Augmenting Password Recovery with Online Profiling

By

Khawla Al-Wehaibi, Tim Storer and Brad Glisson

From the proceedings of

The Digital Forensic Research Conference

DFRWS 2011 USA

New Orleans, LA (Aug 1st - 3rd)

DIGITAL INVESTIGATION 8 (2011) S25–S33

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/diin

Digital Investigation

Augmenting password recovery with online profiling

Khawla Al-Wehaibi [a,*], *Tim Storer* [b], *William Bradley Glisson* [c]

[a] *Department of Computer and Information Sciences, Prince Sultan University, P.O. Box 53073, Riyadh 11586, Saudi Arabia*
[b] *School of Computing Science, University of Glasgow, United Kingdom*
[c] *Humanities Advanced Technology and Information Institute, University of Glasgow, United Kingdom*

Keywords:
Password recovery
Password cracking
Online profiling
PRTK
Dictionary attacks
Digital forensics
Computer forensics

ABSTRACT

In investigations involving password-protected artefacts, password recovery can be a daunting task, consuming resources and causing delays in investigations. This paper describes work conducted to examine whether using online publicly available information to generate individual–related electronic profiles reduces automated password recovery time. In order to accomplish this, a web crawler was developed to capture profiles, which were then processed by Password Recovery Tool kit 'PRTK'. During an exploratory study, four passwords of 18 password-protected Word documents were recovered faster using this technique.

1. Introduction

Computer forensics involves the collection of electronic data, examining it, interpreting and analysing the digital evidence, and finally presenting the findings (Sammes and Jenkinson, 2007). During their investigations, computer forensic professionals may encounter encrypted artefacts. Encryption is the process of transferring data from its readable format 'plain text' to an illegible format 'cipher text' by using cryptographic algorithms. Modern algorithms are parameterised by cryptographic keys, often represented to the user as passwords (Shinder and Tittel, 2002). Encryption can be used to prevent unauthorised or illegal access, disclosure and/or alteration of data (Shinder and Tittel, 2002; HomeOffice, 2007), but can also be used for illegal purposes (Denning and Baugh, 1997; HomeOffice, 2007).

Encryption is a recognised problem for computer forensic professionals during investigations. Rogers and Seigfried (2004) conducted a pilot study and surveyed 60 participants involved in the field of computer forensics to identify the five most common problems they deal with. Encryption was reported as the third most frequently mentioned issue (24 from a total of 173). The challenge of encryption is discussed in many computer forensics resources and can be summarised as follows:

- Encryption is easily available to individuals who can misuse it to hide or secure illegal information. The use of encryption has increased during the past few years (HomeOffice, 2007), with more sophisticated algorithms being employed and more widely available (Sammes and Jenkinson, 2007). The end-user is presented with many user-friendly applications which provide easy-to-use encryption, including whole drive encryption (Sammes and Jenkinson, 2007). For example, individuals can use encryption software such as 'PGP'[1] and 'TrueCrypt'.[2] Some applications, such as Microsoft Office, provide their own default encryption tools which can be used to encrypt files. As a result, more encrypted artefacts are being encountered during investigations.
- During investigations involving encrypted artefacts, legal issues related to data acquisition and analysis such as the

* *Corresponding author*. Tel.: +966 5 04266749.
E-mail addresses: kwehaibi@yahoo.com, k_wehaibi@pscw.psu.edu.sa (K. Al-Wehaibi).

[1] Pretty Good Privacy 'PGP' is a public encryption tool. http://www.pgp.com/.
[2] TrueCrypt is open-source encryption software. http://www.truecrypt.org/.

1742-2876/$ – see front matter
doi:10.1016/j.diin.2011.05.004

admissibility of the evidence retrieved may arise (Altheide et al., 2008). Computer forensic professionals should be aware of related laws and the risks of live acquisition.

- Decrypting encrypted artefacts can be challenging. In order to decrypt encrypted artefacts, it is usually necessary for computer forensic professionals to recover passwords to view the encrypted content as plain text. Attempting to recover passwords can be a lengthy and daunting task, causing delays in investigations (Denning and Baugh, 1997), even with the help of password recovery toolkits, such as AccessData's Password Recovery Tool Kit 'PRTK'.[3]

Regardless of passwords' theoretical strength, some individuals tend to use information relating to them when creating passwords (Marshall, 2003, 2004). The purpose of this exploratory study is to test the plausibility of enhancing automated password recovery by using individuals' online public information.

2. Literature review

2.1. Cryptography and encryption in modern information technology

Encryption can be applied to network communications, individual files, a volume on a hard drive or an entire hard drive. An example of whole drive encryption is 'BitLocker', which is a default feature available in Windows Vista Ultimate and Enterprise (Microsoft, 2008) and Windows 7 Ultimate and Enterprise (Microsoft, 2009). In Windows 7, 'BitLocker' can be simply activated by a right-mouse button click on the selected volume or drive. Other operating systems provide default encryption features, for example Macintosh OS X's 'FileVault' is a built-in feature used to encrypt some or all files in a user's home folder (AppleInc, 2011). Furthermore, the New Technology File System 'NTFS' version 5 includes an Encrypting File System 'EFS' which can be used to encrypt files and/or folders (Shinder and Tittel, 2002). Whole disk encryptions can prevent the investigator from accessing huge amounts of data which might represent valuable evidence.

In an early paper, Denning and Baugh (1997) expressed their concerns about the negative impacts of encryption on computer forensic professionals and law enforcements. They argued that encryption usage will increase, it will be easier to use, more available to users, difficult to break and it will cost time and money. They demonstrated some statistics regarding the number of cases involving encryption. For example, the authors stated that according to the FBI's Computer Analysis Response Team 'CART' forensics lab, the number of cases which involved encryption in 1996 was four times higher than the number of cases in 1994. The paper lists more than ten cases involving encryption. These cases include terrorism, child pornography, drugs smuggling, theft and fraud (Denning and Baugh, 1997).

Encrypted or password-protected artefacts can be problematic for computer forensic professionals, particularly if the key is not known, or the encryption algorithm is strong or unidentified (Reyes et al., 2007).

2.2. Passwords and data acquisition

In the field of computer forensics, data acquisition is defined as the process of collecting electronic data for later analysis i.e. making an exact bit-by-bit duplicate/image of the digital evidence (Reyes et al., 2007). There are two methods to acquire data which are 'dead' or 'post-mortem' acquisition and 'live' acquisition. 'Dead' acquisition is performed if the system is shut down, or after powering it down, whereas 'live' acquisition is carried out while the system is powered on and running (Reyes et al., 2007). It is important to highlight the impact passwords have on data acquisition. Reyes et al. (2007) stated that "*encryption presents a variety of problems for the traditional forensics examiner*." As a solution to this problem, they recommended conducting a live acquisition. This is helpful, provided that the system is running and the encrypted artefact has been temporarily decrypted i.e. it is in its decrypted form. Furthermore, when performing a live acquisition, computer forensic professionals may find passwords saved in RAM 'Random Access Memory'(Reyes et al., 2007). Therefore, it is advisable that computer forensic professionals identify whether the system contains any encrypted files, volumes or drives before powering it down. Nevertheless, an investigator should be aware of the general risks associated with live acquisition, particularly legal issues relating to the digital evidence integrity and procedures' repeatability (Reyes et al., 2007).

In situations where encrypted containers are discovered and two images of the evidence are acquired, a live image and a dead image, then it might be vital to find/recover the passwords of the encrypted artefacts. This is crucial, since a live image of open password-protected documents and/or containers cannot be later verified against a dead image of the same password-protected documents and/or containers unless the password is known. With a password, the artefacts can be opened and the two images can be compared and verified.

Altheide et al. (2008) noted that there is a lack of research for proper procedures for the verifiable acquisition and examination of encrypted containers. They proposed a repeatable procedure to acquire and examine the contents of encrypted containers given that the password/key is known or obtained.

2.3. Finding passwords for encrypted artefacts

There is no standard fixed procedure to be followed when encrypted files and/or volumes are encountered. Nevertheless, some methods have been mentioned in computer forensics literature.

The first step to be attempted is simply to ask individuals for their passwords (Denning and Baugh, 1997; Craiger et al., 2005; Volonino et al., 2006). Sometimes, when suspects are taken into custody they refuse to be cooperative and give passwords (OUT-LAW, 2008). However, governments and law enforcement organisations often demand access to data where criminal activities are suspected.

[3] http://www.accessdata.com/decryptionTool.html.

In the UK on October 1st 2007, Part III of Regulation of Investigatory Powers Act 2000 'RIPA 2000' was introduced and activated as a law (OPSI, 2007). Part III of RIPA 2000 deals with encryption and entitles authorised individuals such as law enforcement members the power to demand suspects reveal their encryption keys or disclose data in its decrypted form (HomeOffice, 2007). Disclosing passwords does not violate a person's rights in relation to self-incrimination and refusal to disclose passwords is considered a criminal offence (OUTLAW, 2008).

However, in the US, individuals can use their Fifth Amendment Constitution rights and choose to refuse to reveal their passwords or decrypted data. An example is the case of Sébastien Boucher who eventually agreed to disclose his password in return for a plea bargain and was sentenced to three years in prison and five years of probation (CBC, 2010).

The second technique for password recovery is 'social engineering' (Shinder and Tittel, 2002; Craiger et al., 2005). An example of social engineering is pretending to be a member of I.T. personnel, contacting employees and asking for their passwords to conduct a certain activity. Nevertheless, investigators should be aware of the limitations imposed by law and security regulations.

Computer forensic professionals can also search a crime scene for passwords that have been written down (Casey, 2002; Craiger et al., 2005; Reyes et al., 2007). Passwords can be found written on notes posted on computers' monitors, placed under keyboards or attached inside drawers. Passwords can be found in journals, diaries or PDAs 'Personal Digital Assistants' (Sammes and Jenkinson, 2007). It is possible to find passwords saved, in a digital format, in files saved on a user's computer or on removable media (Hargreaves and Chivers, 2008).

If passwords cannot be found in clear, then computer forensic professionals can try guessing passwords according to the information they have regarding the suspect and the crime (Shinder and Tittel, 2002; Craiger et al., 2005; Sammes and Jenkinson, 2007). Examiners can use different pieces of findings including names, birth dates, graduation dates, addresses, telephone and/or mobile numbers, pets' names, sports teams' names or players' names. Volonino and Anzaldua (2008) suggested building custom dictionaries with more emphasis on data relevant to suspects, such as pets' and teams' names. Moreover, AccessData Forensic Tool Kit 'FTK' can be used in conjunction with PRTK. FTK can be used to generate word lists of relevant indexes for any specific case. The word lists then can be used in PRTK as dictionaries (Casey, 2002; Craiger et al., 2005).

Another approach that computer forensic professionals should consider is searching the seized medium for any encryption software (Casey, 2002; Volonino et al., 2006). They should also check the application used to create the encrypted files, for example, Microsoft Word. This may assist in narrowing down the most appropriate tool that can be used to recover passwords, for example, specific Microsoft Word password recovery software (Volonino et al., 2006). An investigator can also contact the application developers, who may aid in recovering the password (Volonino et al., 2006).

Other methods include using key loggers, snooping software and wiretaps to capture passwords (Denning and Baugh, 1997; Casey, 2002; Shinder and Tittel, 2002). Computer forensic professionals must ensure that all interceptions are legal and lawful prior to using such software 'RIPA 2000'.

In some cases when applicable, 'key recovery systems' can be used (Denning and Baugh, 1997). In 'key recovery systems' or 'key escrows', the key can be recovered by using information saved with the cipher text and information held with a trusted third party. This technique is mostly used by businesses, although it has been suggested that criminal organisations are also capable of applying the same technique. In situations like these, law enforcement agencies may have to rely on criminals' cooperation in disclosing the key (Denning and Baugh, 1997).

In some situations, parts of the encrypted document can be found in the machine as plain text (Casey, 2002; Craiger et al., 2005). Cache memory can also be searched for any saved passwords or encrypted data in a decrypted form (Casey, 2002). An example of retrieving keys from RAM is the work of Hargreaves and Chivers (2008). The authors tested their work on a TrueCrypt encrypted container created on a virtual machine.

Password recovery software can be used to perform brute force attacks. There is a wide variety of password recovery and decryption tools available offline and online. These tools differ in terms of their functionality, purpose, specifications and implementations. Some examples of these are: PRTK, John the Ripper,[4] L0phtCrack,[5] Cain & Abel[6] and Paraben's Decryption Collection.[7]

An 'Application Specific Integrated Circuit' 'ASIC' can be employed for password cracking. This special chip can be solely programmed to decrypt specific encryption algorithms, thus saving time (Volonino and Anzaldua, 2008). In addition, a distributed network of computers can be used to recover passwords. Nonetheless, this requires many resources and can increase the costs of investigations (Denning and Baugh, 1997).

2.4. *Users' behaviours towards password creation and usage*

Individuals sometimes share their passwords with colleagues or use the same password for different applications (Schneier, 2008). The latter is particularly important for computer forensic professionals because some applications are easier to crack. Once this is accomplished, the same password, or different variations of it, can be tested for the remaining applications (Casey, 2002; Craiger et al., 2005).Schneier suggested that with all advice given to users, they are becoming more aware of and educated in the criteria for 'strong' passwords. Schneier analysed data exposed by a MySpace phishing attack,[8] which targeted users' login names and passwords (Schneier, 2006). His findings revealed that 65% of the passwords were eight-character long or less, 81% of the passwords

[4] http://www.openwall.com/john/.
[5] http://www.l0phtcrack.com/.
[6] http://www.oxid.it/cain.html.
[7] http://www.paraben-forensics.com/password-recovery.html.
[8] In electronic communications, phishing is the act of attempting to obtain individuals' sensitive data by pretending to be a legitimate party. MySpace is a social networking website. http://www.myspace.com/.

were alphanumeric and less than 4% were dictionary words. Schneier commented that "*passwords are getting better*" after comparing his findings with two other studies conducted by Klein (1990) and Spafford (1992), who both examined UNIX passwords (see Table 1).

A survey conducted by 'Infosecurity Europe' in 2003 at Waterloo Station showed that 65% of the 152 surveyed employees use their passwords for more than one application. In terms of what kind of passwords people use, 16% of the employees stated that they use their own names in their passwords and 11% use their football team names. Also, 8% of the employees stated that they use their date of birth in their passwords (Marshall, 2003). Another study conducted by Microsoft in 2004 showed that 20% of British citizens use their mother's maiden name in their passwords (Marshall, 2004). Similar findings were reported in the studies of Klein (1990) and Spafford (1992). In Klein's work, from 24.2% sampled passwords, 2.7% consisted of UNIX accounts' names and users' names. 7% of the passwords consisted of common names, females'/males' names, places' names or sports terms. Spafford reported that 3.9% passwords consisted of UNIX accounts' names, users' names or telephone numbers.

The findings regarding users' behaviour towards password creation and usage motivated this study, i.e. testing the impacts of utilising the wealth of information posted online on finding individuals' passwords.

2.5. *Related studies*

Fragkos and Tryfonas (2007) proposed a cognitive model for assisting computer forensic professionals during the process of recovering passwords. They discussed the psychology behind choosing passwords, explaining that passwords' strength depends on the importance of what is being protected. For example, for some users, an online forum password is not necessarily as important as a bank account password. From this concept, the authors developed a Level of Difficulty 'LoD' for passwords. They mentioned that each password can fall into one of six LoDs. According to the LoD, investigators can assume some passwords' criteria, such as length and characters' type i.e. uppercase/lowercase letters, digits or symbols, therefore refining passwords search criteria, hence saving time. Nevertheless, the model was neither implemented nor tested and the authors acknowledged that the model is susceptible to inaccurate assumptions since individuals' psychology, behaviour, experience, and preferences are different. For example, some individuals constantly use strong passwords for all applications.[9]

Spafford (1992) conducted a study to enhance passwords choices by examining each newly created password and rejecting weak passwords. A software was developed to collect users' UNIX passwords from the computers in the Department of Computer Sciences and the Computing Center at Purdue University. The experiments lasted for approximately 10 months during which a total of 13,787 passwords were collected. To test passwords' strengths, dictionary attacks were performed using four distinct dictionaries (see Table 2). The drawbacks of this work are the process of updating the dictionaries since these need to be sorted each time a new word is added, as well as, storage limitations due to dictionaries' sizes.

Klein (1990) conducted a similar work to demonstrate passwords weaknesses. His work lasted for a year, during which he exposed 13,797 passwords collected from colleagues to dictionary attacks. Klein performed six attacks by using different dictionaries and permutations. Only 24.2% passwords from the sample set were recovered. 2.7% of the passwords were recovered by using accounts' names and users' names including their initials. 7% of the passwords were recovered by using common names, females' names, males' names, places' names, sports terms or teams' names.

Both studies Klein's (1990) and Spafford's (1992) are similar to the work proposed in this paper in that dictionary attacks are performed to recover passwords. The findings in these studies encourage the idea of using online profiling since it has been demonstrated that some individuals use information related to them when creating passwords. Nevertheless, the two studies focused on revealing passwords weaknesses and vulnerabilities, while this work aims to reduce password recovery time.

3. Experiment design

3.1. *Research aim*

This study investigates and analyses potential profiling success rates through publicly available information from participants' websites. The purpose of this research is to test whether online information can speed up password recovery compared with a dictionary attack by an industry standard tool. Rather than solely focussing on the absolute password recovery time, the focus of this work is to conduct a preliminary investigation to determine if password recovery becomes faster using online profiles.

3.2. *Participants recruitment*

Participants consisted of faculty members and postgraduate students from the School of Computing Science and Humanities Advanced Technology and Information Institute 'HATII' at the University of Glasgow in the summer of 2010. These two departments were considered in this exploratory study because of their accessibility to the authors.

An email was sent and candidates who were willing to participate were contacted in person. 23 candidates gave their consent and agreed to participate. Six members did not have useful online information, therefore only 17 candidates were eligible for this study (13 members were from the School of Computing Science and 4 members were from HATII).

[9] The authors did not elaborate on the issue of users' psychology but this leads to an interesting argument of individuals' reactions and whether they would change their behaviour if they become aware of such tool. It would be noteworthy to investigate the outcomes of implementing such software and whether using it would be beneficial.

Table 1 – Passwords comparison between three studies.

	No. of passwords	Percentages of eight-character long passwords	Percentages of only letters passwords
Klein (1990)	13,797	23.4%	–
Spafford (1992)	13,787	41.9%	38.1%[a]
Schneier (2006)	34,000	25%	9.6%

a 38.1% of passwords consisted of both uppercase and lowercase characters and 28.9% of passwords consisted of lowercase characters only.

3.3. *Research instrument and method*

A programme 'a web crawler' was implemented as the research tool. The web crawler was used to crawl and collect participants' online public information, which was then used to generate the individual–related electronic profiles. These profiles were then fed into PRTK and used to generate possible passwords. PRTK version 6.4 was used throughout the experiments. In addition for it being an industry standard tool used in computer forensics, it was chosen because of the wide variety of built-in dictionaries and permutations, the flexibility of choosing passwords attacks and modifying the dictionaries employed in the passwords attacks.

The general procedures of this study can be summarised as follows:

- Ethical approval was acquired.
- The web crawler was developed and tested.
- Candidates who agreed to participate were asked to encrypt a Microsoft Office 97–2003 Word document with a password they use or have used. No restrictions were imposed on the type of password chosen. In order to protect participants' privacy and encourage them to use realistic passwords, they were assured that none of the recovered passwords will be recorded.
- The web crawler was run on participants' web pages hence creating individuals' profiles.
- Participants' Word documents were added to PRTK and password dictionary attacks were performed.

3.4. *Experiment description and design*

The experiment was conducted in HATII's laboratory. It was conducted using three computers. A password-protected laptop[10] was used for running the web crawler and generating participants' profiles. The same laptop was also used for creating and encrypting participants' Word documents. Two password-protected desktop computers with the same specifications[11] were used to run PRTK and perform the password attacks on participants' Word documents. The experiment was divided into three stages: implementing the web crawler, creating the profiles and testing the profiles in PRTK.

3.4.1. *The web crawler implementation*

Several open source web crawlers were evaluated. Eventually, an open source code written in Python was selected, tested and segments of it were used in the web crawler developed for this study. The code was written by 'James Mills' and is available in 'ActiveState Code Recipes'.[12] Python 2.6 was selected because of the built-in libraries that support the functionality of web crawlers.[13] BeautifulSoup 3.0.8.1[14] was the library used for parsing web pages. The open source code was chosen because it complies with the purposes of this study, mainly the use of BeautifulSoup library. After modifying and integrating the required part of the open source code, the remainder of the web crawler was developed. The web crawler's main functionality can be described as two steps.

The first step was extracting URLs 'Universal Resource Locators' from web pages and saving them in lists. The web crawler starts crawling from a root URL given in the code and crawls for a depth of 'n', extracting URLs in the process. 'n' distinct lists were used to save the URLs from each level. This was implemented to allow the flexibility to test whether the crawling depth would be a factor in recovering passwords.

The second step was to retrieve web pages' source code. Each URL in a list was traversed and for every URL the website source code was extracted. This was performed for all URLs in every list. For each web page's source code, the text was extracted, filtered and each word from the filtered text was written in a separate line in an output file. The text was extracted from web pages' source code by using a recursive function which traverses the parsed HTML tree, ignoring tags and only extracting text. The text was then filtered by removing 'non-stop' words,[15] using a regular expression. To further filter the text, words' significance was considered by comparing words extracted from web pages to words retrieved from an English language corpus,[16] which is a list of English words and their frequencies. A word was considered significant in two situations. The first case, if the extracted word does not exist in the English corpus. The second case, if the extracted word exists in the English corpus, then its frequency was calculated and compared against the frequency of the same word found in the English corpus. If the extracted word's frequency is greater than the frequency of the same word in the English corpus multiplied by an 'UpperLimitFactor', then the word is considered significant.

[10] 32-bit Operating System, Intel(R) Core(TM) 2 Duo CPU P7550 @2.26 GHz 2.27 GHz, 4.00 GB RAM.

[11] 32-bit Operating System, Intel(R) Xeon(R) CPU W3503 @2. 40 GHz 2.39 GHz, 8.00 GB RAM.

[12] http://code.activestate.com/recipes/576551-simple-web-crawler/.

[13] http://docs.python.org/release/2.6.5/library/index.html.

[14] http://www.crummy.com/software/BeautifulSoup/.

[15] The list for non-stop words was obtained from http://www.textfixer.com/resources/common-english-words-with-contractions.txt.

[16] The British National Corpus was used. The list of words and their frequencies is available in http://ucrel.lancs.ac.uk/bncfreq/lists/1_2_all_freq.txt.

Table 2 – Passwords recovered using four different dictionaries.

Dictionary used	Number of matched passwords
Words available in the/etc/passwd file including users' names, accounts' names and telephone numbers	592
Words available in the standard dictionary in the/usr/dict/words file	620
A set of 11 dictionaries of different languages	1271
Words such as movies' names, sports teams' names and mythology characters	2498

The 'UpperLimitFactor' is a constant whose value should be greater than or equal to one. It is used to indicate which words are used significantly more than the average word use. In this study, the value of the 'UpperLimitFactor' was set to be one.

3.4.2. *Creating the profiles*

After the web crawler was developed and tested, it was run to crawl participants' web pages, which are available in the University of Glasgow website. The crawling depth for the experiments was chosen to be three since as the depth increases, more time is required to parse the websites and filter the words. The web crawler was executed 17 times, once for each participant, and the output profiles for each participant were saved separately. The total time for running the web crawler to all 17 websites was approximately 11 h, 14 min and 46 s. The least time was around 5 s and the longest time was around 2 h, 33 min and 34 s. The average time for crawling all web pages was approximately 40 min. The total number of words in all files was 151,830 and the average was approximately 8931 words. The maximum word count was 21,763 and the minimum was 33 words (see Table 3). These variations in time and word count may have occurred because participants have different numbers of links and text in their websites. Further examinations were carried out on profiles numbers 1, 2 and 4 (see Table 3) to explain the lack of words in certain depths. It was noted that the websites associated with these profiles did not contain any further links to be followed, thus no text to extract.

Table 3 – Number of words per depth.

Profile No.	Number of words/depth			Total number of words
	Depth (1)	Depth (2)	Depth (3)	
1	33	0	0	33
2	94	0	0	94
3	46	47	46	139
4	99	57	0	156
5	57	445	1371	1873
6	52	819	2210	3081
7	75	685	4737	5497
8	407	739	4621	5767
9	117	1199	6244	7560
10	41	1287	9997	11,325
11	199	1300	11,535	13,034
12	625	2503	10,021	13,149
13	132	1995	13,088	15,215
14	140	2266	14,379	16,785
15	519	2268	14,533	17,320
16	265	2454	16,320	19,039
17	110	2758	18,895	21,763

3.4.3. *Testing the profiles in PRTK*

After the web crawler was executed and the profiles produced for every participant, the second step was to test the profiles by using PRTK. As previously mentioned, 17 candidates were eligible for the experiment and they were asked to protect a Word document with a password they use or have used. One participant provided two Word documents, each protected with a different password. This resulted in 18 Word documents to be tested.

The module used for the Word documents' password recovery was 'Microsoft Office Encryption Module'. This is the default recommended module provided by PRTK (AccessData, 2008). Four types of attacks are associated with this module; only password dictionary attacks were performed because passwords are the focus of this study. In addition to module types and attack types, PRTK provides a range of 'rules' which can be used in password recovery. Rules are applied to the words in the dictionaries to generate potential passwords by using different permutations. Throughout the experiments, PRTK's default rules and their default ordering were used with no modifications.

Each document was added to PRTK and three dictionary attacks were performed:

- One attack was performed using PRTK's dictionaries. There are various dictionaries in different languages available in PRTK. In this work, PRTK's five default English dictionaries were used. Table 4 lists the word count for these dictionaries.
- The second attack was performed using participants' profiles created by the web crawler. This was done by adding the profiles to PRTK as dictionaries.
- The third attack was performed by using both PRTK's default dictionaries and the profile dictionaries. This was performed out of interest to check whether this will have an impact on the findings.

The three attacks were performed on the 18 Word documents, resulting in 54 attacks. Three Word documents were added to each computer and a time-out of 24 h was set for each

Table 4 – Word count for PRTK's default English dictionaries.

English dictionary	Word count
Common-en	11,354
Miscellaneous-en	54,922
Names-en	746,706
General-1-en	671,442
General-2-en	671,441

attack. Therefore, the attacks were conducted in nine working days between August 12th 2010 and August 24th 2010.

3.5. Limitations and considerations

Due to time constraints, the depth of the web crawler was limited to three because as the depth increases, more time is required to parse web pages, extract the text and filter the words. Another consideration is the number of times websites are crawled. In this study, the crawling was done once. Nevertheless, it must be taken into account that users may update their web pages. Moreover, crawling was limited to participants' websites available under the University of Glasgow website. However, other websites should be taken into account. Another limitation associated with BeautifulSoup is that only HTML web pages could be parsed. It is crucial to consider crawling other types of websites, as well as, other online documents such as Microsoft Word and Portable Document Format 'PDF' documents.

After the experiments were completed, two offline tools were found 'Wyd-a password profiler[17] and CeWL-a custom word list generator'.[18] While Wyd does not execute exactly as needed for this study, after slight modifications, the experiments could be re-run using one or both of these tools.

Another consideration is having a larger study sample. The sample size of this study is too small to generalise the findings. Nevertheless, it offers some evidence that password recovery time can be improved using the approach highlighted in the paper.

Due to time constraints, only 24 h were allocated for each attack. More time should be given to examine whether this constraint affected the findings reported in this study.

Although not recording the recovered passwords might be considered a limitation of this study, this was done to encourage participants to provide realistic passwords. Examining the recovered passwords may have resulted in a more detailed discussion of the findings.

4. Findings and discussions

As stated earlier in Section 3.4.3, in this exploratory study, three attacks were performed on each Word document and there was a 24-h limit for each attack. Within this time constraint, four passwords were recovered:

- Two passwords were recovered only by the profile dictionary attack i.e. using the files of online information generated by the web crawler.
- One password was recovered only by using the combined profile dictionary and PRTK's dictionaries attack.
- One password was recovered twice using two distinct attacks. One attack used PRTK's dictionaries and the other attack used the combined profile dictionary and PRTK's dictionaries.

[17] http://www.social-engineer.org/framework/Computer_Based_Social_Engineering_Tools:_Who%27s_Your_Daddy_Password_Profiler_%28WYD%29.

[18] http://www.digininja.org/projects/cewl.php.

Table 5 illustrates the dictionaries used and time consumed to recover the four passwords.

The results do suggest that password recovery is augmented by profiling. However, a measure of the improvement cannot be given from these results, since none of the passwords recovered by the profile dictionaries were also recovered by PRTK's dictionaries.

The limited time permitted for each attack means these results are not directly comparable to other studies such as Klein's (1990) and Spafford's (1992), who both spent more than 10 months conducting their experiments. However, the recovery rate (4/18) or 22.2% is to some extent similar to both. Klein recovered 24.2% passwords and Spafford recovered 20% of the passwords.

Not having access to the recovered passwords restricted any further analysis of the findings. However, there are some reasons which may have contributed to not recovering more passwords. One reason which can be assumed is passwords' types and strength. Individuals with low technical background are under-represented in this study and, consequently, the passwords that were examined may be more complex than those of typical users with low technical background. Additionally, participants were aware that their passwords would be subjected to attack and so may have supplied stronger passwords. As discussed earlier, as passwords' strength increases, more time is required to recover passwords.

Furthermore, for the purposes of this study, it was not essential to consider or limit the scope of participants' demographics. Some participants have English as their second language. Although it was assumed that they would provide English passwords, they were not explicitly informed to do so. There is a possibility that phonetics was used, thus yielding non-English passwords. Further work is necessary to study the effect of profiling participants by nationality and language.

Another reason which should be considered is the quantity and quality of the online information. The quality of the files produced by the web crawler greatly depends on the information available online. If there is no sufficient online information, then the files produced by the web crawler will not be as valuable. Another factor is the type of information posted on the web pages. It is possible that the online information is not useful. This also has a negative impact on the quality of the produced files.

Moreover, the crawler was only run on participants' websites which are available under the University of Glasgow website. However, other websites should be considered and

Table 5 – Experimental results for recovering passwords for the four documents.

File No.	Dictionary used	Time
1	Profile dictionary	10 h, 24 min & 17 s
2	Profile dictionary	5 h, 16 min & 23 s
3	Profile dictionary and PRTK's dictionaries	4 h, 53 min & 14 s
4(a)	Profile dictionary and PRTK's dictionaries	6 h, 37 min & 53 s
4(b)	PRTK's dictionaries	6 h, 53 min & 43 s

crawled. This is important for two reasons. First, this will yield more information. Second, it may improve information variety, assuming that participants would post more useful information on websites other than their university's web pages. Further work is required to explore the extent to which profiling can provide useful information for dictionary attacks, or whether 'excessive' information can be generated from the technique which hinders a default dictionary attack.

In this research, the scope of parsing websites was limited to HTML web pages and this may have affected the results. If other types of websites and documents were considered, then this may have changed the outcomes.

5. Conclusions and future work

5.1. *Summary and conclusions*

The purpose of this exploratory study was to test whether using online publicly available information to generate individual-specific electronic profiles speeds up password recovery. To test this, a web crawler was developed. The crawling starts from a root URL until a depth of three. The web crawler is designed to extract text from the crawled websites. The text is then filtered according to words' significance and only significant words are written to output files. These output files were later added to PRTK and used for password recovery. A total of 18 password-protected Word documents were received from participants. To test the effects of using online information on password recovery, PRTK was used to execute three different attacks on each Word document. For each attack, a different dictionary was used. For one of the attacks, only PRTK's default dictionaries were used. Another attack was performed by using the profile dictionaries i.e. the files produced by the web crawler. The third attack used both PRTK's dictionaries and the profile dictionaries. In total, 54 attacks were performed with a time limit of 24 h per attack. The experiments were performed in nine days, during which four passwords were recovered. Two of these passwords were recovered by using the profile dictionaries of online information. One password was recovered by using the profile dictionary and PRTK's dictionaries. The fourth password was recovered twice; once by using PRTK's dictionaries and the other time using the profile dictionary and PRTK's dictionaries.

5.2. *Future work*

Initial results from the experiment indicate that online profiling appears to have a positive impact on password recovery. The findings from this research warrant further investigation through the development of optimized code, customized password generators and fewer constraints in reference to the amount of time allocated to breaking passwords and access to passwords that are discovered. The research would also benefit from the use of an expanded participant base, i.e. participants from a variety of environments and backgrounds. In the future, the experiment should be expanded to include social networking sites, blogs, and wikis. It could also be executed with offline tools for a comparison of results.

REFERENCES

AccessData. Password recovery toolkit (PRTK) user guide. Lindon, Utah, USA: AccessData Corp; 2008.

Altheide C, Merloni C, Zanero S. A methodology for the repeatable forensic analysis of encrypted drives. In: Proceedings of the 1st European workshop on system security. Glasgow, Scotland: ACM; 2008. p. 22–6.

AppleInc. Mac OS X has you coverd. Available: http://www.apple.com/macosx/security/; 2011 [accessed 20.01.11].

Casey E. Practical approaches to recovering encrypted digital evidence. International Journal of Digital Evidence 2002;1:1–26.

CBC. Quebec man sentenced in U.S. child porn case. CBC News. Available: http://www.cbc.ca/canada/montreal/story/2010/01/22/qe-child-porn-sentence.html?ref=rss; 2010 [accessed 20.01.11].

Craiger JP, Swauger J, Marberry C. Digital evidence obfuscation: recovery techniques. In: Proceedings of SPIE, the society of photo-optical instrumentation engineers. Orlando, FL, USA: SPIE; 2005.

Denning DE, Baugh WE. Encryption and evolving technologies: tools of organized crime and terrorism. Trends in Organized Crime 1997;3:84–91.

Fragkos G, Tryfonas T. A cognitive model for the forensic recovery of end-user passwords. In: Proceedings of the 2nd international workshop on digital forensics and incident analysis (WDFIA 2007). Samos. Washington, DC, USA: IEEE Computer Society; 2007. p. 48–54.

Hargreaves C, Chivers H. Recovery of encryption keys from memory using a linear scan. In: Proceedings of the 3rd international conference on availability, reliability and security (ARES 2008). Barcelona, Spain. Washington, DC, USA: IEEE Computer Society; 2008. p. 1369–1376.

HomeOffice. Investigation of protected electronic information: code of practice. Available: http://www.opsi.gov.uk/si/si2007/uksi_20072196_en_1; 2007 [accessed 20.01.11].

Klein DV. "Foiling the cracker": a survey of, and improvements to, password security. In: Proceedings of the 2nd USENIX UNIX security workshop. Portland: Citeseer; 1990. p. 5–14.

Marshall BK. Password research study (Infosecurity Europe 2003 information security survey). Available: http://passwordresearch.com/stats/study55.html; 2003 [accessed 20.01.11].

Marshall BK. Password research statistics (Microsoft UK password survey). Available: http://passwordresearch.com/stats/statistic191.html; 2004 [accessed 20.01.11].

Microsoft. Windows vista product guide. Available: http://www.microsoft.com/downloads/details.aspx?FamilyID=bbc16ebf-4823-4a12-afe1-5b40b2ad3725&DisplayLang=en; 2008 [accessed 20.01.11].

Microsoft. Windows 7 product guide. Available: http://www.microsoft.com/downloads/details.aspx?displaylang=en&FamilyID=b3c68ec2-e726-4830-ac89-31c71d6be5f3; 2009 [accessed 20.01.11].

OPSI. The Regulation of Investigatory Powers Act 2000 (Commencement No. 4) Order 2007. Available: http://www.opsi.gov.uk/si/si2007/uksi_20072196_en_1; 2007 [accessed 20.01.11].

OUT-LAW. Court of appeal orders men to disclose encryption keys. Available: http://www.out-law.com/page-9514; 2008 [accessed 20.01.11].

Reyes A, O'Shea K, Steele J, Hansen JR, Jean BR, Ralph T. Cyber crime investigations: bridging the gaps between security professionals, law enforcement, and prosecutors. Rockland, MA, USA: Syngress; 2007.

Rogers MK, Seigfried K. The future of computer forensics: a needs analysis survey. Computers & Security 2004;23:12–6.

Sammes T, Jenkinson B. Forensic computing: a practitioner's guide. London: Springer; 2007.

Schneier, B. Real-world passwords. Blog: Schneier on Security. Available: http://www.schneier.com/blog/archives/2006/12/realworld_passw.html; 2006 [accessed 20.01.11].
Schneier, B. Passwords are not broken, but how we choose them sure is. The Guardian. Available: http://www.guardian.co.uk/technology/2008/nov/13/internet-passwords; 2008 [accessed 20.01.11].
Shinder DL, Tittel E. Scene of the cybercrime computer forensics handbook. Rockland, MA, USA: Syngress; 2002.
Spafford, E. Observing reusable password choices. In: Proceedings of the 3rd USENIX UNIX security symposium. Baltimore: Citeseer; 1992. p. 299–312.
Volonino L, Anzaldua R. Computer forensics for dummies. Hoboken, NJ, USA: Wiley; 2008.
Volonino L, Anzaldua R, Godwin J. Computer forensics: principles and practices. Upper Saddle River, NJ, USA: Pearson/Prentice Hall; 2006.

Khawla Al-Wehaibi received a B.Sc in Computer Science from Prince Sultan University in 2006. Following her graduation, she was hired to work as a research assistant on a research project funded by King Abdulaziz City for Science and Technology in the period between 2006 and 2007 during which she participated in international conferences and presented research findings. Al-Wehaibi joined Prince Sultan University in 2007 as a lab and teaching assistant. After being granted two scholarships, she continued her higher education and received her M.Sc. in Computer Forensics and E-Discovery from the University of Glasgow in 2010.

Tim Storer is a lecturer in Software Engineering at the University of Glasgow. His research interests include the modelling and simulation of socio-technical systems, software quality and system dependability.

Brad Glisson is the Director of the Computer Forensics and E-Discovery Program in the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow as well as a lecturer in the program. Brad has ten years industrial experience which includes working for U.S. and UK Global Fortune 500 financial institutions. This corporate interaction provides valuable insight into the technical aspects of the representation of data, the use of different technologies along with practical project management experience. Brad received his doctorate in Computing Science from the University of Glasgow in 2008. His primary research interests are pervasive digital forensics, corporate management and impact of security and forensics policies, along with the accurate representation, preservation, forensic integrity and recovery of this information.