias

We should emphasize that data gathering was not the main goal of the component: there are many things that we would do differently if we were to design a component purely for measurement. There are certain limitations of the study and the data. Users may type passwords from more than one machine, and thus we will miss a potentially large fraction of their password behavior. Conversely, more than one person might be signing in to various online accounts using the same Windows session, and thus we would get higher password counts in the PPL.

The client generates a PRE whenever a password previously associated with another site is typed. If a user chooses a password that is a common word a PRE will be generated every time they type that word. Since, passwords of strength < 20 bits are not entered in the PPL, we view this as a minor source of error.

Since the client was included as part of a toolbar download from Microsoft it is likely that there is a bias toward the sites maintained by Microsoft and away from those of its competitors. For example users of Yahoo toolbar might be expected to be more likely to have accounts at Yahoo sites than MSN or Google sites. Thus the Table 1, which contains relative frequencies, should be interpreted as having some bias. Finally, there is selection bias: we have data only from users who downloaded the toolbar. These users can be expected to be far more active than the general web using population.

3. RESULTS

We now present some of the findings from the data, beginning with some basic quantities about the nature of clients.

3.1 Number and Nature of Clients

Figure 1 shows the cumulative number of clients as a function of time. There were approximately 6400 activations per day; and we had 544960 clients installed after 85 days. This is the client base and time period over which we did the bulk of our analysis. Except when indicated otherwise it can be assumed that we are referring to this set.

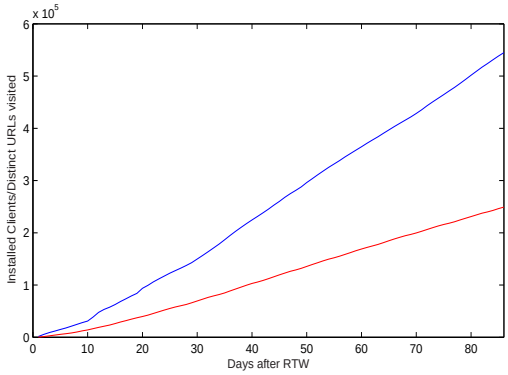

Figure 1: Number of installed clients as a function of time (upper curve). Number of distinct canonicalized password URLs visited by the whole population as a function of time (lower curve).

3.1.1 Number of Passwords Per Client

To estimate the number of passwords per client we use the NPwds field in the PRE report. Since there is no overall unique identifier per client we have to be careful to ensure that we do not count clients more than once in the calculations. While we do not have a unique identifier we do know that each client as it files PRE reports will report a given value for NPREs only once. To get an idea of how many passwords a client has, and how it evolves with the time since instal we calculated the average NPwds used by a client, as a function of the age of the client in days (at the time of the report). We performed this calculation at NPREs = 10; the results are shown in Figure 2. As can be seen clients seem to add passwords rapidly the first few days, but this levels off after a month or so. The average client appears to have about seven distinct passwords that are actively used, and five of them have been used within three days of installing the client.

There can be some over counting here. Recall from Section 2.1, that the HTML password locator adds an entry to the PPL when a non-empty password field is found on the page when the browser `BeforeNavigate2` event fires. There is no verification that a successful login occurs. Thus a user who mistypes a password and submits it would generate a spurious entry in the PPL. However, any typing mistake that is not submitted would not (the `BeforeNavigate2` event fires only when the user clicks or submits).

3.1.2 Number of sites per password *vs.* Age

To calculate the number of times the average user re-uses a password we count the number of secondary URLs accompanying the primary URL in each PRE report. Since some sites show up as a different URL on each visit we canonicalize URLs (as in Section 2.2.1) before counting. Thus a PRE report that contains `https://www.paypal.com` as the primary URL and `http://www.yahoo.com/foo1` and `http://www.yahoo.com/foo2` as secondaries would be counted as being re-used at one site rather than two.

Again, we must avoid over counting. We calculate the averages over all available lists, using the list identifier calculated in Section 2.2.3. The resulting number of sites per password is shown in Figure 3. As can be seen the average password appears to be eventually used at just under 6 distinct login sites, but it takes on the order of two months to reach that point. The first four or so sites that share a password are all visited in the first week. Note that passwords used at only one site generate no PRE reports. Again there will be some over-counting caused by users forgetting passwords and trying several passwords before logging in.

3.1.3 Number of sites per password *vs.* Strength

We also measure the number of sites that share a password on average. This is shown in Figure 4, where we compute the fraction of passwords (across all clients) that are shared by two sites, three sites *etc.*. Recall that unique passwords do not show up. This is because PRE reports are sent to the server only when a password re-use event has occurred. A client who uses a password once will have that entered in the PPL, but if it is never re-used the server will never receive a report on it. Since it takes some time for the client's PPL to fill, we have used data only from reports filed by clients at least 30 days old at the time of the report.

The average number of sites sharing a password (averaged across all passwords on which we received PRE reports) is 5.67; this is the mean of the histogram shown in Figure 3. We also show the histogram for weak passwords, those with bitstrength < 30 bits, and strong passwords, those with bitstrengths > 60 bits. Observe that the strong passwords are used at fewer sites on average (the mean is 4.48 sites). Weak passwords are used at more sites on average (the mean is at 6.06 sites). This accords well with our expectation that users employ weak passwords at multiple sites when the password rules are lax. Sites that impose password strength rules make it harder to share, and users probably have fewer such accounts. There were 24k, 118k, 7.2k passwords used for the weak, average and strong password distributions respectively (all from clients more than 30 days old).

There is also a possibility of over counting in our numbers. Occasionally a user with an account at, say, Yahoo will forget which of his k passwords is used for that account. Suppose his Yahoo password is p_0, but he initially tries p_2 and p_1 before logging in correctly. In using passwords that are already in his PPL the user causes PRE reports to be sent, and for Yahoo to be added to the list of sites sharing passwords p_1 an p_2. Thus this site now appears to share 3 of the passwords in the user's PPL. Since we count the number of secondary URLs with the number of sites sharing a password (without verifying that an actual login has occurred) this causes some over counting of the number of sites per password. We examine bitstrength by site in Section 3.3.1. We examine password forgetting in Section 3.3.3.

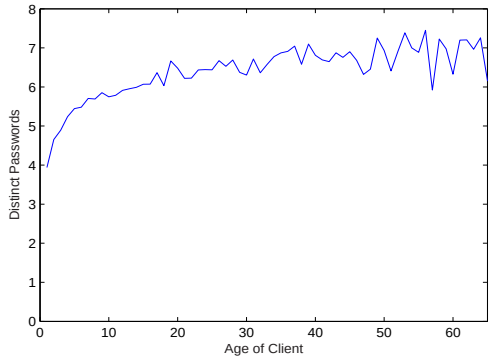

Figure 2: Number of distinct passwords used by a client *vs.* age of client in days. The average client appears to have about 7 passwords that are shared, and 5 of them have been re-used at least once within 3 days of installing the client.

3.1.4 How many passwords does a user type per day?

To estimate this we examined the first 24 hours of the life of a client. Recall, each PRE report carries information about the password being re-used, as well as information about the client. The average number of distinct secondary URLs (canonicalized as in Section 2.2.1) reported by each PRE in the first 24 hours, plus one (for the primary URL) is a good estimate of the number of times that password is used per day. To avoid overcounting we take only the last PRE per list (using the list ID generated in Section 2.2.3). This estimate may be somewhat high, since users may be more active, and more inclined to login at multiple sites the day they install the client. There is also a bias against infrequent internet users. For example a user who goes online only once

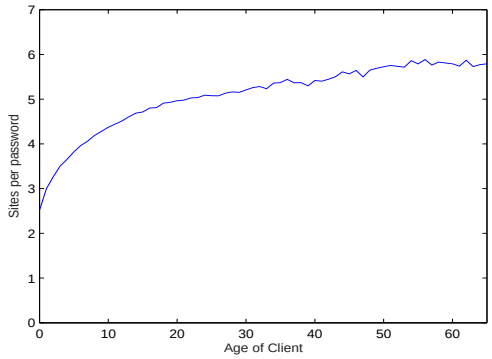

Figure 3: Number of sites per password *vs.* age of client in days. The average password appears to be used at about 6 different sites.

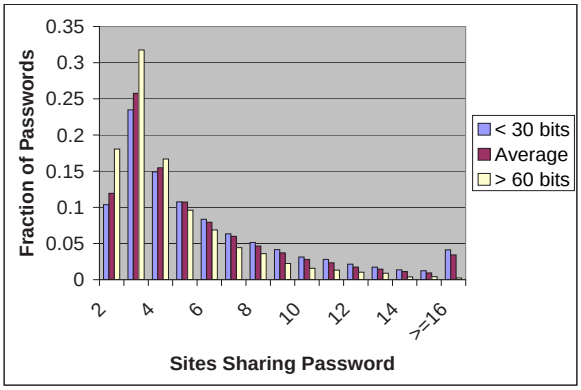

Figure 4: Fraction of the number of sites that share a password for weak passwords (bitstrength < 30 bits), strong passwords (> 60 bits) and the overall average. Observe that weaker passwords tend to be shared at more sites, and stronger ones at fewer.

a week, and checks multiple accounts that day would bias this estimate upwards, since they would have no activity for the rest of the week. Since infrequent users are unlikely to be early adopters of toolbar downloads we view this as a minor source of bias. The number of distinct passwords used by a client is given in the report. Thus averaging we get (for 24 hours)

$$\text{Pwds per Client} \times \text{Sites per pwd} = 2.519 \times 3.2202 \approx 8.11,$$

passwords that a user types per day. There is also a certain measure of under counting in this estimate. A user who checks web email several times a day and retypes the password will have that counted only once in this estimate.

3.1.5 How many password accounts does the average user have?

Passwords per client times sites per password is a good estimate of the number of passwords a user types per day if we measure both those numbers in the first 24 hours, and assume that few passwords are typed more than once per day. If we wait until the client PPL has filled the same product, measured for clients older than 30 days, allows us to estimate the number of password accounts that the average user has:

$$\text{Pwds per Client} \times \text{Sites per pwd} = 6.5 \times 3.9 \approx 25,$$

accounts. We use the estimate of sites per password at 7 days to avoid bias toward passwords that still generate PRE reports after a long period of time. We point out that there may be some over counting caused by typos, and users trying more than one of their passwords (thus boosting the sites per password count).

3.2 Functions of time

3.2.1 Frequency of logins at popular sites

Here we estimate how often users login at different sites. This data is not directly available, since we only generate a report the first time a site is visited. To obtain an estimate, we use a field in the PRE report that indicates how much time has elapsed since last visit to each of the sites in the reuse list. This is generated only when a user types the same password at a different site. If the visits to the sites were independent, the average of the elapsed time would be half the mean time between visits.

The problem with that is that users tend to group their web usage. So, if they go to web sites in random order, data would show that about half the time, they have just visited each other site. To complicate things further, the ordering of visit to sites is probably not random: for example, users may tend to check their e-mail accounts first thing when they login.

So, in trying to obtain an unbiased estimate of how often users login at several web sites, we get a slice in time of the PRE reports, ignore the reports with less than 5 minutes, and use the median of the time elapsed from the remaining reports. Table 1 the estimates for the 15 most popular sites. Worth mentioning is the fact that sites that allow you to keep logged for a long time (e.g., passport) have a reasonably long time between login events. Casual sites, like youtube or paypal have a longer times between login events than, say e-mail sites. Similarly, addictive sites, like social networking sites, or adult sites seem to have a smaller time between logins.

3.2.2 Logins by hour of day, day of week

We evaluated the distribution of login times. Remember logins differ from normal browsing, as they require an account relationship with the hosting site. Figure 5 shows a distribution of a few representative sites across the week. Interestingly, sites like `4kids.tv` have an increased traffic during the weekend, while most other sites present generally higher number of logins during the weekdays. Note how logins correlate with the nature of the site; *e.g.* note the similarity between `Chase.com` and `BankOfAmerica.com`. Figure 6 shows the login distribution along the day. Here the match between the banks is not as perfect, what can be attributed to differences in the time zones of each bank costumers, for example. In fact, by looking at merchants that use different domains in different countries, we can get a feeling for the influence of the time zone. For example, Figure 7 shows the time of the day distribution for 3 different sites of Amazon: `amazon.co.uk`, `amazon.co.jp`, and `amazon.com`.

3.3 Password Strength Analysis

3.3.1 Bitstrength analysis

We next examine the strength of passwords that users

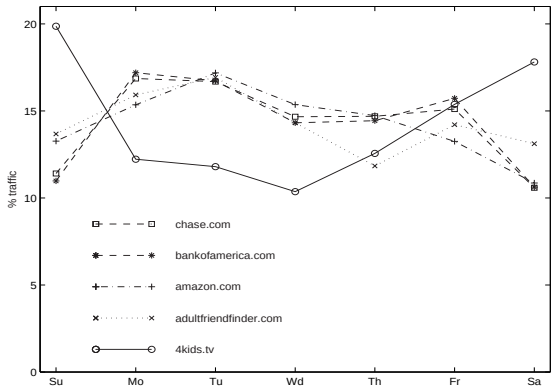

Figure 5: Distribution of logins during the week.

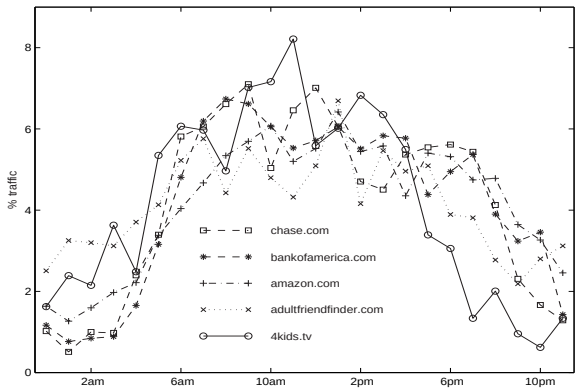

Figure 6: Distribution of logins along the day.

choose and how it varies with site. Recall that each PRE contains a quantized bit strength, PwdStr, for the password being re-used. This was calculated as $\log_2((\text{alph. size})^{len})$. The alphabet size is the sum of the sizes of the different types of characters. These types and sizes are lowercase (26), uppercase (26), digits (10) and special (22). Thus a 9 character password that contains both upper and lower case characters and digits would have bitstrength $\log_2(62^9) \approx 53.59$.

Figure 8 shows the distribution of password strengths for a number of sites. These are the New York Times, PayPal, Fidelity and Microsoft's Outlook Web Access (a web interface that allows Microsoft employees access to email and calendar information). Observe that there is a huge difference between the NY Times passwords (average strength 37.86) where users are merely protecting a newspaper subscription, and Microsoft OWA (average strength 51.36) where employees are forced to choose strong passwords to protect the corporate network.

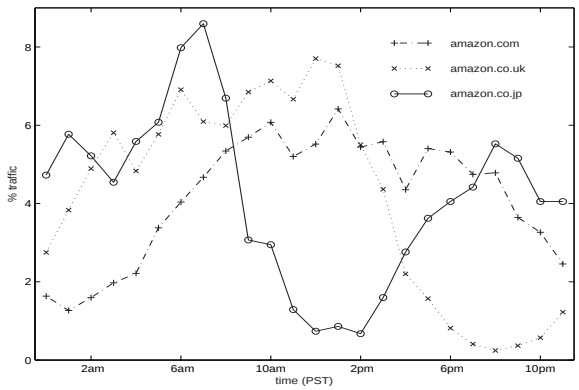

Figure 7: Distribution of logins for amazon at different countries: USA, Japan and UK.

Site	# PRE reports (thousands)	Avg. mins between logins
aol.com	4.5	7565
bebo.com	4.8	6633
ebay.co.uk	5.8	6961
ebay.com	7.8	8043
google.com	21.0	6749
hi5.com	5.2	5558
live.com	120.2	5937
match.com	3.7	9351
msn.com	9.0	11290
myspace.com	36.2	4060
passport.net	10.9	13770
paypal.com	13.6	8447
yahoo.co.jp	8.6	4310
yahoo.com	91.1	4232
youtube.com	5.6	12482

Table 1: The 20 most commonly used login sites order by frequency.

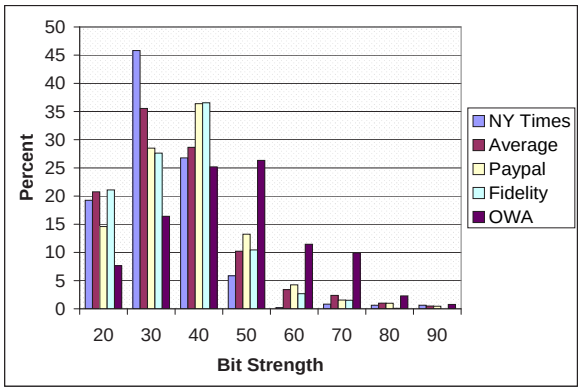

Figure 8: Histogram of password bitstrengths at various sites. The strength appears to be a function of the users' perceived importance of the site. New York Times subscription passwords (average bitstrength 37.86) are noticeable weaker than the average over all sites (average bitstrength 40.54), while PayPal and Fidelity (average bitstrengths 42.04 and 39.92) are stronger. Microsoft OWA, which mandates strong password rules has the highest (average bitstrength 51.36).

3.3.2 *Password Type Analysis*

The password bitstrengths allow us to determine the length of the password and the number of character types it includes. By tabulating all possible password lengths from 7 to 16 (no passwords shorter than length 7 generate entries in the PPL, and the realtime locator FIFO is length 16) and all combinations of the 4 different types we find a unique mapping between the quantized password strength and length, type. This allows us to analyze not merely the strengths of passwords but also the types.

Figure 9 shows the percent of passwords that are of a particular type as a function of length (averaged across all

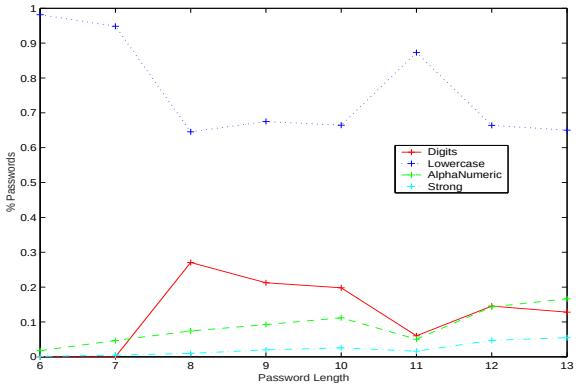

Figure 9: Different types of passwords as a function of length averaged across all sites. Observe that a clear majority of all passwords are lower case letters only. PIN's, or passwords that are purely numeric, account for about 20 % of passwords (note that we did not record numeric passwords of length 7 or less). Alphanumeric passwords, consisting of upper case, lower case and digits constitute small portion. A tiny minority of passwords are strong, in the sense of containing upper, lower, digits and special characters.

login sites). We already know from Section 3.3.1 that users choose weak passwords, but Figure 9 breaks this down even further. We plot, as a function of length, the fraction with lowercase only, lowercase and digits, lowercase, uppercase and digits and all four types. Note the fractions do not precisely sum to one as we have omitted rare combinations for clarity (*e.g.* passwords that contain only digits and special characters but no upper or lower case letters are extremely rare). What is remarkable is that passwords containing only lowercase letters dominate at all lengths. Even as users perceive the need, or are forced, to use stronger passwords, it appears that they use longer lowercase passwords and use uppercase and special characters hardly at all.

Figure 10 repeats the calculation, but restricts to Paypal passwords only. Again the trend is very pronounced: lowercase only passwords dominate overwhelmingly. Even for passwords that are 13 characters long, lowercase-only accounts for 78% of the cases. The situation appears to be changed only when a site forces password policies that use a greater number of types. Figure 11 show the plot for Microsoft Outlook Web Access. Like many corporations, Microsoft forces employees to choose strong passwords and codifies certain passwords strength requirements. In this case passwords that employ both upper and lower case and special characters account for non-negligible fractions of the passwords. This accords well with the finding of [9] where very few users used a special character unless instructed to do so.

3.3.3 *Password Forgetting*

People forget passwords a lot. In the case of Yahoo logins to user accounts generally occur at `https://login.yahoo.com` while forgotten passwords, new registrations etc occur at `https://edit.yahoo.com`. Table 2 shows the number of PRE reports listing various Yahoo edit URLs as the primary URL. If we take the numbers at `change_pw`, `eval_forgot_pw` and `forgot_pwex` as a reasonable approximation to the number of users who forgot passwords we get 2149, out of a total of 50.1k PRE reports against the main login site. This

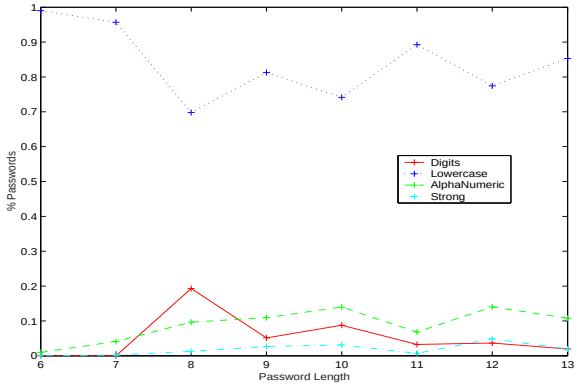

Figure 10: Different types of Paypal Passwords. The most common types of passwords follow the average trend (as shown in Figure 9) quite closely, with the exception that numeric passwords are less common.

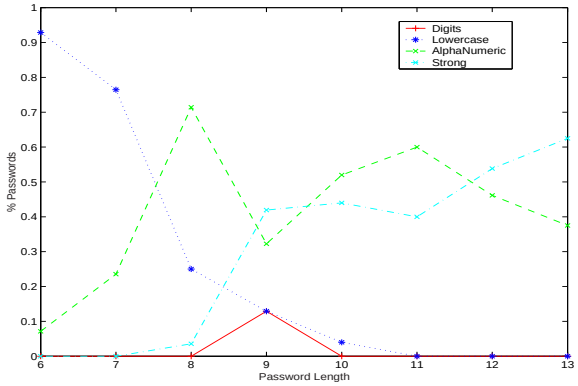

Figure 11: Different types of Microsoft Outlook Web Access Passwords. The most common types of passwords follow the average trend (as shown in Figure 9) quite closely, with the exception that numeric passwords are less common.

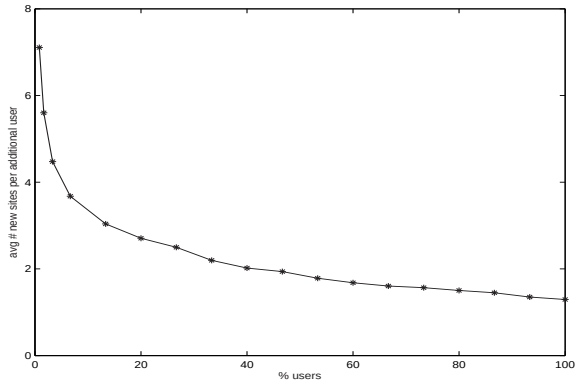

Figure 12: Number of previously unseen password URLs visited by new clients as a function of time. Observe that the number of new sites plateaus relatively quickly.

implies 4.28% of Yahoo users forgot their passwords over a three month period. Also observe that combined visitors to the password change and reset pages (2149) amounted to 60% of the visitors to the new registration page `eval_register`. Observe that all of the instances of users typing passwords to perform administrative functions amounts to 15% of the number of logins (*i.e.* sum of all lines of Table 2 except the first as a percent of the first).

https://login.yahoo.com	50147
https://edit.yahoo.com/config/eval_register	3524
https://edit.yahoo.com/config/change_pw	1888
https://edit.yahoo.com/v/full	668
https://edit.yahoo.com/config/register	546
https://edit.yahoo.com/config/login_verify2	274
https://edit.yahoo.com/config/eval_forgot_pw	218
https://edit.yahoo.com/config/id_check	127
https://edit.yahoo.com/v/recv	123
https://edit.yahoo.com/config/forgot_opt	117
https://edit.yahoo.com/config/delete_user	100
https://edit.yahoo.com/config/login	73
https://edit.yahoo.com/config/forgot_pwex	43
https://edit.yahoo.com/config/child_register	13
https://edit.yahoo.com/config/child_trap	7
https://edit.yahoo.com/config/mail	5
https://edit.yahoo.com	3
https://edit.yahoo.com/config/mail_mev	2
https://edit.yahoo.com/config/parent_verify	2
https://edit.yahoo.com/v/send	2
https://edit.yahoo.com/config/set_profile	1
https://edit.yahoo.com/config/forgot_pw.html	1
https://edit.yahoo.com/config/edit_account	1

Table 2: Users appear to forget passwords and perform other administrative functions a lot. For the example of Yahoo, password change operations occurred 15 % as frequently as sign-in operations.

3.4 Interesting Sites

3.4.1 Phishing Sites

We had access to a list of phishing sites that were active during a three week period toward the end of the study. These sites were determined and verified to be phishing by a third party vendor. There was an average of 436k clients during this three week period. We recorded 101 PRE reports listing one of the verified phishing sites as the primary URL. This implies that the client has typed at the phishing site a password previously used at another site on the user's PPL, which is a fairly good indication that the user has been "phished." We can use this to get an estimate of the annualized fraction of the population being phished as

$$\frac{101 \times 365}{436000 \times 21} \approx 0.00403.$$

Thus the data indicates that on the order of 0.4% of the population falls victim to a phishing attack a year.

3.4.2 Self-maintained Routers

We found a total of 1203 PRE reports listing `http://192.168.0.1`, `http://192.168.1.1` or `http://192.168.2.1` as the primary URL. These addresses are used almost exclusively for router setup pages. Since we have 544k clients, this indicates that roughly 0.2% of users maintain their own router, have a non-empty password (since addresses of the form `http://192.168.*.*` are not accessible from the outside network this is a common practice), *and logged in at least once* during the period of our study. Since many users will configure a router once, and seldom need to login again, the actual percent who maintain their routers is probably much higher.

3.4.3 How many login URLs are there?

A partial list of the frequency with which URLs appeared as the primary URL in a PRE report was given in Table 1. If we compile the full table of frequencies we find a huge range of login frequencies. From `login.live.com` which occurs 120k times to 103978 distinct URLs each of which occurred once. Call N_r the number of distinct URLs that appear r times as primary URL of a PRE report. Since we have observed a certain set of clients for a limited period it is natural to wonder how accurate a picture this data gives us of the entire universe of login URLs. It is reasonable to infer that the most popular sites, listed in Table 1, would remain the most popular if we had 2, 10 or 100 times as many clients. But, how many more new login URLs might we find?

The standard means of estimating the probability mass of unseen species in a limited observation is the Good-Turing estimate [8]. From the full version of Table 1 we have $N = \sum rN_r = 1320515$ and $N_1 = 103978$. Thus the Turing-Good estimate [8] of the amount of the probability mass missing is $N_1/N \approx 0.079$ or 7.9%. This tells us that our estimate of the distribution of login frequencies is reasonably accurate, in that the bulk of the mass has been captured.

So, if the probability mass of the unseen urls is around 7.9%, can we estimate *how many* sites are unseen? A similar question is *"how many new sites are added for each new user we add?"*. Figure 12 shows the number of previously unseen password URLs visited by new clients as a function of number of users considered. The figure was obtained by considering only a (varying size) partial subset of the users. Observe that the number of new sites plateaus relatively quickly: at the user base we had in the experiment, each of the last few users add only about 1.3 new sites, down from more than seven for the first 1% of the clients.

Extrapolating from the existing users to obtain the actual number of login sites is much trickier than estimating the overall probability mass. Efron and Fisher [4] use a Poisson distribution of the frequencies and derive an estimate

$$\Delta(t) = \sum_{r=1}^{\infty} (-1)^{r-1} N_r t^r,$$

where t is the size of a new sample relative to the size of the existing sample. For example $\Delta(1/2)$ would be the estimate of the number of unseen URLs to be expected in sample that included half as many PRE reports as our existing sample. Using our PRE report date we find $\Delta(1) \approx 80811$, with a standard deviation of 428, meaning that a second sample would be expected to turn up 80k new login URLs previously unseen. Unfortunately the estimate for $\Delta(t)$ is unstable for $t > 1$, hence we cannot estimate beyond that with any accuracy.

4. RELATED WORK

The subject of the use of passwords, and alternatives, for authentication has been analyzed at length in the computer security literature. However, web users' password habits have received far less attention. An early study of users'

password habits on a UNIX time sharing system is [12]. Morris and Thompson examined various attacks on password systems, and compile a study of 3289 passwords gathered from many users. They found that 86% of the passwords were extremely weak: being too short, containing lowercase letters only, digits only or a combination of the two, or being easily found in dictionaries or lists of names. The results we present in Section 3.3 in some sense update and extend this work with much more data. While much has changed since 1979 (*e.g.* a minimum of 6 characters is very common) it is just as true that many users appear to choose the weakest possible password, unless forced to do otherwise.

An experiment by Grampp and Morris [6] found that weak passwords, such as names followed by a single digit, were in widespread use in a number of machines they examined in a corporate network. Klein [5] reported being able to crack about 25% of passwords in use, again on a Unix system, by brute force attack. Adams and Sasse [3] surveyed users about password memorability, and also conclude that choosing secure passwords that are memorable is proving a difficult task for many users. Our findings on password strength reinforce and extend these reports.

Yan *et al.* [9] performed a more recent study of password memorability and security. The survey involved 288 students; a third were asked to choose a password (given certain password rules), a third were assigned random passwords, and a third were asked to choose a password using a mnemonic based on a phrase. Among their findings were that the randomly assigned and phrase based passwords were similarly to crack by dictionary attacks, but the phrase based passwords were significantly easier to remember.

There have been numerous surveys of user password habits that employ questionnaires. A good synopsis of recent surveys is [2]. This is a very useful compendium of user responses on questions of their password use, re-use, and forgetting habits as well as a source of password policies at major institutions. Most of the data in [2] is obtained by surveying users. By contrast our study measures what they actually do, rather what they say they do. Further, at 544k participants, we have more than 100 times more data than most of the existing surveys.

5. CONCLUSION

The data allows us to measure for the first time average password habits for a large population of web users. Many facts previously suspected, can be confirmed using large scale measurements rather than anecdotal experience or relatively small user surveys. The results particularly confirm the conventional wisdom about the large number and poor quality of user passwords. In addition passwords are re-used and forgotten a great deal. We are able to estimate the number of accounts that users maintain, the number of passwords they type per day, and the percent of phishing victims in the overall population.

Acknowledgements: the authors thank Geoff Hulten, Raghava Kashyapa, Steve Miller, Anthony Penta and Steve Rehfuss for enormous help in acquiring and analyzing the data.

6. REFERENCES

[1] `http://www.rsasecurity.com`.

[2] `http://www.passwordresearch.com`.

[3] A. Adams and M. A. Sasse. Users are not the Enemy. *Comm. ACM*, 1999.

[4] B. Efron and R. Thisted. Estimating the number of unknown species: How many words did Shakespeare know? *Biometrika*, 1976.

[5] D. V. Klein. Foiling the Cracker: A Survey of, and Improvements to, Password Security. *Usenix Security Workshop*, 1990.

[6] F.T.Grampp and R. H. Morris. UNIX Operating System Security. *Bell System Tech. Jorunal*, 1984.

[7] E. Gaber, P. Gibbons, Y. Matyas, and A. Mayer. How to make personalized web browsing simple, secure and anonymous. *Proc. Finan. Crypto '97.*

[8] W. Gale. Good-Turing Smoothing Without Tears. Statistics Research Reports from AT&T Laboratories 94.5, AT&T Bell Laboratories, 1994.

[9] J. Yan and A. Blackwell and R. Anderson and A. Grant. Password Memorability and Security: Empirical Results. *IEEE Security & Privacy*, 2004.

[10] Jefferson Wells Inc. Microsoft Phishing Filter Feature in Internet Explorer 7 and Windows Live Toolbar. 2006. `http://www.jeffersonwells.com/client_audit_reports/Microsoft_PF_IE7_IEToolbarFeature_Privacy_Audit_20060728.pdf`.

[11] Anti-Phishing Working Group. http://www.antiphishing.org.

[12] R. Morris and K. Thompson. Password Security: A Case History. *Comm. ACM*, 1979.

[13] B. Ross, C. Jackson, N. Miyake, D. Boneh, and J. C. Mitchell. Stronger password authentication using browser extensions. *Proceedings of the 14th Usenix Security Symposium, 2005.*

[14] M. E. Russinovich and D. A. Solomon. *Microsoft Windows Internals*. Microsoft Press, fourth edition, 2005.