Jim Blythe
Sven Dietrich
L. Jean Camp (Eds.)

LNCS 7398

Financial Cryptography and Data Security

FC 2012 Workshops, USEC and WECSR 2012
Kralendijk, Bonaire, March 2012
Revised Selected Papers

Springer

Lecture Notes in Computer Science 7398

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Jim Blythe Sven Dietrich
L. Jean Camp (Eds.)

Financial Cryptography and Data Security

FC 2012 Workshops, USEC and WECSR 2012
Kralendijk, Bonaire, March 2, 2012
Revised Selected Papers

Springer

Volume Editors

Jim Blythe
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292, USA
E-mail: blythe@isi.edu

Sven Dietrich
Stevens Institute of Technology
Computer Science Department
1 Castle Point on Hudson
Hoboken, NJ 07030, USA
Email address: spock@cs.stevens.edu

L. Jean Camp
414 E First St
Bloomington, IN 47401, USA
Email address: ljeanc@gmail.com

ISSN 0302-9743
ISBN 978-3-642-34637-8
DOI 10.1007/978-3-642-34638-5

e-ISSN 1611-3349
e-ISBN 978-3-642-34638-5

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012950705

CR Subject Classification (1998): C.2, K.4.4, K.6.5, D.4.6, E.3, J.1

LNCS Sublibrary: SL 4 – Security and Cryptology

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers from the two workshops held along with the 16th International Conference on Financial Cryptography and Data Security, in Bonaire on March 2nd, 2012.

USEC 2012: Workshop on Usable Security

The goal of the workshop on Usable Security was to engage on all aspects of human factors and usability in the context of security. Many aspects of data security combine technical and human factors. If a highly secure system is unusable, users will move their data to less secure but more usable systems. Problems with usability are a major contributor to many high-profile security failures today.

However, usable security is not well aligned with traditional usability for three reasons. First, security is rarely the desired goal of the individual. In fact, security is usually orthogonal and often in opposition to the actual goal. Second, security information is about risk and threats. Such communication is most often unwelcome. Increasing unwelcome interaction is not a goal of usable design. Third, since individuals must trust their machines to implement their desired tasks, risk communication itself may undermine the value of the networked interaction. For the individual, discrete technical problems are all understood under the rubric of online security (e.g., privacy from third parties' use of personally identifiable information, malware). A broader conception of both security and usability is therefore needed for usable security.

USEC 2012 brought together researchers already engaged in this interdisciplinary effort with others from areas such as HCI, artificial intelligence, theoretical computer science, law, and industry experts.

There were 13 submissions. Each submission was reviewed by at least 2, and on average 3, program committee members. The committee decided to accept 8 papers. Our thanks to the members of the program committee, the indefatigable chair of FC Angelos Keromytis, the IFCA board, participants, and all who submitted their works.

July 2012

Jim Blythe
Jean Camp

USEC 2012 Program Committee

Sadia Afroz	Drexel University
Ross Anderson	University of Cambridge
Matt Bishop	UC Davis
Pamela Briggs	Northumbria University
Tamzen Cannoy	PGP
Rachna Dhamija	Usable Security Systems
Chris Demchak	US Naval War College
Neil Gandal	Tel Aviv University
Seymour Goodman	Georgia Tech
Peter Gutmann	University of Auckland
Raquel Hill	Indiana University
Tiffany Hyun-Jin Kim	Carnegie Mellon
Brian LaMacchia	Microsoft
William Lehr	MIT
Andrew Patrick	Office of the Privacy Commissioner of Canada
Angela Sasse	University College London
Daniel Schutzer	Financial Services Roundtable
Mark Seiden	MSB Associates
Hovav Shacham	UC San Diego
Sara Sinclair	Google
Sean Smith	Dartmouth College
Gene Spafford	Purdue University
Frank Stajano	University of Cambridge
Sid Stamm	Mozilla
Douglas Stebila	Queensland University of Technology
Nicholas Weaver	ICSI Berkeley
Tara Whalen	Carleton University

WECSR 2012: Workshop on Ethics in Computer Security Research

The third Workshop on Ethics in Computer Security Research (WECSR 2012, http://www.cs.stevens.edu/ spock/wecsr2012/), organized by the International Financial Cryptography Association (IFCA, http://www.ifca.ai/), was held in Kralendijk, Bonaire, Dutch Antilles, on March 2, 2012. It was part of the third multi-workshop event co-located with Financial Cryptography 2012.

The goal was to continue searching for a new path in computer security that is acceptable for institutional review boards at academic institutions, as well as compatible with ethical guidelines for professional societies or government institutions. One such major step is the publication of the Menlo Report in the United States Federal Register in Fall 2011, the equivalent of the Belmont Report for this domain.

We mixed the two papers and one panel selected from five submissions with a keynote talk and one invited panel. Each submission was reviewed by at least 5 program committee members. The program committee carefully reviewed the submissions during an online discussion phase in fall 2011. I would like to thank the program committee for their work and suggestions. We like to thank all submitters for their papers and efforts.

The workshop brought together about 15 participants, including computer security researchers, practitioners, policy makers, and legal experts. We joined efforts with the co-located USEC 2012 workshop for the keynote talk by Ross Anderson and the afternoon panel on the ethics of data sharing moderated by Lenore Zuck. The relaxed Bonaire atmosphere allowed for many continued discussions beyond the day itself, including the evening island bus tour.

I would like to thank Angelos Keromytis, Rafael Hirschfeld, Burton Rosenberg, Tyler Moore, and Moti Yung for their hard work and help in organizing this workshop. A special thanks goes to Ross Anderson for a timely intervention. *Masha danki* (thank you very much) to Sara Matera for her support in making the local arrangements. Last but not least my gratitude also goes to the participants, who traveled to this remote island in the Netherlands Antilles close to Venezuela, where Papiamento is spoken. I look forward to many more discussions at future instances of the workshop.

July 2012 Sven Dietrich

WECSR 2012 Program Committee

John Aycock	University of Calgary
Michael Bailey	University of Michigan
Elizabeth Buchanan	University of Wisconsin-Stout
Aaron Burstein	UC Berkeley
Jon Callas	Indiana University
Nicolas Christin	Carnegie Mellon University
Michael Collins	RedJack, LLC
Marc Dacier	Symantec Research
Rachna Dhamija	Usable Security Systems
Sven Dietrich	Stevens Institute of Technology
Roger Dingledine	The Tor Project
David Dittrich	University of Washington
Kenneth Fleischmann	University of Maryland
Maritza Johnson	Columbia University
Erin Kennneally	sdsc / caida / elchemy
Engin Kirda	Intitut Eurecom
Christian Kreibich	ICSI
Howard Lipson	CERT, Software Engineering Institute, CMU
John Mchugh	RedJack LLC and University of North Carolina
Perry Metzger	University of Pennsylvania
Angelos Stavrou	George Mason University
Michael Steinmann	Stevens Institute of Technology
Lenore Zuck	University of Illinois in Chicago

Table of Contents

The Workshop on Usable Security (USEC 12)

The Workshop on Ethics in Computer Security Research (WECSR 12)

Linguistic Properties of Multi-word Passphrases

Joseph Bonneau and Ekaterina Shutova

Computer Laboratory
University of Cambridge
{jcb82,es407}@cl.cam.ac.uk

Abstract. We examine patterns of human choice in a passphrase-based authentication system deployed by Amazon, a large online merchant. We tested the availability of a large corpus of over 100,000 possible phrases at Amazon's registration page, which prohibits using any phrase already registered by another user. A number of large, readily-available lists such as movie and book titles prove effective in guessing attacks, suggesting that passphrases are vulnerable to dictionary attacks like all schemes involving human choice. Extending our analysis with natural language phrases extracted from linguistic corpora, we find that phrase selection is far from random, with users strongly preferring simple noun bigrams which are common in natural language. The distribution of chosen passphrases is less skewed than the distribution of bigrams in English text, indicating that some users have attempted to choose phrases randomly. Still, the distribution of bigrams in natural language is not nearly random enough to resist offline guessing, nor are longer three- or four-word phrases for which we see rapidly diminishing returns.

1 Introduction

Despite decades of research on the vulnerability of human-chosen passwords to guessing attacks [17], passwords continue to dominate web authentication. Passwords' familiarity and extremely low implementation costs are believed to be key reasons for their persistence [9], particularly given failures in the market for web authentication which discourage radical changes [4].

Given these constraints, multi-word passphrases may be a promising improvement, as they require few implementation changes and offer a similar user experience. Requiring multiple words in a password is a natural extension of usability findings which have suggested that mnemonic-phrase passwords[1] are considerably more difficult to guess while still easily memorable [22]. Recent research has also suggested that increasing the minimum length of passwords is the most effective means of increasing security in place of requirements to include character classes like numbers or symbols [12].

[1] Mnemonic-phrase passwords are formed by condensing a natural language sentence like "George Michael and Ann went to the protest on Friday" into a relatively-strong password like `GM&Aw2tpoF`.

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 1–12, 2012.

Specific usability studies of passphrases [11] have found them to be just as memorable as passwords, subject to an increased rate of typographical errors. Several proposals have been made reduce the rate of errors, either by storing multiple hashes of a passphrase to recognise entry of nearly-correct strings [16,2] or by providing visual feedback to allow a user to notice typos when they are made [18]. Passphrases may in fact be more usable in the context of mobile phones, which have input interfaces optimised for natural language and not for pseudorandom character strings [10]. Passphrases are already deployed in widely-used PGP software to protect private keys on disk [23] which has led to speculative research on hardware brute-forcing attacks [21].

Still, the security gains of moving from simple passwords to passphrases are unknown. The few published usability studies of passphrases estimate security either by naive calculations of the total space of possible character strings [11] or rely on Shannon's decades-old estimates of the entropy of characters in English text [20]. Experience from password guessing suggests that the only valid methods of estimating security of human-chosen secrets like passphrases are to run cracking software against real choices [17] or to collect sufficient data that the frequency of common choices can be predicted statistically [3]. Kuo et al. assembled a dictionary of phrases to evaluate the strength of mnemonic-phrase passwords [14], but we are unaware of any attempt to conduct a guessing attack on real human-chosen passphrases.

In this work we study passphrase choices using data collected from the Amazon PayPhrase system. Launched in 2009 for customers in the USA only, this system allows users to register a passphrase to make web purchases and is one of the few passphrase schemes widely deployed on the Internet. While we don't have access to the entire corpus of registered phrases, we can identify general linguistic patterns in passphrase selection which have important implications for future research on passphrases.

2 Data Collection

In the Amazon PayPhrase system, users register a multi-word phrase (with a minimum of two words) to authorise payments. A user can link multiple PayPhrases to the same underlying Amazon account, which is protected by a traditional password. Each PayPhrase is linked to a specific shipping address and payment card, allowing users to purchase items simply by typing in their phrase and a 4-digit PIN. Resistance to guessing attacks is expected to be provided both by the payphrase and the PIN.

Because no username is required, all PayPhrases must be unique. This prevents inferring the distribution of passphrases that humans will choose with no uniqueness restriction, which is common policy for password systems and necessary when passwords are used to protect private key files. This design choice allows us to study user selection of phrases simply by querying the publicly-accessible registration interface. As seen in Figure 1, the registration interface provides feedback to the user when attempting to select a phrase which has

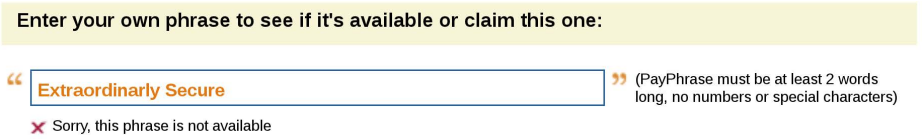


Fig. 1. The selection interface for passphrases deployed by Amazon

already been selected. We tested the registration status of over 100,000 possible passphrases using an automated script which queried this publicly-accessible interface While we found no evidence of rate-limiting, we limited our query rate to 1 Hz.

PayPhrases may only contain the space character and letters in the ASCII character set (the sets {a–z} and {A–Z}). No numbers, punctuation characters, or non-Latin letters are allowed. While PayPhrases must contain at least one space character at registration, spaces and capitalisation are ignored during verification. We will list all phrases we tested in a canonical lowercase form such as bases loaded.

3 Dictionary Attack

Our first experiment was to simulate a dictionary attack by assembling a number of lists of phrases that English-speaking users might be expected to pick. We chose categories in part based on previous research on password guessing dictionaries [13,14], though this is an inherently subjective process.

Our first step was to query a large number of proper nouns of various categories, as summarised in Table 1. All of our proper nouns were taken from "top x" lists on Wikipedia,[2] except for lists of top movies and movie stars, which we took from the film-specific website IMDB.[3] We filtered the items in each list to comply with passphrase requirements, stripping punctuation and converting numbers and non-ASCII characters, as well as removing items which only contained one word. Overall, we tried more than 15,000 proper nouns.

We supplemented our list of proper nouns with a number of idiomatic phrases, summarised in Table 2. We obtained our sports phrases from Wikipedia, common English idiomatic phrases from the English teaching website English Language Learning Online,[4] and a list of the most popular slang expressions from the the online slang website Urban Dictionary.[5]

Our goal is to estimate the underlying probability of a user selecting an individual phrase from each category we identified. We first must approximate the

[2] In some cases, the Wikipedia pages represented objectively collected lists, such as the largest cities in the world. In other cases, they were subjectively collected by Wikipedia editors as lists of notable items.

[3] `www.imdb.com`

[4] `www.usingenglish.com`

[5] `www.urbandictionary.com`

total number of phrases selected. Based on a press release issued two months after our data collection experiments which claimed that now over a million users had registered a phrase, we take $N = 10^6$ as a rough estimate for the total number of phrases registered.

Given a set of n phrases, of which k were selected, we wish to approximate the probability p of any individual phrase in the list being selected. We make a key assumption that within each of our identified lists, all phrases have an equal probability of being selected. We further assume that each user who decides to register a phrase from our list picks randomly from the list. If the phrase the user picks is already selected, they then pick some other phrase not in the list.

Given that we've observed k selections from a list, the expected number of attempted selections k' is an instance of the *partial coupon collector's problem*. The first user attempting to select from our list will always succeed, the second user will succeed with probability $\frac{n-1}{n}$, the second with probability $\frac{n-2}{n}$, and so on. The expected number of attempts before the j^{th} phrase is selected is $\frac{n}{n-j}$ as a Bernoulli trial with $p_{\text{success}} = \frac{n-j}{n}$. Thus, the total number of attempts expected before k phrases are taken is:

$$\mathbf{E}[\#_{\text{attempts}}] = \prod_{j=1}^{k} \frac{n}{n-j} \tag{1}$$

Given that we observed k selections in a list of n from N total trials, we can then compute the maximum-likelihood probability of each item $\hat{p} = \frac{\mathbf{E}[\#_{\text{attempts}}]}{N \cdot n}$. In Table 1, $\hat{p}$ is listed for each category we tried.

3.1 Comparison to Passwords

We estimate that our cumulative dictionary of 20,656 phrases covers the choices of about 1.13% of users. This level of security is equivalent to randomly-chosen strings of length $\lg\left(\frac{20{,}656}{0.0113}\right) \approx 20.8$ bits. For comparison, just 2 passwords (`123456` and `12345`) were chosen by 1.14% of users in the large dataset leaked from RockYou in 2009, equivalent to just 7.5 bits of security. Thus, passphrases appear to provide a significant boost in security over basic passwords against an attacker looking to compromise about 1% of accounts.

In another comparison, an optimal 20,656 word dictionary would cover 26.3% of passwords in the RockYou dataset. In an academic study, Klein manually assembled a dictionary in 1990 which covered over 9% of passwords with just 7,639 passwords [13]. These figures are equivalent to 16.3 or 16.4 bits of security, respectively. Thus, passphrases provide a security boost against attacks with small dictionaries by about 5 bits.

Our security estimates are slightly lower than those for mnemonic-phrase passwords by Kuo et al. [14], who found a 400,000 phrase dictionary which covered about 4% of choices, equivalent to 23.25 bits of security. Efficiency inherently declines with larger dictionaries, which partially explains this result. Additionally, Kuo et al. had to convert each phrase into a password. This can often be done in multiple ways, further making a dictionary attack less efficient.

Table 1. Success rates of phrase dictionaries based on proper nouns

word list	example	list size	success rate	$\hat{p}$
	arts			
musicians	three dog night	679	49.5%	0.0464%
albums	all killer no filler	446	56.5%	0.0372%
songs	with or without you	476	72.9%	0.0623%
movies	dead poets society	493	69.6%	0.0588%
movie stars	patrick swayze	2012	28.1%	0.0663%
books	heart of darkness	871	47.0%	0.0553%
plays	guys and dolls	75	70.7%	0.0093%
operas	la gioconda	254	17.3%	0.0048%
TV shows	arrested development	836	46.3%	0.0520%
fairy tales	the ugly duckling	813	13.3%	0.0116%
paintings	birth of venus	268	11.2%	0.0032%
brand names	procter and gamble	456	17.3%	0.0087%
total		7679	38.5%	0.4159%
	sports teams			
NHL	new jersey devils	30	83.3%	0.0056%
NFL	arizona cardinals	32	87.5%	0.0070%
NBA	sacramento kings	29	93.1%	0.0085%
MLB	boston red sox	30	90.0%	0.0074%
NCAA	arizona wildcats	126	56.3%	0.0105%
fantasy sports	legion of doom	121	71.1%	0.0151%
total		368	71.7%	0.0542%
	sports venues			
professional stadiums	soldier field	467	14.1%	0.0071%
collegiate stadiums	beaver stadium	123	12.2%	0.0016%
golf courses	shadow creek	97	6.2%	0.0006%
total		687	12.7%	0.0094%
	games			
board games	luck of the draw	219	28.8%	0.0074%
card games	pegs and jokers	322	27.6%	0.0104%
video games	counter strike	380	28.4%	0.0127%
total		921	28.2%	0.0306%
	comics			
print comics	kevin the bold	1029	29.5%	0.0361%
web comics	something positive	250	16.8%	0.0046%
superheros	ghost rider	488	45.3%	0.0295%
total		1767	32.1%	0.0701%
	place names			
city, state (USA)	plano texas	2705	33.8%	0.1117%
multi-word city (USA)	maple grove	820	79.0%	0.1283%
city, country	lisbon portugal	479	35.7%	0.0212%
multi-word city	ciudad juarez	55	69.1%	0.0066%
total		4059	43.7%	0.2677%
total		15481	38.1%	0.8479%

Table 2. Success rates of phrase dictionaries based on idiomatic phrases

word list	example	list size	success rate	$\hat{p}$
sports phrases	man of the match	778	26.1%	0.0235%
slang	sausage fest	1270	45.0%	0.0761%
idioms	up the creek	3127	43.6%	0.1789%
total		5175	41.3%	0.2785%

Table 3. Success rates of different classes of natural-language phrases taken from the British National Corpus [15]

bigram type	example	list size	success rate
adverb-verb	probably keep	4999	5.0%
verb-adverb	send immediately	4999	1.9%
direct object-verb	name change	5000	1.2%
verb-direct object	spend money	5000	2.4%
verb-indirect object	go on holiday	4999	0.7%
nominal modifier-noun	operation room	4999	9.8%
subject-verb	nature explore	4999	1.3%

4 Generated Phrases

After exhausting simple dictionaries of the kind utilised in Section 3, a brute-force attack would require generating phrases according to a model of the underlying natural language. Given our online access to the Amazon oracle, we were unable to conduct a realistic brute-force search with millions of possible phrases. Instead, we conduct several experiments with randomly-generated phrases to evaluate linguistic tendencies in passphrase selection.

4.1 Phrases Created Using a Syntactic Parser

Our first linguistic question is, broadly, what type of syntactic constructions are most popular as passphrases? To address this, we evaluated random samples of naturally-occurring 2-word phrases of varying syntactic relation, extracted from the 100-million word British National Corpus [15] parsed by the Robust Accurate Statistical Parser [7,1]. All of the syntactic relations we tested were two words, except for indirect object relations where a preposition is required (e.g. pay in cash). We found vanishingly small numbers of longer phrases to be registered, preventing research on longer passphrases with this data source.

The list of grammatical relations we examined and the summary of results are presented in in Table 3. Of immediate interest, nominal modifier-noun phrases (e.g. bedtime story) were the most likely to be registered by nearly a factor of two. The next most popular list was adverbial-modifier verb relations (e.g. never leave), again twice as popular as any other list. This suggests that users prefer phrases which represent as a single object or a single action, rather than a verbal phrase containing an action and a subject or object.

Table 4. Success rates of bigrams taken from the Google n-gram corpus [6]

bigram type	example	list size	success rate
adjective-noun	powerful form	10000	13.3%
noun-noun	island runner	10000	4.4%

4.2 Phrases Created Using the Google n-gram Corpus

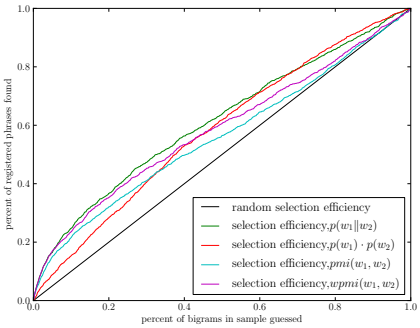
Our second linguistic question is, what factors predict how likely a given natural-language phrase is to be selected as a passphrase? We focus specifically on noun phrases using the much larger Google n-gram corpus which consists of over 10^{15} words of text harvested from the World Wide Web in 2006 [6]. Because this corpus contains counts for n-grams (sequences of n consecutive words) of only up to 5 words, sentence-level parsing is impossible. We instead relied on a much cruder classification of words as adjectives and nouns based on their most common part-of-speech tag in the RASP parsing of the BNC corpus [1].

We chose two random lists of 10,000 bigrams from the Google n-gram corpus, one consisting of adjective-noun bigrams and one of noun-noun bigrams. Basic statistics are given in Table 4. To evaluate how users may be selecting passphrases, we compared several potential models to rank each phrase in order selection probability. In Figure 2, we plot the percentage of registered phrases found against the percent of phrases guessed when proceeding in ranked order according to each model.

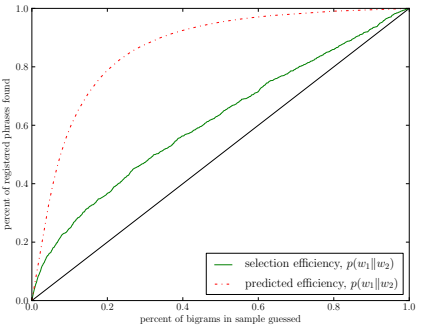
As a baseline, a random model considers users equally likely to pick any phrase from the list. This model produces a 45° degree diagonal line when plotted. We compare this to several other models:

- $p(w_1||w_2)$: bigrams are ranked by their overall probability. This simulates users generating passphrases exactly as pairs of words are generated in natural language.
- $p(w_1) \cdot p(w_2)$: bigrams are ranked by the product of the probabilities of each constituent word. This simulates users selecting each word in their phrase independently.
- $pmi(w_1, w_2)$: bigrams are ranked by the point-wise mutual information [8] of w_1 followed by w_2: $\lg \frac{p(w_1||w_2)}{p(w_1) \cdot p(w_2)}$. This simulates users having a tendency to pick words which are strongly associated with each other and hence occur together much more frequently than would be expected by random chance.
- $wpmi(w_1, w_2)$: bigrams are ranked by the point-wise mutual information of w_1 followed by w_2, multiplied by $p(w_1||w_2)$. This is a blended model.

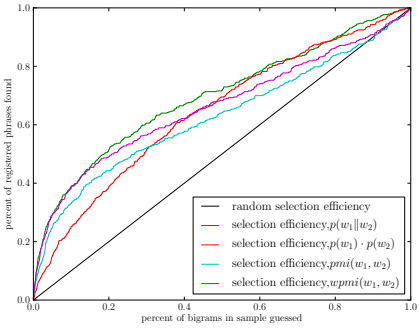
As seen in Figures 2a and 2c, the overall bigram probability is the best model for passphrase selection, though for the least-likely phrases, the independent probability model is just as accurate. Neither model based on pointwise mutual information provides additional predictive power. This leads us to conclude that users don't stray far from natural language patterns when choosing passphrases.



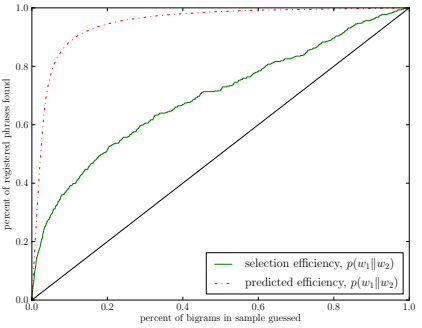
(a) Comparison of selection predictors (adjective-noun bigrams)



(b) Comparison of bigram probability to actual selection (adjective-noun bigrams)



(c) Comparison of selection predictors (noun-noun bigrams)



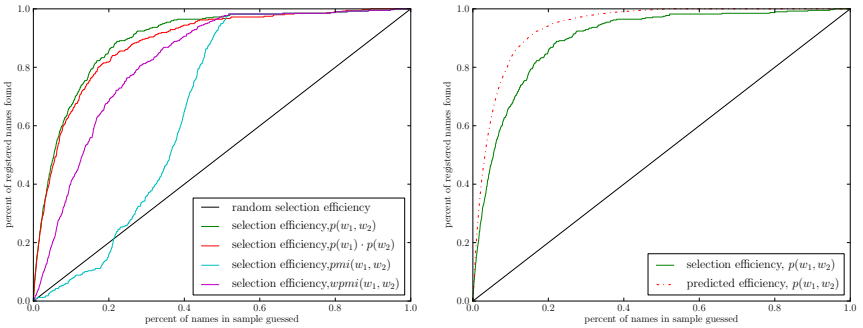
(d) Comparison of bigram probability to actual selection (noun-noun bigrams)

Fig. 2. The influence of different factors on the likelihood of individual bigrams being selected as passphrases. In Figures 2a and 2c, four different models are compared against a random-selection model: the overall bigram probability in the Google n-gram corpus, the product of the individual word probabilities, the pointwise mutual information of the bigram, and pointwise mutual information weighted by the overall bigram probability. In both cases, overall bigram probability is the best model. In Figures 2b and 2d, the expected efficiency of the overall bigram probability is compared to the observed effiency. In both cases, actual selection is considerably closer to random than predicted by the model.

However, this model is far from complete. In Figures 2b and 2d, we plot the expected efficiency if users perfectly followed the bigram probability model against our observed results. The large gap shows that users are considerably more random when choosing passphrases than when speaking naturally.

4.3 Phrases Created from Personal Names

A special class of phrases we identified are those based on a personal name, e.g. ekaterina shutova. Using 10,000 random names from a large corpus crawled from Facebook's public index of users in 2010 [3], we found 4% to be registered, a rate exceeding many of the types of natural language phrases as shown in Table 3. This is consistent with user preference for noun phrases.



(a) Comparison of selection predictors for personal names

(b) Comparison of name probability to actual selection

Fig. 3. The influence of different factors on the likelihood of personal names (e.g. joseph bonneau) being selected as passphrases. The selection models are equivalent to those used defined in Section 4.2 and Figure 2.

We again tested several models for user selection of names as phrases as in Section 4.2, using the frequency of each name in the Facebook corpus as the overal "bigram probability" and the product of the frequencies of the first and last name from the Facebook corpus to simulate creating a random name, as plotted in Figure 3. In this case, these two models are nearly equivalent, as first and last names have relatively low mutual information compared to bigrams occurring in natural language; that is, being given the first name or last name of a person's name doesn't greatly help in guessing the other component. Still, as seen in Figure 3, guessing names in order of overall probability is the most effective model, with no indication that a name's point-wise mutual information influences user choice. As seen in Figure 3b, the model of users choosing a name for their passphrase at random according to the population-wide distribution of names produces very close results to our observed data.

4.4 Security Implications

Given the evidence that user choice is partially predicted by the frequencies of phrases in natural language, it is natural to ask what security can be achieved if users in fact chose passphrases exactly in accordance with their distribution in natural language. We can examine this using the Google n-gram corpus to estimate of the probability distribution of multi-word phrases in English.

We use the *marginal guesswork* model to measure the guessing difficulty of a distribution [19,3,5]. The metric $\tilde{\mu}_\alpha$ measures the effective strength of a distribution in bits against an attacker desiring a α probability of guessing a user's passphrase correctly. It has been shown that no single metric can accurately measure guessing difficulty against attackers with different values of α [3]. Thus it is necessary to plot $\tilde{\mu}_\alpha$ across a range of values for α.

Figure 4 plots $\tilde{\mu}_\alpha$ for a phrases of 1–4 words, as well as randomly-chosen 2-word phrases, randomly-chosen names, and passwords. The results are somewhat discouraging for the passphrase concept, as 2-word phrases provide slightly less guessing resistance than existing text passwords. There is some gain from moving to 3-word phrases, but only a very small gain from 4-word phrases after that.

Given that we found users choose phrases more randomly than their natural language distribution, these findings should be considered a lower bound for security. Many of the most common phrases in natural language are purely functional, such as as well as, and would be unlikely to be chosen as passphrases. Additionally, the Google n-grams corpus contains many artifacts of the web, with the most common 3-word phrase being all rights reserved and the most common 4-word phrase being property of their respective. Still, these findings suggest that multi-word phrases, if chosen naively according to natural language tendencies,

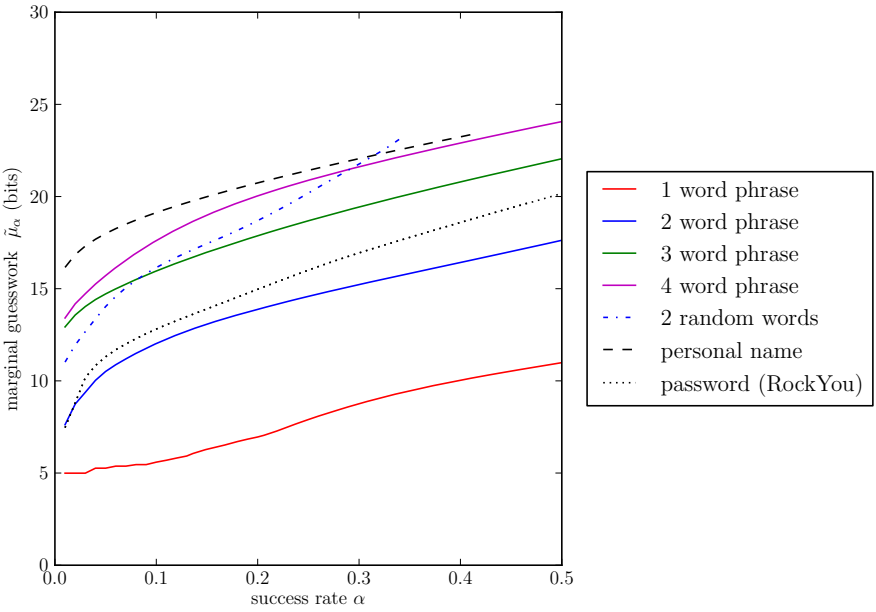


Fig. 4. The security provided by natural-language phrases of 1–4 words, based on estimated probabilities from the Google n-gram corpus. Also plotted is the difficulty of guessing a 2-word phrase if the words are selected independently, the difficulty of guessing a personal name based on the population distribution of names, and the difficulty of guessing a user-chosen password based on the leaked RockYou corpus.

are not as effective at mitigating guessing attacks as alternate choices, such as choosing 2 random words or choosing a personal name at random.

5 Concluding Remarks

We consider our work preliminary due to the limitations of our dataset. In particular, without a full list of registered phrases, we can only test predicted selection strategies and there may be large classes of passphrases which we have not considered. Additionally, the unusual setup of the Amazon PayPhrase system may not encourage users to choose a difficult to guess password, as additional security is provided by a random PIN.

Our work suggests that multi-word passphrases have some promise as a means to improve security over traditional passwords. Even 2-word passphrases may be able to raise the security of the weakest selections from below 10 bits to over 20 bits which could be sufficient to make online attacks impractical. However, our results suggest that users aren't able to choose phrases made of completely random words, but are influenced by the probability of a phrase occurring in natural language. Examining the surprisingly weak distribution of phrases in natural language, we can conclude that even 4-word phrases probably provide less than 30 bits of security which is insufficient against offline attack. Our results are a caution against optimistic security estimates arising from Shannon's esimates of entropy [10] in place of of probabilities of whole phrases from modern corpora of natural language.

We recommend further collaboration between the security and linguistics research communities to explore what is possible in multi-word passphrases. In particular, user testing for longer phrases is necessary to determine the extent to which users will tend to choose passphrases with natural-language-like properties as more words are required and not resort to easier-to-remember patterns like repeated words, idioms, or well-known titles. We also suggest exploring random multi-word phrases in place of users-chosen ones, which our results suggest may allow improved guessing resistance with much shorter phrases.

Acknowledgements. Joseph Bonneau is supported by the Gates Cambridge Trust. Ekaterina Shutova is supported by the EU FP7 PANACEA project. We thank Diarmuid Ó Séaghdha for sharing his database of noun compounds.

References

1. Andersen, Ø., Nioche, J., Briscoe, E.J., Carroll, J.: The BNC Parsed with RASP4UIMA. In: Proceedings of LREC 2008 (2008)
2. Bard, G.V.: Spelling-Error Tolerant, Order-Independent Pass-Phrases via the Damerau-Levenshtein String-Edit Distance Metric. In: ACSW 2007: Proceedings of the 5th Australasian Symposium on ACSW Frontiers, vol. 68, pp. 117–124. Australian Computer Society, Inc., Darlinghurst (2007)
3. Bonneau, J., Just, M., Matthews, G.: What's in a Name? Evaluating Statistical Attacks against Personal Knowledge Questions. In: Sion, R. (ed.) FC 2010. LNCS, vol. 6052, pp. 98–113. Springer, Heidelberg (2010)

4. Bonneau, J., Preibusch, S.: The password thicket: technical and market failures in human authentication on the web. In: WEIS 2010: Proceedings of the 9th Workshop on the Economics of Information Security (2010)
5. Bonneau, J., Preibusch, S., Anderson, R.: A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In: Keromytis, A.D. (ed.) FC 2012. LNCS, vol. 7397, pp. 25–40. Springer, Heidelberg (2012)
6. Brantz, T., Franz, A.: The Google Web 1T 5-gram corpus. Technical Report LDC2006T13, Linguistic Data Consortium (2006)
7. Briscoe, T., Carroll, J., Watson, R.: The second release of the RASP system. In: COLING-ACL 2006: Proceedings of the COLING/ACL on Interactive Presentation Sessions, pp. 77–80. Association for Computational Linguistics, Stroudsburg (2006)
8. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16, 22–29 (1990)
9. Herley, C., van Oorschot, P.C., Patrick, A.S.: Passwords: If We're So Smart, Why Are We Still Using Them? In: Dingledine, R., Golle, P. (eds.) FC 2009. LNCS, vol. 5628, pp. 230–237. Springer, Heidelberg (2009)
10. Jakobsson, M., Akavipat, R.: Rethinking Passwords to Adapt to Constrained Keyboards (2011), `www.fastword.me`
11. Keith, M., Shao, B., Steinbart, P.J.: The usability of passphrases for authentication: An empirical field study. International Journal of Human-Computer Studies 65(1), 17–28 (2007)
12. Kelley, P.G., Mazurek, M.L., Shay, R., Bauer, L., Christin, N., Cranor, L.F., Komanduri, S., Egelman, S.: Of Passwords and People: Measuring the Effect of Password-Composition Policies. In: CHI 2011: Proceedings of the 29th ACM SIGCHI Conference on Human Factors in Computing Systems (2011)
13. Klein, D.: Foiling the Cracker: A Survey of, and Improvements to, Password Security. In: Proceedings of the 2nd USENIX Security Workshop, pp. 5–14 (1990)
14. Kuo, C., Romanosky, S., Cranor, L.F.: Human Selection of Mnemonic Phrase-based Passwords. In: SOUPS 2006: Proceedings of the 2nd Symposium on Usable Privacy and Security, pp. 67–78. ACM (2006)
15. Leech, G.: 100 million words of English: the British National Corpus. Language Research (1993)
16. Mehler, A., Skiena, S.: Improving Usability Through Password-Corrective Hashing. In: Crestani, F., Ferragina, P., Sanderson, M. (eds.) SPIRE 2006. LNCS, vol. 4209, pp. 193–204. Springer, Heidelberg (2006)
17. Morris, R., Thompson, K.: Password Security: A Case History. Communications of the ACM 22(11), 594–597 (1979)
18. Perrig, A., Song, D.: Hash Visualization: a New Technique to Improve Real-World Security. In: International Workshop on Cryptographic Techniques and E-Commerce, pp. 131–138 (1999)
19. Pliam, J.O.: On the Incomparability of Entropy and Marginal Guesswork in Brute-Force Attacks. In: Roy, B., Okamoto, E. (eds.) INDOCRYPT 2000. LNCS, vol. 1977, pp. 67–79. Springer, Heidelberg (2000)
20. Shannon, C.E.: Prediction and entropy of printed English. Bell System Technical Journal 30, 50–64 (1951)
21. Shimizu, K., Suzuki, D., Tsurumaru, T.: High-Speed Search System for PGP Passphrases. In: Franklin, M.K., Hui, L.C.K., Wong, D.S. (eds.) CANS 2008. LNCS, vol. 5339, pp. 332–348. Springer, Heidelberg (2008)
22. Yan, J., Blackwell, A., Anderson, R., Grant, A.: Password Memorability and Security: Empirical Results. IEEE Security & Privacy Magazine 2(5), 25–34 (2004)
23. Zimmermann, P.R.: The Official PGP User's Guide. MIT Press (1995)

Understanding the Weaknesses of Human-Protocol Interaction

Marcelo Carlos* and Geraint Price

Royal Holloway University of London,
Egham, Surrey, TW20 0EX, United Kingdom
{marcelo.carlos.2009,geraint.price}@rhul.ac.uk

Abstract. A significant number of attacks on systems are against the non-cryptographic components such as the human interaction with the system. In this paper, we propose a taxonomy of human-protocol interaction weaknesses. This set of weaknesses presents a harmonization of many findings from different research areas. In doing so we collate the most common human-interaction problems that can potentially result in successful attacks against protocol implementations. We then map these weaknesses onto a set of design recommendations aimed to minimize those weaknesses.

Keywords: human-protocol interaction, human factors, usable security.

1 Introduction

Sometimes, even some of the most secure and robust protocols are vulnerable to attacks when implemented. The reason for this is that a significant number of these attacks are against the non-cryptographic components, such as the human-protocol interaction.

In protocol design and analysis, the human interaction is usually part of the assumptions and not specifically included in the description. However, user behaviour is often unpredictable, making the assumptions not precise enough [11,16]. Despite that unpredictable nature, there are some common design errors and weak-assumptions that can be avoided if previously known.

We conducted a thorough study of the existing work, among different areas, to learn and understand the most common characteristics and behavioural patterns of human-protocol interaction. Based on the results of this study, we propose in this paper a unified set of human-protocol weaknesses that merges the findings from different research areas into a harmonised taxonomy. In particular, we focus on human characteristics that are usually overlooked during the protocol design process. Based on this set, we then discuss ways we can tackle these weaknesses and minimize their impacts. Our analysis is partly based on related research findings, but is also based on our own proposals which evolved from our taxonomy of weaknesses. Ultimately, this allows us to also present a set of

* Supported by CNPq/Brazil.

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 13–26, 2012.

design recommendations to address the problems inherent in the human-protocol interaction. What is interesting is that this set is not simply a linear evolution of the taxonomy of weaknesses. What we can see is that a second independent layer of structure emerges when we separately consider the categorisation of solutions to the problems we collate and identify in the taxonomy of weaknesses.

In Section 2 we present and discuss related work. In Section 3 we present our proposed set of overlooked components of human-protocol interaction. Based on the that set, we discuss, in Section 4, ways to minimize the impact of those weaknesses. In Section 5 we present the design recommendations derived from the discussion in the previous section. Our conclusions and some opportunities for further research are discussed in Section 6.

2 Overview

Human computer interaction is a topic which spans several different areas, including computer science, sociology and psychology. A large portion of research in this subject is related to design and usability of security systems. Each of these research areas independently address different layers of systems security.

Figure 1 gives us an overview of the layers involved. The specification layer, represents the protocol specification; the application layer contains the implementation of the specified protocol in an application; the interface brings a point of interaction between the application and user layers; finally, the user layer represents a user of the application.

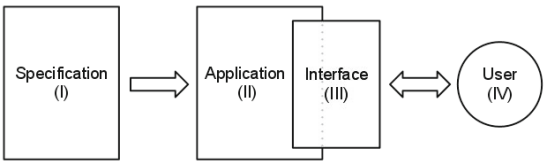


Fig. 1. Human-protocol interaction layers

Software engineering is focused on the application's design and implementation (application layer); HCISec focus on the interface and usability aspects (interface and user layers); computer security design focuses on the design and implementation level (specification and application). As we can see, most of research focus on a specific layer rather than the whole set. This creates a gap specially between the specification and implementation layers where, the human-protocol interaction involved, has received little attention. The human-protocol perspective, which we focus in this paper, goes trough all the layers, such as specification, application, interface and user.

Within the scope of human-protocol interaction we consider any type of action performed by a user that might impact on the security properties of a system. Since the interaction is usually made via a software interface, we have to consider usability issues. Additionally, we need to look at the implementation and specification levels. In a protocol specification, the human-protocol interaction is usually part of the design assumptions, that is, static components where actions are assumed to happen without being explicitly included into the specification. When implemented, these static choices are then replaced by dynamic user-interactions. It is often the case where the assumptions are too strong, that is, it is very unlikely for a implementation to provide the expected security properties. Our work focuses on detecting patterns of problems that occur during the human-protocol interaction and highlight human characteristics that are often overlooked during the protocol design and implementation. We explore the findings of research within these layers to construct a broader point of view.

As mentioned above, different research fields address security issues in different ways. However, we were able to observe overlaps in the definitions. These overlaps occur due to the use of different terminologies and are also due to the distinct research goals. For example, there have been studies of phishing attacks [14,21,29,7,6]; users' susceptibility to attacks [9,24]; factors exploited to allow an attack to be successful [27,8,12,25] and others. Among this wealth of studies, we found similar findings labelled in different ways. We also found that some findings applied to a specific context, which can be extended to different contexts. Each of these existing areas of research independently contributed to the construction of the set of human-protocol interaction weaknesses and recommendations we propose in this document.

3 Frequently Overlooked Components of Human-Protocol Interaction

The interaction between computers is relatively easy to define and the results are, hopefully, predictable. However, defining or predicting human behaviour is a challenging task. It requires a more subtle approach, commonly based on empirical results [16]. Despite this complicating factor, we can create a generic (but not perfect) human-protocol interaction model by using empirical and statistical information and use it to improve the human-protocol interaction process.

Therefore, it is necessary to study and analyse the most common characteristics and behavioural patterns regarding human-protocol interaction. There is a significant amount of research which maps human characteristics (or principles, weaknesses, etc). This existing body of work offers important and relevant insights that constitute the basis of the set of overlooked components of human-protocol interaction we discuss during this section. By analysing existing work and detecting the common components among them, we could define our list of five main overlooked components of human-protocol interaction: user knowledge, authentication capabilities, decision making influencing factors, bounded attention, and inherent characteristics.

3.1 User Knowledge

We define knowledge as familiarity, awareness, experience or understanding of a certain subject. Within the human-protocol interaction context, users' knowledge certainly is an important factor to be analysed. We have seen several situations where this factor (or the lack of it) is exploited by attackers. Therefore, a secure human-protocol interaction should carefully deal with users' knowledge.

Phishing attacks provide a very interesting case study from where we can evaluate human-protocol interaction. The main reason is that, in many cases, the user has to interact with an implementation of the SSL/TLS protocol. Many studies and experiments [8,29,9,24,14] have shown that, indeed, users are not familiar with computer systems, security, security indicators and risks (such as web frauds). When an attacker is able to perceive a weakness in the victim's level of knowledge, it is relatively easy to manipulate and exploit a users' lack of knowledge to successfully attack a protocol. In summary, we can list the knowledge-related issues that are most commonly exploited by attackers:

Lack of knowledge of computing – many people do not have proper understanding of how operating systems, networks and protocols work [8,9,29].

Lack of knowledge of security – users do not have knowledge about digital certificates, cryptography and most of security technologies [8,29,14,9].

Lack of knowledge of security threats – many users do not know they can be attacked; that spoofing websites is possible; and what the techniques used by the attackers are [8,29].

Inaccurate mental models – people frequently construct their own concepts about computing, security and threats. It is often the case these concepts are not correct [1,24].

Human-protocol interaction cannot rely on users' computing/security knowledge and awareness. A secure human-protocol interaction must consider user's knowledge and, preferably not require higher training levels.

3.2 Authentication Capabilities

An authentication performed by a user is a task where the user verifies that the authenticating party is whom it is expected to be. People frequently make use of visual cues as an important authentication tool. However, there are studies [26,27,8,29,14] that show that this visual authentication mechanism is weak and unreliable in some cases. In these specific situations human authentication capabilities should not be used as an important component in the protocol specification. In general, we found four different users' authentication skills that are well known by attackers, but not always correctly addressed by designers:

Users are good at authenticating people they know – in general, users are very good and efficient at authenticating people they know [26].

Users are not good at authenticating objects – usually, users have problems in authenticating objects [26,6,29]. Objects are easy to spoof and when facing a spoofed object, it is likely that a user will perceive it as original.

Users are not good at authenticating strangers – people are not good at establishing whether someone belongs to a designated class (e.g. policeman) [26]. When authenticating strangers, users have to make use of other authenticating factors, such as documents or references given by someone else (e.g. physical attributes). This shifts the authentication type to object-based, which it is not precise enough to be used in security protocols.

Users are not good at authenticating digital objects – in the same way as real objects, digital objects, such as websites, software and email are also not easily authenticated by users. By creating visually identical (or very similar) copies of the original source, attackers can fool users into believing they are contacting the entity they trust [9,14,8].

Directly or indirectly, most scams exploit the false acceptance of a spoofed content by users and almost all scams are forms of deception [27]. Asking a user to authenticate an object (e.g. an online banking website) by checking its elements (e.g. digital certificates, padlocks, etc) does not represent a proper translation from the authentication design goal to its implementation. In fact, it is very likely that the authentication task, despite technically feasible, will not be performed properly, introducing a security breach.

3.3 Decision Making Influencing Factors

There are different factors that need to be taken into account when considering users' decision making and its influencing factors. These aspects include personal and environmental issues. Despite the differences, the core concept behind this component is that users can be influenced to make different (and potentially damaging) decisions to those they would usually make. Thus, when designing human-protocol interaction, security engineers must be aware of which factors may influence the user's decisions and check whether this decision under influence can introduce security breaches or not.

The most common influencing factors found are:

Social conditioning – when people receive commands from strangers, they are unlikely to follow that command without questioning the request. However, when the command comes from a recognized authority (or someone mimicking an authority), people are very likely to obey this command. This happens because people are trained to accept commands from certain people, such as police officers, without further rationalization [27,24,17].

User's principles – victims' principles such as need, greed or dishonesty, make them vulnerable because the attacker can use them to force the victim to behave in a predictable manner [27].

Time constraints – the main idea behind this factor is to push the victim to make a decision without sufficient time to rationalize the decision. Consequently, the actions taken by the victim tend to be more predictable and easier to manipulate [27]. The decision strategies used under time pressure is typically based on affective and intuitive heuristics, rather than on a reasoned examination of all the possible options [12].

Shared risk – this aspect is found in real-world situations where someone accepts a risk because there are many others sharing the same risk [27].

Fear – many techniques such as scareware are effectively used by attackers to scare people and make them fall into attacks (e.g. Mac defender malware[1]).

Users' decision making factors involves many different factors that should be carefully analysed. Even trained users might have their decision strategies shifted under certain circumstances. People will make errors and will eventually make wrong decisions. It is important to identify potential situations where this component might be exploited and make the system insensitive to them.

3.4 Bounded Attention

Users are focused on their main task, and consequently, most of their attention is bound to the activity of performing that task. Security protocols are frequently used as part of a computational system or software. Consequently, from the users' perspective, the protocol used and its security aspects are a secondary concern. As a result users have a tendency to notice only what they are interested in and do not pay attention the fact those security mechanisms were created to protect them from attacks [27,8]. In our research, we found four main factors which can potentially weaken the security aspects of the human-protocol interaction:

Lack of attention to security – user's focus is not on the security aspects of the system. Consequently, security checks are executed less carefully [8,9].

Lack of attention to the absence of security – In the same way that security checks may be dismissed without further rationalization, their absence might not be noticed by the users [8,28].

Security in a secondary workflow – users are more likely to finish their main task rather than stop it due to a security warning. Security checks that occur outside of the main task interrupt the user's focus and are more likely to be dismissed without much consideration. Users will try to finish their main tasks if they believe they are more important then the security tasks, even if there are potential risks [29,5].

Conditioning – an excessive number security interruptions ends up training users to dismiss warnings, pop-up boxes and any other security interruptions in a insecure way because this is the only (or the simplest) way to finish their tasks [2]. A excessive number of warnings, in the course of time, can make users become less inclined to take them seriously in the future [19,5].

As we can see, security should be included in the main workflow to be effective. Simply warning users by stating that something is wrong is not sufficient: they need to be provided with a safe alternative to achieve their goals [29]. Bounded attention may also be affected by user's knowledge. For example, a user may not know what security cues they should look for or whether the operation being performed is insecure or not.

[1] http://support.apple.com/kb/ht4650

3.5 Inherent Characteristics

Human skills is a broad concept. It includes proficiency or ability that is acquired or developed through training or experience. Overlooked inherent characteristics encompasses situations where human skills might not be enough to perform an activity or task as intended. It also includes particularities of human behaviour. We cannot expect that humans behave similarly to a computer, nor believe they share similar skills. By equating these two different components during the human-protocol interaction, a series of security threats may arise. A usable and secure design must consider human abilities and check what people can, and more importantly, what they cannot do well [7]. There are several skills we should consider when designing secure systems:

Memory limitations – human capacity for working memory is limited and decays over time [24]. We cannot expect users to remember large and random keys nor recall dozens or hundreds of different passwords [24,5]. People also cannot "forget on demand". So, even undesired items will remain in memory even when they are no longer needed [24].

Lapses and Slips – lapses and slips errors occur when the plan to achieve a certain goal is correct, but an error happen when a required action is forgotten (e.g. a step in a sequence of actions) or an action performed incorrectly (e.g. pressing the wrong button) respectively [22,5].

Problem solving limitations – some problems can be easily solved by some users but the same problem can be a complex task for others. This limitation can be influenced by many of the previously presented weaknesses, such as lack of knowledge or bounded attention.

Task termination – when users finish their main task, they might leave the subsidiary tasks incomplete. For example, a user accessing a webmail, may leave the computer without logging out. This is ok on a private computer, but is it a security threat in public environments. In a similar manner, users may terminate the interaction if they assume there is no alternative to proceed due to a fault or an unexpected system state [23].

Non-deterministic behaviour – As opposed to the previous limitations, this issue is not related to a limited set of capabilities, in fact it is the opposite situation. According to Ruksenas [23], in any situation, any one of several cognitively plausible behaviours might be taken.

We cannot expect that users have skills or abilities they do not have. The human-protocol interaction should be designed considering the human's inherent characteristics and checking whether the task given to the user is feasible or not.

4 Minimizing the Weaknesses in the Interaction

After merging several research findings into a harmonized and limited set of often overlooked human-protocol interaction components, a set of design recommendations to minimize the effects of these weaknesses is a clear next step. Despite the wide range of users' characteristics and behavioural patterns, we proposed

a set overlooked components of human-protocol interaction, and from them, we could develop a set of design recommendations.

To construct a set of recommendations on how to reduce the impact of these weaknesses, we initially attempted to make a one-to-one association between a weakness and a corresponding recommendation, where, for each weakness, we proposed a design recommendation. We independently analysed each of the influencing factors on the human-protocol interaction weaknesses, and, by making use of our findings and the results of related work we proposed design recommendations for each influencing factor.

What we found was that, for some factors, even those belonging to the same category of weakness, have to be treated in different ways. However, the opposite situation was also found, when factors from different weaknesses could be handled in a similar manner.

Figure 2 represents the associations between our set of frequently overlooked components of human-protocol interaction and our design recommendations. As we can see, almost all of components are linked to two or more recommendations. This happens due to the internal subdivision of each component (described in Section 3). In many cases, each sub-component had to be treated in a different way. Due to space constraints, the description of how each subcomponent is mapped to a design recommendation were left out of this paper.

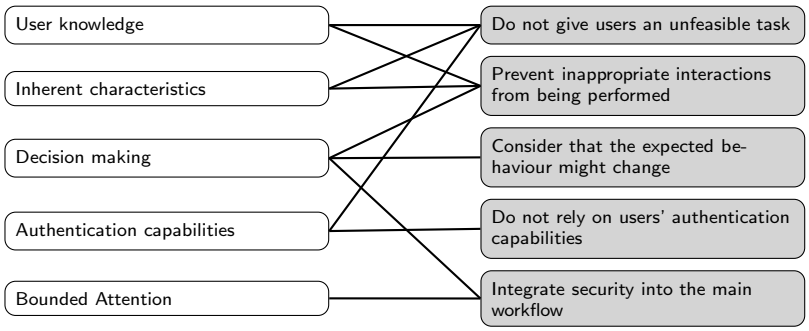


Fig. 2. Mapping human weaknesses into design recommendations

5 Design Recommendations

By analysing the human-protocol interaction weakness we presented in section 3 and by exploring related research findings, we were able to discuss ways to tackle these weaknesses and minimize their impacts. The associations among the weaknesses and solutions to minimize their impacts were a key factor that allowed us to identify the design recommendations we discuss in this section. By proposing a set of design recommendations, we introduce guidelines to help designers to overcome the problems presented in Section 3. Due to space constraints, some examples to illustrate how each recommendation can be applied in real word scenarios were left out of this paper.

5.1 Do Not Give Users an Unfeasible Task

Humans have different levels of knowledge in a wide range of areas. Some people have stronger abilities in subjects such as logic or mathematics and others are better dealing with human sciences and so on. Some protocols might be expected to be used only by specialists and consequently require a higher level of knowledge. However, there are other protocols that are designed for general purpose use and, consequently, used by people that have different levels and areas of knowledge. In both cases, protocols should be designed and implemented keeping in mind the level of knowledge of the person who will interact with it.

An example of a task that depending on the context may be considered unfeasible, is to require the user to generate input data to be used as part of the protocol workflow (e.g. random passwords). Certain inputs can represent fundamental elements of the whole system security and should be carefully analysed. In the Kerberos [18] protocol, a user input (password) represents a fundamental part of the protocol security. If we simply allow the user to create a password, a weak password might be generated and consequently compromise the protocol security [3]. On the other hand, defining password policies (e.g. minimum length) can also generate other types of problems such as information disclosure [13]. Thus, it is necessary to find alternatives to produce input data to a protocol which has sufficient quality to be used as a trustworthy source, as well as make its generation and use feasible to ordinary humans.

The recommendations about not giving users an unfeasible task can be summarized in the following list:

- Identify where the security conditions of the protocol relies on a task performed by users and identify the level of knowledge and skills of the target audience.
- Check whether the task requires specific types of knowledge or skills. If it does require, check whether the target audience attend possesses prerequisites. The more generic the audience, the lower level of understanding and skills should be required.
- Avoid using user input as a main part of the establishment of security properties of the protocol. If a user input is required (e.g. user password used as the seed to a key) and check if it is necessary to create policies to guarantee the good quality of the input.

5.2 Do Not Rely on Users' Authentication Capabilities

People are very good at recognizing people they already know, but they are not good when authenticating strangers or objects [27]. Thus, except for particular cases, such as human-human interaction between people that know each other, we cannot rely on human's authentication capabilities and consequently should not include this task in the human-protocol interaction.

During the SSL/TLS protocol handshake, when an unknown server certificate is presented to the client's browser, users are asked whether they trust that

certificate or not. By being asked that question, users are receiving an authentication task. However, asking users to authenticate an object (a digital certificate in this case) is not recommended because humans are not capable of authenticating digital objects properly, and therefore, this authentication process becomes insecure. In this specific cause, user's knowledge is also not properly considered, since this authentication task a high level of knowledge.

The designer, to avoid security failures due to authentication mistakes, should:

- identify where the security properties of a protocol relies on an authentication task performed by humans [27].
- check whether the authentication task includes authenticating unknown people or objects.
- verify if the authentication task given to the user is feasible for a ordinary verifier, not requiring specific technical knowledge [27].

5.3 Integrate Security into the Main Workflow

When security is a secondary activity for the user, it tends to be ignored or underrated. Warnings, messages and prompts asking users whether to accept a certain change in the security context tend to be ignored by users, compromising the protocol security [25,29]. Moreover, most current implementations are plug-ins or amendments to existing designs which are attempts to overcome inherited design problems. Security concerns about human-protocol interaction should be part of the design and included into the main path of the protocol's flow.

The following recommendations summarize some considerations that should be used during the protocol design:

- If a decision is critical to the security of the protocol, integrate the security concerns into the critical path of their tasks. By doing it, users will be forced to interact with it, and will not be able to ignore it [29].
- Use active interruption other than passive warnings. However, consider the usability impact of the new design to avoid an excessive use of warnings, which may reduce the attention given to the them over time [29].
- Incorporate security decisions into the users' workflow, and, whenever possible, infer authorization from acts that are already part of their primary task [30].
- Respect user intentions. Warning users that something is wrong and advising them not to proceed (while still giving them the option to continue) is not the right approach [29].

In the SSL/TLS protocol implementation, we could apply these recommendations by changing the message presented to the user regarding the untrusted server certificate. Currently a message from the browser to the user is sent via an active warning (a window asking whether the user wants to accept the certificate). Despite some recent changes in the implementation of these warnings (which made them more effective [10]) we still believe that once users learn how to dismiss these warnings, the efficiency of the this type of warning tend to be

reduced. Thus, a third warning type might be needed. We call "interactive warning" a new type of warning that instead of informing or interrupting users, it makes the user interact with the protocol.

In the SSL/TLS certificate warning, an interactive warning could, for example, be implemented by asking the user to input a web address confirmation. Consequently, the user would only be allowed to access the website if the "Common Name" field in the server's certificate matches the address typed by the user. By making this change, attacks that exploits users' bounded attention, for example, would be less effective. Additionally, the effects of deceptive URLs and also the limited authentication problems would be reduced. This solution needs further analysis, however, the idea behind it is to remove the decision (passive/active interruption) from the user by asking him a piece of information (interactive interruption) that allows the system to make the decision. In this case, the security will be integrated in the main protocol's flow. Finally we will be converting a complex activity into a task that a ordinary user can perform.

5.4 Consider That the Expected Behaviour Might Change under Different Circumstances

Human behaviour is likely to change under different circumstances. The ability to influence the users' decision making, discussed in Section 3, includes several factors that might influence users' behaviour. Factors such as social conditioning, user's principles, time constraints and shared risk are efficiently exploited by attackers. It is necessary to avoid situations where a user interaction might be made under influenceable conditions. In general, protocol designers should:

- check whether external and internal changes might influence user decisions.
- avoid asking users for decisions when they might be under influence.

In the web browser implementation of SSL/TLS protocol, for example, the dialog (or screen) that asks the users whether they want to accept a certificate or not is implemented in a way that external factors can easily influence users' final decision. If they are under time pressure, for example, this dialog will probably be dismissed with less reasoning. To avoid this situation, warnings should require further checks, such as the domain name confirmation presented in the Section 5.3. By implementing those changes, the user would be forced to "authenticate" the URL, and consequently, avoiding some attacks. An attack exploiting time pressure, in this case, would be less effective.

5.5 Design Should Prevent the User from Performing an Inappropriate Interaction

The system should prevent the user from performing an inappropriate interaction. Norman [20] introduced the concept of a "forcing function", which aims to prevent a user from behaving in any other way than the correct way. Basically, the forcing function prevents users in progressing in their task until they perform an action which must be taken to avoid a failure. Additionally, the forcing function will only enabled the "safe" options when an action is being performed.

Forcing functions prevent errors where a user skips an important step and condition users to progress with the correct (safe) behaviour. To be effective, efforts (cognitive and physical) required to follow the forcing function must be less than the effort required to circumvent it [15]. Thus, protocol designers should:

- attempt to provide only safe options to users, and avoid giving unnecessary (and unsafe) options when not needed.
- avoid drastic changes in the usability due to use of forcing functions. If the impacts are too high, users will try to find ways to avoid the "safe paths".

Following the previous examples, in the web browsers' implementation of SSL/TLS protocol, the way that the decision of accepting a certificate is implemented does not protect users from making an inappropriate decision. In the case of a spoofed website presenting a certificate, the invalid option (accepting the fake certificate) is still available. On the other hand, predicting users' intentions is, for obvious reasons, infeasible. However, it is possible to "ask" users for their intentions and then check if the actions match the intentions (Brustolini and Salomon implemented such mechanism in [4]) before proceeding. The domain name confirmation presented in the Section 5.3 is an example of a forcing function in this case. The user would only be able to proceed if the certificate presented by the server matches with the server the user wants to have access.

As we can see, this function is very useful. However, we must take care with the usability impacts and trade-offs, otherwise users will attempt find ways of dismissing this feature, whenever possible.

6 Conclusions and Future Work

We have shown that there are many factors that should be taken into account when considering human-protocol interaction. At the same time, there is a wealth of research involving human behaviour analysis and detecting human characteristics that might be exploited by attackers in specific contexts, such as phishing scams and authentication systems. However, despite the existence of similar findings, there is a lack of harmonization regarding the definitions of human characteristics and weaknesses. In this paper, we proposed a set of human-computer interaction overlooked components weaknesses that merges different research findings into a well defined set.

From the first set, we built a set of recommendations to assist designers in the complex task of minimizing security threats from user interaction. The recommendations are based on our findings, related work, empirical analysis and extrapolation from the set of weaknesses presented earlier.

Despite the fact that most of the examples used are based on experiments developed in the key pieces of research carried out to date, further validation of the human-protocol interaction weaknesses and design recommendations against real world systems is an important next step. This will allow us to verify and improve the findings of this work and also the associations among the two sets (weaknesses and recommendations) we propose.

References

1. Adams, A., Sasse, M.A.: Users are not the enemy. Communications of the ACM 42, 40–46 (1999)
2. Anderson, R.: Security Engineering: A Guide to Building Dependable Distributed Systems, 2nd edn. Wiley Publishing (2008)
3. Bellovin, S.M., Merritt, M.: Limitations of the kerberos authentication system. ACM SIGCOMM Computer Communication Review 20, 119–132 (1990)
4. Brustoloni, J.C., Villamarín-Salomón, R.: Improving security decisions with polymorphic and audited dialogs. In: Proceedings of the 3rd Symposium on Usable Privacy and Security, SOUPS 2007, pp. 76–85. ACM, New York (2007)
5. Cranor, L.F.: A framework for reasoning about the human in the loop. In: Proceedings of the 1st Conference on Usability, Psychology, and Security, pp. 1–15. USENIX Association, Berkeley (2008)
6. Dhamija, R., Tygar, J.D.: The battle against phishing: Dynamic security skins. In: Proceedings of the 2005 Symposium on Usable Privacy and Security, SOUPS 2005, pp. 77–88. ACM, New York (2005)
7. Dhamija, R., Tygar, J.D.: Phish and HIPs: Human Interactive Proofs to Detect Phishing Attacks. In: Baird, H.S., Lopresti, D.P. (eds.) HIP 2005. LNCS, vol. 3517, pp. 127–141. Springer, Heidelberg (2005)
8. Dhamija, R., Tygar, J.D., Hearst, M.: Why phishing works. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2006, pp. 581–590. ACM, New York (2006)
9. Downs, J.S., Holbrook, M.B., Cranor, L.F.: Decision strategies and susceptibility to phishing. In: Proceedings of the Second Symposium on Usable Privacy and Security, SOUPS 2006, pp. 79–90. ACM, New York (2006)
10. Egelman, S., Cranor, L.F., Hong, J.: You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 1065–1074. ACM, New York (2008)
11. Ellison, C.: Ceremony Design and Analysis. Cryptology ePrint Archive, Report 2007/399 (October 2007)
12. Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M.: The affect heuristic in judgments of risks and benefits. Journal of Behavioral Decision Making 13(1), 1–17 (2000)
13. Inglesant, P.G., Sasse, M.A.: The true cost of unusable password policies: password use in the wild. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 383–392. ACM, New York (2010)
14. Jakobsson, M.: The human factor in phishing. In: Privacy & Security of Consumer Information 2007 (2007)
15. Karlof, C., Tygar, J., Wagner, D.: Conditioned-safe ceremonies and a user study of an application to web authentication. In: Sixteenth Annual Network and Distributed Systems Security Symposium, NDSS 2009 (February 2009)
16. Martina, J.E., Carlos, M.C.: Why should we analyse security ceremonies? In: First CryptoForma Workshop (May 2010)
17. Mitnick, K.D., Simon, W.L.: The Art of Deception: Controlling the Human Element of Security. John Wiley & Sons, Inc., New York (2003)
18. Neuman, C., Yu, T., Hartman, S., Raeburn, K.: Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. RFC 4120 (Standards Track) (July 2005)

19. Norman, D.A.: Design rules based on analyses of human error. Commun. ACM 26, 254–258 (1983)
20. Norman, D.A.: The design of everyday things. Basic Books, New York (2002)
21. Oppliger, R., Gajek, S.: Effective Protection Against Phishing and Web Spoofing. In: Dittmann, J., Katzenbeisser, S., Uhl, A. (eds.) CMS 2005. LNCS, vol. 3677, pp. 32–41. Springer, Heidelberg (2005)
22. Reason, J.: Understanding adverse events: human factors. Quality in Health Care 4(2), 80–89 (1995)
23. Ruksenas, R., Curzon, P., Blandford, A.: Modelling and analysing cognitive causes of security breaches. Innovations in Systems and Software Engineering 4, 143–160 (2008)
24. Sasse, M.A., Brostoff, S., Weirich, D.: Transforming the 'weakest link' - a human/computer interaction approach to usable and effective security. BT Technology Journal 19, 122–131 (2001)
25. Schechter, S.E., Dhamija, R., Ozment, A., Fischer, I.: Emperor's new security indicators: An evaluation of website authentication and the effect of role playing on usability studies. In: Proceedings of the 2007 IEEE Symposium on Security and Privacy, SP 2007, pp. 51–65. IEEE (May 2007)
26. Stajano, F., Wilson, P.: Understanding scam victims: seven principles for systems security. Technical Report 754, Cambridge (August 2009)
27. Stajano, F., Wilson, P.: Understanding scam victims: seven principles for systems security. Communications of the ACM 54(3), 70–75 (2011)
28. West, R.: The psychology of security. Communications of the ACM 51, 34–40 (2008)
29. Wu, M., Miller, R.C., Garfinkel, S.L.: Do security toolbars actually prevent phishing attacks? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2006, pp. 601–610. ACM, New York (2006)
30. Yee, K.-P.: Aligning security and usability. IEEE Security and Privacy 2, 48–55 (2004)

High Stakes: Designing a Privacy Preserving Registry

Alexei Czeskis and Jacob Appelbaum

University of Washington, Seattle, WA, USA
{aczeskis,ssladmin}@uw.edu

Abstract. This paper details our experience designing a privacy preserving medical marijuana registry. In this paper, we make four key contributions. First, through direct and indirect interaction with multiple stakeholders like the ACLU of Washington, law enforcement, the Cannabis Defense Coalition, state legislators, lawyers, and many others, we describe a number of intersting technical and socially-imposed challenges for building medical registries. Second, we identify a new class of registries called *unidirectional, non-identifying* (UDNI) registries. Third, we use the UDNI concept to propose holistic design for a medical marijuana registry that leverages elements of a central database, but physically distributes proof-of-enrollment capability to persons enrolled in the registry. This design meets all of our goals and stands up in the face of a tough threat model. Finally, we detail our experience in transforming a technical design into an actual legislative bill.

1 Introduction

Washington State, like fifteen other US states and the District of Columbia, has legalized marijuana for medical use [1]. However, Washington State is the only one that does not yet have a medical marijuana registry [16]. This paper details our experiences in helping multiple stakeholders design a legal framework and the technology behind a privacy preserving medical marijuana registry. Additionally, we believe our design to be broadly applicable for many other kinds of registries.

We began by directly and indirectly gathering information from multiple stakeholders like the ACLU of Washington, law enforcement, the Cannabis Defense Coalition, state legislators, lawyers, and many others. Each group had their own goals and agendas, which often conflicted with the goals and agendas of other groups. These interactions, generated many complex design goals, technically and socially imposed challenges, among which was the need to function in the face of a very strong adversary.

As a result, the exercise drove us to study a new class of databases or *registries* that we believe have not previously been discussed in literature or deployed in the wild. Specifically, our proposed registry design does not store any Personally Identifiable Information (PII) – either in digest or encrypted form. Instead, we delegate limited information out to *proof-tokens*, which are given to enrollees (people enrolled in the registry). Enrollees can use the proof-token to prove their enrollment in the registry. Additionally, because it is impossible to indentify enrollees by having access to the registry, enrollees can deny that they're enrolled by hiding or destroying the proof-token.

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 27–43, 2012.

We begin by giving a background of registries. Next, we motivate the need for a new type of registry – a unidirectional, non-identifying (UDNI) registry. We then outline the goals and challenges for a successful UDNI registry design. Next, we provide several example architecture designs and explain why they fail to meet all of the required UDNI goals and break in the face of our threat model. We then present our proposed design and examine it in the context of a detailed case study that covers each aspect of our design in depth. The case study focuses on the proposed medical marijuana registry in Washington State and is grounded in actual facts and concrete discussions. Finally, we discuss what it means to put this type of technology into law, give some pointers on careful implementation, and finish by examining a couple of other interesting, relevant topics.

2 Background

Most modern societies maintain records about people – who they are, where they live, what they are allowed and not allowed to do. These records often manifest in the form of databases or *registries*[1] and are often crucial to how certain aspects of law and order are enforced. For example, in many countries people legally drive only if the state driver's licence database says so and every driver must carry proof in the form of a driver's license while operating a vehicle. As another example, entries in common medical prescription systems tell pharmacies which pharmaceuticals to fill, to whom, and when. As a final example, infectious disease registries record appearances of certain diseases and may be used to detect potentially dangerous outbreaks and epidemics.

Some registries claim to be purely statistical, privacy preserving, or even anonymous. They try to achieve this goal by utilizing a variety of well-studied techniques like *k*-Anonymity [19] or Differential Privacy [9]. Indeed, the "privacy in databases" community is quite rich with literature that provides models and metrics for achieving some *reasonable* level of anonymity for entries in a database. The majority of these techniques give a heuristic for how much PII a database maintainer must trim (or how much noise they should add) in order to get a set of data that is lean to a point where a person can be mapped into a large set of entries (instead of just one). Other techniques deal with how many and what kind of aggregate queries should be permitted on a PII database in order to maintain anonymity and privacy for any registered person's data. Dwork gives a good survey of database privacy techniques in [10]. Nevertheless, attacks against these metrics do occur [15], and statistical registries can morph into identifying registries overnight.

However, a portion of registries (currently implemented as identifying registries) only require a uni-directional link to function – that is, the registry's sole use is for people to prove that they are "enrolled" in the registry. These registries are not designed to enumerate members or to store any information about the enrolled members other than the fact that they are enrolled. We call such registries *uni-directional non-identifying registries* or UDNI registries for short.

[1] Note that the terms *database* and *registry* are often interchangeable. However, *registry* is often used to refer to a holistic system (including people that use it); this is why we default to using this term.

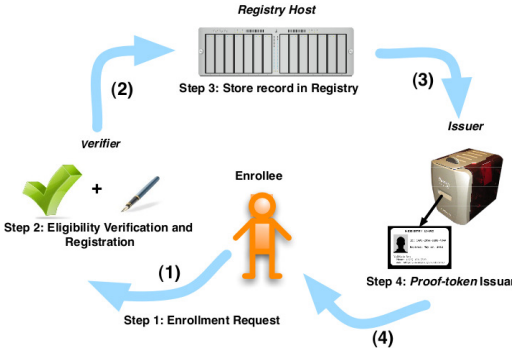


Fig. 1. Generic enrollment process

Fig. 2. Proof of enrollment

TERMS AND DEFINITIONS

Before diving deeper into the registry sea, we first present some common terminology. A *verifier* verifier checks whether a person is entitled to be listed in the registry. If so, the verifier registers or enrolls the person in the registry, which is stored by the *registry-host*. Once enrolled, the person is referred to as an *enrollee*. If supported by the registry, an *issuer* provides the *enrollee* with proof-of-enrollment, called *proof-token*. While proof-tokens can come in a variety of form factors, we assume that it will be physically manifested in the form of an ID card. Figure 1 demonstrates the enrollment process. Post enrollment, an *enforcer* can ask an individual to provide the proof-token or face the legal consequences of not being enrolled. This process is illustrated in Figure 2.

3 Motivation and Goals

In this section, we'll discuss examples of registries that could be made to be UDNI, what challenges arise with providing security along with functionality and privacy, and outline some of the desired goals for a privacy preserving UDNI registry along with a threat model it must protect against.

3.1 Motivating Examples

We provide two brief motivating examples of how UDNI registries can be used and where designs can fail.

EXAMPLE 1

Bob has finished his undergraduate degree in political science and has just been accepted to law school. Unfortunately, Bob has also been diagnosed with cancer. Bob begins chemotherapy (chemo), but the chemo-induced nausea and vomiting make it difficult for Bob to study. Since Bob doesn't seem to respond to standard antiemetic drugs, Bob's doctor suggests that Bob try Medical Marijuana, as some studies have shown it to work in such cases [17]. Bob is afraid of being arrested – he has seen news articles about patients who are mistakenly arrested by police and are only able to prove

their innocence much later. Bob's doctor recommends that Bob register with the state's Medical Marijuana Registry (MMR), which will issue Bob a card that he can carry in his wallet and present to police in case there is ever a question. Bob's doctor says this registry is also private and secure. Bob agrees to try Medical Marijuana and follows the doctor's advice to join the registry.

A couple years later, Bob's cancer is gone, his career has taken off, and he decides to run for public office. Based on an anonymous tip, the opposition hacks the server that hosts the database for the state's MMR and releases the database anonymously online. The opposition then issues an ad saying that Bob is a drug addict and is probably currently using other drugs. Bob is shocked and tries to explain to the voters that he was using marijuana by his doctor's recommendation, but the opposition's tactic of shock and awe have won – before the voters are able to logically think through all of the facts, the election is over and Bob has lost.

DISCUSSION

In this example, Bob followed his doctor's suggestion and used a controversial, recommendation-only medicine in order to stem his chemotherapy induced symptoms. Bob also registered in the state-provided registry in order to receive a state-issued card that would protect him from arrest if he were ever stopped with possession of the medicine. Unfortunately, the state registry stored enough information so that when the registry was compromised the attackers were able to uncover Bob's name from the registry and cause irreparable harm to his reputation.

EXAMPLE 2

Sue loves the wilderness, especially fishing. Her sister Mary, however, finds fishing and hunting distasteful – so much so, in fact, that Mary and Sue have had numerous arguments about this issue. Sue likes to hunt and fish, but she doesn't want to feud with Mary either. Sue decides to try to keep mum this season and avoid confrontation with Mary. This strategy seemed to work so well, that when Mary one day did ask Sue whether she still hunted and fished, Sue automatically said "no".

Sue decided to go fishing one last time and had great success – she caught a huge fish. On the way back, however, she was stopped by the local park ranger and asked to present her fishing license. Sue's fishing license had her name on it, which the ranger noted and included in his daily report. The ranger submitted his report at the end of the day to the office assistant, Mark. As Mark typed up the report, he noted Sue's name. Mark was friends with Mary and he thought it would be amusing to let her know that he had come across Sue's name in the papers. Mary was furious; not only had Sue continued to fish and hunt, but she also lied to Mary.

DISCUSSION

In this example, Sue wanted to keep her hobby private, but she had to enroll in a state registry because her hobby required a state issued license. Unfortunately, the license contained Sue's name, which was recorded by a ranger during a routine check. In this manner, Sue's hobby was disclosed against her wishes; this is a kind of misclosure that could be avoided by design.

3.2 Goals and Challenges

Using these examples, we now derive goals for a UDNI registry. These goals are the result of consulting with multiple stakeholders during an actual medical cannibis bill [3] drafting process that was later signed into law[2].

FUNCTIONALITY GOALS
First and foremost, the registry must be functional. The registry should support at least the following features:

- **[G1]** *Controlled enrollment* – Only those persons that actually belong in the registry can be enrolled.
- **[G2]** *Provable enrollment* – If they so desire, the enrollee can provide proof of enrollment. This "proof of enrollment" must pass police "muster" (the police must accept this as valid proof). Furthermore, no non-enrollees should be able to claim enrollment in the registry (the proof should be reasonably hard to fake or forge).
- **[G3]** *Deniable enrollment* – Enrollee may deny enrollment in the registry if they so desires.
- **[G4]** *Revocable enrollment* – An enrollee may be removed from the registry upon a valid request.
- **[G5]** *Expiring enrollment* – The registry must support the expiration of entries after a certain fixed period of enrollment. Note that this goal differs from *revocable enrollment* because the former refers to revoking an enrollment based on name, while the later refers to revoking enrollment based on enrollment date.

SOCIALLY IMPOSED GOALS
Multiple stakeholders are involved in the registry system and often, they can have conflicting goals and complicated relationships. This results in subtle, but architecturally interesting goals and tensions. In order to maximally satisfy this criteria, the UDNI registry must support:

- **[G6]** *Inexpensive implementation and maintenance* – Ultimately, the registry-host role will most likely be filled by a government organization. Furthermore, the registry itself will likely generate little or no income. Consequently, the registry must be inexpensive to implement and maintain.
- **[G7]** *No new proof-of-enrollment hardware or software* – Proof-of-enrollment should be possible without specialized hardware or software for the enforcer. Practically speaking, law enforcement is reluctant to add hardware or software to their existing tools for officers in the field. This will not only facilitate quicker adoption, but will also help satisfy G6.
- **[G8]** *No social stigma or unintended consequences* – The enrollment status should be non-obvious to casual onlookers (e.g., proof-token color and features should be considered in the context of other identification systems in current deployment).

[2] Most of the registry system was line item vetoed with a suggestion for it to return in a bill by itself.

SECURITY GOALS

Adding further dimensions to the design, is the possibility of a powerful attacker. We assume the attacker is able to:

- *Mount Network attacks* – The attacker can perform all known network based attacks. For example, the attacker can hijack DNS or perform man-in-the-middle attacks.
- *Steal the Registry* – We assume that the attacker will be able to steal the registry. Once database is stolen, the attacker can execute very powerful brute force attacks. Additionally, we assume the attacker knows the full domain of all possible entries in database (e.g., the attacker knows the names of people residing in a particular region).
- *Employ Social Engineering* – The attacker can threaten or socially engineer the maintainers of the registry into accessing the registry and reading back values.

Given the broad abilities attacker, we claim that the registry should have the following security goals:

- **[G9]** *Minimal PII required to enroll; destroyed after use* – The registry may require certain linkable pieces of PII, but such information must not be kept beyond the time needed to produce and distribute a proof-token.
- **[G10]** *No PII in registry* – The registry must not store any PII. This will assert that a compromise of registry does not reveal identities of enrollees.
- **[G11]** *No external identification requirements* – Enrollees must not be required to carry or produce any additional documents in order to prove enrollment in the registry. This will assert that no PII is ever transmitted during the proof-of-enrollment phase and will prevent the network attacker from gathering any useful information.
- **[G12]** *Positive verification does not produce PII* – Positive Verification should not create additional PII details. Note that a negative verification may, however, produce PII related data in the context of fines court proceedings, or other law enforcement actions.

4 Architecture

While many registry architectures are possible, very few actually meet all of our desired goals and stand up to our threat model. In many cases, it may be very subtle or seemingly unintuitive why a certain design may fail. To this end, we first discuss some promising, yet flawed architectures before finally proposing our design.

DESIGN 1: PII DATABASE

The first registry design to consider is one that stores all possibly relevant data. For example, the registry could store the names of enrollees, which entity served as the verifier and when. To prove enrollment, an enrollee could present any acceptable identification (e.g. a driver's license) to the enforcer (e.g. policeman), who would then verify enrollment by calling the registry maintainer or visiting a portal and entering the ID information. This will clearly enable some functional goals like `G1`, `G2`, `G4`, and perhaps `G5`/`G6`/`G7`, but it will fail to meet `G3` in a very serious way. Socially imposed goals such

as G8 and G9, G10, G11 and G12 are nearly impossible to satisfy with such a simple design.

If the database or the verification mechanism were ever compromised, then confidential enrollee PII would be obtained by the attacker[3]. Moreover, this system is subject to the whims of insiders – an employee may be coerced into disclosing sensitive data or could confirm the presence (or absence) of a specific individual in the database based on personal or financial interests [6, 7, 21].

Similarly, this class of designs encourages extremely unsafe practices such as the collection of large amounts of PII during registration and proof-of-enrollment. This creates an environment where large amounts of PII is collected, transfered, and/or stored by an unknown number of parties – any of which can record or expose it in unauthorized manners. Furthermore, this design relies on traditional ID cards, which contains a large amount of PII that is irrelevant to the registry. Finally, this type of a design carries with it a negative "big brother" social stigma.

DESIGN 1.5: SPRINKLE IN ENCRYPTION/BLINDING

A better decision would be to encrypt stored PII. One approach may be to encrypt the database using a single key. While at first glance, this may appear to help protect against an attacker who is able to "steal" the database, this design still likely fail because we assume the attacker will likely be able to gain physical access to the database machine and thus compromise the decryption key [14]. Additionally, the encrypted database will also not survive in the face of a malicious or coerced employee (who could access the entire database).

A more sophisticated approach may be to encrypt each database entry with a different key and store each key (or a password used to derive the key) on the *proof-token* that's issued to the enrollee. This approach would indeed distribute the encryption keys in such a way, that each enrollee's information would only be accessible if given access to their proof-token. In order to verify an enrollee, an enforcer could either transmit their encryption key to the registry maintainer and receive decrypted data or fetch particular encrypted data from the registry and decrypt it locally using the proof-token decryption key. In the first case, transmission of the decryption key makes it vulnerable to recording by a malicious registry employee. In the later case, the enforcer would need specialized equipment – directly violating our goal G7. Furthermore, in both cases, the enforcer could record or photocopy the proof-token (or the enrollee could lose it) – completely revealing the enrollee's data. The same logic holds for a system that relies on enrollee-remembered passwords, except with the added complexity of enrollees forgetting passwords (especially in times of distress). Extensions such as using one-time-pad encryption or re-encrypting the enrollee's data also fail (either because of complexity and cost of implementation or because of the same reasons as plain encryption).

These solutions still require the extensive collection of PII for verification and enforcement, and still carry a negative stigma of having the enrollee's data stored in a database. Although technically-savvy people often understand the protections offered by encryption, others don't and forgo the benefits of such systems because of perceived

[3] Interestingly, as a possible feature creep some states have actually been known to offer information in such databases up for sale [12, 13].

privacy concerns – we found this to be true in our conversations with people who work closely with existing registries.

DESIGN 2: HASH DATABASE

Instead of storing enrollee PII in a database, the registry could store a one-way digest of enrollee PII. For example, the registry may store a hash of the enrollee name or driver licence number. Note that the hash must also include a per-enrollee secret, otherwise a stolen database can be brute-forced by an attacker who can easily discover the domain of all possible enrollees. In any case, this class of designs faces the same problems as *Design 1.5* above: in order to verify an enrollee's, the enforcer would again require special hardware or would need to send enrollee PII to the registry maintainer for verification. Both options are unacceptable in the context of our goals and threat model.

DESIGN 3: NO DATABASE

Having run into fundamental problems using a central database, we now turn to exploring fully distributed approaches (completely lacking a central database). One approach in this space could be to issue proof-tokens containing encrypted data to enrollees. This could be in the form factor of a card with a hexadecimal string on it. The tokens would at the least have to encode the enrollee's identity (to prevent forgery and impersonation) and an expiration date (to support goal `G12`). In order to access the decoded data, an enforcer would need to decrypt the data. As a variation of the above, proof-tokens could contain unencoded data along with an authenticating signature. The cryptographic signature would cover all of the data on the proof-token. To verify the authenticity of the token, an enforcer would need to verify the signature of the data.

This design is attractive, but unfortunately, verifiers would be required to carry specialized cryptographic equipment. Additionally, in order to mint proof-tokes, PII would need to be collected – making this a nonviable class of designs.

DESIGN 4: UNLINKABLE TOKEN/DATABASE HYBRID

In order to eliminate the need for the enforcer to collect/transfer PII or carry additional devices, and to remove PII from the database, we propose a hybrid token/database system. In this design, enrollees will be issued a proof-token in the form of a card with their photo, a random nonce, and an expiration date (see Figure 3). The registry database will store the issued nonce and the associated expiration date – it will *not* store the associated photograph. The card will be printed with the same anti-forgery techniques (e.g., lenticular printing or watermarking) that are deployed for other government issued IDs. In order to verify the validity of an ID, an enforcer must check that it contains all of the required anti-forgery signs, that it has a valid expiration date, and that the photograph matches the enrollee in question. Note that law enforcement, park rangers, and club bouncers are accustomed to doing all of these steps already. For extra verification, and to check for revocation, the enforcer could contact the registry (either via a phone or via and verify that the nonce has not been revoked).

Note that this design does not require the database to store any PII – stored data could be made public without any negative consequences to enrollees. Also observe that the enrollee can prove their enrollment by presenting the proof-token or deny enrollment by hiding or destroying the token. The storing of the random nonce in the database allows for token revocation and the presence of the expiration date on the proof-token

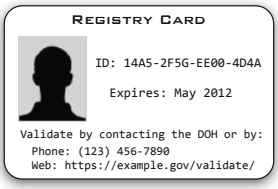


Fig. 3. A proof-token as per *Design 4*

	G1	*G2*	*G3*	*G4*	*G5*	*G6*	*G7*	*G8*	*G9*	*G10*	*G11*	*G12*
Design 1	**Y**	**Y**	N	**Y**	M	M	M	N	N	N	N	N
Design 1.5	**Y**	**Y**	M	M	M	M	**Y**	N	N	N	N	N
Design 2	**Y**	**Y**	**Y**	**Y**	N	**Y**	N	**Y**	**Y**	**Y**	N	N
Design 3	**Y**	**Y**	**Y**	**Y**	N	**Y**	N	**Y**	**Y**	**Y**	N	N
Design 4	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**

Fig. 4. How each design meets the system goals (Y = yes, N = no, M = maybe)

allows for easy expiration checking. Also note that enrollment only requires a photo (which will not be stored, but only used during the issuance phase). Finally, enforcers will not be required to carry additional equipment or collect PII to check an enrollee's proof-token.

We go further in-depth regarding this design in the following section, where we present a case-study of an actual system currently being developed.

5 Case Study

We now analyze our system as it would fare if adopted by the medical marijuana registry (MMR) currently being considered (but not yet not implemented) in Washington State.

5.1 Background and Assumptions

BACKGROUND

In the United States, fifteen states and the District of Columbia have approved marijuana for a variety of medical uses [1, 16]. However, because cannabis is not approved for medical use on a federal level, it cannot be regulated through the regular prescription system like other controlled substances such as hydrocodone or morphine. This means that doctors cannot issue prescription for qualifying patients, patients cannot obtain medical marijuana at pharmacies, and police have difficulty determining whether a person is a criminal or a patient in pain when in possession of marijuana. However, doctors can talk to patients about medical marijuana, make recommendations and record them in patient records. The doctor can provide a copy of this recommendation to the patient, who can then use this medical documentation as part of a legal defense against prosecution in court. Nevertheless, medical marijuana patients could still be arrested and detained by police – a fairly large inconvenience to sick people (especially if they are in pain) and a matter of public record.

In order to clear the regulatory haze, fifteen out of fourteen states (and the District of Columbia) have begun to design and deploy medical marijuana registries [16]. These registries enable law enforcement a way, sometimes a quick way, to verify a patient's legal status and offer patients protection against unwarranted arrest, search, and seizure.

These registries are born into an interesting ecosystem. They enable enforcers to identify persons possessing marijuana legally under State law, but at the same time they

identify people who may be breaking Federal law, which does not exempt medical use of marijuana from criminal liability. Enrollees want to be able to prove enrollment in some cases, but be able to deny it in others. Law enforcement wants easy verification of enrollment, but no additional equipment to do it. The State wants cheap and quick implementation and maintenance. Moreover, because Federal seizures of State material have occurred, the threat of physical or legal removal of the registry database [5] and arrest of enrollees is quite real.

ORGANIZATIONAL ASSUMPTIONS

We assume the proposed registry system will be run by the state Department of Health (DOH). The DOH will issue cards that stand as proof of enrollment in the registry. Additionally, the DOH will confirm doctors as eligible to make medical recommendations to patients and will keep on file the confirmed physical mailing address of the authorized doctors. We do not assume that the DOH is trustworthy in every way and we assume that they may even be subject to a subpoena or a National Security Letter (NSL) supported by a gag order that prevents them from disclosing the receipt of such a subpoena.

Note that every state DOH subject to Federal legislative action and is not willing to expose state workers to a large level of controversy; the DOH may be willing to issue privacy preserving cards but they may not wish to take on the liability of having a database of PII under their purview.

Additionally, we assume that physical mail will not be subject to constant monitoring and inspection beyond cursory recording of source/destination addresses. Specifically, we believe that the Federal government will not seize all outgoing mail for the DOH and record the contents of every letter starting at the first day of operation. We additionally assume that a doctor's office is not automatically subject extralegal action such as a document retrieving raid without due process. A doctor's records are a likely target but the actual doctor who writes a recommendation is decoupled from the registry entry after it is issued.

5.2 Registry Enrollment and Card Issuance

The enrollment phase consists of several steps. We examine each in turn below in the context of a fictional patient Robert and his physician, Jane.

STEP 1: DOCTOR-PATIENT RECOMMENDATION

Robert is an elderly man with serious pain management issues. Dr. Jane suggests that Robert enroll in the medical marijuana registry run by the State. Dr. Jane explains the system to Robert and affirms that she considers it to be a fairly safe system with only a small number of tradeoffs. Robert decides to enter the registry based on the advice of his doctor.

STEP 2: REGISTRY ENROLLMENT

Dr. Jane connects to the DOH registry website and authenticates with her account credentials as issued by the DOH. To further reduce the PII held by the Department of Health, Dr. Jane uses The Tor Browser bundle [2, 8] to access the website[4]. She requests

[4] This prevents local observers from noticing that she is connecting to the DOH – it also prevents the DOH from having an IP address in their logs that is meaningful after her session has ended.

that the DOH issue a valid card, and she submits the only required piece of information – a photograph of Robert that is to the same standard as the state driving license.

The whole enrollment process is shown in Figure 5. Note that it differs from the enrollment process shown in Figure 1 with the addition of using the doctor as a privacy preserving *proxy*[5].

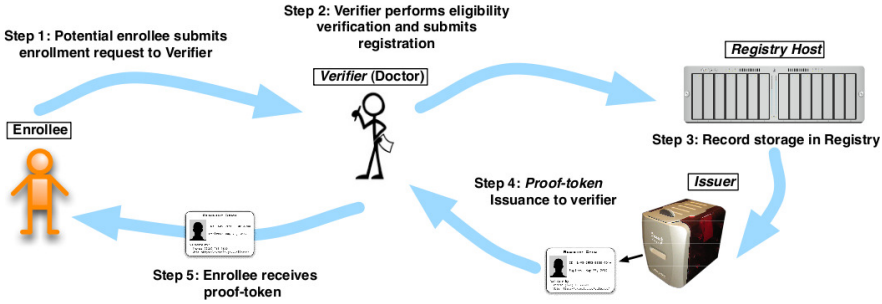


Fig. 5. Privacy preserving enrollment process

Step 3: Card Issuance

The DOH takes care to not store this photo during or after the card production process. This is performed in the following manner:

1. The DOH computer system generates a random nonce that is unique for Dr. Jane's session and it automatically creates an expiration date one year in the future.
2. The DOH system stores the random nonce and the associated expiration date in the DOH MMR database.
3. The DOH system simultaneously prints a plastic card with Robert's image, the random nonce, and the expiration date.
4. The DOH system also prints an addressed envelope (if one hasn't already been printed that day) and a receipt that tells an operator to place the card with a particular nonce into a particular envelope (at the end of the day, the envelope is sealed and mailed to Dr. Jane's office).

Note that the above operations are performed as an atomic, blocking transaction – they either all succeed or all fail and Dr. Jane (or her nurse) must wait for the process to complete. If the process completes, the DOH effectively assures Dr. Jane that the stated registry system processes have completed as expected. Finally, Dr. Jane is presented with a receipt number, which she writes down in Robert's file. Robert does not have the protections provided by the registry until the card arrives and until that time, he has whatever protections are provided by Dr. Jane's recommendation letter.

A week passes and Robert returns to the office and meets with Dr. Jane as he would during any office visit. Dr. Jane has received the letter from the DOH and opens the envelope; inside she finds other envelopes with the appropriate receipt number for Robert.

[5] We did not include the *proxy* in the previous figure because it is a generic explanation of enrollment roles; the proxy is a privacy-adding, non-standard role.

Dr. Jane hands Robert his card in private and Robert examines the envelope to notice that it is sealed and appears in an untampered state. Robert breaks the DoH seal and together with Dr. Jane they confirm that this is his valid card. For her records, Dr. Jane records the number on the card in Robert's file. The receipt number is no longer retained and Robert inspects the card. Robert inspects the card and notices important key features:

- Robert's photo under a Lenticular coating and other anti-forgery features such as a holograph of the State Seal
- An easy to read registry number with an expiration date that is set to expire in a year
- A secure (HTTPS) URL for the DOH verification website and a toll free number to call for verification
- It states that it is "State Issued Photo ID"

The card otherwise blends in with the other cards in his wallet.

5.3 Enforcement and Proof-of-Enrollment

DECIDING TO DECLARE PROOF OF REGISTRY MEMBERSHIP
On the way home from the doctors office, Robert stops at a local medical marijuana dispensary in order to purchase his recommended medication. Following standard procedure, the dispensary operator challenged Robert for proof that he was a qualifying patient. While looking around the shop it became clear to Robert that he had no intention of revealing his name to complete strangers and he opted to use the privacy preserving registry card (instead of his patient records).

The dispensary operator verified the photograph visually and as a final step of the verification process, the operator launched a copy of the Tor Browser Bundle and visited the secure website for the Department of Health[6]. The dispensary operator entered the registry card number and submitted it for verification. The DOH website verified that this card was valid, not expired, and not revoked. The dispensary operator was now fully satisfied that Robert was currently eligible to purchase goods from the dispensary and warmly welcomed a new customer.

REQUESTED TO DECLARE PROOF OF REGISTRY MEMBERSHIP
On the way home, a police officer observes Robert produce a marijuana-like substance from his bag. Following standard procedure, the officer decides to check if Robert is carrying an illegal substance and prompting Robert for an explanation. Robert produces his state issued registry card. The officer verifies that Robert matched the photograph on the registry card, that the card carries all of the required anti-forgery features, and then proceeds to call the dispatch to verify the details on the card.

Robert is entirely compliant and waits while the police officer receives confirmation over the radio. While Robert has no idea if the police radio is encrypted or if the phone call from dispatch to the Department of Health is somehow secure, he feels content that

[6] The operator also made sure that the HTTPS certificate matched the well known certificate for the DOH.

none of his private information is being transmitted since he did not give the officer any PII to transmit. Meanwhile, the police officer reads off the details on the card, waits, and hears the dispatch officer respond that the card is indeed valid, non-expired, and non-revoked. The officer thanks Robert for his compliance, tells Robert that he's free to go, and wishes him well with his treatment plan.

5.4 Registry Database Compromise

THE DEPARTMENT OF HEALTH HAS A MALICIOUS INSIDER
Shortly after Robert enrolls in the registry, a DOH employee gains access to the Department of Health registry database and sends it to the local newspaper. However, the newspaper is only able to extract and print the total number of valid cards in the registry, the number of cards that were revoked and the dates of expiration.

THE DEPARTMENT OF HEALTH IS INVESTIGATED
After the high profile leak there is a surge in enrollment by many people who previously feared entering into the registry. This surge attracts the attention of the Federal government. Law enforcement agents raid the state DOH and seize the registry database[7].

The Federal agents involved are unable to extract any specific patient names. Furthermore, no state employees are interrogated or prosecuted as there is no information that could be gained from them besides what is already known. Such an activity is, by its very nature, disruptive to ongoing card issuing attempts and no further information collection is possible after service is disrupted.

5.5 Renewal and Unenrollment

RENEWAL OF REGISTRY MEMBERSHIP
Robert finds that his treatment plan has worked well for him and after one year he asks Dr. Jane to renew his enrollment status in the registry. He schedules an appointment and goes through the enrollment process with his doctor as before.

CANCELING REGISTRY MEMBERSHIP
Robert decided that he no longer needs to use medical marijuana for pain management and consequently does not require protection provided by the registry. Even though the card will expire, Robert cuts his card in half and mails the number half of his card, without his photograph, to the DOH from a public post office. He does so without a return address.

6 Discussion

MISTAKING CRYPTOGRAPHY FOR A PANACEA
A privacy preserving registry system ensures that many privacy properties that were once a function of policy become a key part of the actual technical design. Privacy

[7] Across the United States there are currently legal battles and law enforcement raids whereby the Federal government is attempting to force the disclosure of the list of enrolled patients.

as a function of design is an absolute necessity when deploying a system in a legally hostile environment. While cryptography can often turn a policy goal into a technical reality, it's not always feasible to deploy because of confusion, cost sensitivity, lack of trust in perceived to be complex systems, and fear of serious legal or physical consequences. For example, a system without a photograph with binding to a name, with a per name secret is an entirely reasonable security system – it is also a system that only an expert can understand and is nearly impossible to deploy in a way that will not enable coercive disclosure of real names by verifying parties. Furthermore the world may someday be ready for fully anonymous credentials but the first deployments will be extremely difficult and world-shifting for law enforcement and enrollees alike. The system we propose makes a small anonymity compromise at the level of enrollee tokens by including an image. However, this compromise neither enables easy privacy violating attacks nor adds PII to a central system – it does, on the other hand, make a system that meets socially imposed restraints. We believe that this is a great improvement over the status quo.

DRAFTING LAWS FOR PRIVACY BY DESIGN REALITIES

A group or a person wishing to write a privacy preserving law would be well served to carefully and specifically phrase certain design goals in a registry creation bill. A concrete example is to ensure that the bill will not permit collection of unneeded PII and to ensure that any such data is kept in a one-way, non-reversible format. Encryption is simply not enough for many high risk registries – disclosure of cryptographic keys may be accidental or forced through any number of means (e.g., rubber hose cryptanalysis) – the stakes may simply be too high for designs that allow for both forward and reverse queries of the dataset.

Some stakeholders completely reject the concept of a registry at all costs [4] because of privacy concerns. Indeed, there may be very compelling reasons to ask if a registry is really the step that society wishes to take, especially given the concerns that a poorly designed registry may pose to otherwise lawful citizens. However, when a registry must be deployed, we believe that it is imperative to reduce the total PII to the absolute minimum level possible.

PRACTICAL IMPLEMENTATION ISSUES

The devil is often in the details and practical implementation decisions can often make or break the privacy properties of a registry. For example, in the case study we presented above, the following details need to be taken into account:

- The unique registry identifier must be chosen from a uniformly random set and must be globally unique; we assume that the token does not need to be easy for a human to remember and so Zooko's triangle [20] is not a problem for any of the parties involved.
- To slow any verification process that may become a registry identifier oracle (allowing a forger to guess valid registry identifiers), we strongly suggest only allowing queries to be performed on an ID in conjunction with an expiration date. A forger would have to guess both in order to receive a valid answer. Additional rate limiting would also help to curb abuse.
- Data retention in such a registry is an extremely important issue – while we discuss not retaining PII, including IP addresses, of the enrollee, we need to also stress the

importance of erasing data as soon as it is no longer needed. By not having data, the registry prevents a "data valdez" [11] incident from occurring.
– Any database field that is unique per-person must be stored in a one-way, non-reversible manner. It may make sense to protect some non-enrollee PII data with a scheme like scrypt [18].

PRACTICAL COMPROMISES

Often in system design, and especially in security, one has to make tradeoffs and compromises. For example, one fundamental property of an anonymous registry is for the ability of some enrollees to have duplicate entries. Observe that since enrollment is deniable, there is no way to verify whether or not a particular person has already been enrolled in a database or not. As another example, consider how revocation is impacted by providing deniable enrollment in a system. Let's observe how revocation would work in the context of our case study. Proof-token IDs can be marked as revoked – making the proof-token cards invalid. However, in order to actually revoke an ID, one has to discover which ID number to revoke. Because the system we present provides deniable enrollment, an enrollee can always hide the fact that he's enrolled and never give up his ID number – making it hard to revoke. Note that doctors may also have the enrollee's ID number, on file, but fishing for that a particular enrollee's doctor is a difficult, privacy violating, and legally murky endeavour.

Taking a step back, this is a traditional tension between privacy and security where more information causes privacy concerns for the enrollee, but more data may (or may not) provide greater security. Often a suitable security tradeoff is nevertheless possible without sacrificing much privacy – for example, we don't currently require full body scans or retina scans to have a drivers license – a close match is good enough for verification by a law enforcement officer.

LAST BITS OF ADVICE

Writing technology and privacy solutions into a bill is a challenging topic and requires meeting with people from many different groups and with a varied set of incentives. It is not possible to make every stakeholder happy as some stakeholders hold on to opinions based on idealogical grounds and are not willing to compromise.

7 Conclusion

In this paper, we discuss some of the key privacy properties offered by existing registries and discover that a fundamental space exists for unidirectional, non-identifying (UDNI) registries. Through conceptual investigation, extensive discussions with multiple stakeholders, and empirical analyses of current designs, we outline the goals for a holistic UDNI registry that takes into account not only technology, but also the people that will be using it. For example, some key goals that we meet include *provable, but also deniable enrollment* and we have *no PII stored in registry*. We also mandate that the registry not require any additional hardware or software beyond what it will take to store and operate the central registry system (i.e., the police won't have to carry any additional hardware). We complicate the design space further by supporting a sophisticated and strong adversary, who can not only interpose on all network traffic, but can also physically steal the servers on which the registry resides.

Next, we generate and systematically analyze various candidate systems and examine why and how they fall short of our goals and threat model. Interestingly, we find that complicated cryptographic techniques are insufficient to solve our problem. Instead, we propose a hybrid system that leverages elements of a central database, but physically distributes proof-of-enrollment capability to persons enrolled in the registry. This design meets all of our goals and stands up in the face of our threat model.

We explore our design further in the context of an actual case study focused around the medical marijuana registry currently being discussed in the State of Washington. Finally, we discuss how we translated our technical design into legal language, which we then helped incorporate into a Washington State bill that was recently signed into law.

This paper contributes a deep exploration of privacy and anonymity issues in certain types of registries, serves as a case study in holistic system design, and provides our experience in transforming a technical design into legalize for inclusion in a bill.

Acknowledgements. This publication was made possible in part by Grant Number HHS 90TR0003/01. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS.

A giant thanks to everyone who knocked down our early designs, gave us their stakeholder feedback that we could not ourselves see or understand, and read our early drafts. A special thank you to Alison Holcomb, Brian Alseth, and others at the ACLU of Washington for their insight, support, and guidance. Also, a great thank you to John Gilmore for his brilliant outlook and deep understanding of all of the underlying issues. Thank you to Zooko Wilcox-O'Hearn of Least Authority Enterprises, Andy Isaacson & Leif Ryge from Noisebridge. We're especially grateful to Tadayoshi Kohno from the University of Washington, Kelly Caine from the School of Informatics and Computing at Indiana University, Dr. Nadia Heninger from Princeton University, and Phillip Mocek and Ben Livingston of the Cannabis Defense Coalition. Thank you to everyone else we forgot to mention.

References

[1] RCW 69.51A.010, Section 4, http://apps.leg.wa.gov/rcw/default.aspx?cite=69.51A.010

[2] The Tor Browser Bundle, https://www.torproject.org/projects/torbrowser.html

[3] WA Senate Bill 5073, http://apps.leg.wa.gov/documents/billdocs/2011-12/Pdf/Bills/Session%20Law%202011/5073-S2.SL.pdf

[4] Hands off Washington Patients (2011), http://cdc.coop/registry

[5] ACLU of Washington. Medical marijuana patient records are private, court rules (2007), http://bit.ly/lPODeY

[6] Auckland Stuff.co.nz. Staff pry into files of celebrity patients (2009), http://www.stuff.co.nz/auckland/local-news/130205

[7] Ornstein, C.: Fawcett's cancer file breached (2008), http://articles.latimes.com/2008/apr/03/local/me-farrah3

[8] Dingledine, R., Mathewson, N., Syverson, P.: Tor: The second-generation onion router. In: Proceedings of the 13th USENIX Security Symposium (August 2004)
[9] Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
[10] Dwork, C.: Differential Privacy: A Survey of Results. In: Agrawal, M., Du, D.-Z., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008)
[11] EFF. Aol's data valdez violates users' privacy, https://www.eff.org/deeplinks/2006/08/aols-data-valdez-violates-users-privacy
[12] Essig, C.: Illinois makes millions selling personal information (2010), http://www.thesouthern.com/news/article_0a5fd6a0-4b6b-11df-a353-001cc4c03286.html
[13] Estus, J., Monies, P., Off, G.: State profits from residents' data (2010), http://www.tulsaworld.com/news/article.aspx?subjectid=11&articleid=20100404_11_A1_Thesta994848
[14] Halderman, J.A., Schoen, S., Heninger, N., Clarkson, W., Paul, W., Calandrino, J., Feldman, A., Appelbaum, J., Felten, E.: Lest we remember: Cold boot attacks on encryption keys. In: Van Oorschot, P. (ed.) Proceedings of the 17th USENIX Security Symposium, pp. 45–60. USENIX (July 2008)
[15] Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: International Conference on Data Engineering (2007)
[16] Marijuana Policy Project. Grid: A comparison of key aspects of state medical marijuana laws (2011), http://www.mpp.org/assets/pdfs/library/MMJGrid15StatesMarch2011.pdf
[17] National Cancer Institute. Marijuana Use in Supportive Care for Cancer Patients (2010), http://www.cancer.gov/cancertopics/factsheet/support/marijuana
[18] Percival, C.: Stronger key derivation via sequential memory-hard functions, http://www.tarsnap.com/scrypt.html
[19] Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 557–570 (2002)
[20] Wilcox-O'Hearn, Z.: (2003), http://en.wikipedia.org/wiki/Zooko's_triangle
[21] WLWT News 5. IRS Worker Admits Snooping In Celebrities' Files (2008), http://www.wlwt.com/news/17015370/detail.html

Protected Login

Alexei Czeskis[1] and Dirk Balfanz[2,⋆]

[1] University of Washington, Seattle, WA
[2] Google Inc., Mountain View, CA

Abstract. Despite known problems with their security and ease-of-use, passwords will likely continue to be the main form of web authentication for the foreseeable future. We define a certain class of password-based authentication protocols and call them *protected login*. Protected login mechanisms present reasonable security in the face of real-world threat models. We find that some websites already employ protected login mechanisms, but observe that they struggle to protect first logins from new devices – reducing usability and security. Armed with this insight, we make a recommendation for increasing the security of web authentication: reduce the number of unprotected logins, and in particular, offer opportunistic protection of first logins. We provide a sketch of a possible solution.

1 Introduction

The overwhelming majority of user authentication on the web today uses passwords. Perceiving passwords as weak and hard-to-use, the research community has in the past focused their efforts on replacing them with "better" authentication mechanisms [2,5,6]. Recently, Herley and van Oorschot [7] have suggested that such research is misguided because passwords, they argue, present a sweet spot in usability and security that is hard to match. Instead, they call for a research agenda that embraces passwords. We agree with much of the premise of Herley and van Oorschot's paper. In particular, we acknowledge that from a human-computer interaction point of view, passwords are hard to beat, and are likely to stay with us. We also agree that much confusion around password usage still exists that warrants further research.

We disagree, however, with the idea that "fixing" passwords is (probably) not necessary because the overall harm done by using them is (probably) small.[1] While the overall harm of account hijackings may be small, the blow to an individual's life when their account is hijacked can be devastating [4]. We take the opinion that every hijacked account is one hijacked account too many. Passwords should be fixed *now*, no matter how small the global harm done is due to their insecurity.

In fact, some websites have already begun to "protect" password-based logins in ways such that passwords alone are not sufficient to authenticate. Interestingly, this happens without affecting the user experience: users still simply enter a password, but their browser submits a cookie along with the password to protect the login.

⋆ The opinions expressed here are those of the authors and do not necessarily reflect the positions of Google.

[1] To be fair, their main point is that we can't currently quantify the harm done, but the implication is that if little harm is done then no corrective action is needed.

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 44–52, 2012.

In this paper, we study this phenomenon and suggest a generalization: we formally introduce the notion of "protected login", which is a class of authentication mechanisms that add credentials to the authentication flow that are invisible to the user. Because they are invisible, the user cannot be phished for them.

We have chosen the name "protected login" consciously: just like in physical human relationships, users can engage in "protected" or "unprotected" logins, and the two mechanisms are essentially identical: adding protection does not fundamentally change the way login is performed. When protection is unavailable, login still works, but the risks may be higher.

The rest of the paper is organized as follows: After reviewing real-world threat scenarios in Section 2, we formally introduce the notion of protected login in Section 3 and examine how websites use protected login today. We describe that the "Achilles' heel" of contemporary protected login mechanisms is the first login from a new device, both in terms of security and usability. In Section 4 we therefore explain how opportunistically adding protection to first logins can address this problem. We conclude in Section 5.

2 Practical Threats to Authentication on the Web

Based on our experience operating the authentication infrastructure of a site with hundreds of millions of users, we start by providing what we believe to be a real-world threat model for the majority of users authenticating on the web today. Most notably, we assume the attacker is capable of stealing user credentials (*i.e.,* passwords) through phishing or through compromising poorly protected web servers. Users are known to re-use (or share) passwords across websites, therefore a credential stolen from a poorly protected, unimportant web server may, for any given user, very well turn out to be also a credential for that user's banking or email provider [9].

Throughout this paper, we assume that the attacker is not controlling malware on the user's machine (although the careful reader will notice that some of the authentication mechanisms discussed here are secure against certain kinds of malware, such as keyloggers).

We'll furthermore assume that the authentication protocol runs over a secure connection, and the attacker is not able to steal cookies, passwords, or other credentials by eavesdropping on the messages between a user agent and a web server (*e.g.,* by compromising the Public Key Infrastructure [11] and becoming an "SSL man-in-the-middle"). It is possible to relax this assumption and design non-bearer-token-based authentication protocols that are secure against such attackers, but outlining this in detail goes beyond the scope of this position paper.[2] For the purposes of this paper, we'll assume that the attacker obtains possession of the user's credentials by phishing for them, or by breaking into poorly-protected websites.

[2] A hint as to how this might work: one way to make stolen credentials useless to eavesdroppers is to *channel-bind* [10] them to an *origin-bound client certificate* [1]. Coincidentally, this can also protect the credentials from malware theft, especially if the client's TLS private key is protected by hardware.

3 A New Paradigm: *Protected Logins*

A class of login mechanisms that we call "protected login" is well-positioned to mitigate against the threats outlined above. We begin with several definitions:

Definition 1. User-supplied credentials *are passwords or other secrets that users input into client devices in order to authenticate to web servers. A login to a web server that is authenticated only through user-supplied credentials is called an* **unprotected login**. *Any login that is not an unprotected login is a* **protected login**. *A login mechanism that allows the web server to distinguish between protected and unprotected logins is a* **protected login mechanism**.

Several observations follow:

- User-supplied credentials are subject to phishing attacks and theft from poorly managed servers.
- A protected login involves credentials beyond just user-supplied credentials. Because these additional credentials are never supplied by the user (and presumably not even known to the user), the user cannot be phished for them.
- A credential thief must always perform an unprotected login to gain access to a victim's account (because the attacker is only ever able to obtain – through phishing or server compromise – user-supplied credentials[3]).
- A legitimate user may or may not have to perform unprotected logins.

From the point of view of a web application, protected logins are less risky than unprotected logins, which can be initiated by an attacker after credential theft. As such, it seems natural to require additional user-supplied credentials (such as mother's maiden name, one-time-PIN, *etc.*) during unprotected logins, while requiring only passwords during protected logins. Observe that the additional user-supplied credentials are still phishable and subject to theft.

We often assume that more "secure" mechanisms must be less "usable". Note the seemingly counter-intuitive consequence of our definition: risky unprotected logins, which require manual entry of additional user-supplied credentials, tend to be, in practice, less user-friendly than safer protected logins, which use supplementary credentials such as special cookies, but require from the user at most a password. We will provide examples for this below.

3.1 Bootstrapping Protected Login: Current Best Practices

Although they may not use this terminology, some web applications are already using protected logins today. We will now examine a few examples. All of them *bootstrap* protected logins using different types of unprotected logins. This is a problem that we will later return to.

[3] This assumes a minimum of competence on behalf of the web server - they need to design their non-user-supplied credentials such that they are distinguishable from *other* servers' credentials. If they do this, then the credentials stolen from another web server won't be usable for a protected login.

Facebook Login Notifications

Facebook allows users to opt into a mechanism called "Login Notifications" [3]. When users have Login Notifications turned on, the first time they log in from a new device they are asked to "name" that device. The user is then notified of the (unprotected) login via SMS or e-mail which contains the device's name. Facebook associates the user-agent's HTTP session with that user and device name (presumably by setting a cookie[4]).

Observe that the *first* login from a new device is an unprotected login because it uses only user-supplied credentials. Because this is risky, Facebook asks users to provide a device name and sends them a login notification. In certain circumstances, the user will have to answer even more challenges, such as identifying a known person from a given image. However, *subsequent* logins from the same device require the user to only supply his username and password and do not generate notifications. Facebook is able to give subsequent logins a higher trust rank because, under the hood, subsequent logins include the user's password and a browser cookie (for which the user could not have been phished, and which could not have been stolen from a non-Facebook web server). Therefore, subsequent logins are protected logins, and a credential thief will always cause at least one unprotected login when accessing a victim's account.

Google 2-Step Verification

Google allows users to opt into a mechanism called "2-Step Verification" [8], which is a form of two-factor authentication. Users obtain a one-time code (OTC) through SMS or from a smartphone app, and must enter this short code during login (in addition to their password). Just as with Facebook Login Notifications, users must perform this step the *first* time they log in from a certain device. *Subsequent* logins don't require an OTC – they only require a password. Google can do this because, just as with Login Notifications, subsequent logins are protected by a cookie that is set during the two-factor login and is sent along with all subsequent logins.

Again, observe that the *first* login is unprotected according to our definition[5] because an attacker can phish the user for their password and OTC. As before, *subsequent* logins are protected. Similarly, after a credential theft, an attacker will always have to perform an unprotected login to access a victim's account.

Quora Login

Previous examples showed protected logins that required only passwords, and unprotected logins that required additional user action. Quora makes a different trade-off between protected and unprotected logins. The first time users log into Quora from a new device, the (unprotected) login requires a username and password. This is an unprotected login because a phisher can perform the same login once he obtains the user's credentials.

[4] We haven't identified the particular cookie responsible for maintaining this state, but verified that cookies remain set after logout, and that removing all cookies results in the user having to "name" their device again.

[5] This doesn't mean that 2-Step Verification is a bad idea. In fact, it in practice affords vastly improved security to Google account holders, in part because it protects against password sharing.

On the login page, Quora shows the user a checkbox that says "let me login without a password on this browser". Once checked, subsequent logins will not require a password – instead, the login page shows the user their profile picture, and a single click on that picture logs the user back in. Note that this login page with the user's profile picture is shown *after* the user has specifically clicked the logout button. Even though this login page doesn't require a password at all, this is a protected login (presumably affected by a cookie that was saved on the user's machine), because a phisher cannot cause this kind of login simply by stealing user-supplied credentials (*i.e.,* the user's Quora password) – he would first have to go through an unprotected login.

Some threat models would consider the protected Quora login less secure than the unprotected login (*e.g.,* if the threat is that of an attacker walking up to the user's terminal). In our threat model however, which we argue reflects real-world threats for the majority of internet users, this attack is not an issue.

3.2 Problems with Current Login Mechanisms

In the above examples attackers can bootstrap protected logins by first performing an unprotected login with stolen user-supplied credentials. Web sites have naturally sought to increase the security of such unprotected logins. The techniques we examined above (sending login notifications and using two-factor authentications) are examples of this.

Other popular methods of bootstrapping protected login include questions such as "what's your mother's maiden name?", "when was your father born", and "did you recently take out a mortgage?". These approaches are common and are used by many banks (*e.g.,* Bank of America and ING Direct) and credit history agencies (*e.g.,* Equifax).

The general approach of bootstrapping an easy-to-use, protected login with a more onerous, unprotected login, unfortunately has several problems:

- First, unprotected first logins lack *security* because they are fundamentally phishable. Even if the users are not phished, answers to security questions can often be found by determined attackers through publicly available records or social networks – Sarah Palin's personal Yahoo! e-mail account was "hacked" in this way [12].
- Second, current login flows also do not provide *availability* as users can easily be locked out of their accounts. For example, users often forget the answers to secret questions (what was the name of my favorite song 3 years ago?) As another example, second factor authentication users are at the mercy of the second device's availability. If they misplace it or let the device's battery die, users may be locked out of their account.[6]
- Most notably, all of the existing approaches to strengthen unprotected login sacrifice *usability* by introducing changes that slow down, inhibit, or otherwise worsen the user experience (UX). The degraded UX may, in turn, lead to user discontent and accordingly, companies like Google and Facebook are cautious to make many of the login protections mandatory – rather leaving them in the "opt-in" arena for

[6] Google 2-step verification users are asked to print out a set of "backup" codes that can be used in this scenario. Users are urged to carry these codes with them. Users can also provide backup telephone numbers where codes can be sent in case of a lockout.

more tech-savvy and security conscious users. This, in turn, leaves the vast majority of users without the protection provided by the disabled-by-default additional security measures.

Given that strengthening authentication for unprotected logins seems to severely degrade the user experience, how can we improve on the current best practices around web authentication?

4 Reducing Unprotected Logins

We believe the key to improving web authentication without sacrificing usability is to significantly reduce the number of unprotected logins. This would increase *security* because unprotected logins are by definition phishable, and hence less of them means less opportunity for passwords and other credentials to get compromised. If unprotected logins are very rare, websites can afford to "raise alarms" whenever an unprotected login occurs and legitimately treat those sessions as less secure. Websites would also be able to notify the users with more conspicuous messages, perhaps require them to take action, taint sessions initiated with unprotected logins, or even have infrastructure to revert changes caused by a potential attacker. Additionally, having less unprotected logins would allow immense *usability* benefits for many users. As we showed, unprotected logins tend to be onerous; therefore, fewer unprotected logins means fewer onerous logins.

Clearly, reducing unprotected logins would be beneficial from a security and usability perspective. However, can it be done without affecting availability? If so, how and where in the web authentication flow this be done?

4.1 Protecting Subsequent Logins

We noted earlier that some websites already offer protected logins to their users. They do this by setting a special cookie during the first, unprotected, login from a new device, and then checking that cookie on subsequent logins.

This much is simple: one way to reduce the number of unprotected logins is to implement this pattern in more places around the Internet: set a cookie when users first log in, and use the presence of that cookie as an additional signal to increase the trust rank of (subsequent) logins.

This leaves us with the problem of unprotected first logins, the "Achilles' heel" of current web authentication, both in terms of security and usability. Achieving a protected login during the *first* authentication from a new device has been elusive. Why? What makes protecting the first authentication so difficult? Is there really nothing we can do?

4.2 Protecting First Logins

To understand why protected login during first authentication is hard to achieve, let's consider what that process must involve. By definition, during a protected login, the

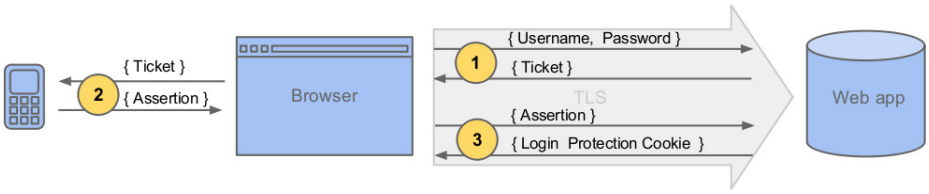


Fig. 1. Possible protocol for a protecting *first* login from new devices

user's client must transmit to the server a piece of data, for which the user cannot be phished, but which will convince the web server that a legitimate login is in progress. Where can this data come from? For secondary logins, the browser can send secret cookies, but on first login, the user's browser hasn't yet established user-specific secrets with the server and is unable to do so. Could the data come from the user? Unfortunately, any secret the user knows is fundamentally phishable. This indicates that for a first login to be protected, it must involve some second factor device. However, all second factor authentication mechanisms seem to degrade users' experience. Even smartcards, which can be used in ways that don't alter the authentication user experience, require users to carry additional hardware and are henceforth seldom used outside of the corporate and government settings. Therefore, we believe the biggest impediment to providing protected login during the first authentication is the lack of a second-factor based protected login mechanism that is largely transparent for users.

We believe that asking users to carry additional devices is unacceptable, so we are focusing our attention on mobile phones – the only additional piece of hardware that many users consistently carry with them. But even phones are often unavailable or nonexistent, and relying on them is a recipe for failure.

We now offer a key insight and propose a security compromise: what if protected login is provided opportunistically? That is, the website will always ask the user for his password, but if it's possible and all of the "stars align", the authentication flow will involve an additional operation (that's transparent to the user), which will result in a protected login. However, if the protected login mechanism fails to successfully complete, the authentication will result in an unprotected login. In either case, the user will be logged in, and would have done no work beyond entering his password. This compromise will allow the system to maintain good *availability*. [7]

Of course, several questions immediately arise: How does one make the protected login mechanism transparent to users? How reliable will the protected login mechanism be and how often will users be forced into an unprotected session? What protocol should the protected login mechanism use and what type of data should it send – a certificate, a token? Finally, is it possible to provide a framework that's usable by any site on the web and does not involve significant developer effort?

These questions are complicated and have non-trivial tensions between availability, security, usability, accessibility, and privacy. We are in the beginning stages of building a system that attempts to navigate these various constraints.

[7] Nevertheless, opportunistic protected login may be insufficient to meet the security needs of some organizations; they may choose to deploy a policy that will enforce mandatory protected login for some or all transactions.

SKETCH OF A POSSIBLE SOLUTION

We now provide a high-level glimpse of our work-in-progress – an authentication system that opportunistically allows users to achieve a protected login from new devices without altering the user experience. We omit the majority of details as well as the discussion of why certain tradeoffs were made – those issues are complicated, non-obvious, and are still in slight fluctuation.

Our design assumes that users have a smartphone, which various web services can leverage to protect user logins. We also assume that at some point, the phone and web service were able to perform a key exchange. The core idea of our design is described in Figure 1 and works as follows. First, the user navigates to a website of his choice and is presented with a login page (as usual). He enters his username and password. These user-supplied credentials are sent to the server, which authenticates the user and responds with a *login ticket*.

Next, the web browser sends the ticket to the user's phone over a wireless distance-bound protocol. The phone verifies the ticket, generates an assertion and hands it back to the browser. The browser then sends the assertion to the server, which can verify it based on the previous key exchange with the phone. The server then responds with a cookie[8] and marks the HTTP session as protected. If the user's phone is not reachable within a small delta of time, the browser cannot send the phone's assertion. Instead, it sends an error message to the server, which will mark the session as unprotected, but still responds with a cookie.

We leave all further discussion of this proposal as future work, but would like to note that a successful protected login under this protocol protects users against phishing and password theft. Further refinements that utilize channel-bound assertions and cookies also protect against SSL men-in-the-middle attacks and cookie theft [1,10].

5 Conclusion

Passwords have been the primary authentication mechanism since before the web was born, and they don't seem to be going away. At the same time, attackers have become adept at stealing passwords through a variety of attacks. In this paper, we made several key contributions. First, we presented a real-world threat model for web authentication and password use. Second, we introduced a new paradigm, *protected login*, that is useful for analyzing various web authentication techniques. Third, we used the protected login paradigm to examine current web login flows and find that many web applications already deploy protected login, but still make themselves vulnerable by forcing users to bootstrap protected login through unprotected login from new devices. Fourth, we gave an agenda for improving web authentication: reduce unprotected logins by opportunistically protecting *first logins* with a protocol like the one outlined above and by protecting *subsequent logins* with mechanisms similar to what we already see in the wild today.

[8] This cookie is what protects subsequent logins, for which presence of the phone is no longer required.

References

1. Balfanz, D., Smetters, D., Upadhyay, M., Barth, A.: TLS Origin-Bound Certificates (Working Draft) (July 2011), http://tools.ietf.org/html/draft-balfanz-tls-obc
2. Everitt, K.M., Bragin, T., Fogarty, J., Kohno, T.: A comprehensive study of frequency, interference, and training of multiple graphical passwords. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, pp. 889–898. ACM, New York (2009)
3. Facebook. What are Login Notifications? (2011), https://www.facebook.com/help/?faq=162968940433354
4. Fallows, J.: Hacked! (2011), http://www.theatlantic.com/magazine/archive/2011/11/hacked/8673/
5. Forget, A., Chiasson, S., Biddle, R.: Shoulder-surfing resistance with eye-gaze entry in cued-recall graphical passwords. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 1107–1110. ACM, New York (2010)
6. Gajek, S., Schwenk, J., Steiner, M., Xuan, C.: Risks of the CardSpace Protocol. In: Samarati, P., Yung, M., Martinelli, F., Ardagna, C.A. (eds.) ISC 2009. LNCS, vol. 5735, pp. 278–293. Springer, Heidelberg (2009)
7. Herley, C., van Oorschot, P.: A Research Agenda Acknowledging the Persistence of Passwords. IEEE Security & Privacy Magazine (2011)
8. Google Inc. Getting started with 2-step verification (2011), http://goo.gl/5r8Za
9. Leyden, J.: Anonymous hack showed password re-use becoming endemic (2011), http://www.theregister.co.uk/2011/02/password_re_use_study/
10. Williams, N.: On the Use of Channel Bindings to Secure Channels. RFC 5056, RFC Editor (November 2007), http://www.ietf.org/rfc/rfc5056.txt
11. Zetter, K.: Diginotar files for bankruptcy in wake of devastating hack (2011), http://www.wired.com/threatlevel/2011/09/diginotar-bankruptcy/
12. Zetter, K.: Sarah Palin E-mail Hacker Sentenced to 1 Year in Custody (2011), http://www.wired.com/threatlevel/2010/11/palin-hacker-sentenced/

Enabling Users to Self-manage Networks: Collaborative Anomaly Detection in Wireless Personal Area Networks

Zheng Dong

School of Informatics and Computing, Indiana University, Bloomington, IN, USA
zhdong@indiana.edu

Abstract. Personal area networks such as home or small office LANs are usually more vulnerable to cyber-attacks than those with dedicated support staff and the ability to invest consistently in security defenses. In this paper I propose leveraging physical characteristics of these personal area networks in order to enable non-technical individuals to secure their networks or at least be aware that their devices have been compromised. This proposal leverages records of location for mobile devices, proximity authentication, and individual homophily. In this work, I summarize previous studies on securing personal networks, proximity authentication, and software attestation. I then present a preliminary design for the detection of and recovery from infection for personal area networks. Limitations and future work are also discussed.

1 Introduction

With the improvement in affordability of many electronic devices, small-scale networks are commonly constructed in home and small office environments. The term "personal area network" (or "PAN") usually refers to a local, connected group of personal devices. These devices may include laptop computers, personal digital assistants (PDA), palmtops, and cell phones [1]. The boundary of PAN is the area physically covered by the wireless network and/or the central server. Previous network members (e.g. laptops and phones) may leave and re-enter the network multiple times.

Unlike larger networks that are dominated by wired connections, devices in personal area networks are normally connected by wireless protocols, such as Wi-Fi (802.11) or Bluetooth (802.15). It has been documented that wireless networks tend to be more vulnerable than wired networks [2].

While significant research has been performed on inter-PAN network security, little attention has been paid to security issues inside a personal area network. PAN environments are unique for two reasons.

First, most personal area networks have clear physical boundaries and basic access controls. For instance, all participating personal devices are at some time in the home, so that only the residents (and possibly a few guests) have physical access to the devices.

Second, the ownership of personal area networks (including all devices) is unitary. As a consequence, the device owner has an incentive to protect the security of the

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 53–67, 2012.

network. At the same time, the owner is unlikely to be skilled in computer networking. Therefore techniques designed for personal area networks need to be highly automated with minimum human interference embedded in the interactions, and should fully leverage the geographical information inherently provided by a personal area network.

The purpose of this work is to design a security protocol specifically for personal area networks. This protocol incorporates proximity authentication, collaborative rating, and software attestation, but it is not a simple combination of the above techniques. Protocol phases are carefully designed and adjusted to meet the security needs of personal area network devices. This proposal builds on physical location, particularly co-location and proximity authentication of the devices. Following proximity authentication, the proposed design also uses Bluetooth, leveraging the inherent distance limitations of Bluetooth. Other proximity authentication methods are equally applicable.

The rest of the paper is organized as follows. Section 2 introduces related work on personal network, proximity authentication, and software attestation. Section 3 defines the threat model, enumerating the threats which the proposal is designed to mitigate. Section 4 discusses the assumptions underlying the protocol design. Section 5 provides an overview of the components of the design, including participants. Section 6 provides details of the proposed protocol with an example consisting of three devices. Section 7 summarizes the findings, and concludes the paper.

2 Related Work

2.1 Personal Networks

As electronic devices become even more affordable, it is common for homes to contain an increasing number and diversity of digital devices with a small-scale network shared amongst them. Bisdikian et al. [1] introduced the notion of "wireless personal area network (WPAN)", which includes various types of personal wearable or handheld devices such as laptop computers, personal digital assistants (PDA), palmtops, cell phones, etc. These authors also pointed out that WPAN differs from traditional wireless local networks (WLAN) in network size, implementation cost, usability, and power consumption. The definition of "personal area network" assumes that all networked devices are within a short distance, typically within 10 meters. Furthermore, IEEE 802.15 [3] defines the characteristics of WPAN.

To connect personal area networks (PANs) that are geographically distributed, a personal network (PN) can be established [4]. The definition of PN relies on the idea of pervasive computing. The design of PN, however, is not merely an extension of PAN. Mechanisms such as addressing, routing, and authentication need to be implemented. Extending PN, Hoebeke et al. proposed a "personal network federation" (or "PN-F"), which enables device linkage between different personal networks [5]. PN-F addresses secure communication needs within a common interest group, such as family members, classmates, and colleagues. In addition to the proposed scheme, the author also discussed several designed challenges, such as membership management, application support, and system maintenance.

Network security and user privacy concerns are also increasing with the proliferation of personal networks. These concerns generally focus on untrusted inter-PAN web traffic. Jacobsson et al. proposed a secure PN mechanism that ensures anonymity by encryption and MAC or IP address change after certain intervals [6]. Social activities, such as device lending or sharing, were also considered. Patrikakis et al. analyzed typical threats in personal networks and introduced a trust model over personal networks [7,8]. In their design information needed for authentication was treated differently than sensitive data like user preference. A central server was established in this scheme for device registration and group key distribution. In addition, networked device status needs to be reported so that malicious users and devices can be detected. This work is different from their design for a much smaller network range, and therefore different assumptions and techniques are proposed. For example, since all networking devices are within a certain physical range, I do not consider inter-PAN network traffic, which requires further encryptions. I also consider issues of usability, social context and social engineering.

2.2 Proximity Authentication

Significant research has been performed on authentication between network devices within a short distance. These approaches typically rely on inherent physical constraints. McCune et al. [9] proposed a mechanism that establishes a trusted channel between camera-phones by integrating public keys in 2D barcodes. Rasmussen et al. [10] introduced a proximity-based protocol to authenticate remote access for medical devices that are implanted in patients' body. This approach relies on the speed of sound, which is a constant. In addition, Cai et al. [11] proposed a mechanism that verifies when communication devices are co-located. This approach requires more than one antenna in the verifier, and is based on the relationship between signal parameters and distance.

The proximity-based authentication system, Amigo [12], relies on a mechanism that verifies co-located mobile devices by generating digital signatures from wireless radio strength, and then comparing the remote signature with the local one. Similar radio strengths indicate that two devices are within a short distance. Based on a similar idea, another proximity-based authentication system, Ensemble [13], relies on variation in radio signal strengths to determine physical proximity; trusted third parties (e.g. MP3 players, laptop computers) are included in this approach to monitor the security channel establishment and help verifiers prove authentication.

2.3 Software Attestation

Ensuring software execution on untrusted platforms is not the research contribution of this work. I recognize this as a distinct research challenge while building on the advances of others. There are two fundamental approaches to software attestation with the difference being the assumption of the (non) existence of a TCP. Seshadri et al. introduced Pioneer [14], a software attestation protocol that validates the execution of codes on an untrusted platform, even though malicious codes may be on

the machine. For embedded systems such as smart phones, SWATT [15] was proposed to detect malicious memory changes in embedded systems caused by viruses, Trojan horses, etc. The SWATT technique does not require prior authentication on the verified phone memory. Two types of attacks against these software-based attestation protocols were suggested [16]. To conquer these attacks, Jakobsson et al. [17] designed a new attestation protocol that evaluates both active applications in the memory and inactive programs that have been swapped out.

3 Threat Model

The primary threat to personal area network security is infection by malicious software (Malware). A number of malware types have been reported. Malware can be characterized by its payloads, targets, and mechanism for propagation.

Malware payloads refer to the primary actions taken [18] by active malware or the damage caused by malicious code [19]. Different types of malware may vary significantly in their payloads. For example, certain computer viruses (such as the well-known Melissa [20] and Iloveyou [21] viruses) were designed to tamper with users' files and/or operating systems. An example of the most destructive computer viruses would be CIH [22], which is capable of overwriting the BIOS on victims' computers. In addition to unauthorized modification on file systems, some malware steal data from the victims' computer. As an example, Schlegel et al. proposed Soundcomber [23], a context-aware sound Trojan that steals sensitive information from smartphones. Additionally, adware and spyware are often included as part of a software installation package [24]. Adware displays commercial advertisements and spyware monitors system surreptitiously, forwarding the collected information to third-parties. Botnet is an important type of malware. Instead of infecting a single machine, the botnet master can control thousands of bots. By directing a large number of infected machines, attacks originated from a botnet are often powerful. Typical malicious activities from botnets include DDoS attacks [25], email spams [26], etc.

Propagation mechanisms have also been used to categorize malware. Among all malware types, computer viruses and worms attract the most public attention. Generally, when a computer virus is executed, it replicates itself and spreads to uninfected files. Compared to viruses, worms are more active in propagation. In addition to self-replication, worms are capable of automatically detecting system vulnerabilities and infecting victim machines autonomously [27]. This characteristic leads to different propagation media for viruses and worms. According to recent studies, removable storage (such as CD, DVD, flash disks), emails and online downloads are the primary entry points [28] for viruses, while online transmission is a critical part in the propagation of worms.

It is much easier than many people would believe for malware infection. In fact, malware threats to mobile devices, especially smartphones, arise with the enhancement on device functionalities. It has been documented that the capabilities of web browsing, online messaging (e.g. send and receive multimedia emails or instant messages), reading flash-memory cards, or communicating by Bluetooth radios may all lead to vulnerabilities [29]. In other words, every machine faces a unique and wide range of possible malware attacks.

It has been reported that malware targeting on mobile devices has increased in recent years. According to the malicious mobile threat report published by Juniper Networks on May 2011 [30], the number of unique malware variants targeting the Android platform has increased by 400% since summer 2010. Malware detected on Nokia Symbian and Windows Mobile still dominate mobile malware according to the Jupiter sample database.

In addition to malware propagation, public recognition of malware threats remain insufficient. As shown in many forum posts, smartphone users do not realize that their phones need antivirus software just like computers. In fact, mobile devices are often more vulnerable to attackers than desktop computers. First, the mobile users are often considered more economically valuable targets. As the mobile applications and functionalities proliferate, more information is stored on the phones. Greater incentives are therefore created for malware development and distribution. Second, antivirus software is less well developed for mobile devices. Compared to antivirus programs developing for PCs, software functionality is preliminary or limited on phones. Furthermore, more malware is run in the background, which makes it difficult to detect, without the help of antivirus software.

In order to understand malware distribution, some researchers focused on the scale of machine subversion. In [31], Eeten et al. proved by a large-scale experiment and argued for Internet Service Providers (ISPs) as good control points for botnet mitigation. I agree that this is necessary but it is not sufficient. In this work, I propose that the malware mitigation could be augmented within personal area networks. I argue that this goal is achievable by incorporating collaborative rating and software attestation into the protocol design. The design problem is different for a PAN. I describe in Section 5 the technical heterogeneity and user homophily.

4 Assumptions

I made the following assumptions in this work.

- A1. Machines in a PAN are not infected simultaneously.

The proposed solution can only apply if the infection of a machine is not determined by the location of that machine. That is to say, in a home or personal network with x devices, the likelihood of subversion for these devices is independent. This is particularly the case in a typical home environment where there is significant heterogeneity in device models in the home. For example, individuals are less likely than firms or organizations to dispose of a machine simply because it has non-standard or dated capacities. PAN networks may include phones, laptops, desktops, eReaders, and a single router or server.

- A2. Power limits are not a concern in the home itself.

That is, when a mobile device is at its home, it is easy to plug in. Power consumption is a constraint in most cases when security protocols are designed for mobile networks. Because I am focusing on the home, the consumption of power is not such a limiting factor.

- A3. There is a transport layer that is shared to some degree. In other words, each device has the knowledge of other devices.

In the case there is not a shared transport layer, I assume the ability of a machine to sense the behavior of other participants through interactions during a re-introduction phase of the protocol. Devices are also required to share state.

- A4. There is a pattern or patterns of interaction generally on a daily or weekly basis.

Please note I assume that there are at least two devices in the constructed personal area network. The interactions among devices roughly follow certain patterns, particularly if proximity is considered an interaction. The interaction patterns between the devices can create or predict the context. For example, if an individual's daily schedule ends at 10pm, then a device login at 1am is particularly suspicious. If there are two devices, there is usually also a management device (router or bus). Notice that this depends on the colocation of the devices. For instance, when none of the mobile devices is present, the desktop should be inactive.

- A5. There is limited human capacity but there is the incentive enough to motivate set-up, interaction, and recovery.

An initial configuration is needed when constructing personal networks, while very little human interaction is expected afterwards. Each device in the personal network would be incorporated into the network with human interaction. I do not want authentications to run automatically when a new participant is added. Introductions are based on proximity authentication. Authentication is automatic when a known participant returns to the network. Humans engage in introduction and recovery only. Re-introductions and evaluations are handled by the machines.

- A6. Mobile devices are aware of their own locations and reintroduce themselves when returning to the home area network.

Considering the mobile nature of some devices in personal networks, it is necessary that mobile devices are given unique IDs so that linking authentication requests is possible. However, depending on time disconnected and probability of connection to external networks, the investment in authentication may change when a mobile device leaves and returns.

- A7. There is a limited period upon initial introduction during which devices are either trustworthy or can be made trustworthy with self-audit.

I argue that self-recovery is possible and can be automated once initiated by a human. The recovery task may be accomplished by a third-party recovery service. In this protocol, the self-recovery is executed from the central server which I will introduce later.

- A8. The central server is trusted.

Comparing to mobile devices, security measures on servers are more common. In addition, given the fact that the central server is responsible for proximity authentication, collaborative rating, and possible device recovery, the individual would have a strong incentive to protect the security of the central server. Note that I begin with a central server design and move to a distributed solution.

5 Protocol Design

5.1 Central Server Model: Participating Parties

Three parties are included in this protocol, the central server, the claimant and the verifier(s). Given that personal area networks are often implemented within relatively small ranges and with clear physical boundaries (e.g. home or office), I assume that communications among mobile devices and servers are trusted.

The central server is in charge of mobile device management. Specifically, a database is maintained on the server. It contains devices' physical addresses (for example, MAC addresses of WLAN or Bluetooth adapters), presence information of mobile devices (for example, records on entering and leaving the network), and collaborative rating results. Considering the importance of data transmission and storage to authentication, I recommend that the central server is located near the mobile devices. In addition, due to security concerns and data transmission rates, multiple personal networks should not share a central server.

The claimant is a mobile device that is being verified by other mobile participants. Each mobile device can be distinguished by its physical address, and it is also possible to include a secret message in the identification process. Note that being a claimant in a verification transaction does not exclude a device from being a verifier in another transaction. In this design, the data integrity of a mobile participant will be verified when the device enters a personal network, and on a pre-set frequency (for example, every two hours) afterwards.

In this design, the verifier(s) refer to one or more mobile participants that examine the identity and data integrity of the claimant. I require that at least one verifier presents in the network before the verification process starts. By observing the amount of the claimant's inbound and outbound data, deviation from historical patterns, the response to attestation challenges, each verifier submits a score to the central server, indicating the level of confidence that the claimant device has been subverted. No further action is taken until a final verification result is generated on the central server.

5.2 Protocol Phases

For the purpose of simplicity and clarity, I begin with a proposal that includes a central server. I then propose that authentications could also be accomplished without the central server. There are four components to the protocol: an introduction phase, a run phase, a reintroduction phase, and a recovery phase.

In the introduction phase I choose a proximity authentication. This authentication process ensures that the device is actually located within the house range. Specifically, a challenge is generated by the central server, and passed to the mobile device. The mobile device then responds to the central server. The mobile device will not be granted full network access unless it passes the test. Normally, these challenges rely on physical constraints and/or mathematical hardness. I therefore argue that it is infeasible for an outside attacker to pass this test.

In the reintroduction phase, mobile devices need to prove to the central server that they have been registered before. I base this phase on the design that each mobile device keeps historical keys for a period of time under a proper key management protocol. It is therefore possible to identify an old device by validating previous authentication information. Specifically, the central server generates a historical challenge such as a previous assigned key index. To pass this test, the mobile device searches for a previous key with the index, and sends the hash value of the key back to the server. After validation of the previous communication key, the server continues with a proximity authentication. After the device passes both history and proximity tests, a new communication key will be assigned by the central server, and the device database will be updated accordingly.

In the run phase the mobile devices audit each other in two ways. First, each device attests to the other that it has not changed state. Second, since historical activities have been recorded for each participating device, the transmissions of the devices could then be compared to past transmissions and states after reintroduction. If significant and sudden deviations are detected, then the recovery phase is entered. I base this phase on software attestation. Specifically, an application is installed on each mobile device, and performs scheduled verification tasks even if malicious programs are executed. The application is dedicated to check the memory status, as well as inbound and outbound network traffic. Results from the application will be shared with verifiers and the central server and be considered as a strong indication of whether a claimant has been subverted.

The recovery phase is focused on the repair of machine or malware infection. In the first implementation of this protocol I assume that this is a central server in the network to assist in recovery. In later instantiations recovery is addressed as a socio-technical challenge when the human is directed to implement recovery using a set of hard-wired external systems for that process. Majority voting by devices is required with per device risk assessment of other devices. Malicious reports on other devices' behaviors initiate automatic or human-driven recovery.

6 Example of Implementation

In this section, I propose a sample implementation of this security protocol. Please note that there are other possible technologies that may be utilized to achieve the security goal of the protocol. For example, in proximity authentication, the 2D challenge could be substituted by technologies such as Bluetooth pairing.

Two sets of cryptographic keys are utilized in this implementation: the public/private keys used to initialize symmetric keys and short-term symmetric keys. Each participant (verifier or claimant) holds a long-term public key. At the beginning of this protocol, the central server and the new mobile device need to exchange their public keys with the SSL or MQV protocol. While short-term symmetric keys are used in attestation message encryptions, long-term public keys are needed in both symmetric key generation and digital signatures.

In the introduction phase, the mobile device first submits its MAC address to the central server over the Wi-Fi connection. Upon receipt of the message, the server starts with a proximity authentication algorithm such as the so-called "Seeing-is-believing" algorithm, which was designed by McCune et al. [9]. In this example the server generates a 2D challenge (or displays an unchanged 2D bar code), and the mobile device uses its camera to capture the 2D code. Regarding the mechanism of proximity authentication, I propose that a nonce and the hash value of the server's public key should be included. For 2D the mobile device would then respond to the server with a message 'hiding' in the 2D bar code. To prevent fake responses from eavesdropping attackers, the new device also attaches the hash value from the response message, which can be generated by a message authentication code (MAC) algorithm with the long-term public key of the mobile device. After proximity authentication, the central server adds a new entry to the device database, and assigns the symmetric key to the device with a MAC result of that message to ensure information integrity. The entire process of this phase is illustrated in Figure 1. Notations that I use in the figures are summarized in Table 1.

Table 1. Notations in the Protocol Figures

Notation	Explanation	Primary Purpose
$K_{pub\text{-}serv}$/ $K_{pub\text{-}dev}$	Public key of the server/ a mobile device	Symmetric key allocation
$K_{prv\text{-}serv}$/ $K_{prv\text{-}dev}$	Private key of the server/ a mobile device	Symmetric key allocation
$Adrs_{dev}$	Physical address of a device	Device identification
$K_{dev,time}$	Symmetric key which allocates to a device at a certain time	Attestation message encryption
Hash	Hash function	Prevent message forgery
Nonce	Cryptographic nonce	Ensure message freshness
Timestamp	Current system time	Ensure message freshness

In the reintroduction phase, the central server first searches in the database for previous keys which have been assigned to the device in the past. The server then asks the device to send back the hash value of a key that was assigned at a particular point in time. The mobile device then looks up the particular symmetric key indicated in the challenge and attaches a hash value of the key in its response. The server compares the hash value with the previously stored information and decides if the mobile device has entered into the network before. In addition, similar to the introduction phase, the reintroduction phase then performs a proximity authentication

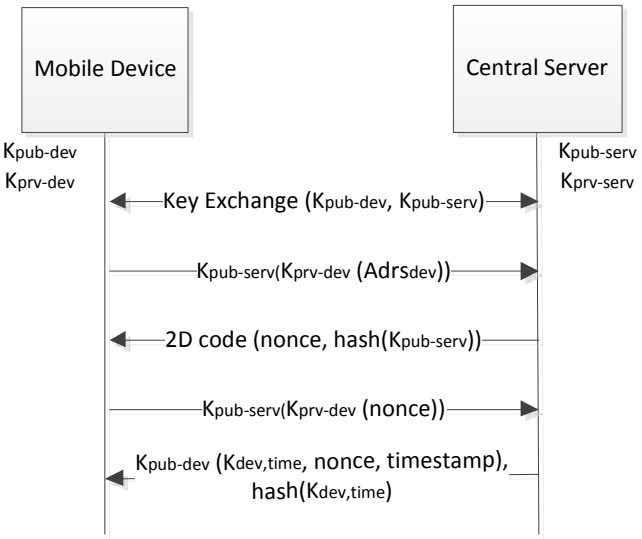


Fig. 1. Introduction Phase

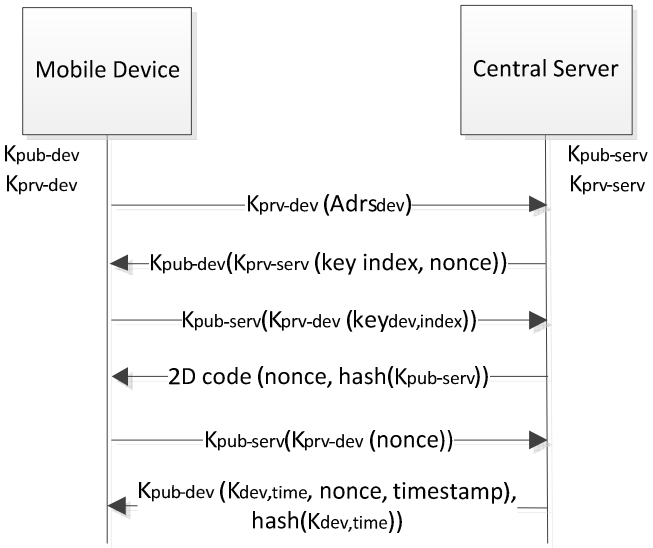


Fig. 2. Re-introduction Phase

and verifies the response regarding the 2D bar code. A new key is generated and assigned when both historical and proximity authentications have completed. The reintroduction phase is shown in Figure 2.

Figure 3 shows the run phase in this recommended implementation. Specifically, I apply a collaborative rating algorithm [32] on the run phase of the protocol. Each mobile device provides the attestation message, which may include recent application activities, inbound and outbound network traffic information to verifiers (by default, all other mobile devices in the network) and a timestamp. The device to be tested also needs to send a hash value of the attestation message generated by MAC with its symmetric key. The verifiers compare the provided information with previous 'experience' with the device. A score indicating the probability of a device being subverted is then sent to the central server. The hash value of the attestation message also needs to be forwarded to the central server to prevent masquerade and replay attacks. After evaluating all ratings from devices, the central server decides if a recovery phase is entered. In the recovery phase a special application is sent to the suspicious device to remove possible malware on the mobile device. This package should be transferred to the target machine by appropriate transport layer protocols, such as SSL. I argue that this self-recovery phase is feasible, and the execution of the recovery application could be guaranteed. While this application is protected by software attestation, and therefore can run without interference of malware, a self-recovery application can be sent to suspicious devices. This application should scan the suspicious device and determine the nature of repairs that needs to be undertaken before the actual work.

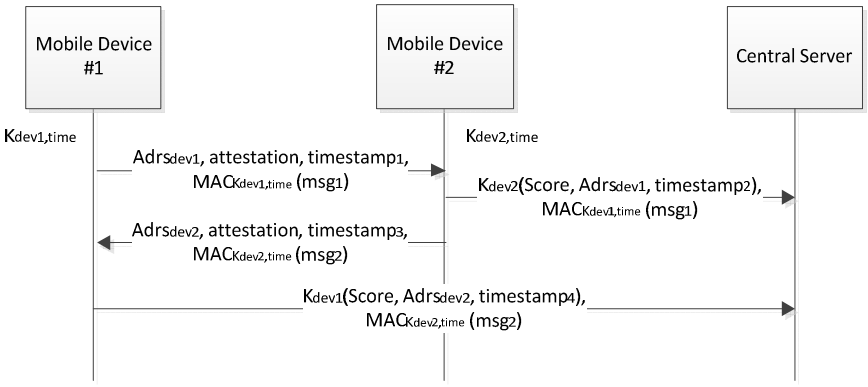


Fig. 3. Collaborative Rating (with server)

In addition, I propose that mobile device verification can also be performed without a central server, as shown in Figure 4. This process is based on the assumption that each device knows public keys for all other devices. This phase also starts with sending out attestation messages from one device to the rest of the PAN. Instead of sending scores to the central server, the verifier then sends the score with its digital signature on a hash value of the message to all other mobile participants. After this sharing process, scores may be generated by any one of the other devices.

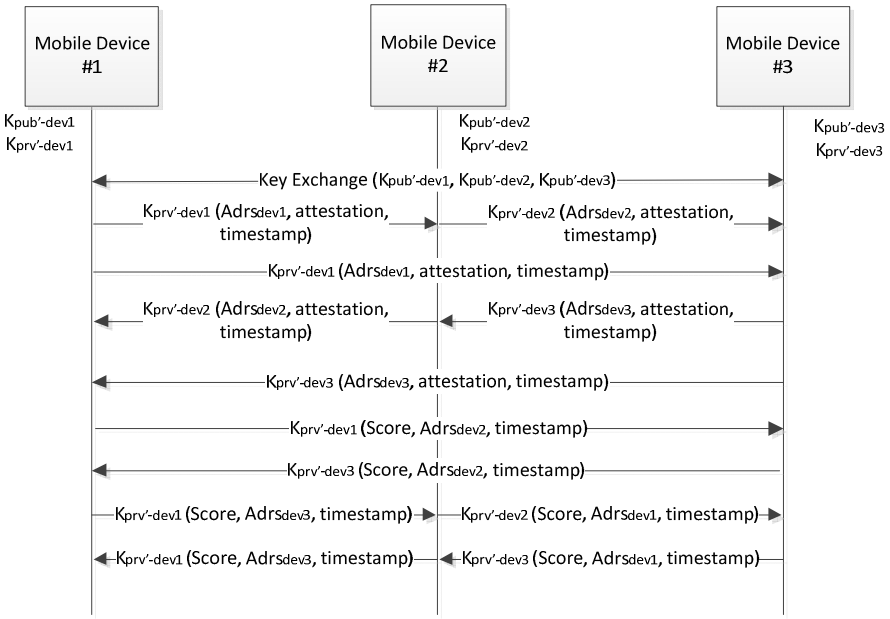


Fig. 4. Collaborative Rating (without server)

In this protocol anomaly detection relies on collaborative ratings from other mobile devices. I base this method on the observation that mobile devices owned by a single individual (or friends in a social network) tend to be similar in many ways (for example, the applications installed, the web browsing patterns, etc.). I analyzed the browsing history of over 1,000 college students that live in the same dormitory [33]. The subjects were selected for their homogeneity in order to mimic a social network. I finally showed that for a highly homogeneous network (with more than 5 participants), more than 95% of websites have been visited in the past. Therefore, if the previous activities of a claimant device are not available, the verifier may still generate a score by comparing its previous pattern with that of the claimant, while the predictability of human behavior has previously been seen as an obstacle. (e.g. in password generation [34]). In this case I leverage predictability and seek randomness or change as identifiers.

7 Conclusion

As the rapid development of portable electronic devices, it is common that people own more than one mobile device. While it is potentially vulnerable in small-scale computer networks, little research has focused on the security and privacy in this particular area. In this paper I proposed a preliminary security protocol for personal area networks. In this protocol, new mobile devices such as laptop computers, PDAs, cell phones are first introduced into the network after a proximity authentication; returning devices need to pass an additional history check before being added into the

network; each participating device performs collaborative anomaly detection with a pre-set frequency.

There are a few limitations in this work. First, this protocol only works for personal area networks. In other words, the assumptions and protocol design would be completely different if the mobile devices were geographically distributed. Additionally, I did not evaluate the performance of the protocol implementation, and I leave this as part of my future work. Certain types of attacks, such as denial-of-service, are not discussed in this work. Further, this protocol design relies on the assumption that the central server is always trusted. I realize that there are possible attacks against the central server during authentication and collaborative rating processes and plan to further the study in this general direction, with an eye towards DoS attacks on this protocol in particular.

Acknowledgements. The author would like to thank Professor L. Jean Camp for her valuable comments and suggestions and John McCurley for his editorial comments.

References

1. Bisdikian, C., Bhogwat, P., Golmie, N.: Wireless personal area networks. IEEE Network 15(5), 10–11 (2001)
2. Rogers, D.: Why Wireless Networks Are More Vulnerable Than Wired Networks, http://www.articlesbase.com/computers-articles/why-wireless-networks-are-more-vulnerable-than-wired-networks-886434.html (accessed 2009)
3. IEEE. IEEE 802.15 Working Group for WPAN, http://www.ieee802.org/15/
4. Niemegeers, I., Heemstra De Groot, S.: Research Issues in Ad-Hoc Distributed Personal Networking. Wireless Personal Communications 26(2-3), 149–167 (2003)
5. Hoebeke, J., Holderbeke, G., Moerman, I., Jacobsson, M., Prasad, V., Wangi, N., Niemegeers, I., Groot, S.: Personal Network Federations. In: Proceedings of the 15th IST Mobile and Wireless Communications Summit, Myconos, Greece (2006)
6. Jacobsson, M., Niemegeers, I.: Privacy and anonymity in personal networks. In: Pervasive Computing and Communications Workshops, pp. 130–135 (2005)
7. Patrikakis, C., Kyriazanos, D., Prasad, N.: Establishing Trust Through Anonymous and Private Information Exchange Over Personal Networks. Wireless Personal Communications 51(1), 121–135 (2009)
8. Patrikakis, C., Kyriazanos, D., Voulodimos, A., Nikolakopoulos, I.: Privacy and resource protection in Personal Network Federations. In: Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments, Corfu, Greece, pp. 29:1–29:5 (2009)
9. McCune, J., Perrig, A., Reiter, M.: Seeing-Is-Believing: using camera phones for human-verifiable authentication. In: IEEE Symposium on Security and Privacy, Oakland, CA, pp. 110–124 (2005)
10. Rasmussen, K., Castelluccia, C., Heydt-Benjamin, T., Capkun, S.: Proximity-based access control for implantable medical devices. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, Chicago, IL, pp. 410–419 (2009)

11. Cai, L., Zeng, K., Chen, H., Mohapatra, P.: Good Neighbor: Ad Hoc Pairing of Nearby Wireless Devices by Multiple Antennas. In: Proceedings of the 18th Annual Network & Distributed System Security Conference (NDSS 2011), San Diego, CA (2011)
12. Varshavsky, A., Scannell, A., LaMarca, A., de Lara, E.: Amigo: Proximity-Based Authentication of Mobile Devices. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) UbiComp 2007. LNCS, vol. 4717, pp. 253–270. Springer, Heidelberg (2007)
13. Kalamandeen, A., Scannell, A., de Lara, E., Sheth, A., LaMarca, A.: Ensemble: Cooperative Proximity-based Authentication. In: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, San Francisco, CA, pp. 331–344 (2010)
14. Seshadri, A., Luk, M., Shi, E., Perrig, A., van Doorn, L., Khosla, P.: Pioneer: Verifying Code Integrity and Enforcing Untampered Code Execution on Legacy Systems. In: Proceedings of the Twentieth ACM Symposium on Operating Systems Principles, Brighton, United Kingdom, pp. 1–16 (2005)
15. Seshadri, A., Perrig, A., Doorn, L., Khosla, P.: SWATT: SoftWare-based ATTestation for Embedded Devices. In: Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, p. 272 (2004)
16. Castelluccia, C., Francillon, A., Perito, D., Soriente, C.: On the difficulty of software-based attestation of embedded devices. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, Chicago, IL, pp. 400–409 (2009)
17. Jakobsson, M., Johansson, K.-A.: Retroactive detection of malware with applications to mobile platforms. In: Proceedings of the 5th USENIX Conference on Hot Topics in Security, Washington, DC, pp. 1–13 (2010)
18. Kanellis, P. (ed.): Digital Crime And Forensic Science in Cyberspace. Idea Group Publishing, Hershey (2006)
19. Malware Wiki, http://malware.wikia.com/wiki/Payload (accessed 2011)
20. CNN. Clone of 'Melissa' virus infects the Internet, http://articles.cnn.com/2001-04-19/tech/virus.matcher_1_melissa-bug-windows-address-original-melissa-virus?_s=PM:TECH
21. CNN. Destructive ILOVEYOU computer virus strikes worldwide, http://articles.cnn.com/2000-05-04/tech/iloveyou.01_1_melissa-virus-antivirus-companies-iloveyou-virus?_s=PM:TECH
22. CNN. CIH virus may hit on Monday, http://www.cnn.com/TECH/computing/9904/23/cihvirus.idg/index.html?iref=allsearch
23. Schlegel, R., Zhang, K., Zhou, X., Intwala, M., Kapadia, A., Wang, X.: Soundcomber: A Stealthy and Context-Aware Sound Trojan for Smartphones. In: Proceedings of the 18th Annual Network & Distributed System Security Symposium (NDSS 2011), San Diego, CA, pp. 17–33 (2011)
24. Stafford, T., Urbaczewski, A.: Spyware: The Ghost in the Machine. Communications of The AIS (2004)
25. Mirkovic, J., Prier, G., Reiher, P.: Attacking DDoS at the Source. In: Proceedings of the 10th IEEE International Conference on Network Protocols, Washington, DC, pp. 312–321 (2002)
26. Levy, E.: The making of a spam zombie army. Dissecting the Sobig worms. In: Proceedings in IEEE Security & Privacy, Oakland, CA, pp. 58–59 (2003)
27. Pfleeger, C., Pfleeger, S.: Security in Computing, 4th edn. Pearson Education Inc., Boston (2006)
28. Skoudis, E., Zeltser, L.: Malware: fighting malicious code. Prentice Hall PTR, Upper Saddle River (2003)

29. Lawton, G.: Is It Finally Time to Worry about Mobile Malware? Computer 41(5), 12–14 (2008)
30. Juniper Networks Malicious Mobile Threats Report 2010/2011, http://www.juniper.net/us/en/local/pdf/whitepapers/2000415-en.pdf (accessed May 2011)
31. Eeten, M., Bauer, J., Asghari, H., Tabatabaie, S.: The Role of Internet Service Providers in Botnet Mitigation: An Empirical Analysis Based on Spam Data. In: Proceedings of The Ninth Workshop on the Economics of Information Security (WEIS 2010), Cambridge, MA (2010)
32. Kinateder, M., Rothermel, K.: Architecture and Algorithms for a Distributed Reputation System. In: Nixon, P., Terzis, S. (eds.) iTrust 2003. LNCS, vol. 2692, pp. 1–16. Springer, Heidelberg (2003)
33. Dong, Z., Camp, L.: The Decreasing Value of Weak Ties in Recommended Networks. ACM SIGCAS Computers and Society 41(1) (2011)
34. Burr, W., Dodson, D., Polk, W.: Electronic authentication guideline: Recommendations of the National Institute of Standards and Technology (2006)
35. Jansen, W., Gavrila, S., Korolev, V.: Proximity-based Authentication for Mobile Devices. In: Proceedings of the 2005 International Conference, Las Vegas, NV, pp. 398–404 (2005)

A Conundrum of Permissions: Installing Applications on an Android Smartphone

Patrick Gage Kelley, Sunny Consolvo[2], Lorrie Faith Cranor, Jaeyeon Jung[1], Norman Sadeh, and David Wetherall[2]

Carnegie Mellon
[1] Microsoft Research
[2] University of Washington
{pkelley,lorrie,sadeh}@cs.cmu.edu, sunny@consolvo.org, jjung@microsoft.com, djw@cs.washington.edu

Abstract. Each time a user installs an application on their Android phone they are presented with a full screen of information describing what access they will be granting that application. This information is intended to help them make two choices: whether or not they trust that the application will not damage the security of their device and whether or not they are willing to share their information with the application, developer, and partners in question. We performed a series of semi-structured interviews in two cities to determine whether people read and understand these permissions screens, and to better understand how people perceive the implications of these decisions. We find that the permissions displays are generally viewed and read, but not understood by Android users. Alarmingly, we find that people are unaware of the security risks associated with mobile apps and believe that app marketplaces test and reject applications. In sum, users are not currently well prepared to make informed privacy and security decisions around installing applications.

Keywords: privacy, security, android, applications, smartphone, permissions, information design.

1 Introduction

Since the launch of the first Android phone in October 2008 the rise of the platform has been meteoric. Android phones accounted for over half of all smartphone sales as of Q3 2011 [6]. With each smartphone sold, more users are downloading applications from the Android Market. As of May 2011, Google reported that over 200,000 applications were available in the Android Market and that those applications had been installed 4.5 billion times in total [2].

Applications are not pre-screened, instead users are given the opportunity to decide which software to install on their phone. Android app rating and recommendation site AppBrain reports that there are now 310,000 applications in the Android market, and that 33 percent of those are rated at "low quality."[1]

[1] http://www.appbrain.com/stats/number-of-android-apps

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 68–79, 2012.

Additionally, according to a 2011 Juniper Networks report, and follow up press release, they found "a 472% increase in Android malware samples since July 2011 [to November 2011]" [8]. Similar studies from McAfee [11], Kaspersky Lab [12], and Symantec are all reporting continued exploits.

Juniper attributes this rise to the ease of posting Android applications to the market, as they state: "all you need is a developer account, that is relatively easy to anonymize, $25 and you can post your applications. With no upfront review process, no one checking to see that your application does what it says."

While some believe this openness is harmful to users, Google has promoted it. In one of Google's many tributes to openness, Senior Vice President of Product Management, Jonathan Rosenberg wrote, "At Google we believe that open systems win. They lead to more innovation, value, and freedom of choice for consumers, and a vibrant, profitable, and competitive ecosystem for businesses" [13]. As such, there has been no certification process for Android developers, nor pre-review of applications before they enter the Android Market, though applications reported as malicious have been later removed.

The market requires users to make two choices when reviewing potential applications for their device.

1. Do I believe this application will compromise the security and function of my phone if I install it?

2. Do I trust this developer and their partners with access to my personal information?

This leaves users left to leverage word-of-mouth, market reviews and ratings, and the Android permissions display to assist users in making decisions that protect their mobile privacy and security. We conducted a series of 20 semi-structured interviews to better understand how users navigate the Android Market, install and use third-party applications, and comprehend the decisions they make at install time.

In the remainder of this paper we will detail related work on users' understanding of privacy and access control concepts as well as the current state of Android security/permissions, our interview methodology, the demographics and expertise of our participants, and finally a collection of participant responses that qualitatively detail their ability to make decisions in the Android ecosystem.

2 Related Work

While Android has only existed publicly since 2008, a significant amount of work has been conducted on studying the Android permissions/security model. Much of this work focuses on creating theoretical formalizations of how Android security works or presents improvements to the system security, and is largely out of scope. Enck et al.'s work with TaintDroid has bridged the gap between system security and user-facing permissions, focusing on analyzing which applications are requesting information through permissions and then sending that data off phone [4].

Follow up work by Hornyack et al. detailed a method for intercepting these leaked transmissions and replacing them with non-sensitive information [7]. This functionality would allow users post-installation privacy-control. In their investigation they detailed the current permission requests of the top 1100 applications in the Android Market as of November 2010. However, our work, which tests users' understandings of the most common of these permissions, finds users have great difficulty understanding the meaning of these terms. Thus, giving users the ability to limit on a case-by-case basis would likely be ineffective without assistance.

Work by Vidas et al. has also studied how applications request permissions, finding prevalent "permissions creep," due to "existing developer APIs [which] make it difficult for developers to align their permission requests with application functionality" [15]. Felt et al., in their Android Permissions Demystified work, attempt to further explain permissions to developers [5]. However, neither of these papers explore end-users understanding of permissions. In our own work we find users attempt to rationalize why applications request specific permissions, trying to understand the developers' decisions, even if their understanding of these requests is flawed.

Others who have looked at Android permissions have attempted to cluster applications that require similar permissions to simplify the current scheme [3] or have attempted a comparison of modern smartphone permission systems [1]. Their work finds that Android permissions provide the most information to users (compared to other modern smartphone OSs such as Symbian, Windows Phone 7, and iOS), however our interviews show that much of the information provided is not understood.

Research in privacy policies, financial privacy notices, and access control have all similarly shown that privacy-related concepts and terms are often not well understood by users expected to make privacy decisions [9,10,14]. Our earlier work specifically investigated how the information display of privacy policies could influence understanding, focusing on standardized formats, terms, and definitions. While the Android ecosystem uses a standard format and terms, clear definitions are not readily available to users.

3 Android Permissions and Display

Android app permissions are displayed to users at the time they decide to install any third-party app through the Android Market on the web or on the phone. Apps downloaded from third-party app stores (e.g., onlyAndroid, the Amazon Appstore for Android, etc.) do not necessarily show full permissions on their websites, however upon installing the application package (APK) the user is presented with a permissions screen variant.

Permissions are shown within the Android Market as detailed in the following diagram, Figure 1. A user browses applications using the view shown in Screen 1. Here there is a truncated description, information about ratings, reviews, screenshots, etc. If a user decides to install they click the button labeled with the price of the application, here **FREE.** This brings them to Screen 2, where they are

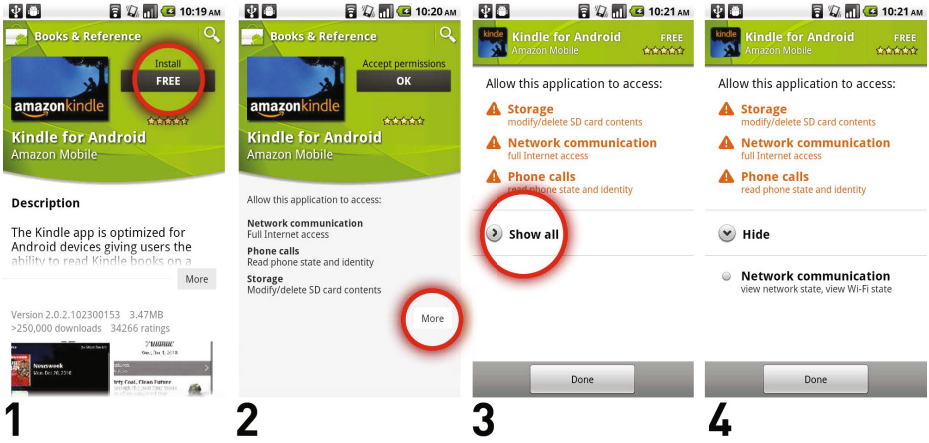


Fig. 1. The figure above shows the workflow for installing applications and viewing application permissions. Screen 1 shows the Amazon Kindle application as displayed in the Android Market. If a user were to click "FREE," circled in red, they are shown Screen 2, which allows them to Accept permissions and install the application, or to click the "Show" button which leads the user to Screens 3 and 4.

given a short list of permissions. If users double tap the **FREE** button on Screen 1, they skip Screen 2 and essentially approve the permissions without reading. Though Screen 2 serves the sole purpose of an interstitial permissions display between the market and a purchase decision, the complete list of permissions is not displayed.

To explore the full permission request they would click the **More** expander, bringing them to Screen 3. Here they would see a more complete list of permissions with some permissions shown in red and a **Show all** button, which displays the entire list if toggled.

At no point in this process is there an explicit way for users to cancel. The only way for users to not install the application after viewing the permissions is to use the physical back or home buttons on their phone.

The default permissions and groups in the Android SDK are detailed at Android's developer site.[2] The human readable terms are not included in the Android documentation.

4 Methodology

To reach a deeper and more nuanced understanding of how people navigate the current Android ecosystem, we conducted semi-structured interviews in Summer

[2] `http://developer.android.com/reference/android/Manifest.permission.html` and `http://developer.android.com/reference/android/Manifest.permission_group.html`

2011 with 20 participants from Pittsburgh and Seattle. The interviews were exploratory in nature, seeking broad understanding of participants' interactions with their smartphones as well as diving deeply into issues surrounding the display of permissions, the safety of the Android Market, and possible harms of information sharing.

We recruited participants through flyers around each city and local Craigslist postings. Each candidate filled out a short pre-survey online before the interview, which allowed us to confirm they did use an Android-enabled smartphone. Those participants who opted into the subsequent interview arrived at our labs and completed our consent form allowing us to make an audio recording of their interview. Following the interview participants were given the opportunity to opt-in to share their application information with us, collected through a script running on a local laptop, which we connected their phone to via USB while they watched.

Participants' quotes throughout the remainder of the paper are taken from transcriptions made from the audio recordings of the interviews. Participants were paid $20 for successful completion of the interview, in the form of their choices of Target, Starbucks, or Barnes & Noble gift cards.

5 Demographics and Survey Responses

Our online survey was completed by 77 participants, 20 of whom completed the lab interview. The remainder of this paper will discuss solely those 20 users, whose demographic information and survey responses are summarized in Table 1. Participants P1-P6 are from Seattle, P7-P20 from Pittsburgh. 10 participants are female, and 10 are male. The ages of our participants range from 19 to 48, with an average of 29. Six of our participants were in tech-related fields, the other fourteen were not. Fourteen of our participants have been using Android for less than a year, five participants reported up to two years of use, and only one reported having used Android for more than two years.

6 Results and Discussion

The following sections detail our findings and participants' thoughts on various parts of the Android ecosystem. We begin with the responses to six of the ten permissions we asked participants to explain. These responses highlight the broad range of often inaccurate knowledge around the human-readable terms Android provides to users at application install. Next, we discuss general concerns, response to Android in the media, and awareness of malicious applications.

6.1 Permissions Display Understanding

Half of our participants mentioned the existence of the permissions display before being prompted. When a participant did mention the display, we immediately showed a paper example of one (using the Facebook, Pandora, or Amazon

Table 1. Overview of our 20 survey participants. Columns 2-4, list their age, gender, and industry. Columns 5-8 list their phone provider, phone model, Android OS version, and the amount of time they have primarily used Android devices. Columns 9 and 10 show the number of apps they have downloaded and the number they report frequently using. All information is self-reported.

Participant overview

#	Gender	Age	Occupation	Phone provider	Phone model	OS version	Time Using Android	# Apps downloaded	# Apps really used
1	Female	24	Education	Verizon	LG Ally	I am not sure	1-6 months	1-10	A few 1-5
2	Male	48	Other	Verizon	HTC Incredible	Froyo	1-6 months	11-25	A few 1-5
3	Male	44	Agriculture	T-Mobile	Motorola Cliq	Cupcake	1-2 years	101+	A ton 20+
4	Male	19	Food Service	T-Mobile	Galaxy S	Eclair	1-6 months	11-25	A bunch 6-20
5	Female	45	Legal	Sprint	HTC EVO 4G	Honeycomb	1-6 months	1-10	A bunch 6-20
6	Female	26	Retail	Sprint	Samsung Replenish	I am not sure	1-6 months	1-10	A bunch 6-20
7	Female	34	Engineering	T-Mobile	LG Optimus	Eclair	7 months-1 year	11-25	A few 1-5
8	Male	23	Computers	Verizon	Motorola Droid X	Gingerbread	7 months-1 year	26-100	A ton 20+
9	Female	25	Other	Verizon	Motorola Droid X	I am not sure	Less than 1 month	1-10	A few 1-5
10	Male	32	Engineering	T-Mobile	HTC G2	Eclair	7 months-1 year	11-25	A bunch 6-20
11	Female	21	Entertainment	Sprint	Something Samsung	I am not sure	1-6 months	1-10	A few 1-5
12	Female	22	Other	T-Mobile	HTC MyTouch 4G	I am not sure	7 months-1 year	11-25	A few 1-5
13	Female	21	Don't work	Sprint	HTC Evo Shift	Gingerbread	1-2 years	1-10	A few 1-5
14	Male	20	Real Estate	Verizon	Motorola Droid X	Gingerbread	1-2 years	101+	A bunch 6-20
15	Male	36	Media / Publishing	Verizon	Motorola Droid 2	Froyo	7 months-1 year	1-10	A few 1-5
16	Male	22	Engineering	Sprint	HTC EVO 4G	Gingerbread	1-6 months	26-100	A bunch 6-20
17	Male	22	Don't work	Verizon	Motorola Droid 2	I am not sure	1-2 years	26-100	A bunch 6-20
18	Female	23	Other	T-Mobile	HTC G2	Gingerbread	More than 2 years	26-100	A bunch 6-20
19	Male	46	Engineering	AT&T	Google Nexus One	Gingerbread	1-2 years	26-100	A bunch 6-20
20	Female	21	Engineering	AT&T	Galaxy S II	Gingerbread	Less than 1 month	1-10	A few 1-5

Kindle permissions, Screen 3 of Figure 1). Many reported reading, or at least skimming, these displays with some regularity, though also admitted they did not necessarily understand all of the terms used.

Participants were able to identify these screens, recognized them immediately, and occasionally felt very strongly about them. When asked if he read these screens frequently, one such participant said, "Yeah, all the time. It is just so easy for those apps to do whatever they want, it's a way to protect yourself I guess. Call me paranoid."

Some participants stated that they were not sure how trustworthy the permissions display was. One said of it, "Is it a requirement to be on there [the market] that the software tells you what it is accessing ... Are they required to notify me or not, I don't know."

Unfortunately, most participants do not believe they understand the terms used and have not gone out of their way to learn what they mean. We showed a list of ten permissions with the permission group label, in the fashion they would be shown in the permissions display, to each user and asked them to explain to us their understanding of each term (as if they were explaining it to a relative or friend who was less tech-saavy). Participants reacted to this task with consternation.

Here we present a selection of common, surprising, and strained responses that we received on six of the ten terms we tested.

– **Network communication: full Internet access**

  Of the 1100 applications reported on in Hornyack's work [7], full Internet access is by far the most requested permission, requested by 941 of the 1100 applications, or 85.5% of those surveyed. Our participants were aware of what the Internet is and understood why applications needed it. However how applications have access to it, why they would need to specify it, and how applications would function without it were often unclear.

  - "That you can have access to all kinds of websites, even the protected ones." –P1
  - "I would say, this just requires a data plan, and you would need to have Internet access." –P6
  - "Any app that needs to get information from somewhere other than that is local on the phone." –P7
  - "For this game to be active, it require Internet access, I cannot play it offline." –P11
  - "I would guess that this means, no I don't know. I just assume that it is like taking up data plans. Using stuff with your data plan." –P12

– **Phone calls: read phone state and identity**

  Read phone state and identity is a compound Android permission which leads to participants only correctly anticipating part of the functionality granted. While most of our participants correctly identified functionality related to phone state, the idea that that the phone has unique IDs that are

also being revealed with this permission was lost on most users (P18 notes a phone ID, but adds an incorrect ability, location). While some applications are requesting this permission to actually detect phone state, many current advertising packages require IDs.

- "I would assume it would probably be along the lines of, it knows when my phone is sleeping or in use or in a phone call, and the type of phone" –P2
- "Phone state whether it is on or off, and identity I would assume it is like my telephone number." –P3
- "So it knows whether or not I am in the middle of a call? I don't really know what that part [identity] means." –P13
- "Know where you are, and what phone ID you are on, what type of phone it is." –P18
- "If you are on the phone maybe it shuts itself off. ... Maybe like your carrier? Hopefully not like *who* you are." –P19

– **Storage: modify/delete SD card contents**

Modification and deletion rights themselves were reasonably well understood (largely using metaphors to computers or thumb drives), however what was stored on the phone itself, compared to the external SD card was often misunderstood or simply not disambiguated.

- "That I am about to reach my capacity, or I need to get a new one." –P1
- "Basically, just saving on your memory card or harddrive." –P6
- "That is for games and things to save your play, store information as needed." –P10
- "It can see what is on my SIM card and on the phone itself." –P13

– **Your location: coarse (network-based) location**

While we showed participants both types of location that can be collected within Android, participants largely understood that "fine (GPS) location" meant their exact position. It was the coarse location that seemed to confuse more participants. They all understood it was location related, but there was large deviation on how exact that location was.

- "No, I don't. I haven't the foggiest idea of what that means." –P3
- "Your network based location, I don't know the difference between the GPS, but basically where you are at." –P6
- "This is essentially just where your network is located, based on maybe I guess cellphone tower triangulation." –P10
- "I would guess that this is like the source of your data, like a satellite of some sort." –P12
- "Is coarse location, does that have anything to do with like, when you have phone service and are in range or roaming?" –P13

– **Your personal information: read contact data**

Nearly all participants understood that this permission was requesting their address book, or full contact list. Some gave examples of purposes why this was needed, citing apps that could use this (P7, P18). A few participants were confused due to the permission group label "your personal information." As a result, like P11, they thought it was reading only data about themselves.

- "I would think that would mean my contacts list." –P2
- "Like Facebook, and if it was syncing with contacts." –P7
- "My phone number." –P8
- "My personal information can reach them, my name, address, phone number, email address." –P11
- "Your phone number. They go into your phone, your contacts, and then on Skype they get the number, and he is your friend in your phone. I guess that is what this is." –P18

– **Your accounts: act as an account authenticator**

This permission was rarely correctly identified (P3, while being unsure, has the right idea), and often described as scary. P12 explicitly said it "freaked" her out. The accounts that participant thought could be "authenticated" or, controlled, were frequently not associated with the application itself, with many participants believing applications that asked for this permission would have much wider ranging abilities.

- "Controlling the account? I don't know. I have zero idea." –P2
- "That I don't like, I don't know what it means, ... my impression is that instead of me being able to authorize something, that application is saying it can." –P3
- "That freaks me out. What does that mean exactly, cause I am not quite sure." –P12
- "I dunno is that associated with my T-mobile account?" –P13
- "I don't know, I guess it is in charge of whatever accounts you open up." –P18

As seen above, for each of the permissions we received answers that we would grade as a misunderstanding. For some of the more obscure permissions, participants simply admitted they didn't know, or gave up. None of our participants correctly understood all of the permissions, and most participants simply repeated the words given in the human readable description, a sign they may not have had complete understanding of the of the concepts.

Participants asked questions throughout about why applications needed the access they requested. Participants frequently asked the interviewer for examples of applications that requested the permissions we listed, as well as why they were needed. The relationship between the applications and the permissions they requested seemed, without assistance, unknowable.

One participant, when asked if she thought others understood these permissions said, "No. I mean for me to have to think as much, and I have been using these things, and have been sort of a tech-geek for years. Yeah, thats concerning." With Vidas and Felt finding that developers are misunderstanding permissions, and often applying them without need, and self-proclaimed "tech-geeks" finding the terms difficult, common users are left near helpless. The system and terms as they currently stand have not been created or explained for the average user.

6.2 Application Selection

While permission information is one vector to assist users in selecting which applications to install, many of our participants reported heavy reliance on star ratings, full text reviews, and word of mouth. These other sources of information were better understood and more trusted.

While reading through the reviews was seen as time-consuming, word of mouth was a trusted way to find high quality applications. One participant recounted his frustrations with simply searching the store and why he trusted others' opinions: "I feel it is very much a trial and error exercise. And that, I don't know whether that app is a piece of crap or whether it works. So when I know somebody that tells me that this app is good, that really means a lot to me."

Participants also reported hearing about apps, largely of services and products they already used, through advertisements. One participant described his experience with seeing Android app ads, "I have seen magazines and billboards. The phones and the applications. For instance Time Magazine, they have written you can also download the application."

While most of our participants said they do not purchase apps at all, others said in certain cases they would. P6 said, "I try to look for the free ones first, and if I can't find any free ones I will go ahead and buy it."

6.3 Concern over Malicious Applications

We asked participants if they had heard anything about Android phones or Android applications in the news, media, or on the Internet. Participants told us about Android's increasing market share, comparisons between iOS and Android, and about a few well advertised apps.

When asked a follow up, to specifically inquire on their awareness of malicious applications in the Android Market, our participants were largely unaware of any such activity. While some said they had meant to, or were intending to install anti-virus applications on their phones, most were unconcerned about the threat of malware.

We attribute this lack of concern to two strands we picked up throughout the interviews. The first is an expected coping mechanism that many participants admitted to, a lack of trust in new technology. For example, participants reported an unwillingness to do banking from their phone. One participant said "I don't do banking online through my phone because that doesn't seem particularly safe to me.... I prefer an actual desktop for that because I am paranoid."

The second part of this lack of concern towards malicious apps shows a deeper misunderstanding of the Android ecosystem. All of our participants, without exception, believed (or hoped) that Android, the entity, was pre-screening applications before entrance into the market. Participants elaborately described the reviews that they thought were taking place, screening not just for viruses or malware, but running usability tests (on users!), blocking applications that were too repetitive, or even screening out applications not enough people would want. They believed Android was checking for copyright or patent violations, and overall expected Android to be protecting their brand.

Additionally, people were unaware of who was actually running Android. They saw it as a vague entity, that they could not attribute to any specific parent company. Some knew and some guessed it was Google, others realized they had never stopped to think about that before and were simply unable to attribute the OS to any other company.

7 Conclusion

Users do not understand Android permissions.

Specifically, the human-readable terms displayed before installing an application are at best vague, and at worst confusing, misleading, jargon-filled, and poorly grouped. This lack of understanding makes it difficult for people, from developers to nontechnical users, to make informed decisions when installing new software on their phones. Largely, the permissions are ignored, with participants instead trusting word of mouth, ratings, and Android market reviews.

Users also are largely uninformed about the existence of malware or malicious applications that could be in the Android market. They have difficulty describing the possible harm that could be caused by applications collecting and sharing their personal information. While participants stated they try to find good applications in the market, they believe they are protected by oversight processes which do not exist.

Overall, users are not currently well prepared to make informed privacy and security decisions around installing applications from the Android market.

Acknowledgments. The authors would like to thank Intel Labs Seattle for their sponsorship of this work. We acknowledge our colleagues at Intel Labs Seattle, Microsoft Research, the University of Washington, and Carnegie Mellon University, including Seungyeop Han, Peter Hornyack, Jialiu Lin, Stuart Schechter, and Tim Vidas. Additional support was provided by the National Science Foundation under Grants CNS-1012763 (Nudging Users Towards Privacy) and DGE-0903659 (IGERT: Usable Privacy and Security). Additional support was provided by NSF grants CNS-0905562 and DGE 0903659, by the CMU/Portugal ICTI Program, by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office as well as Google.

References

1. Au, K.W.Y., Zhou, Y.F., Huang, Z., Gill, P., Lie, D.: Short paper: a look at smartphone permission models. In: Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, SPSM 2011 (2011)
2. Barra, H.: Android: momentum, mobile and more at Google I/O. The Official Google Blog (2011), http://googleblog.blogspot.com/2011/05/android-momentum-mobile-and-more-at.html
3. Barrera, B., Kayacik, H.G., van Oorschot, P.C., Somayaji, A.: A methodology for empirical analysis of permission-based security models and its application to android. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS 2010 (2010)
4. Enck, W., Gilbert, P., Chun, B., Cox, L.P., Jung, J., McDaniel, P., Sheth, A.: TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In: Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation, OSDI 2010 (2010)
5. Felt, A.P., Chin, E., Hanna, S., Song, D., Wagner, D.: Android Permissions Demystified. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011 (2011)
6. Gartner: Gartner Says Sales of Mobile Devices Grew 5.6 Percent in Third Quarter of 2011; Smartphone Sales Increased 42 Percent (2011), http://www.gartner.com/it/page.jsp?id=1848514
7. Hornyack, P., Han, S., Jung, J., Schechter, S., Wetherall, D.: These aren't the droids you're looking for: retrofitting android to protect data from imperious applications. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011 (2011)
8. Juniper Networks. Mobile Malware Development Continues To Rise, Android Leads The Way (2011), http://globalthreatcenter.com/?p=2492
9. Kelley, P.G., Bresee, J., Cranor, L.F., Reeder, R.: A "nutrition label" for privacy. In: The 5th Symposium on Usable Privacy and Security, SOUPS 2009 (2009)
10. Kleimann Communication Group, Inc. Evolution of a Prototype Financial Privacy Notice (2006), http://www.ftc.gov/privacy/privacyinitiatives/ftcfinalreport060228.pdf
11. McAfee Labs. McAfee Threats Report: Third Quarter 2011 (2011), http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q3-2011.pdf
12. Namestnikov, Y.: IT Threat Evolution: Q3 2011 (2011), http://www.securelist.com/en/analysis/204792201/IT_Threat_Evolution_Q3_2011
13. Rosenberg, J.: The meaning of open. The Official Google Blog (2011), http://googleblog.blogspot.com/2009/12/meaning-of-open.html
14. Smetters, D.K., Good, N.: How users use access control. In: Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS 2009 (2009)
15. Vidas, T., Christin, N., Cranor, L.F.: Curbing Android Permission Creep. In: W2SP 2011 (2011)
16. Wetherall, D., Choffnes, D., Greenstein, B., Han, S., Hornyack, P., Jung, J., Schechter, S., Wang, X.: Privacy Revelations for Web and Mobile Apps. In: HotOS 2011 (2011)

Appendix: Interview Questions

The entire interview guide, as well as additional quotes and some coded data, can be found online at http://patrickgagekelley.com/research/android.

Methodology for a Field Study of Anti-malware Software

Fanny Lalonde Lévesque[1], Carlton R. Davis[1], José M. Fernandez[1], Sonia Chiasson[2], and Anil Somayaji[2]

[1] École Polytechnique de Montréal, Montréal Canada
{fanny.lalonde-levesque,carlton.davis,jose.fernandez}@polymtl.ca
[2] Carleton University, Ottawa Canada
{chiasson,soma}@scs.carleton.ca

Abstract. Anti-malware products are typically evaluated using structured, automated tests to allow for comparison with other products and for measuring improved efficiency against specific attacks. We propose that anti-malware testing would benefit from field studies assessing effectiveness in more ecologically valid settings. This paper presents our methodology for conducting a 4-month field study with 50 participants, including discussion of deployment and data collection, encouraging retention of participants, ethical concerns, and our experience to date.

Keywords: anti-malware testing, field study, user study.

1 Introduction

How should the effectiveness of anti-malware software be assessed in practice? Current strategies typically involve automated testing against standard datasets, sometimes with automated user profiles to imitate user interaction with security messages [1]. Even with the more advanced tests that include user profiles, this assumes that users' behaviour and all of the variables affecting their computing environments can be predicted and reflected in these automated profiles.

We suggest that many infections are due to direct or indirect user actions that allow malware to infect a system. These actions (or inactions) may occur immediately prior to infection, weeks or months prior to infection, or may even occur over time so that a combination of actions lead to a vulnerable system state. These situations would not be accurately reflected in automated testing, nor would they be identifiable through traditional lab-based user testing.

One alternative is to conduct long-term field studies of anti-malware software with real users in more ecologically valid settings. By monitoring real usage over time, one can gain a better understanding of how anti-malware systems are used and how external factors influence their effectiveness. However, a large number of confounding variables exist which significantly complicates the data analysis. In our study, we wanted to provide a common and controlled "clean slate" to begin the experiment, but somehow allow users to take ownership over their

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 80–85, 2012.

system and use it as they would normally, while we monitor the system for signs of infection.

In this paper, we present our approach to conducting a field trial of an anti-malware product. Section 2 summarizes related work in anti-malware testing and conducting field studies of security products. Section 3 describes our methodology for conducting the trial, including our approach to selling laptops and reimbursing participants for their purchase throughout the study. Section 4 discusses how we addressed ethical and privacy issues, while Section 5 highlights our experience with this ongoing study. The paper concludes with a discussion of our anticipated analysis and brief description of a larger scale follow-up study.

2 Background

Although there are currently several methods for evaluating anti-malware products [7], they do not reflect the performance of products in real life. Typical evaluation methods are based on scanning collected or synthesized malware along with legitimate programs. While such approaches can measure raw detector accuracy, they cannot take into account factors such as user interactions, evolving threats, and different environments. One major issue is that the sample collection is often too small, inappropriate, and unvalidated [8,9]. Even with a well-maintained malware collection, testing against such data sets has become unreliable due to the increased dynamic nature of malware. To partially address this issue, Vrabec and Harley [13] proposed emulating user interaction with the system and creating user-specific testing scenarios.

In the broader security community, field studies of computer security are frequently advocated but are still relatively uncommon in the literature, likely due to the costs, time demands, and potential security and privacy risks to users. Recent field studies of security software have mostly involved evaluating the use of authentication mechanisms [3,6,4]. In 2009, Somayaji et al. [11] introduced the concept of computer security clinical trials. The conceptual proposal was to evaluate security products using methods and controls similar to those used in clinical trials of medical products, but no studies have been conducted thus far.

Ethnographic studies examine usage of security systems in the field, but use qualitative methods such as interviews, diaries, and observation to understand how and why participants interact with computer systems. Botta et al. [2] conducted an ethnographic study of security professionals, Rode [10] examined parental behaviour in protecting children's online safety, Wash [14] used interviews to understand users' mental models of security, and De Luca et al. conducted a field observation of ATM usage to evaluate PIN usage [5].

While some of the above studies mention anti-malware usage as a security measure taken by users, it is not the focus of these studies. To our knowledge, there are no published user studies focusing specifically on anti-malware usage.

3 Study Description

The goals of this study are to: (1) determine how phenomena such as the configuration of the system, the environment in which it is used and user behavior can affect the probability of infection of a system; (2) develop an effective methodology to evaluate anti-malware products in real-world environment; (3) determine how malware infects computer systems, and identify sources of malware infections. The study includes monitoring real-world computer usage through diagnostics and logging tools, monthly interviews and questionnaires, and in-depth investigation of any potential infections.

We are conducting a 4-month field study with 50 participants that were recruited through posters and newspaper advertisements on campus. A short online intake questionnaire was used to collect initial demographic information. Using these profiles, we categorized interested volunteers and randomly chose a sample from each category in order to have a diverse and representative sample of users that include students and employees from various fields.

3.1 Equipment

We supplied laptops with identical configuration to the participants. The following software was installed: Windows 7; the antivirus (AV) product to be evaluated; diagnostics tools, such as HijackThis, ProcessExplorer and Autoruns; and custom Perl scripts which we developed. We utilised the scripts to automate the execution of the tools as well as for compiling statistical data regarding the system configuration, the environments in which the system is used, and the manner in which the system is utilised. The AV product is centrally managed on our server. An AV client installed on the laptops sends relevant information to the server about any malware detected or suspected infections as they occur.

Before deployment, we benchmarked the laptops by running the diagnostics tools and recording the output. The information included: a hash of all files plus information about whether the files were signed; a list of auto-start programs; a list of processes; a list of registry keys; and a list of browser helper objects.

3.2 Procedure

The study consisted of 5 in-person sessions: an initial session where participants received their laptop and instructions, followed by monthly 1-2 hour sessions where we performed analysis to determine if the laptop was infected.

Participants initially purchased the laptops from us at a reduced rate; it was theirs to keep after the study. To encourage the participants to remain in the study, we paid them to attend the monthly in-person sessions. If participants complete all required sessions, the entire cost of the laptop would be reimbursed, along with an additional honorarium. We encouraged participants to configure their laptop as they desired and use it as they would normally use their own computer. The only restrictions applied during the experiment were that the participants do not format the hard drive, do not replace the operating system, and not install any other AV product on the laptop.

Each month, participants booked an appointment via an online calendar system hosted on our website. During these monthly sessions, participants completed an online questionnaire about their computer usage and experience, while the experimenter collected the local data compiled by the automated scripts. The questionnaire was intended to assess the participant's experience with the AV product and gain insights about how the laptop was used.

The data compiled by our scripts included, but was not limited to, the list of applications installed, the average number of hours per day the laptop is connected to the Internet, and the number of web sites visited. Diagnostics tools were also executed on the laptop to determine if infection was suspected. If the AV product detected any malware over the course of the month, or if our diagnostics tools indicate that the laptop may be infected, we requested additional written consent from the participant to collect data that will help us identify the means and the source of the infection.

Before the last visit, participants completed an online survey about their experience during the study. The aim of this exit survey was to identify activities or mindset that may have unduly influenced the experimental results. We chose to administer the survey apart from the in-person session in case participants were more comfortable revealing such information while not in the presence of the experimenter. In the last session, we requested that participants keep the experiment data stored on their laptops for an additional three months, so that if we discover that further analysis is necessary, we can contact them and seek their permission to collect and analyse the relevant data. Nonetheless, we provided a procedure for deleting the diagnostic tools and the scripts, as well as the experiment data stored on their laptop. We also explained that residual data may still remain on the laptop even after the experiment data is deleted. If they wanted to completely remove all traces of the experiment from their laptop, we referred them to external resources for re-imaging the laptop.

4 Ethical and Privacy Considerations

This study was reviewed and approved by the Computer Risks Evaluation Board (CREB) and the Research Ethics Board (REB) of École Polytechnique de Montréal. Ethical and privacy guidelines were of particular concern because the experiment involved the collection of personal data over an extended period of time.

To preserve anonymity, each participant was assigned an identification number such that the identity of the participant was not linked to any data during analysis. No personally identifiable information, such as usernames and passwords were collected, content of personal documents stored on the computer were not examined, and no exact URLs were collected (only aggregate data about categories of web sites such as "social networking" and "gaming").

Because the study involved malware, necessary precautions were taken to protect the university's infrastructure as well as that of the users. For example, in the event that an infection could not be cleaned by the AV product, we relayed

the relevant details to the AV company. The company developed and provided a product update to detect and remove the infection. This update was applied to participants' laptops as part of regular automated software updates.

5 Experience to Date

The study officially started in November 2011. The first step was to configure the laptops and meet all 50 participants individually to provide instructions, have them sign the consent form, and pay for their laptop. As noted earlier, the full cost of the laptop will be reimbursed, with an additional honorarium, provided that the participant attends all four monthly visits. Partial reimbursement will be provided if only some of the sessions are completed.

The study is ongoing. Participants have their laptops and the AV clients have been communicating with our AV server. Thus far (November 2011), a total of 18 malware incidences have been reported on eight of the laptops. Also, we know that at least one incidence is an actual infection; the participant informed us that a program on his laptop requested that he pay money to upgrade his AV software. The responsible program was confirmed to be a known malicious scareware that pretends to be legitimate security software [12]. All incidences will be explored when these participants return for their first monthly session.

The number of incidences in the first month is much higher than anticipated. This initial spike may be because participants were installing software and customizing their computer. We will be closely exploring and analysing the data collected during the first monthly sessions. It remains to be seen whether this rate of infection will persist throughout the rest of the study.

Most participants have expressed a high level of willingness to collaborate and some have even shown scientific interest in the study. Surprisingly, some participants asked us how they should act to get their laptop infected, to which we responded that they should use their laptop normally. In the event that participants need assistance, we provided them a telephone number and an email address that they could use to contact us. Other than the participant whose laptop has been infected with malware, only a few participants have contacted us via email to obtain support, and none of them has contacted us via telephone.

6 Conclusion and Future Work

This study is intended to demonstrate what we believe is a more effective way of evaluating anti-malware products: the main conjecture being that it is imperative that actual users be involved in the evaluation process, and that they use the products in realistic environment over an extended period of time. Our study is in progress, but results so far point to a rich data set that will provide evidence for how user behaviour and environments of use affects incidences of malware.

We will be analyzing data on a monthly basis to ensure that we are collecting appropriate data and to determine if any modifications are necessary. Overall, we intend to perform in-depth statistical analysis to determine whether there

is a correlation between user behaviour and incidences of infection, as well as probing specific incidences to fully understand the causes of infection.

We will use our findings to inform a second, larger study examining specific variables and confirming the results from the first study. This second study will be designed to determine which factors (i.e. type of AV product or user behaviour) has the most impact on incidences of infection of a system. It will compare multiple AV products, more participants will be involved, and the study will take place over a longer period of time, likely over 6 to 12 months. We hope the results of this follow-on work will help inform the design of future consumer-level security products.

Acknowledgement. This project has been funded by the NSERC Internetworked Systems Security Network (ISSNet), MITACS and Trend Micro.

References

1. Anti-Malware Testing Standards Organization: AMTSO testability guidelines. Tech. rep. (May 2011), `http://www.amtso.org/documents.html`
2. Botta, D., Werlinger, R., Gagné, A., Beznosov, K., Iverson, L., Fels, S., Fisher, B.: Towards understanding it security professionals and their tools. In: ACM Symposium on Usable Privacy and Security (SOUPS). ACM (2007)
3. Brostoff, S., Sasse, M.: Are Passfaces more usable than passwords? A field trial investigation. In: British Human-Computer Interaction Conference (HCI) (2000)
4. Chiasson, S., Biddle, R., van Oorschot, P.C.: A second look at the usability of click-based graphical passwords. In: ACM Symposium on Usable Privacy and Security (SOUPS) (2007)
5. De Luca, A., Langheinrich, M., Hussmann, H.: Towards understanding ATM security: a field study of real world ATM use. In: ACM Symposium on Usable Privacy and Security (SOUPS) (2010)
6. Florencio, D., Herley, C.: A large-scale study of WWW password habits. In: ACM World Wide Web Conference (WWW) (2007)
7. Gordon, S., Ford, R.: Real world anti-virus product reviews and evaluations - the current state of affairs. In: 19th National Information Systems Security Conference (NISSC) (1996)
8. Harley, D., Lee, A.: Who will test the testers? In: 18th Virus Bulletin International Conference (2008)
9. Košinár, P., Malcho, J., Marko, R., Harley, D.: AV testing exposed. In: 20th Virus Bulletin International Conference (2010)
10. Rode, J.A.: Digital parenting: designing children's safety. In: British HCI Conference (BCS-HCI) (2009)
11. Somayaji, A., Li, Y., Inoue, H., Fernandez, J.M., Ford, R.: Evaluating security products with clinical trials. In: Workshop on Cyber Security Experimentation and Test (CSET) (2009)
12. Stone-Gross, B., Abman, R., Kemmerer, R.A., Kruegel, C.: The underground economy of fake antivirus software. In: Workshop on the Economics of Information Security (WEIS) (2011)
13. Vrabec, J., Harley, D.: Real performance? In: EICAR Annual Conference (2010)
14. Wash, R.: Folk models of home computer security. In: ACM Symposium on Usable Privacy and Security (SOUPS) (2010)

My Privacy Policy: Exploring End-user Specification of Free-form Location Access Rules

Sameer Patil[1], Yann Le Gall[2], Adam J. Lee[2], and Apu Kapadia[1]

[1] School of Informatics and Computing, Indiana University, Bloomington, IN 47408
[2] Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260
{patil,kapadia}@indiana.edu, {ylegall,adamlee}@cs.pitt.edu

Abstract. The increasing inclusion of location and other contextual information in social media applications requires users to be more aware of what their location disclosures reveal. As such, it is important to consider whether existing access-control mechanisms for managing location sharing meet the needs of today's users. We report on a questionnaire ($N = 103$) in which respondents were asked to specify location access control rules using free-form everyday language. Respondents also rated and ranked the importance of a variety of contextual factors that could influence their decisions for allowing or disallowing access to their location. Our findings validate some prior results (e.g., the recipient was the most highly rated and ranked factor and appeared most often in free-form rules) while challenging others (e.g., time-based constraints were deemed relatively less important, despite being features of multiple location-sharing services). We also identified several themes in the free-form rules (e.g., special rules for emergency situations). Our findings can inform the design of tools to empower end users to articulate and capture their access-control preferences more effectively.

1 Introduction

The popularity of online social networks has resulted in an unprecedented amount of sharing of personal information. Furthermore, the extensive use of mobile devices enables and encourages broadcasting *contextual* information wherever one happens to be. For instance, location-sharing systems, such as Facebook Places, Google+, and Foursquare, allow users to share their current location with friends. Recent technologies like Cenceme [10] can determine the current activity (e.g., "running" or "dancing") from a smartphone's onboard sensors. With the growing availability of ways to share personal contextual information, personal privacy management has become increasingly important and also more difficult.

Several studies have examined location-sharing preferences of end users. However, most prior work has focused on user specification of simple rules for controlling location disclosure. For example, many location-sharing systems — including commercial systems mentioned above as well as those in the research literature [4, 10, 13, 17, 18] — allow users to set up access-control rules based only upon *who* is accessing their location, or *when* this information is being accessed. Given the increasing adoption of location sharing, whether these types of simple rules are sufficient for capturing the access-control preferences of today's social media users is an open question.

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 86–97, 2012.

Toward this end we set out to understand (i) which contextual factors are deemed important by users when developing rules for controlling access to their location, and (ii) how users express access-control rules in everyday language. Understanding the importance of various contextual factors in location-sharing decisions can help guide the design (in terms of both features and user interface) of frameworks for authoring structured personal policies for location sharing. Further, understanding how users express location access-control rules using everyday language can provide insight into how tools for rule specification should be realized: e.g., imprecise free-form rules support the case for designing more structured editors to capture user intent, while high precision statements motivate natural language (i.e., 'Siri-like') interfaces.

We report on an online questionnaire conducted to explore these issues. The questionnaire asked respondents about preferences for access to their location. In particular, respondents rated and ranked the importance of a variety of contextual factors (such as the recipient of location information, the time of day, and the disclosure specificity) in making location-sharing decisions. We also collected free-form natural-language statements in which respondents described how they wish to manage access to location. Some interesting findings from our data include the following:

- The recipient of the location information was the most highly rated and ranked factor. This finding echoes prior research. However, the time and the day of location disclosure exhibited the lowest ratings and rankings, which was unexpected.
- Respondents found it difficult to express complex or even complete location-sharing rules in everyday language. Further, the factors mentioned in these statements often did not reflect the relationships observed in numeric ratings and rankings of the same factors.
- Participants did not seem to consider social nuance and technical limitations of their policy statements, such as the social implications of denying access to someone or the inability to revoke location disclosures that had already taken place.
- The rules included several recurring themes and factors. For example, many respondents desired means to facilitate location access during emergency situations and to exercise manual controls, such as the ability to apply temporary blocks on location tracking.

These findings can inform how location-sharing systems could be made more privacy-sensitive. For instance, in addition to recipient-based access control, location-sharing systems often offer settings based on temporal considerations. Therefore, it is notable that *time* and *day* were rated and ranked lower than other contextual factors. This result suggests that the usability of privacy controls in current location-sharing systems might not be well aligned with user preferences; systems rarely provide the ability to specify privacy preferences for other factors indicated as being important, e.g., frequency of access or one's current location. Also, the low expressivity of the free-form access control statements suggests several potential interpretations and implications. It might be the case that people are generally not able to articulate their privacy preferences. If so, designers can aid the creation of policy statements using structured rule specification interfaces. On the other hand, users may not be adequately motivated to specify details.

2 Related Work

Prior work on access control in location-sharing applications falls into two broad categories: factors influencing sharing and idioms for privacy-policy expression. We briefly survey key prior work and indicate differences with our study.

Factors influencing sharing: Lederer et al. [8] conducted a study to determine the relative importance of two factors: the recipient of location information and the user's current situation. They found that the recipient had a larger influence on privacy preferences. Consolvo et al. [4] conducted an experience-sampling study in which 16 participants responded to simulated location requests. They found that location disclosure was influenced by the recipient of location information, the reason for the request, and the level of detail revealed. Tsai et al. [18] discuss field deployment of *Locyoution*, a location-sharing application integrated with Facebook. They discovered that user comfort with sharing increased when given feedback regarding who accessed their location. However, Locyoution users were limited to time-based location-disclosure rules. Toch et al. [16] engaged in a four-week field study of the *Locaccino* system using a statistical approach to examine the relationship between locations and corresponding privacy preferences. Their data showed that users tended to feel more comfortable sharing in public places visited by many people. Wagner et al. [19] further carried out a 16-participant study in which subjects drawn from a university population were trained in the use of the Locaccino system and questioned about sharing preferences. They found that highly-granular location information was shared only when there was a perceived need and that subjects preferred not to broadcast location. Benisch et al. [1] collected location data from the phones of 27 people for 3 weeks. They observed that participants were more comfortable with location- and time-based policies to share with friends, family, or advertisers. A recent study by Schlegel et al. [15] found that individual perceptions of privacy loss varied greatly according to who was accessing the individual's location and how often this location was accessed.

Idioms for policy expression: Brodie et al. [3] examined the utility of natural-language policy authoring in the SPARCLE policy workbench. Employees from various organizations were tasked with crafting organizational policies, which were converted to XACML using shallow parsing. This work demonstrated the viability and utility of allowing people to specify certain types of policies using free text. Sadeh et al. [14] studied privacy concerns in the context of *PeopleFinder*, a location-sharing system for laptops and mobile devices. Lab experiments and field studies showed that PeopleFinder users were often dissatisfied with the location disclosures that their rules permitted, even after revising initial rules. Users were, however, consistent in their (dis)satisfaction feedback regarding location disclosures; the authors propose using this feedback to bootstrap machine learning techniques to generate and refine disclosure rules. Kapadia et al. [5] studied the use of usable metaphors such as *virtual walls* to control access to contextual data and showed that such metaphors were easy to understand and use. Their work, however, did not address user policies for using such metaphors.

Our work differs from previous work in a number of important ways. Most prior location-sharing studies relied on sampling a couple dozen participants from university populations (largely students). On the other hand, we recruited over one hundred adults

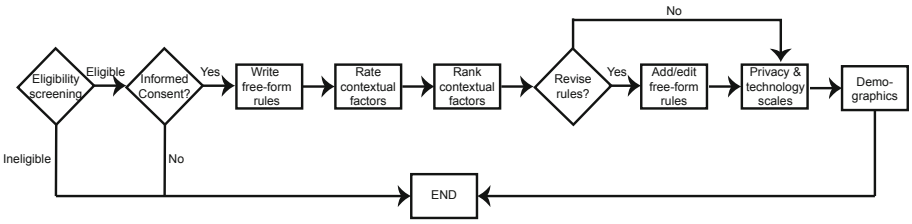


Fig. 1. Flow diagram of the questionnaire

spanning a wide age range and geographical area. Furthermore, previous studies typically focused on a small set of location-sharing factors sufficient for 'write-once' static disclosure rules. Our study analyzes the absolute and relative importance of a superset of these and other factors. We also performed a detailed analysis of the characteristics of over 200 rules written in free-form everyday language.

3 Method

We used an online questionnaire to investigate the research questions outlined above.

3.1 Questionnaire Structure

Figure 1 shows the flow of the various parts of the questionnaire. The questionnaire asked respondents to write free-form statements describing rules for allowing (or disallowing) access to information about their location via a location-sharing service. We provided four sample free-form rules as illustrative examples. To avoid priming respondents with respect to location-access rules, the example rules dealt with controlling access to an electronic health record (e.g., "Allow a nurse to view my EHR only when I am present in front of her and limit access to the record to the duration of my clinic visit."). Respondents were asked to specify such rules for location sharing. We did not limit the number of rules a respondent could specify. Collectively, these rules formed the respondent's privacy policy for a location-sharing service.

After specifying these rules, respondents were asked to *rate*, on a scale of 1 (Not at all important) to 5 (Very important), the importance of the following factors in determining whether to grant access to location information: (1) *who* will receive the location information, (2) the *reason* for the access, (3) the *time* of the day, (4) the *day* of the week, (5) the user's *present location*, (6) the *specificity* with which location is revealed, and (7) the *number of accesses* within a given period. The factors were presented in random order. We selected these factors because prior studies identified them as important for location-sharing decisions (see Section 2). We were also interested in how these factors are ordered relative to one another. Therefore, we next asked the respondents to *rank* the factors in the order of perceived importance for controlling access to location information.

In order to examine how various individual characteristics of respondents affected location-sharing preferences, the questionnaire also included assessments of the following measures: (1) *Online privacy concern*, which was measured using Internet Users' Information Privacy Scale (IUIPC) [9], and (2) *Interpersonal privacy concern*, which was measured using a scale from prior studies [7,12]. In addition, the respondents were asked about their experience using the Internet and smartphones. The questionnaire concluded by collecting demographic information.

3.2 Respondents

The questionnaire was advertised to a subject pool maintained by a university in Pittsburgh,[1] as well as in the Et Cetera Jobs category of the widely-used advertisement site Craigslist. To ensure broad geographical reach across the U.S., we advertised using the Craigslist sites for the cities of Los Angeles, Chicago, Atlanta, and Boston. As compensation, respondents were entered in a drawing for one of five rewards of $15.

Since privacy is culture-dependent, we chose a culturally-homogeneous sample by limiting participation to those who had lived in the U.S. for at least 5 years.[2] Prior research suggests that privacy attitudes and practices of undergraduate students are often different from those of older adults [11]. Therefore, we ensured that no more than 35% of respondents were in the 18–22 age group (i.e., the typical age range of undergraduates). An initial screening questionnaire was used to enforce these criteria.

As a check for detecting whether respondents completed the questionnaire attentively, we included eight 'verification' questions interspersed inconspicuously among other questions. These required the respondents to perform basic mathematical operations (e.g.,"What is 2 + 7?") or follow simple instructions (e.g., "Select option five."). We eliminated from consideration the responses of 31 respondents who did not answer all eight verification questions correctly. We also set browser cookies to reduce the likelihood of multiple submissions from the same respondent.

We received 103 valid questionnaire responses with 21 of these (20.4%) from respondents in the 18–22 range. In the sample 41 (40%) of the respondents were males and 60 (58%) were females.[3] The sample captures a broad age range; the ages of the respondents ranged from 18 through 61 years (median: 28, mean: 32, standard deviation: 12). The respondents were well-educated; 92% ($N = 94$) reported having attended college with 61% ($N = 62$) holding Bachelor's degrees or higher. The respondents also indicated being familiar with technology; 92% ($N = 95$) reported using the Internet for more than 7 years and 68% ($N = 70$) owned smartphones.

3.3 Coding of Free-form Access Rules

The free-form statements written by respondents were coded to mark whether or not the text was a rule for controlling access to the respondent's location. The first three

[1] The pool contains a diverse set of individuals from the community and not just university students.

[2] Prior research indicates that sufficient cultural assimilation can be assumed after 5 years [6].

[3] Two respondents did not provide gender information.

authors acted as three independent coders. The coders also marked whether any of the seven factors that the participants rated and ranked were present in the rules. Further, during the first coding pass, the coders individually identified common themes among the rules. These themes were labeled and agreed upon, and a second independent coding pass was made to mark whether any of these were present in each of the specified rules. The intercoder agreement was high (approximately 82%). All coding differences were collectively resolved until full intercoder agreement was reached. During this process the coders identified 5 respondents who seemed to have misunderstood the instructions (e.g., they wrote rules regarding health records instead of location). The responses of these individuals were removed from the set of valid responses. In the next section we describe the findings from the analysis of valid responses.

4 Findings

We analyzed our coding of the rules the respondents wrote and examined the numeric rating and rankings the respondents attached to the contextual factors we provided.

4.1 Analysis of Free-form Rules

In total the respondents wrote 321 free-form statements. Of these, 234 (73%) were judged as valid rules that could be used for managing access to location information. Notably, 15 (14.6%) respondents did not write a single valid rule (this includes 2 respondents who did not write any rules at all). On the other hand, all of the statements written by 63 (61.2%) were marked as valid rules. However, the number of rules written by most respondents was very small. Of the 88 respondents who wrote at least one valid rule, almost 80% wrote no more than three, with 32 (36.4%) writing one, 22 (25%) two, and 16 (18.2%) three, respectively. The average number of valid rules among the 88 respondents was 2.66/respondent, with a relatively large standard deviation of 2.12.

In addition, the coders identified a few common themes in the 234 valid rules beyond the seven factors we provided (see Section 3). These were:

- **Emergencies:** 26 (11.1%) rules specified permissions for emergency situations.
- **Manual control:** 24 (10.3%) rules reflected a desire for manual control over location sharing. Two types of manual controls were noted: deciding how to handle *each* access for location as it came in (17/234 = 7.3%), and deciding to share location only when explicitly 'checking in' (7/234 = 3%).
- **Do not track:** 37 (15.8%) rules reflected a desire not to have one's locations known or tracked at all. These were further split roughly equally into rules for complete and permanent disabling of location tracking under all circumstances (20/234 = 8.5%) and those for going 'offline' temporarily when desired (17/234 = 7.3%).
- **Current activity:** 15 (6.4%) rules pertained to the activity (e.g., shopping, partying, etc.) the person was engaged in when their location was accessed.

We also noted that two types of recipients — family/friends and the government — were mentioned frequently in the rules, but in contrasting ways. Rules were created to *allow* access to family/friends and to *deny* access to the government. Many respondents did, however, grant location access to the police during emergencies.

Table 1. Descriptive Statistics for Ratings and Rankings of Contextual Factors

Factor	Ranking					Rating					Rules
	N	Mean	SD	Median	Mode	N	Mean	SD	Median	Mode	
Recipient	103	2.07	1.69	1	1	103	4.79	0.68	5	5	175
Where one is when location is accessed	103	3.24	1.73	3	2	103	4.28	0.98	5	5	12
Specificity of disclosure	103	3.68	1.79	4	4	102	4.18	1.01	4	5	12
No. of times location is accessed in a given time	103	4.05	1.61	4	3	103	4.06	1.16	4	5	2
Reason for accessing location	103	4.31	1.70	4	3	102	4.53	0.96	5	5	45
Day of the week	103	5.13	1.51	5	6	103	3.36	1.31	3	3	2
Hour of the day	103	5.52	1.81	6	7	103	3.62	1.23	4	5	9

4.2 Ratings and Rankings of Contextual Factors

Table 1 provides descriptive statistics for the ratings and rankings of the seven contextual factors we provided. The table presents the factors ranked by their mean ranking score. It can be readily observed that most of the factors were rated as highly important when making decisions about location sharing, with modes of 6 out of the 7 factors being 5 (the highest value of importance). However, examining the frequency distributions of the ratings (see Fig. 2) suggests that the ratings did vary. This is also reflected in the differences in the rating means. Pearson's Chi-square test confirmed that the differences were statistically significant ($\chi^2 = 158$, df = 24, $p < 0.001$).

An exploratory statistical factor analysis suggested a four-factor solution with the following components: (1) recipient and reason for accessing location (*purpose*), (2) one's current location at the time of location access and the specificity with which location is revealed (*location*), (3) time of the day and day of the week (*time*), and (4) the frequency of location accesses (*frequency*). With the exception of the two temporal ratings (i.e., time and day), the ratings also showed a small positive correlation with the level of Internet privacy concern measured by the IUIPC score. The correlation coefficients for the individual ratings ranged between 0.2 to 0.33 and were statistically significant at the 0.05 level or better. In contrast, only the temporal ratings were correlated with interpersonal privacy ratings for non-professional relationships (i.e., significant other, ex, family, and friends). The correlations coefficients ranged between 0.19 to 0.3 and were statistically significant at the 0.05 level.

The rankings in Table 1 shed more light on the relative importance of these factors. The differences in ranking were statistically significant ($\chi^2 = 395$, df = 36, $p < 0.001$). It can be seen that the recipient of location information and where one is when location is accessed ranked at the top. Moreover, temporal aspects (such as specific times or days) ranked the lowest. The ranking of the factors mostly matched the rank order of mean ratings with one notable exception: the reason behind location access was ranked lower at 5 compared to its rank order at 2 in terms of rated importance.

As mentioned in Section 3, we also coded whether each of these factors was mentioned in the rules written by the respondents. It is seen in Table 1 that roughly 3/4th of

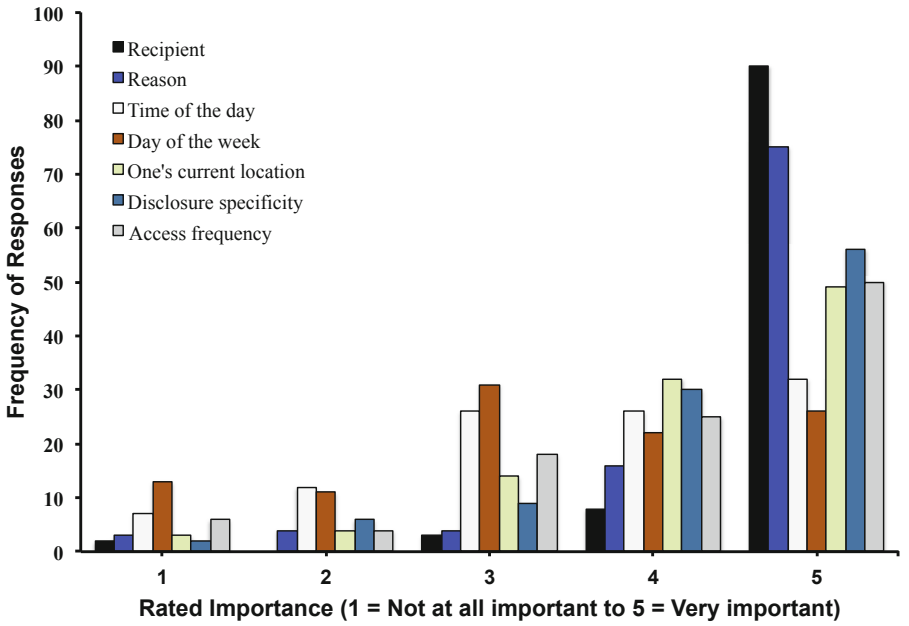


Fig. 2. Frequency Distribution of Ratings for the Importance of Contextual Factors

all rules (175/234 = 74.8%) were based on specific recipients. Almost 1/5th of the rules (45/234 = 19.2%) included specific reasons for location accesses. However, the rest of the factors were mentioned in only a handful of rules, despite being rated and/or ranked high in importance in terms of making location-sharing decisions.

The ratings and rankings did not exhibit any notable impact of smart phone and Internet use, or other demographic factors such as age, income, and education. However, we found that, compared to males, females assigned slightly more importance to the location recipient ($p < 0.05$) and specificity of disclosure ($p < 0.01$).

5 Discussion and Implications

Free-form rule specification: Our findings indicate that people find it challenging to articulate rules describing how access to their location should be controlled. A small but notable group of respondents (14.6%) was not able to do it at all, while most others could only specify one or two rules. As a result, it is likely that the set of rules of a respondent (i.e., the individual's privacy policy for location access) underspecified his or her location-sharing preferences. In other words, most of the contextual factors that were rated and ranked highly by respondents for making decisions regarding location sharing were not captured in their rules. Consider, for instance, the "frequency of

location access," which, despite being ranked higher than the "reason for the location access," was only mentioned in 2 rules out of the 234.

We suspect that the difficulties of articulation could be attributed to one or more of the following reasons:

- **Difficulty of 'recall':** It is conceivable that the respondents could not think of all requisite rules in one go. This is reflected in the small number of rules specified by most respondents.
- **Inability or unwillingness to articulate:** Respondents may not have been able to articulate their preferences in the form of a rule and/or may have been unwilling to do so due to the burden imposed by the specification effort. This is also suggested by the respondents choosing not to revise or add to their initially specified rules even though we offered them the opportunity to do so. Only 5 of the 103 respondents revised their earlier rules or specified new rules.
- **Lack of incentive:** It is possible that the respondents lacked sufficient incentive to specify rules because their location information was not at risk during the study or because the compensation offered for study participation was insufficient motivation for putting in the effort.

These considerations point to several possibilities for design explorations to enhance privacy management in location-sharing systems to mitigate the impact of these issues. Users could be provided with lightweight and quick ways to add and revise rules *in situ* at the time of incoming location accesses. The rule set then grows into a comprehensive location privacy policy over time instead of requiring the user to think of every necessary rule at the outset. Moreover, it allows the policy to adapt to situations that the user may not initially have thought of.

The effectiveness of rule specification could also be elevated by an interface that presents important contextual factors for controlling access to location information along with various ways of combining these factors. Such interfaces are typically utilized by email programs for end-user specification of filters for incoming email. Using similar techniques for access-control rules could provide greater flexibility and control than is offered by the typical privacy options in current systems. This may also mitigate the burden of articulation imposed by free-form specification. Templates of important rules can also be included not just to handle commonly expressed desires (e.g., dealing with emergencies) but also to serve as useful initial examples. These rules could be chosen by conducting studies in which users rate and rank various given rules.

Caller ID or Recipient privacy: The dominant importance of the recipient of location information (see Table 1) suggests that it might be useful to provide an incoming 'location call' feature with caller ID. Revealing location in response to a call could then be automated based on pre-specified rules or handled manually by choosing to accept or deny the call. This does, however, present a privacy dilemma: identifying the recipient matches the desires and expectations of those whose location is being accessed (and is aligned with the principle of reciprocity), but hinders the ability of the

recipient to anonymously or covertly consume location information.[4] More studies of actual user practices in real-world location-sharing services could shed light on the impacts of enabling or disabling privacy for the recipient.

Temporal factors: The low relative importance attached to temporal factors is somewhat surprising, as prior research noted the importance of temporal boundaries [13], albeit in a professional context. However, the correlation of temporal factors with desires for privacy from non-professional relations suggests that temporal considerations could be of particular use to those who wish to maintain somewhat distinct personal and professional lives. Traditionally these two spheres have often been temporally separated.

Social, technical, and societal considerations: It is noteworthy that many of the rules seemed to ignore considerations of social nuance (e.g., the connotations of the recipients knowing that they were denied access) as well as technical details (e.g., the possibility of the service provider's records being exposed to hacking or leaks). The rules also expressed desires that may not be easily implementable in purely technical ways. For instance, many respondents expressed a desire to share location during emergency situations. Yet, it is not straightforward to determine what *exactly* constitutes an emergency or to detect such situations automatically. Similarly, preventing access by the government is necessarily intertwined with legal and public-policy considerations. These types of rules likely require *socio-technical* solutions.

6 Limitations and Future Work

It should be noted we sampled only the US population. Since privacy attitudes and considerations vary across cultures, generalizability of these findings to other populations requires empirical verification. Although our sample is diverse in terms of age and geographical reach across the US, it still cannot be considered a representative sample of the US, especially since we recruited participants from two specific sources. The sample is also affected by self-selection bias. Further, the sample size of 103 was too small for adequately analyzing the impact of various demographic factors. Collecting data from additional respondents is necessary to investigate these issues.

In terms of methodology, this is an attitudinal study; self-reported preferences regarding privacy do not always match actual user practice [2]. Moreover, semi-structured interviews might have provided richer details regarding access-control rules than free-form text entries. However, it should be noted that our technique was closer to the specification constraints that users encounter in real-world system implementations.

We are pursuing further research to overcome some of these limitations and to shed more light on user preferences and practices in the new landscape of location sharing. We are currently working on gathering additional data in order to strengthen these findings and conduct more statistical analyses. We also plan to apply the insights to the design of a structured access-rule editor. It would be interesting to study whether rules

[4] Note that anonymous or covert location accesses need not be malicious. For instance, an individual's plans for surprising the spouse could require knowing the spouse's location without the spouse finding out that the information was accessed.

created using such a tool capture more of the factors deemed important for managing location access. We further hope to expand our exploration to other cultures.

7 Conclusion

We reported the results of an online questionnaire ($N = 103$) that sought to investigate factors influencing people's preferences for location sharing. Location sharing has only recently started gaining mainstream adoption due to the increasing use of smartphones. Prior work on location sharing, however, has mostly been conducted during the infancy of location-sharing systems. Further, several of the previous user studies were limited in size or scope. In contrast, we reported on a study of a sample of adults in a wide age range (18–61 years) from across the US. We investigated privacy preferences expressed using system-independent, natural language rules. While we confirmed some of the previous findings, we also uncovered new and interesting results that could inform privacy management features of location-sharing systems in today's landscape. For example, we noted that the frequency of accesses is an important factor typically not taken into account by current systems. We also found temporal factors (such as the time of day) to be relatively less important in general, but preferred by those more sensitive to privacy from non-professional social relations. Many contextual factors rated and ranked high in importance consistently failed to show up in free-form access rules. This points to limitations of end-user free-form expression for articulating how access to location information ought to be controlled. The free-form statements did, however, reveal notable insights for managing access to location information. These include special treatment for emergencies, manual control over location disclosure, and turning off location tracking (temporarily or permanently).

Acknowledgements. We acknowledge Kristy Caster, Greg Norcie, and Roman Schlegel for help in the implementation and testing of the questionnaire and Tijana Gonja for comments on the analysis. We thank John McCurley for editorial comments on a draft version of this paper. We also thank the study participants for their time and effort. This research is supported by NSF grants CNS-1016603 & CNS-1017229, and US DHS grant 2006-CS-001-000001, under the auspices of the Institute for Information Infrastructure Protection (I3P). The contents of this paper do not necessarily reflect the views of the sponsors.

References

1. Benisch, M., Kelley, P.G., Sadeh, N., Cranor, L.F.: Capturing Location-Privacy Preferences: Quantifying Accuracy and User-Burden Tradeoffs. Personal Ubiquitous Comput. 15, 679–694 (2011)
2. Berendt, B., Günther, O., Spiekermann, S.: Privacy in e-Commerce: Stated Preferences vs. Actual Behavior. Communications of the ACM 48, 101–106 (2005)
3. Brodie, C.A., Karat, C.M., Karat, J.: An Empirical Study of Natural Language Parsing of Privacy Policy Rules Using the SPARCLE Policy Workbench. In: Proceedings of the Second Symposium on Usable Privacy and Security, SOUPS 2006, pp. 8–19. ACM, New York (2006)

4. Consolvo, S., Smith, I.E., Matthews, T., LaMarca, A., Tabert, J., Powledge, P.: Location Disclosure to Social Relations: Why, When, & What People Want to Share. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2005, pp. 81–90. ACM, New York (2005)
5. Kapadia, A., Henderson, T., Fielding, J.J., Kotz, D.: Virtual Walls: Protecting Digital Privacy in Pervasive Environments. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds.) Pervasive 2007. LNCS, vol. 4480, pp. 162–179. Springer, Heidelberg (2007)
6. Khan, R.M., Khan, M.A.: Academic Sojourners, Culture Shock and Intercultural Adaptation: A Trend Analysis. Studies About Languages 10, 38–46 (2007)
7. Kobsa, A., Patil, S., Meyer, B.: Privacy in instant messaging: An impression management model. Behaviour & Information Technology 31(4), 355–370 (2012)
8. Lederer, S., Mankoff, J., Dey, A.K.: Who Wants to Know What When? Privacy Preference Determinants in Ubiquitous Computing. In: CHI 2003 Extended Abstracts on Human factors in Computing Systems, CHI EA 2003, pp. 724–725. ACM, New York (2003)
9. Malhotra, N.K., Kim, S.S., Agarwal, J.: Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. Information Systems Research 15, 336–355 (2004)
10. Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S.B., Zheng, X., Campbell, A.T.: Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application. In: SenSys 2008: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, pp. 337–350. ACM, New York (2008)
11. Patil, S., Kobsa, A.: Instant Messaging and Privacy. In: Proceedings of HCI 2004, pp. 85–88 (2004), http://www.ics.uci.edu/~kobsa/papers/2004-HCI-kobsa.pdf
12. Patil, S., Kobsa, A.: Uncovering Privacy Attitudes and Practices in Instant Messaging. In: GROUP 2005: Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work, pp. 109–112. ACM, New York (2005), doi:10.1145/1099203.1099220
13. Patil, S., Lai, J.: Who Gets to Know What When: Configuring Privacy Permissions in an Awareness Application. In: CHI 2005: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 101–110. ACM, New York (2005), doi:10.1145/1054972.1054987
14. Sadeh, N., Hong, J., Cranor, L., Fette, I., Kelley, P., Prabaker, M., Rao, J.: Understanding and Capturing People's Privacy Policies in a Mobile Social Networking Application. Personal and Ubiquitous Computing 13, 401–412 (2009)
15. Schlegel, R., Kapadia, A., Lee, A.J.: Eyeing your Exposure: Quantifying and Controlling Information Sharing for Improved Privacy. In: Proceedings of the 2011 Symposium on Usable Privacy and Security (SOUPS) (July 2011)
16. Toch, E., Cranshaw, J., Drielsma, P.H., Tsai, J.Y., Kelley, P.G., Springfield, J., Cranor, L., Hong, J., Sadeh, N.: Empirical Models of Privacy in Location Sharing. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Ubicomp 2010, pp. 129–138. ACM, New York (2010)
17. Toch, E., Cranshaw, J., Hankes-Drielsma, P., Springfield, J., Kelley, P.G., Cranor, L., Hong, J., Sadeh, N.: Locaccino: A Privacy-Centric Location Sharing Application. In: Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing, Ubicomp 2010, pp. 381–382. ACM, New York (2010)
18. Tsai, J.Y., Kelley, P., Drielsma, P., Cranor, L.F., Hong, J., Sadeh, N.: Who's Viewed You?: The Impact of Feedback in a Mobile Location-Sharing Application. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, pp. 2003–2012. ACM, New York (2009)
19. Wagner, D., Lopez, M., Doria, A., Pavlyshak, I., Kostakos, V., Oakley, I., Spiliotopoulos, T.: Hide and seek: Location Sharing Practices with Social Media. In: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI 2010, pp. 55–58. ACM, New York (2010)

Spamming for Science: Active Measurement in Web 2.0 Abuse Research

Andrew G. West[1], Pedram Hayati[2], Vidyasagar Potdar[2], and Insup Lee[1]

[1] Department of Computer and Information Science
University of Pennsylvania, Philadelphia, USA
{westand,lee}@cis.upenn.edu
[2] Anti-Spam Research Lab
Curtin University, Australia
{p.hayati,v.potdar}@curtin.edu.au

Abstract. Spam and other electronic abuses have long been a focus of computer security research. However, recent work in the domain has emphasized an *economic analysis* of these operations in the hope of understanding and disrupting the profit model of attackers. Such studies do not lend themselves to passive measurement techniques. Instead, researchers have become middle-men or active participants in spam behaviors; methodologies that lie at an interesting juncture of legal, ethical, and human subject (*e.g.,* IRB) guidelines.

In this work two such experiments serve as case studies: One testing a novel link spam model on Wikipedia and another using blackhat software to target blog comments and forums. Discussion concentrates on the experimental design process, especially as influenced by human-subject policy. Case studies are used to frame related work in the area, and scrutiny reveals the computer science community requires greater consistency in evaluating research of this nature.

1 Introduction

Spam needs little introduction given estimates that 95%+ of email traffic, 75% of all blog comments, and nearly every medium of human communication has been pervaded by the practice. The growing prevalence of distributed and collaborative models of information dissemination (*i.e.,* Web 2.0 forums, wikis, blogs, *etc.*) has only expanded the battleground. Measurement studies from the end-user perspective have long been the predominant method of examining these phenomena. More recently research has begun to consider the attacker's perspective: What are the motivations? How much money is made? What are the greatest marginal costs? By answering these questions researchers can hope to better understand the spam profit model and how to undermine it.

However, an empirical view of these notions does not come cheaply. The first-person viewpoints that enable such studies raise interesting legal, ethical, and human subject questions. Although formal bodies (*e.g.,* Institutional Review Boards, "IRBs") exist to regulate these matters it appears the computer science

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 98–111, 2012.

community is unfamiliar, or questioning of, their role. As two case studies reveal this yields unfair and inconsistent academic evaluations that satisfy neither authors, reviewers, or program committees (PCs).

This work begins by describing two recent works in this domain (Sec. 2), both actively conducting spamming campaigns with scientific interests. One targeted a collaborative platform (Wikipedia) and the other blog/forum environments. These case studies were subject to institutional review and their approval process is described at length (Sec. 3). This description: (1) sketches the approval process and policies that regulate this kind of research, and (2) outlines the experimental methodologies that brought these studies into compliance.

After the experiments were completed/described, the papers proceeded into the publication process. What followed exemplifies the inability of the community to soundly evaluate research of this kind (Sec. 4). Reactions ranged from applause to outrage; some endorsing IRB approval and others rejecting it entirely. It is unclear if the community is: (1) unfamiliar with the scope/role of the current review process, or (2) informed but dissatisfied with its judgments. Regardless, authors deserve a system by which research can be approved and evaluated under the same criteria. Similarly, reviewers are entitled to one that allows them to judge submissions on technical merit and not personal beliefs. This work continues by surveying literature about the evolving role of ethics, IRBs, and researchers in technical studies – and framing other active measurement work in this context (Sec. 5). Finally, concluding remarks are made (Sec. 6).

It should be emphasized that this work is a case study exemplifying ethical disagreement/issues. While it advocates the need for improvement, it does not endorse any particular mechanism for achieving it. While this remains an open issue for the computer science community, we believe it important to realize that the IRB is the only such regulator in the status quo.

2 Case Study Research

In this section we describe our case studies, two active measurement Web 2.0 link spam experiments which serve as the basis for later discussion. Included here is information about how these studies were conducted, the statistics collected, and the conclusions yielded by analyzing that data.

At a high level the experiments are quite similar with both economic components following the "pipeline" model described by Kanich *et al.* [13]. Summarily, after the spam hyperlinks have been disseminated there are three important measurements. First is the number of *exposures*, the quantity of individuals who view the spam link. Second is the *click-through* rate, the percentage of exposures that result in a visit to the *landing site* (*i.e.,* the webpage at the spam URL, or at the conclusion of that URL's redirection chain). Finally, the ratio of site visitors that actually make a purchase is termed the *conversion* rate.

Both case studies implemented a "payment disabled" store front (as per [13]) in order to collect the latter two statistics. These landing sites operate much like any online store but attempts to "check out" result in a technical failure or other

Fig. 1. Example landing site

Fig. 2. Prominent link display in a *wiki*

complication. In this manner, one can approximate purchase quantity and value without having to fulfill transactions. Both case studies constructed/scraped landing sites that were pharmaceutical in nature (see Fig. 1).

The two experiments[1] differ most in the environments being studied. The first, conducted at the University of Pennsylvania, performed proof-of-concept attacks to show the economic viability of a *novel* spam model against the collaborative encyclopedia, Wikipedia (Sec. 2.1). The second, via Curtin University, used blackhat software to target web forums and blog comments (Sec. 2.2).

2.1 UPenn Research: Wikipedia Spam

Collaborative functionality is becoming increasingly prevalent in web applications and no paradigm embodies this more purely than the *wiki*. The open-editing permissions and massive traffic[2] of some wiki installations (*e.g.,* English Wikipedia) would seem to invite spam behaviors. To the surprise of its authors, a measurement study [25] found status quo spam behaviors to be technically naïve, comparatively infrequent, and ineffective for their perpetrators.

Using their expertise of collaborative security the researchers next sought to identify vulnerabilities of the platform. In doing so they described a novel attack model that exploits the latency of Wikipedia's human-driven enforcement [25]. In this model link placement is characterized by: (1) targeting high traffic pages, (2) prominent link styling (see Fig. 2), and (3) the use of privileged accounts.

To show the viability of the attack model an active measurement study was engineered. The experiments added links to a payment-disabled pharmacy per the proposed strategy. Only seeking to establish a proof-of-concept, just 3 Wikipedia accounts were used. These accounts posted 341 hyperlinks with each surviving for an average of 93 seconds. Public article view statistics show that ≈14,000 individuals were exposed to the link, generating 6,307 click-throughs (*i.e.,* landing site visits) that led to 8 "purchases" for $1940 USD.

The "revenue" generated considerably exceeded the marginal attack costs, suggesting a viable attack model (at least initially). Despite IRB approval, these

[1] The authors of *this* paper are a subset of those conducting the case study research.
[2] Distributed attacks that target low traffic and poorly maintained wikis for search-engine-optimization (SEO) are not uncommon. The research under discussion, however, concentrates only on direct traffic (*i.e.,* click-throughs) in high exposure wikis.

results remain unpublished for reasons discussed in Sec. 4. However, this result did motivate additional research into protections against such vulnerabilities, the suggestions of which have been actively described and implemented [24].

2.2 Curtin Research: Blog/Forum Spam

Relative to the novel proposal on Wikipedia, link spam in blog comments and forums is a pervasive issue. One source estimates that 75%+ of blog comments are spam [4]. This proliferation suggests the status quo attack model is profitable to its perpetrators, motivating research into its economic dynamics. We choose to highlight [10,11], which actively posted links to these environments.

Work began by *harvesting* sites running common software (*e.g.,* phpBB, Wordpress) with circumvent-able protections (*e.g.,* CPU-solvable CAPTCHAs, no registration, *etc.*). In this manner the common structure and weaknesses can enable autonomous link placement at minimal marginal cost. Such functionality has been encoded into blackhat software and the experiment used one such tool: XRumer [3,21]. In addition to posting links the researchers also spoofed the "referrer URL" of the HTTP headers to point to the pharmacy site[3].

The harvesting stage produced a list of ≈98,000 websites of which 66,226 were practically targeted after discounting network errors. From these, 7,772 links (11.7%) were successfully posted to public view, with the remainder being caught by spam filters, manual moderation, *etc.* The month-long experiment produced 2,059 pharmacy visits and 3 "intents to purchase." Minor modifications in link placement strategy also permitted more fine-grained observations. For example, non-English websites produced higher response rates, and referrer spam produced more landing site hits than the actual link placements.

As of this writing the recent research remains unpublished. However, it is the authors' intention for the work to appear as [10].

3 Obtaining Research Approval

Our case studies now summarized, we next describe their formal approval process. More than rote description, this discussion intends to use the approval criteria as an outline for focusing on experimental design and ethical issues. We begin by justifying the need for active measurement (Sec. 3.1). Having decided to use human subjects, we next describe the approval workflow (Sec. 3.2). Then, we handle the talking points of approval: informed consent (Sec. 3.3), maintenance of privacy (Sec. 3.4), and minimization/justification of harm (Sec. 3.5).

3.1 Infeasibility of Passive Measurement

Before engaging in active measurement it should be the case that a passive approach is not feasible. A leading study of email spam economics [13] creatively

[3] This is a technique called "referrer spam", "log spam", or "referrer bombing." Sites that make access logs public will have the spam URLs indexed by search engines.

became a "man-in-the-middle" to a botnet operation and rewrote spam URLs to a payment-disabled pharmacy under their own control. In this manner, no additional spam was sent and the spam they rewrote was less malicious than it would have been otherwise. A similar strategy is difficult to imagine in Web 2.0 environments where attacks are coordinated by software empowered individuals.

Recruitment of cooperative blog/forum owners for research purposes deserves consideration. No additional spam would need to be injected as status quo events could be examined. Visitor logs would quantify exposure and outgoing link clicks could be tracked. However, this presents issues: (1) participating owners are unlikely to form a representative set (poorly maintained sites are likely crucial for attackers), and (2) this result says nothing about conversion rates.

The Wikipedia study has additional complications. Given a single intended target (English Wikipedia), a rejected request for cooperation would raise administrative awareness and bias any subsequent (non-consenting) trials. Moreover, because the strategy is a novel one it is impossible to glean statistics without injecting links per the proposed model.

3.2 Approval Workflow

Having decided to undertake active measurement, formal approvals must be obtained from organizations overseeing: (1) human-subjects/ethics and (2) legality.

Human-Subjects/Ethics: Any experiment involving data collection from humans is required to undergo review. Internationally these groups go by different names but are quite similar in function; the U.S. has the Institutional Review Board (IRB), Australia prefers Human Research Ethics Committee (HREC), and the European Union uses Research Ethics Committees (RECs).

There is ongoing debate over whether human subjects approval is equivalent to an experiment being ethical or whether it is only a component thereof. We challenge readers to imagine any form of ethically interesting research that does not at least indirectly impact humans (or animals) in some way (physically, psychologically, economically, *etc.*). Regardless, some in the computer science community do draw this distinction leading to inconsistency in the evaluation of research (see Sec. 4). Further examination of this controversial issue is beyond the scope of this work, as we prefer to focus on experience-driven analysis.

In the blog/forum case study (Curtin University, AUS), the process began by contacting a department-level ethics coordinator. This individual determined the experiment to be "low risk" and eligible for an expedited review. Per University/AUS policy [1] low risk research is that which "does not pose a greater risk than participants would face in their normal daily routine." Supporting this criterion are the facts that: (1) advertisements and spam are already ubiquitous in blog/forum environments, and (2) statistical collection on the web is omnipresent. After one meeting the study was allowed to continue.

Matters were more complex for the Wikipedia case study (UPenn, USA). After an IRB coordinator found that the research posed "more than minimal risk to subjects" [2] a request for expedited approval was rejected in favor of a

full board review. Seemingly, the concern was that publication of the novel attack model could considerably endanger Wikipedia's operation if the vulnerabilities remained unpatched (see Sec. 3.3). After multiple iterations of clarification and research gathering the protocol was approved in ≈14 weeks time[4].

This "full board review" produced a number of observations which may be interesting to readers. First, the board proceedings are non-transparent and closed-door (except for clarifications), doing little to inform other researchers how to best shape their experiments to the satisfaction of the IRB/HREC/*etc.* Further, the latency and lack of technical expertise among members have previously been identified as weaknesses of the process [6,9].

Legal Approval: The legal approval process was less structured. The Wikipedia study did come to the attention of the University's Office of the General Counsel, who did not object to publication of the study results with IRB approval. The blog/forum research was not required to seek such approval by their coordinator. The legal framework in which this research operates is beyond the scope of this work (see [7]), though it is interesting to consider how differing jurisdictions may affect what is deemed "acceptable" research.

3.3 Regarding Participant Consent

A majority of human subjects studies operate under *informed consent*, whereby a potential subject is informed a priori of the purpose and potential risks of participation. If he/she voluntarily decides to proceed, this removes considerable responsibility from the researcher. It is possible to forego informed consent where it is: (1) technically impractical and/or (2) biasing of results. However, this places stricter requirements on the experimental methodology. Both case studies operated without the prior consent of any participant.

As discussed in Sec. 3.1, contacting site *administrators* would create recruitment bias and/or raise administrative awareness. In the case of *readers*, informed consent also produces numerous issues. Consider that experiments take place on a 3rd-party site where: (1) the consent dialogue alone might constitute spam, and (2) limited control would force that dialogue to be awkwardly adjacent to the behavior being measured. Further, those who choose to ignore the spam messages (the vast majority of exposures) incur minimal disruption. One might imagine that forcing everyone to opt-in/out of the experiments would create more annoyance than the experiments themselves.

Following these arguments, both case studies were approved to proceed without informed consent. Having chosen this course, the anonymity of the exposures/readers becomes paramount (Sec. 3.4). This does not eliminate the

[4] Experiment design was influenced by the ethical norms of the IRB process. However, it should be acknowledged that approval was received in an ex post facto fashion, due in part to an initial miscommunication with the IRB. Such ex post facto scrutiny follows the same workflow and is held to the same standard as a priori review. This occurrence speaks to the unfamiliarity and poor working relationship others have reported between computer scientists and these boards (see Secs. 4 and 5).

possibility of *debriefing* test subjects after their participation. Readers could potentially be notified as they exit the experiment pipeline (navigating off-site; attempting to purchase) but such information could influence how others interact with the experiment. For example, in a wiki setting, a reader who discovers the "spam" to be an academic experiment may not give it treatment consistent with spam links (*i.e.,* removal). Notification en masse after the entire experiment duration is not possible given the decision to preserve anonymity.

In the case studies, one instance of debriefing was present. The administrative community of Wikipedia (the Wikimedia Foundation) was contacted post-experiment. In this email notification the vulnerabilities were described and technical assistance was offered towards mitigating the exploit.

3.4 Privacy and Data Security

Given that we have collected the behavior of non-consenting users, there is a responsibility to protect that data: its release could lead to embarrassment or other harm. One way to prevent this is by severing the mapping between experiment events and real persons (*i.e.,* an anonymous experiment).

In the Wikipedia experiment under the IRB system, information capable of identifying real persons is called *personally identifiable information* (PII). The IRB required that data collected by the checkout system be immediately destroyed (it never left the client machines) with the exception of the items being purchased and their value. To uniquely identify click-through and purchasing users, a hash of the IP address was stored[5]. The IP addresses themselves were considered PII due to static IPs, the possibility of geo-location, *etc.*

The Australian notion of privacy had a very different interpretation. In the blog/forum case study server logs were maintained and authors geo-located their landing site visitors. In their setup, registration was required to checkout, with relevant fields including: (1) first name, (2) last name, and (3) email address. These fields were manually *inspected* to ensure the registration attempts were legitimate, before the data was destroyed. The Australian body seems to operate on the logic that "since normal spam sites would view registration data, it is permissible for the researchers to do so."

Just as participant data must be secured there is a need to protect the identities of the researchers and their institutions mid-experiment. Thus, we discuss the computing framework in which these studies operated. Consider that it is desirable to use non-institutional IP addresses to launch the experiments and host the landing site. This avoids experimental bias (*e.g.,* `*.edu` sites might not trip filters) and protects the institution from ill consequences (*e.g.,* blacklisting). One case study launched experiments from a large cloud provider and hosted their landing site via a 3rd party service (whose data retention was vetted). The other study purchased a dedicated Internet connection (outside the University network) for hosting and used proxy servers and VPN for outbound traffic.

[5] As one reviewer pointed out, the finite nature of IP space makes it feasible to reverse these (now destroyed) hashes – a consideration not foreseen in experiment design.

3.5 Minimizing and Justifying Harm

Harm in any experiment should be both *minimized* and *justified*. We now extend previous discussion about risk minimization to include experiment elements that were not major "talking points" of the approvals process.

Experiment Scale: Rather than assessing risk at the per-subject level, a more pragmatic approach is to consider the cumulative cost to all participants, making the *scale* of experiments a significant factor. The blog/forum case study was approved without any conditions on the size of the experiments (though ethical approval is valid for 12 months, after which re-evaluation is required). In those experiments 66,000 sites were targeted, a scope seemingly justified by the large number of sub-experiments and need for statistically significant data. Consider that while the number of targets/exposures is large, relatively few engage in the interesting behaviors (click-through, conversion) being measured.

Generally one should carefully weigh the need for statistical significance against human costs. Showing the viability of novel theories should require less iterations than measurement studies (although the Wikipedia study produced 14,000+ exposures in just 3 trials). Also consider that long running or narrowly focused experiments could target the same individual(s) multiple times.

Deceptive Advertising: Ethical review boards tend to be sensitive to *deception* of test subjects. At the same time, attackers are by their very nature deceptive agents and accurate simulations need to reflect this nature. In the case studies one potential source of deception is hyperlink presentation, *i.e.,* the *hooks* or description that is associated with links. Among several strategies it was the alluring and deceptive hooks (*e.g.,* "click to collect your prize") which proved most controversial. However, hooks of this type far outperformed more mundane approaches, speaking to the effectiveness of such social engineering tactics.

Others might question the choice of landing site genre. One case study sold a wide range of pharmaceuticals while the other focused only on the "male enhancement" subset. A previous study [14] showed it is precisely these products which dominate online spam revenue. Moreover, care was taken to make sure the sites were free of any harmful/suggestive imagery and descriptions.

Just because harm is minimized (while maintaining experiment integrity) does not mean it is *justified*. For that to be true, experiments must produce a net benefit which exceeds any harm (a *consequentialist* approach [8]). This is a particularly unsatisfying condition given that the outcomes of the research cannot be known a priori. Nonetheless, in the case studies the: (1) novelty of the work, (2) daily exposure of readers/administrators to spam behaviors, and (3) projected understanding of the spam ecosystem indicated a foreseeable benefit.

Project benefit can be more accurately assessed in an ex post facto fashion. Wikipedia active measurement showed the attack model viable, motivating the authors to create a spam detection engine for wikis [24]. A live implementation of the technique has already assisted in the removal of far more spam instances than placed during active measurement. Moreover, circumstances arising during the experiment encouraged further study into "redacted revisions" [26].

4 Community Response and Discussion

Once case study research was completed the logical next step was to submit the findings to academic journals and conferences. The inconsistent treatment of ethical issues in reviewer feedback was not expected. More importantly, it raises questions about how the community and publication process can better accommodate research of this kind.

Reviewer Response: The response to Wikipedia active measurement was extremely mixed[6] (all submissions noted the IRB approval). Some reviewers applauded the study, finding the methodology appropriate and necessary proof of an earlier hypothesis. Others took a more neutral approach, pointing to possible ethical implications but stating that the IRB approval was evidence of reasonable conduct. Others still assaulted the methodology, questioned the social conscience of the authors, and were prepared to reject the paper on ethical grounds alone. Excerpts from some of the critical responses are in Appendix A.

Several iterations of submission followed. In one attempt, several pages were dedicated to ethical justifications (pages that could have been dedicated to technical content). In the end, it was decided to omit the active measurement results from the paper due to these complications and concerns over statistical significance/stability. In the published version [25] there is only a numerical estimation of attack viability. Though only just beginning the publication process the blog/forum study is experiencing similar reactions.

Discussion: A major issue is why some reviewers are not satisfied with ethical review decisions and make it their own responsibility to regulate the matter. An IRB is best-equipped to make these decisions, being armed with experience and precedent, and having seen supporting documents to which reviewers are not privy[7]. This may be partially explainable by the unfamiliarity many computer scientists have of these organizations, as described by Garfinkel [9].

One such example can be seen in [5] where the author suggests IRBs are insufficient and bases this on a flawed argument. He cites two experiments that "do not involve human subjects...": (1) an experiment that congested residential Internet networks to learn about their characteristics and (2) a study that de-anonymized packet traces, linking them to physical addresses. We believe strongly these are human subjects issues that fall under IRB/committee jurisdiction. The first example would affect Internet QoS for users and the second has obvious privacy implications.

We acknowledge that the IRB (and its equivalents) may be a *logistically* imperfect system (see [9]). However, in the absence of an alternative, this is no reason not to respect its findings. Allowing PC chairs or reviewers to interject

[6] It is difficult to quantify the weight these ethical disagreements had in accept/reject decisions (although one reviewer did make the connection explicit, see Appendix A). We prefer to focus solely on the qualitative feedback given about active measurement.

[7] Submitted versions included a footnote indicating that reviewers/PC-members could be contacted to obtain a copy of the approval documents (*e.g.,* via the conference chair to preserve anonymity). No such requests were made.

their beliefs only lends greater subjectivity to evaluation. Consider that research similar to the case studies has been published in spite of complaints and *without* IRB approval (Sec. 5.1). Such a situation is unimaginable in many fields, where IRB approval is considered a gold standard of approval (see again, Sec. 5.1).

For those who advocate a more responsive and technically-staffed IRB-like organization it is clear it should come into force *before* research is conducted. The current situation creates awkward situations where research has been performed (along with any harm) but cannot be released to the community. Such an organization also faces practical challenges in creating an objective and level playing field. For example, how does one integrate the legal frameworks of international researchers? Will the organization supersede or operate alongside the human subject boards (adding bureaucracy)? Who writes the policies?

Such an organization could be an asset to the community but is far from being realized. Focusing on the status quo, human-subjects boards are the only organizations properly equipped to handle these matters (and arguably, already do so at the appropriate scope). This division-of-labor allows reviewers/PCs to concentrate on their area of expertise: technical merit. Although imperfect, these boards are the most satisfactory regulators of research ethics at this time. As such, respect for their decisions is the greatest hope the community has for fairly evaluating ethically interesting research.

5 Related Work

Discussion of related literature begins by looking at other spam and electronic abuse research that has employed ethically notable active measurement techniques (Sec. 5.1). Then, we look at writings about the formal review process and issues specific to computer science research (Sec. 5.2).

5.1 Similar Abuse Research

As other active measurement studies are surveyed, we encourage readers to think about experimental design and the potential risk posed. We divide our review into: (1) studies that have engaged in spam-like behaviors, and (2) studies that involve payment to spammers or spam-support services. To permit discussion our literature selection is both non-exhaustive and narrow. Readers are encouraged to see the survey of Moore and Anderson [18] for a broader look at ethically-interesting security research, particularly of the empirical and behavioral variety.

Studies Conducting Abuse: The most similar work to the wiki case study is [22] wherein the authors befriended 942 popular individuals on a social networking site. Then, they posted a "comment" including a 1×1 pixel image hotlinked from their own server (the image(s) were sometimes 50+MB in size). In their 12-day experiment their server recorded 2,598,692 hits, indicating the feasibility of conducting DDOS attacks in this fashion. Though readers' attention

was not particularly affected (their bandwidth was), administrator workload was non-trivial. Communication with the authors revealed the work did *not* have IRB approval and some reviewers raised complaints, yet the paper was still published.

Another work looked at the topic of "social phishing" [20]. There, the researchers mined social network data in order to write personalized phishing emails, sent to 600 university students. Relative to a control these customized emails produced higher "success" rates (with 72% of students providing their university credentials). This research had IRB/institutional approval as the paper discusses at length, along with test-subject reactions to the work.

An interesting cross-domain perspective comes from [17] where the authors posed as prospective students, emailing 6,600 university professors requesting a meeting. Student names were chosen to imply gender/race, with the research measuring the varying response rates. While not commercial spam, one could imagine the cost per participant was quite high (reading the email, responding, meeting scheduling/cancellation). This study had the IRB approval of multiple institutions but drew considerable criticism in Internet communities. In an interesting contrast to the case studies much condemnation was directed not at the researchers, but the IRBs involved. Similarly, communication with an author indicated no reviewer had raised ethical complaints.

Finally, [13] deserves mention for inspiring much research on active measurement of electronic abuse. Therein the authors became a coordinating node in a spam botnet. From this position they instruct worker nodes to send the same spam emails they would have otherwise, but change the hyperlink URL to one under their control (a payment-disabled pharmacy). However, because the study "strictly reduced harm" (no new spam; made sent spam less malicious) it lies on less tenuous ethical footing than the other work described herein.

Studies Aiding Abusers: Another frame of reference into spam economics can be achieved by becoming a consumer of spam services. Just like one of the case studies, [21] purchased blackhat spamming software (at $400+ USD) in order to analyze its operation. Another work [19] spent hundreds of dollars to solve 100,000+ CAPTCHAs to study the dynamics of that underground economy. Finally, in [14,15] researchers made 156 purchases from spam-advertised business to make inferences about sale volume and examine financial routing. It is especially hard to quantify the harm that may be indirectly suffered as a result of financially assisting these individuals/services. However, as evidenced by the above papers, this seems to be a tactic generally well-received by the community.

5.2 IRB/Ethical Discussions

Numerous works have looked at legal, ethical, and human subjects issues in computer science research. None is more relevant than [8], which lends a broader perspective to the experiences shared herein. That work identified the weaknesses/limitations of the status quo to be: (1) an absence of shared community values, (2) lack of familiarity with ethics and review systems, and (3) lack of

consensus on enforcement. They too looked at ways the community could move forward, suggesting self-governance, public discussion, and protocols to reward ethical behavior. Outside the scope of our writing, [8] also considers the roles of professional societies (*e.g.,* ACM, IEEE) and funding organizations.

Other writings have more narrow scope. Focusing on legal issues in particular is [7], emphasizing the collection/sharing of network traces. The IRB has been a point of emphasis, beginning with a look at how the Internet has changed its role [23]. Other works [6,9] criticize the IRBs latency and lack of technical expertise, with the latter claiming that much CS research runs afoul of regulation. Moving beyond the IRB, [5] examines the program committees role in ethical evaluation. Then, there are "best practices" papers like [16], focusing on "vulnerability research". Finally, Kanich [12] writes similarly based on his extensive experience with economic and cyber-crime research.

6 Conclusions

In this work, two case studies guided a discussion of the legal, ethical, and human-subject considerations of active measurement research in spam and electronic abuse. Much discussion was dedicated to how experimental design was shaped by the review process, bringing the controversial methodologies into policy compliance (of the IRB or its international equivalents). We intended this to give some introduction into the role/operation of these review committees and inspire readers to think about increasingly benign ways to gather data.

Paper rejections, negative reviews, and harsh personal criticism are not something many authors are eager to speak about. However, in relaying our own experiences we hope to give exposure to an issue on which the computer science community can improve: evaluating ethically interesting research. Critics condemn the IRBs latency, handling of technical matters, and scope. If this is indeed a widely held view the review stage is a poor place to enforce it, and new bodies need to be assembled to proactively regulate these matters. If no such consensus exists (or until such a body is in place) then the community should respect the current standard. Either way, the status quo is detrimental to authors, reviewers, PCs, and the entire community – and the exposure and elimination of this practical dilemma was our motivating interest in authoring this work.

Acknowledgments. The authors would like to thank their colleagues who contributed to the original case study research: Jian Chang, Krishna Venkatasubramanian, and Oleg Sokolsky from UPenn [25]; Nazanin Firoozeh and Kevin Chai at Curtin [10]. The authors also acknowledge the helpful advice and service of their respective human-subject/ethical/legal coordinators and boards. In particular, Robert R. Terrell of UPenn's Office of the General Counsel is thanked for his guidance. Any opinions expressed in this work do not necessarily reflect the sentiments of those acknowledged here.

References

1. Curtin: Research management, http://research.curtin.edu.au/guides/
2. UPenn: Office of regulatory affairs, http://www.upenn.edu/regulatoryaffairs/
3. XRumer (Blackhat SEO software), http://www.xrumerseo.com/
4. Abu-Nimeh, S., Chen, T.: Proliferation and detection of blog spam. IEEE Security and Privacy 8(5), 42–47 (2010)
5. Allman, M.: What ought a program committee to do? In: USENIX Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems (2008)
6. Buchanan, E.A., Ess, C.M.: Internet research ethics and institutional review boards: Current practices and issues. SIGCAS Computers and Society 39(3) (2009)
7. Burstein, A.J.: Conducting cybersecurity research legally and ethically. In: LEET: Proc. of the Wkshp. on Large-Scale Exploits and Emergent Threats (2008)
8. Dittrich, D., Bailey, M., Dietrich, S.: Building an active computer security ethics community. IEEE Security and Privacy 9(4) (July/August 2011)
9. Garfinkel, S.L., Cranor, L.F.: Institutional review boards and your research. Communications of the ACM 53(6), 38–40 (2010)
10. Hayati, P., Firoozeh, N., Potdar, V., Chai, K.: How much money do spammers make from your website? (Working paper, in submission)
11. Head, B.: Storage bills top $43,000 say spam-busters. ITWire.com (August 2011), http://www.itwire.com/business-it-news/security/49239-storage-bills-top-43000-say-spam-busters
12. Kanich, C., Chachra, N., McCoy, D., Grier, C., Wang, D., Motoyama, M., Levchenko, K., Savage, S., Voelker, G.M.: No plan survives contact: Experience with cybercrime measurement. In: CSET 2011: Proceedings of the 3rd Workshop on Cyber Security Experimentation and Test (August 2011)
13. Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G.M., Paxson, V., Savage, S.: Spamalytics: An empirical market analysis of spam marketing conversion. In: CCS 2008: Proc. of the Conf. on Computer and Comm. Security (2008)
14. Kanich, C., Weaver, N., McCoy, D., Halvorson, T., Kreibich, C., Levchenko, K., Paxson, V., Voelker, G.M., Savage, S.: Show me the money: Characterizing spam-advertised revenue. In: Proc. of the USENIX Security Symposium (August 2011)
15. Levchenko, K., Chachra, N., Enright, B., Felegyhazi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., McCoy, D., Pitsillidis, A., Weaver, N., Paxson, V., Voelker, G.M., Savage, S.: Click trajectories: End-to-end analysis of the spam value chain. In: Proc. of the IEEE Symposium on Security and Privacy (2011)
16. Matwyshyn, A.M., Cui, A., Keromytis, A.D., Stolfo, S.J.: Ethics in security vulnerability research. IEEE Security and Privacy 8, 67–72 (2010)
17. Milkman, K.L., Akinola, M., Chugh, D.: The temporal discrimination effect: An audit study of university professors (Working paper)
18. Moore, T., Anderson, R.: Economics and Internet security: A survey of recent analytical, empirical and behavioral research. Tech. Rep. TR-03-11, Harvard University, Department of Computer Science (2011)
19. Motoyama, M., Levchenko, K., Kanich, C., McCoy, D., Voekler, G.M., Savage, S.: Re: CAPTCHAs - Understanding CAPTCHA-solving services in an economic context. In: USENIX Security Symposium (August 2010)
20. Nathaniel, T.J., Johnson, N., Jakobsson, M.: Social phishing. Communications of the ACM 50(10) (October 2007)
21. Shin, Y., Gupta, M., Myers, S.: The nuts and bolts of a forum spam automator. In: LEET: Proc. of the Wkshp. on Large-Scale Exploits and Emergent Threats (2011)

22. Ur, B.E., Ganapathy, V.: Evaluating attack amplification in online social networks. In: W2SP 2009: The Workshop on Web 2.0 Security and Privacy (2009)
23. Walther, J.B.: Research ethics in Internet-enabled research: Human subjects issues and methodological myopia. Ethics and Info. Technology 4(3), 205–216 (2002)
24. West, A.G., Agrawal, A., Baker, P., Exline, B., Lee, I.: Autonomous link spam detection in purely collaborative environments. In: WikiSym 2011: Proc. of the Seventh International Symposium on Wikis and Open Collaboration (October 2011)
25. West, A.G., Chang, J., Venkatasubramanian, K., Sokolsky, O., Lee, I.: Link spamming Wikipedia for profit. In: CEAS 2011: Proc. of the Eighth Annual Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conference (September 2011)
26. West, A.G., Lee, I.: What Wikipedia deletes: Characterizing dangerous collaborative content. In: WikiSym 2011: Proc. of the Seventh International Symposium on Wikis and Open Collaboration (October 2011)

Appendix A: Reviewer Comments

Below is a sample of reviews received in response to the Wikipedia line of link spam research. Effort has been made to preserve the context of the feedback. Each bullet point represents the comments of a single reviewer.

- "The second measurement study is a bit offensive, but the IRB approval seems to cover this ... While the IRB problem is discussed, I am still not convinced that such experiments with Wikipedia are good from an ethical point of view."
- "I personally am concerned about the ethics of the active link-spamming research ... In particular, a natural guideline is that research should not cause harm or damage to subjects without their informed consent. In this study, it appears that harm or damage may have been done to Wikipedia by this research ... and was done without prior consent of the Wikipedia foundation ... not persuaded that the 'consequentialist' viewpoint is a suitable response to this concern."
- "Although they did get their institution's IRB to approve it, IRB approval is a necessary, but not sufficient, step for justifying such an experiment ... the experiment imposed a substantial cost on the Wikipedia community, both the editors who had to fix the page, and the thousands of users who encountered their spam. Such a cost, which is involuntary to the participants, needs to be justified by a significant gain in scientific understanding."
- "... their active experiment is ethically deficient ... I view each [of multiple issues, the 'ethical deficiency' included] as a deal-breaker ... The ethical standing is dubious enough that it does *not* suffice to simply tell us that you had IRB approval. We need to know the wording of what the IRB approved. In addition, while the text briefly mentions (un)informed consent, there is no mention of *post facto debriefing* ... [this] makes the reviewer wonder to what degree the authors really did obtain IRB approval that was itself informed."
- "... the paper is rather offensive, it seems like Wikipedia actually received negative press related to this experiment ... I find this a bit questionable, the discussion in the appendix is also not very convincing. Actually I had not thought that the authors would receive IRC approval for this kind of study. I suggest to revise the appendix and maybe even publish all IRC documents ... Apart from this aspect, the study is interesting and the authors demonstrate convincingly that Wikipedia is an attractive target for link spam."

A Refined Ethical Impact Assessment Tool and a Case Study of Its Application

Michael Bailey[1], Erin Kenneally[2], and David Dittrich[3]

[1] Computer Science and Engineering, University of Michigan
[2] Cooperative Association for Internet Data Analysis, Univ. of California, San Diego
[3] Applied Physics Laboratory, University of Washington

Abstract. Research of or involving Information and Communications Technology (ICT) presents a wide variety of ethical challenges and the relative immaturity of ethical decision making in the ICT research community has prompted calls for additional research and guidance. The Menlo report, a revisiting of the seminal Belmont report, seeks to bring clarity to this arena by articulating a basic set of ethical principles for ICT research. However the gap between such principles and actionable guidance for the ethical conduct of ICT research is large. In previous work we sought to bridge this gap through the construction of an ethical impact assessment (EIA) tool that provided a set of guiding questions to help researchers understand how to apply the Menlo principles. While a useful tool, experiences in the intervening years have caused us to rethink and expand the EIA. In this paper we: (i) discuss the various challenges encountered in applying the original EIA, (ii) present a new EIA framework that represents our evolved understanding, and (iii) retrospectively apply this EIA to an ethically challenging, original study in ICTR.

1 Introduction

Information communication technology research (ICTR) presents a wide variety of ethical challenges, touching on diverse research topics including botnets, spam, malware, phishing, etc. Examples of interesting ethical questions raised by such studies include: If someone has the ability to take control of a botnet, can they just clean up all the infected hosts? What risks do researchers face when they provide data to the community? How do theoretical exploits and concepts differ from existing vulnerabilities? What impact does the immediacy of an event (e.g., DDoS) have on our response to the event? [4] Unfortunately, the relative immaturity of ethical decision making and a lack of community standards has prompted calls for additional research and guidance [5].

1.1 ICTR

Before delving deeply into the above challenges, it is first instructive to briefly discuss ICTR, its goals and potential risks. Information cannot be separated

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 112–123, 2012.

from the systems in which it is stored, processed, or through which it is transmitted. The umbrella term *Information and Communication Technology (ICT)* encompasses these systems, and implicitly the information (or data) that they store, transmit, and process. Research involving ICT often involves risks centered around the core properties of these systems information – confidentiality, integrity, and availability.

Harm that results from impacts on these properties can manifest in physical, psychological, legal, social, and economic damage. These non-informational risks are typically viewed in light of historical behavioral and biomedical research that involve physical procedures that can cause physical pain, bodily harm, or psychological traumas. Informational risks derive from inappropriate use or disclosure of information, which could be harmful to the study subjects or groups. Both categories of harm must be dealt with in ethical evaluation of research involving ICT, spread across all potentially affected stakeholder populations.

When research focuses primarily on ICT itself, indirect harm (either informational or non-informational) to humans can still occur. As ICT evolves and is more tightly integrated into our lives through process controls and cyber-physical systems such as automobile braking controls, smart energy meters, and embedded medical devices, the use and disclosure risks to ICT will increasingly put humans at risk. This necessitates shift from considering research in terms of human subjects involvement to that of human-harming potential [1].

1.2 The Menlo Report

The Menlo report [6], a revisiting of the seminal Belmont report [8], seeks to bring clarity to this arena by articulating a basic set of ethical principles for ICTR. The effort is the result of an interdisciplinary working group sponsored by DHS which commenced in mid-2009. The goal of this effort was to create an updated Belmont report for the field of ICTR. The report appeared for comment in the Federal Register at the end of 2011.

1.3 The EIA v1.0 and Its Limitations

While the Menlo report describes fundamental principles, the gap between such principles and actionable guidance for the conduct of ICTR is large. In previous work we sought to bridge this gap through the construction of an ethical impact assessment tool (EIA) [10] we will refer to as EIA v1.0. The EIA v1.0 provided a set of guiding questions to help researchers understand how to apply the Menlo principles. While a useful tool, experiences in the intervening years have caused us to rethink and expand the EIA. Specifically, we believe the EIA v1.0 was successful in achieving its goal of *education*, highlighting the specific classes of ethical problems that need to be addressed. However, in spending the intervening years applying the EIA v1.0 ourselves to both our own work and numerous case studies of others work in the field, we feel two alternative goals now warrant attention. Specifically, those of *Consistency* and *Lowering Barriers to Use*.

An EIA that has a *Low Barrier to Use* will make it easier for researchers to use reasoning by analogy, to trend classes of ethical issues, to assure fairness, etc. It must be easy to use and map, in an understandable way, to existing processes and methodologies. In achieving *Consistency* in ethical analysis, researchers will be better suited to develop ethically defensible research protocols from the start, and others will have an easier time evaluating these protocols because of the clarity and consistency with which researchers describe which humans may be at risk, to what extent, and what protective measures researchers have implemented.

The EIA v2.0 we present here embodies the lessons we have learned to date and uses a least common denominator set of stakeholders that we believe makes it suitable for the majority of ICT research of minimal or low-to-medium risk. We wish to be clear that there are some research situations presenting higher risk, such as vulnerability research involving threat to life or real property, or large-scale computer crime situations, where even the EIA v2.0 may not be sufficiently fine grained or comprehensive to address all stakeholders listed in in Table 1, or all case studies documented in relation to the Menlo Report [7].

2 Ethical Impact Assessment (EIA)

In this section we present the EIA v2.0 framework, with special attention to places where it has been expanded or modified from v1.0 as our understanding has evolved.

2.1 Research Lifecycle

One common experience analyzing case studies using the EIA v1.0 framework was that we consistently repeated classes of risk in our analysis. In many cases these similarities were more an artifact of the phase of research, rather than the research methodology itself.

While we find that we are mostly concerned with experimental computer science, theoretical computer science can also pose risks to humans. In experimental computer science, "[t]he key ideas [are] an apparatus to be measured, a hypothesis to be tested, and systematic analysis of the data (to see whether it supports the hypothesis)." [2]. In such studies, we have robust models for thinking about the lifecycle of data (i.e., collection, use, dissemination) [12]. Explicitly examining the data lifecycle, it is evident that the ethical concerns differ by phase and that concerns repeat across studies in various classes.

In the EIA v2.0 framework, three activities are commonly called out: the *collection* of information (i.e, research data), the *use* of information or information systems in research (whether as vehicle for conducting research or as research subject), and the *disclosure* of research data or vulnerability information that could be used to cause harm. In this paper, we use these terms in a broad sense and emphasize that risks from information collection, use, and disclosure are transitive across stakeholder populations. Risk is present even when the only data involved are facts and observations about the functioning of a

cyber-physical device, and in cases when there is no information involved at all yet harm could arise from unintended consequences resulting from the manipulation of information systems that humans are dependent upon. This latter area is the hardest to evaluate with the EIA v2.0 framework as the focus on the data lifecycle does not cleanly accommodate all potential stakeholder populations, nor those risks that are not data related. We believe that continued evolution of the EIA into a richer and finer-grained framework will further enhance its consistency of evaluation and further lower the barriers to use.

2.2 Stakeholders Analysis

One of the major changes in the EIA framework since v1.0 [10] is the integration of a set of stakeholders as columns in the EIA spreadsheet.

Stakeholder Analysis identifies the key players in the situation in terms of their interests, involvement, and their relationship (i.e., producer or recipient) of outcomes such as benefit or harm. In previous case studies [3] we have have adapted the definitions of stakeholders [11] used in other domains for ethical analysis. We also have found that some ICT research, such as studies of botnets and other ongoing computer crime activity, or vulnerability research where publication of research results could be used by malicious actors to cause grievous damage, require consideration of both *Positively Inclined* and *Negatively Inclined* stakeholders in order to fully understand the risk vs. benefit calculus over time [1]. These stakeholders are listed in Table 1.

The problems we seek to address through a comprehensive stakeholder analysis are *indirect harm* and *consideration of intermediaries*. Indirect harm may result from secondary effects, such as disrupting a service provider, which in turn affects the customers of that service provider and the customers of those customers (i.e, in a wholesale vs. retail sales relationship). Or it can be harm that occurs long after publication of vulnerability information as attackers make use of the information for criminal gain before system owners learn of patches and apply them to render services immune to attack. The complexity resulting from the involvement of ICT makes it hard to see what the impacts of ones actions may be. Enumerating the stakeholders helps elucidate the potential harms and benefits. We also find that there are a common set of re-occurring stakeholders, which is reflected in the EIA, however we acknowledge that the full range of Positively and Negatively Inclined stakeholders as depicted in Table 1 must be dealt with effectively in future iterations of the EIA framework.

2.3 Ethical Principles and Their Application

The EIA v1.0 framework was invented at a time when the Menlo report was still in its infancy and well before we had external feedback from reviewers of the document. In the interim, the Menlo Report has matured [6] and the EIA v2.0 framework has been modified to align with the current set of principles and their applications. These include:

Table 1. A complete breakdown of stakeholders for a Botnet research scenario. While both Positively and Negatively Inclined stakeholders are shown here, most ICT research involves neither criminal activity nor vulnerability disclosure and would thus not involve the Negatively Inclined Stakeholders.

Stakeholder Type	**Positively Inclined**	**Negatively Inclined**
Key [*Affect on producing outcome*]	Researchers Programmers Operations Staff Executives Law Enforcement	Criminals (Individuals/Gangs) Malware Programmers Botmasters Criminal Masterminds
Primary [*End users*]	Consumers (product/service) Enterprises (.edu, .com, .org) Manufacturers Government entities	Espionage Consumers Criminal Enterprises
Secondary [*Intermediaries in delivery*]	Service Providers Platform Providers Transit Providers Retailers	"Bullet Proof" Hosting Providers Malware Delivery Providers Malware Obfuscators Sellers of fake goods

- **Identification of Stakeholders.** As research targeting or involving ICT can hide potentially harmed humans, a thorough analysis of stakeholders is a necessary pre-requisite to a comprehensive analysis of risks, benefits, identification of burdens, and mitigation of actualized harms.
- **Informed Consent.** Researchers should obtain informed consent to collect, use or disclose data, or to interact with systems in ways that could have a negative impact on those systems.
- **Harms.** Researchers should consider the full spectrum of harms to both persons and information systems (systems assurance, privacy, reputation, physical, psychological, economic)
- **Benefits.** Researchers should identify benefits to all stakeholder populations, including (but not limited to) benefits to the broader society.
- **Balancing Risks and Benefits.** Research should be designed and conducted not simply to maximize benefits and minimize harms, but to appropriately balance risk and benefits across all stakeholder populations.
- **Mitigation controls.** Researchers should notify appropriate parties if research causes harm and have plans in place to efficiently and effectively resolve problems.
- **Fairness and Equity.** The benefits and burdens of research should be apportioned fairly across all stakeholder populations.
- **Compliance.** researchers should perform due diligence in regards to respecting laws, contracts, etc. in order to protect individuals and organizations.
- **Transparency and Accountability.** Researchers should act in ways that garner trust with the general public by communicating intent, research methodology, risk-benefit analysis, and ethical reasoning.

2.4 Bringing it Together: The EIA

The EIA v2.0 framework (see Figure 1) assists researchers in formulating policies, processes, and methodologies that align with ethical principles throughout three research lifecycle phases. It illuminates all relevant ICT stakeholders, as well as both the benefits and human-harming risk potential of research in order to achieve ethically-defensible methodologies and results. A downloadable version is available at `http://www.eecs.umich.edu/~mibailey/EIA.xlsx`

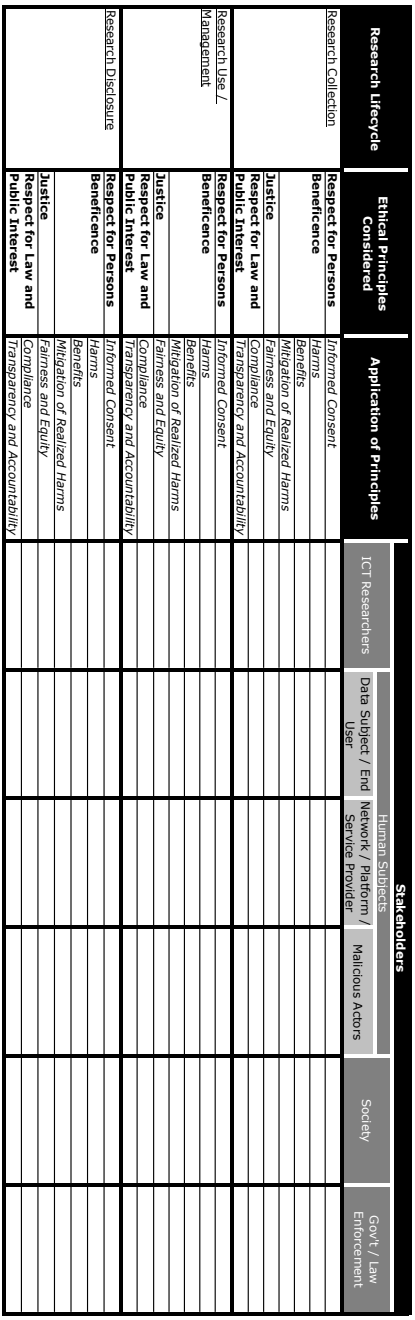


Fig. 1. The EIA worksheet

3 Case Study

We illustrate the evaluative use of the EIA v2.0 framework by retrospectively applying it to a case study that provoked ethical debate within the research community. The Menlo Report and the EIA did not exist at the time, so use of the principles and assessment framework during the fundamental research design, implementation and publication was not possible. The researchers in this case study were advised by one of this paper's authors, who was also substantially involved in the then-parallel Menlo effort. These deliberations influenced the EIA v1.0 framework and the subsequent evolution of both the Menlo Report and EIA framework. The post hoc analysis performed here exposes opportunities where researchers could have made more ethically-defensible decisions.

3.1 Background

Researchers at University of California San Diego (UCSD) undertook an

experiment to measure the conversion rate of unsolicited commercial e-mail as part of an empirical study to understand the quantitative value proposition of spam [9]. Lacking sufficient methods to indirectly measure spam conversion, the methodological challenges stemmed largely from the ethical implications of mimicking real spam campaigns. Specifically, key components of such operations involved building fake e-commerce sites, marketing them via spam, presenting sales transactions for the advertised goods, and distributing the various communications (e-mail marketing, processing recipient responses) via illicit botnets.

To address the obvious ethical and legal problems posed by spamming and botnet activities, researchers sought insight from *non-malfeasance* theory[1] and legal and ethical advisement. This guidance informed the research methodology which involved parasitically infiltrating the command and control infrastructure of an existing spamming botnet by accepting invitations to become proxy bots, or conduits between master servers and worker bots. Researchers then modified a subset of the spam the botnet was already distributing, so respondent users were directed to servers under researcher control, not those of the real spammer. Then researcher servers presented web sites that mimicked those actually hosted by the spammer, however they "de-fanged" them by removing functionality designed to compromise the user's system or that would collect and disclose sensitive user information (e.g., name, address, credit card data).

3.2 Stakeholder Identification

ICT Researchers. In addition to the obvious inclusion of the UCSD research team, it became clear that other researchers were also analyzing the same botnet. The "in vivo" nature of botnet studies warrants consideration of these other stakeholders who may be simultaneously undertaking various empirical studies.

Data Subject or End User. Stakeholders here were the users of computers infected with the Storm bot (a.k.a., worker machines), and recipients of spam email sent through the botnet. This research impacted the collective rights and interests of not only the owners and users of computers that were infected with the Storm bot, but those being tricked by it.

Network, Platform, or Service Provider. Parties to be considered here were network services providers for the botnet proxy hosts and command and control servers, Internet service providers (ISPs) of users with infected computers, webmail platform providers, registrars of mimicked illicit phishing sites, and the network community (the Overnet peer-to-peer platform) used by the botnet to communicate.

Society. Beyond those directly affected by botnet infection, this research impacted the collective rights and interests of all users of computers that are affected by social engineering attacks involving spam and online fraud activities.

[1] Researchers should act in good faith and control risks, exposing end users to no more harm than they would face but for the research activities.

Government or Law Enforcement. As the primary source of funding for the research, the National Science Foundation provided authoritative influence and is thus a stakeholder. Similar to the rationale for considering other bot researchers, the research had the potential to impact law enforcement agencies (LEAs) in multiple countries who were investigating and attempting to enforce various laws against the parties responsible for the botnet's illegal activities.

3.3 Research Collection

Consent – Informed consent was obtained from the network provider for the proxy collector machines, the webmail platform providers, and the domain registrar for the researcher's mimicked phishing sites. Each had an interest in safeguarding the ICT resources it owned, controlled or managed, including the data associated with those resources. The researchers believed they could justify a waiver of informed consent from owners of worker machines and end user subjects of the research. Identifying and providing notice to the owners of thousands of compromised home computers was impracticable, given the scale and scope of the botnet. Informing both worker hosts and end user stakeholders about the research procedure, purpose, risk-benefit analysis, and withdrawal opportunities would negatively impact the scientific integrity of the research by altering the behavior that was attempted to be studied. A determination on waiver of informed consent due to impact on research integrity is often the responsibility of an IRB, not a researcher decision. A Menlo evaluation using the EIA framework raises questions about whether researchers should have debriefed end users who were deceived via the phished sites (fake pharmaceutical and e-card) via some form of pop-up alert.

Compliance – Legal due diligence analysis was performed to address a number of factors. Research activities respected federal and state laws concerning computer fraud (e.g., no unauthorized access to systems or networks, researcher proxy bots were invited to participate in botnet, researchers were authorized to log traffic to their own fake phish website, no exceeding access to webmail platform since Terms of Service were not violated, research action did not cause legally cognizable damage or harm), electronic communications privacy (e.g., no interception of traffic; proxy bots were a party to the communications, although there was possible violation if acquisition of bot communications would be deemed to require two-party consent), intellectual property (e.g., mimicked phished sites did not replicate the images that infringed copyright on the real phished sites, no circumvention of mediating devices), or contract laws (e.g., there were no agreements associated with nodes in the Overnet platform; researcher actions adhered to normal and expected functioning of Overnet protocols; use of webmail did not violate Terms of Service prohibiting sending of spam since those accounts were receiving users' responses to redirects). While researchers did engage ex ante ethical and legal risk analysis, federal regulation

required that they should have consulted with their IRB prior to, rather than after, the completed research.

Harms – Researcher actions (i.e., botnet command rewriting, interposing Spam delivery, interposing user click-through) did not diminish the performance, availability or integrity of the networks or machines in the bot infrastructure. There were no new machines compromised or worker bots created, nor did researchers cause corrective action to be undertaken by systems administrators. Privacy harms were avoided by not collecting, storing or transmitting any private personal information from either worker systems with whom the researcher proxy hosts communicated or from the mimicked sites. There was no reason for the researchers to believe that the study was interfering with LE investigation activities involving the botnet. Researchers minimized potential reputational harm to Webmail providers from spam-advertised product association by obtaining informed consent. With the fake e-card phished sites, researchers presented a benign executable that performed a simple *HTTP POST* to the researcher controlled backend server, and then exited. This could be interpreted as direct intervention with the environment of subjects who have not consented, however the potential for harm here was strictly minimized and there was no malicious intent.

Benefits Considered – This research aimed to enhance understanding of internet criminal activity and thus produces benefits to the broader society by improving user's abilities to safely use ICT in their daily lives.

Mitigation – Researchers mitigated any harm to integrity or functionality of user's systems from the botnet-directed spam by redirecting them to de-fanged fake phished site, only logging the user-agent string to determine if the exploit would have likely worked. The users were always asked to download the file, but where not actually provided with an executable (e.g., presented a 404 error).

3.4 Research Use or Management

Consent – The webmail and network provider's consent to collect information for specific research activities extended to the ongoing use of those platforms for the limited duration of the experiment.

Compliance – Researchers designed their methodology to avoid running afoul of consumer protection laws (e.g., prohibiting the sending of commercial e-mail). Researchers acquiesced to being infected by the botnet and subsequently interposed as proxy bots within the existing bot infrastructure. This positioned researchers as a conduit, passively transmitting and observing the spam-related commands and data between the master servers which initiated and controlled the transmission of spam and the worker bots which carried out the directives. Actions that altered command messages (spam template, dictionary entries) to include researcher-controlled sites arguably did not alter the spam liability evaluation since the primary purpose of the deception employed by researchers was not related to advertising or promoting a commercial product or service,

but rather, to study users' susceptibility to engage these campaigns. Measurements associated with fake phishing sites respected intellectual property rights of legitimate brand owners by not replicating known trademarked or copyrighted material from the legitimate sites. In the event that the cloned phished sites (e-card and pharmacy sites) did include protected intellectual property unbeknownst to researchers, they were well-positioned to exercise a "fair use" defense. As with collection, researchers should have obtained IRB approval prior to engaging in research.

Harms – Researcher's actions did not expose end users to more harm than they would face but for the research activities, and steps were taken to reduce harm from the Storm bot. The probability and magnitude of any harm or discomfort anticipated in the research was not greater than that ordinarily encountered by users in normal use of the Internet. The only sensitive data retained was internet protocol addresses of worker bots as needed for research measurement, and they were discarded immediately after statistics were collated. The measurement infrastructure did not create new qualitative or quantitative harm to other protected computer systems – absent researcher involvement, the same users would have received the same spam e-mails from the same worker bots. Researcher proxies were passive actors that did not initiate the transmission spam e-mail, compromise hosts, or contact worker bots asynchronously. The modification of messages strictly reduced harm to users who followed the embedded links. Additional burden was not placed on hosting network resources. Research proxy nodes did not transmit or distribute any illicit information or program, send e-mail, mount or participate in denial of service attacks, crawl for or scrape e-mail addresses, compromise or otherwise introduce user accounts, or interfere with the ability of users systems to protect themselves or use the network. Researcher nodes acted in accordance with expected P2P infrastructure functions, including respecting communications protocols that maintained topological consistency with the rest of the infrastructure, and receiving and forwarding commands.

Foreseeable harms related to legitimate intellectual property rights holders were addressed in several ways. Researchers did not duplicate the phished sites that were copies of legitimate websites stolen by scraping (i.e., cloning or copying the text, logos, artwork or design templates). Rather, they replicated the general look and feel. Legitimate domain names were not spoofed, forged, or otherwise hijacked. To avert trademark likelihood of confusion harms, researchers did not obtain economic or commercial benefit, nor were not unjustly enriched by mocking the legitimate website design.

Benefits Considered – Research management and use of the measurement infrastructure provided empirical knowledge of end user susceptibility to spam marketing campaigns, botnet structure and function, and un-quantified behavior underlying the spam value proposition. Collateral individual user benefits included thwarting visits to malware-infected phishing sites and further communications with botnet command channel.

Mitigation controls – Researchers were sensitive to possible interruption of network services from retaliatory denial of service against the network hosting the proxy bots and were prepared to discontinue their utilization if that harm manifested.

3.5 Research Disclosure

Harms – Researchers did not disclose any sensitive individual or organizational information, including the internet addresses of infected worker machines or confidential network data. This was done to prevent foreseeable harms to privacy, reputation, and systems assurance associated with botnet victimization and vulnerability. Any relatively small burden borne by recipients of spam was balanced against the larger benefit to society from performing beneficial research. Researchers could have been more mindful of risks to themselves as a stakeholder class, specifically pertaining to probable reputation harms from not adequately disclosing their efforts related to ethics considerations in the design and execution of their research.

Benefits Considered – In addition to previously mentioned benefits, disclosure of research results could enhance understanding of the structure and function of digital criminal enterprises in the interests of law enforcement investigations, take-downs, and prosecutions.

Mitigation controls – While researchers did not have actual and specific knowledge of LE or other research involvement in the botnet study, there was no overt effort made to avoid collision.

Fairness & Equity – The selection and targeting of end user subjects and owners of worker machines was outside researcher control. Similarly, selection of network and application providers was likely a function of the Overnet network.

Transparency – Although the ethical controls were implicit in research design, researchers did not explicitly disclose details about the plethora of ethical considerations that informed their research. While researchers did offer a high level description of ethical undertakings, the EIA suggests that transparency and accountability could have been strengthened by more granular, a priori disclosure of the methodology and results in various publicly-available conference publications and presentations. However, unless conference committees make accommodations in paper length limitations, researchers will be deincentivized from elucidating ethical considerations in their published work.

4 Conclusion

We have described the second iteration of an ethical impact assessment framework that operationalizes the application of principles described in the Menlo Report. We are continuing to evolve this framework and other tools for the ethically-justifiable design and assessment of research involving ICT. It reflects

the iterations and refined collaborative thoughts that occurred between the chosen case study and this paper. We are continuing to improve this tool so that it most effectively assists in ethical design and assessment of research involving ICT that carries a probable risk for human harming activities.

References

1. Carpenter, K., Dittrich, D.: Bridging the Distance: Removing the Technology Buffer and Seeking Consistent Ethical Analysis in Computer Security Research. In: 1st International Digital Ethics Symposium. Loyola University Chicago Center for Digital Ethics and Policy (2011)
2. Denning, P.J.: ACM President's Letter: What is experimental computer science? Commun. ACM 23, 543–544 (1980)
3. Dittrich, D., Bailey, M., Dietrich, S.: Have we Crossed the Line? The Growing Ethical Debate in Modern Computer Security Research. In: (Poster at) Proceedings of the 16th ACM Conference on Computer and Communication Security (CCS 2009), Chicago, Illinois, USA (November 2009)
4. Dittrich, D., Bailey, M., Dietrich, S.: Towards Community Standards for Ethical Behavior in Computer Security Research. Technical Report 2009-01, Stevens Institute of Technology, Hoboken, NJ, USA (April 2009)
5. Dittrich, D., Bailey, M., Dietrich, S.: Building an Active Computer Security Ethics Community. IEEE Security and Privacy 9(4), 32–40 (2011)
6. Dittrich, D., Kenneally, E. (eds.): The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, `http://www.cyber.st.dhs.gov/wp-content/uploads/2011/12/MenloPrinciplesCORE-20110915-r560.pdf`
7. Dittrich, D., Kenneally, E. (eds.): Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Department of Homeland Security Menlo Report (January 2012)
8. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research National. The Belmont Report - Ethical Principles and Guidelines for the protection of human subjects of research, 1978. U.S. Government Printing Office. DHEW Publication No. (OS) 78-0008. Reprinted in Federal Register 44, 23192 (April 18, 1979)
9. Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G.M., Paxson, V., Savage, S.: Spamalytics: an empirical analysis of spam marketing conversion. In: CCS 2008: Proceedings of the 15th ACM Conference on Computer and Communications Security, pp. 3–14 (2008)
10. Kenneally, E., Bailey, M., Maughan, D.: A Tool for Understanding and Applying Ethical Principles in Network and Security Research. In: Workshop on Ethics in Computer Security Research (WECSR 2010), Tenerife, Canary Islands, Spain (January 2010)
11. Mascarenhas-Keyes, S.: Ethical Dilemmas in Professional Practice in Anthropology (July 2008), `http://www.theasa.org/networks/apply/ethics/analysis/stakeholder.html`
12. Vardigan, M., Heus, P., Thomas, W.: Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation 3, 107–113 (2008)

It's Not Stealing If You Need It: A Panel on the Ethics of Performing Research Using Public Data of Illicit Origin

Serge Egelman[1], Joseph Bonneau[2], Sonia Chiasson[3], David Dittrich[4], and Stuart Schechter[5]

[1] University of California, Berkeley
[2] University of Cambridge
[3] Carleton University
[4] University of Washington
[5] Microsoft Research

1 Introduction

In a world where sensitive data can be published to a worldwide audience with the press of a button, researchers are increasingly making use of datasets that were publicized under questionable circumstances. In many cases, such research would otherwise not be possible. For instance, Weir et al. examined over thirty million user-generated passwords in order to observe the effects of entropy on password cracking [10]. All of the passwords in their dataset were obtained from various private databases that were breached by others and then subsequently posted to the Internet, the vast majority of which came from the RockYou breach [9]. Komanduri et al. used this same dataset to examine the effects of password creation policies on entropy [5]. Research on how users generate passwords is important, as passwords are the most common authentication mechanism. The resulting publications help system designers create password policies that balance both security and usability. Such data is only available as the result of an independent party's illegal actions. At the same time, the question exists of whether benefiting from this data makes a researcher a party to the underlying release, and whether the resulting research is ethical. This is a difficult question, especially when similar data could not otherwise be gathered: passwords generated solely for study lack ecological validity and "real" passwords are usually unobtainable due to obvious security concerns. Thus, if the researchers are not personally involved with the illegal acquisition of goods, does their use create an ethical dilemma?

Similarly, some researchers have gone beyond simply using data that others have published. In the course of gathering data, many have likely violated various terms of service—civil contracts. Amitay published an analysis of iPhone unlock PINs that were collected by his app, which mimics the iPhone unlock screen [1,3]. He published a summary of this data with the goal of demonstrating that users choose predictable PINs (e.g., 1234) and hoped that this may prompt them to choose more secure ones. This resulted in Apple removing the app from

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 124–132, 2012.

the app store, alleging a breach of their agreement. Bonneau et al. published a study on the use of so-called "secret questions" used for backup authentication with the goal of making these questions harder to compromise [2]. Part of this research involved compiling lists of common names by crawling Facebook. Others have performed similar research involving crawling various social networking sites [4,7,6]. All of these studies likely violated the sites' terms of service. This raises another ethical question about where the line should be drawn: are there fewer ethical issues involved with gathering data by violating terms of service (i.e., civil law) vs. violating criminal laws?

The use of data of questionable provenance in research is not just limited to passwords. Graphics researchers routinely use a test image featuring a female model known as "Lenna" [11]. The origin of this image was from a November 1972 issue of *Playboy* magazine. Despite being copyrighted, this image routinely appears in journal and conference publications. While Playboy, the copyright holder, has not taken action against any researchers to date, the ethics and legality of this practice—despite being widespread—are still questionable. When an ethical violation has become pervasive, does that lessen its magnitude? Is it no longer unethical if it becomes a social norm?

These examples illustrate how the desire to disseminate knowledge for the greater public good may involve actions that are ethically debatable. Indeed, we are organizing such a debate. Our panel will focus on discussion surrounding the ethics of using stolen data for research purposes. The panel will be moderated and will feature panelists representing the following viewpoints:

- Someone who has used stolen data to conduct research.
- Someone who does human subjects research outside the US.
- Someone who sits on an Institutional Review Board (IRB).
- Someone who is morally opposed to using stolen data in research.

2 Participants and Positions

2.1 Joseph Bonneau

I advocate that we can adapt ethics of "white-hat hacking" to the use of illicit data in research. The research community generally accepts papers which identify vulnerabilities in real software or websites, subject to a few basic principles. I propose that we work to adapt these into a set of ethics for using illicit data. First, we should develop a "do no harm" principle which can be realized by only using illicit data to advance scientific knowledge and not aid any parties in acting maliciously. In many cases there are technical ways to transform illicit data to prevent illicit use while still enabling research, such as stripping usernames out of a leaked password file. Second, we can require responsible disclosure, which is easy to adopt and often superfluous if companies already know that they have lost data. Third, external review of proposed studies, for example by an appropriate institutional ethics board, can help researchers in designing ethical studies. It is important to develop these principles as studies involving leaked data become

more prominent, but I believe the scientific potential of illicit data sets is too large to ignore their use.

Joseph is a PhD student at the University of Cambridge. His forthcoming thesis will focus on the statistics of human chosen secret distributions such as passwords, PINs, and passphrases. This research has included many real-world datasets, both leaked and obtained with permission. Joseph's prior research has included side-channel cryptography, obfuscation, reverse engineering, and white-box cryptography. Prior to his PhD, Joseph worked at Cryptography Research, Inc. He holds MS and BS degrees from Stanford University.

2.2 Sonia Chiasson

I think that as a research community, we need to come up with clear guidelines and minimum ethical standards for what we will accept for publication in international venues. These standards should be upheld regardless of whether the researchers' IRBs (in some cases these are non-existent) or local/national laws are more permissive.

I am not entirely opposed to using publicly available stolen datasets, but the case must be made for no conceivable harm to the victims. Cases where the "greater good" is served at the expense of a relatively small number of victims should not be entertained.

The issue of consent is important here — if we were to conduct a study to collect this same data rather than using a stolen set, would we need informed consent from participants? Should we require researchers to put in a reasonable effort at obtaining consent after the fact (they probably have usernames/email addresses available), if they want to use stolen data? It may be a daunting task, but perhaps this is the most ethical way to deal with the issue.

Sonia Chiasson is an assistant professor in the School of Computer Science at Carleton University in Ottawa, Canada, where she holds the Canada Research Chair in Human Oriented Computer Security. Her main research interests focus on the intersection between human-computer interaction and computer security. Current projects are on user authentication, usable security for mobile devices, and computer games for teaching about computer security. She leads the NSERC ISSNet project on Human Behaviour and Computer Security. Before moving to Ottawa, she was an instructor in the Department of Computer Science at the University of Saskatchewan and a member of the HCI Lab. She has been conducting empirical studies requiring approval from ethics review boards for over a decade.

2.3 David Dittrich

The Common Rule has many definitions and proscribes what research is or is not exempt from IRB review. It is unclear how any given IRB would determine which question is more important: that research is exempt from review because the stolen data is "public" (45 CFR 46.101(b)(4)), or that there is personally identifiable information in the stolen dataset that was obtained illegally under

circumstances where those persons identified reasonably believed their data was not being recorded and would remain private (45 CFR 46.102(f)(2)). I believe it is more important for researchers to always be able to clearly and coherently explain their intent in performing research using stolen data, who the researcher is trying to serve, what measures the researcher is taking to balance benefit to society vs. risk to those identified in the data, and how those individuals identifiable in stolen data will feel about the fact that their stolen data was made public, how it was studied and what about it was published.

David has over 15 years of experience in computer security operations, computer forensics, network forensics, distributed intruder attack tools (also known as "botnets"), and the legal and ethical frameworks for responding to computer attacks. He has co-authored several papers, articles, and book chapters dealing with legal and ethical issues in computer security research and operations. David has served on the University of Washington's IRB Committee K for the past two years, where he provided data security expertise to his Committee and occasionally to PIs.

2.4 Stuart Schechter

Just as one cannot assume that an act that has not been deemed illegal is socially acceptable, one cannot assume that research that is not forbidden by the common rule, and allowed by IRBs, would be considered ethical by greater society. Alas, the ethical debate over the acceptable use of stolen data often ends with a declaration that once the data becomes public, the rules of the game make its use acceptable. Consider, for example, if attackers who had compromised and released email passwords had also harvested emails and posted them publicly. Researchers might be tempted to use the data to determine if certain traits revealed in the emails (e.g., erectile dysfunction) were correlated with other, possibly more embarrassing, traits (e.g., affinity to the music of Barry Manilow). Even if individuals who had written the emails being studied were not identified by the researchers and came to no personal harm, these unwitting research participants might consider it unethical that their personal information be used by researchers without their consent. Such a study could not be ethically justified purely on the willingness of an IRB to approve it. Similarly, it is not sufficient to assume that lists of compromised passwords are fair game so long as criminals have already made the lists sufficiently public. They must imagine all reasons why the owners of this passwords might object to the use of these passwords and argue why they feel justified in going forward despite these objections. Researchers should not treat compliance with rules as a substitute for sufficient ethical consideration, as doing so may lead to these rules causing more harm to participants than protection.

Stuart is a man of few accomplishments and so, the reluctant reader should be pleased to learn, his biography is correspondingly short. Stuart researches computer security, human behavior, and occasionally missteps in such distant topics as computer architecture. Those who have worked with Stuart rave about his "tireless dedication to shooting down any idea that he cannot take credit

for." Institutions that may or may not be re-evaluating their admissions or hiring policies in response to past associations with Stuart include The Ohio State University College of Engineering (B.S.), Harvard's School of Engineering and Applied Sciences (Ph.D.), MIT Lincoln Laboratory (his former employer), Microsoft Research (his current employer), and KAIST (to use a Facebookism, "It's complicated").

2.5 Serge Egelman

Serge Egelman, normally type cast as an instigator, will be in the role of moderator. Expect a lively panel.

Serge is a postdoctoral researcher at the University of California, Berkeley. His research focuses on usable security, with the specific aim of better understanding how people make decisions surrounding their privacy and security, and then creating improved interfaces that better align stated preferences with outcomes. This has included human subjects research on social networking privacy, access controls, authentication mechanisms, web browser security warnings, and privacy-enhancing technologies. He received his PhD from Carnegie Mellon University and prior to that was an undergraduate at the University of Virginia. He has also performed research at NIST, Brown University, Microsoft Research, and Xerox PARC.

3 Post-panel Summaries

3.1 David Dittrich

This panel looked at the question of whether or not it is ethical to use stolen data, made available on public web sites without the consent of the owners of that data or anyone potentially exposed within the data, in research. Just because it is hard to get access to data, does not mean it is okay to use any data a researcher can get their hands on. Nor does it mean a researcher can take short-cuts that may increase risk to individuals who are identifiable in data used in research (regardless of whether or not those identified are the direct subjects of research).

Implicit in the question of the ethics of using publicly available stolen data is a determination of whether such data fits the criteria of "research using publicly available data sets," as well as whether such a determination by itself is sufficient for research to need Institutional Review Board (IRB) review (even expedited review of minimal risk research). Just because data is found on a web page does not make it "public." Researchers have been heard to utter statements like, "I am using public data, which does not require IRB approval, so there is no need for me to even talk to my IRB." Such statements imply the researcher knows best and that no outside review of their actions are necessary. The argument that researchers are capable of deciding for themselves what is or is not subject to external review is belied by stories of failed self-regulation of research in books

like, *The Immortal Life of Henrietta Lacks* [8]. This book is widely read and discussed in the IRB community for its telling of the personal story of a family that suffered multiple medical research abuses in the mid 1900s. Researchers cannot always be trusted to act appropriately in the face of potentially harmful research and self-interests, which is part of the reason why IRBs exist today.

Private data that was obtained through illicit means (e.g., data stolen in an intrusion incident) and put on a public web site is still private data. U.S. Federal Regulation 45 CFR 46.102(f)(2) defines "identifiable private information" as including:

> "Information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (for example, a medical record). Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects."

Therefore, some data made publicly available, such as the Statfor subscriber database stolen by LulzSec/Anonymous in December, 2011, would fit the definition of "identifiable private information" and would likely require IRB review of use in research, regardless of whether that data is available on a free and open public web site like Pastebin. Data sets, such as the RockYou password file, may also fit this definition.

To a large extent, IRBs at each institution in the United States function independently and have some leeway to interpret/apply the elements of the "Common Rule" as they see fit. Each federally funded research institution in the United States operates under something known as their Federal Wide Assurance (FWA). The FWA is the institution's commitment to the Department of Health and Human Services (HHS) that it will comply with HHS rules for human subjects protection under 45 CFR 46. Some institutions may choose to require IRB review for all research at the institution, regardless of the funding source, while others may only require that federally funded research go before an IRB. The IRB committee is there to evaluate the risk to subjects from the research subjects' perspectives, in a way acting as their representative.

Those who own the data, and those who are identified within the data, may have an expectation of privacy in that data. When stolen data is made public, and a private individual decides to archive that data, they are likely operating outside the purview of an IRB and may be taking no consideration of the risks to identifiable individuals that an IRB would. The researcher wanting to use that data may, however, be operating within the IRB's purview and must conform with institutional requirements for IRB review of proposed research. A situation in which researchers bypass IRB review by asserting the "public data" exclusion

may create an environment where individuals purposefully steal data in order to make it available to researchers, which violates both the spirit and letter of the law regarding human subjects protection via IRBs.

The identifiability of individuals within the data may be of greater importance in evaluating whether an IRB committee must approve research using public data than simply answering the question, "is the data available to anyone on the internet?" There may be instances when publicly available data may be partially de-identified, but can be combined by a researcher with other data sources, re-identifying individuals within the data. The act of re-identifying individuals and exposing them publicly can be harmful to those individuals. For this reason, many bio-repositories that make de-identified data available to researchers without necessitating IRB review, in order to safeguard the identifiability of subjects, will require the researcher to sign an agreement that includes a clause that prevents the researcher from taking steps to re-identify the individuals whose bio-samples are being studied. While it may show cleverness on the part of a researcher to identify an individual from de-identified or anonymized data, a researcher could be sanctioned by their IRB for doing so.

The University of Washington's Human Subjects Division publishes guidance/policy on use of public data sets.[1] UW's policy defines public data sets as being, "data files prepared by investigators or data suppliers with the intent of making them available for public use" and discusses usage restrictions, access agreements for restricted datasets, and data protection mechanisms that must be applied to ensure no unauthorized disclosure of individuals who are identifiable within data sets. They also define what "publicly available" and "de-identified" mean. A list of over two dozen data sets that have been evaluated by the UW IRB office are on a list of approved data sets that require no IRB review. Researchers who want to use other data sets that are not on the pre-approved list can nominate the data set for evaluation. If a funding agency does not require IRB review for publicly available data, researchers can provide documentation to that effect and the IRB will make a determination about whether any IRB review is required. For all other data, the IRB evaluates the proposed use of the data.

It is not a researcher's right to decide whether their research is exempt from IRB review, or whether data they wish to use does or does not conform with the definition of "public data" under the Common Rule. The researcher is obligated to confirm their interpretation with the IRB, who is the arbiter of how the Common Rule is interpreted as specified in their FWA. The researcher may risk sanction if they bypass or ignore the IRB's determination, which can vary by institution and by IRB. OHRP is relatively silent on the parameters of non-compliance. If an IRB determines a researcher acted unethically, or failed to submit research or data use to review when it should have been evaluated, the IRB may have the authority to do any/all of the following: (1) Halt current research and/or any further research; (2) Ask for publication of results to be halted, withdrawn, or modified to note researcher non-compliance; (3) Cite the

[1] http://www.washington.edu/research/hsd/docs/1125

researcher for serious non-compliance; (4) Require that all future research by that researcher be reviewed.

In other words, when it comes to performing research using stolen data, the catch phrase should be "researcher beware."

3.2 Stuart Schecter

Stuart argued that exemption four in the Common Rule, which states that all research using publicly available sources need not be reviewed by IRBs, gives researchers the freedom to perform studies that a great majority of the public might consider objectionable and unethical. He cautioned that researchers should not "turn off their ethics caps" and assume a study will be considered ethical simply because it qualifies for exemption from IRB reviews. To support this position, he provided three examples of research that qualifies for exemption four, but for which the social costs may outweigh the benefits.

In the first example, he explained example passwords from a compromised password data set may be traceable back to the accountholder even if no other information about the accountholder is present. It may be a password that contains data about the accountholder or even a password that appears random, but that contains a string that others may associate with that accountholder. For example, part of the password may be a password that the user shares with a significant other.

In the second example, Stuart described the implications if researchers were to come across a publicly-available repository of thousands of stolen medical records. He described how researchers might use these records to create a machine learning algorithm that could predict the likelihood that a patient suffered from a degenerative mental illness that would cause increasingly erratic behavior. The consequences of such research is those patients who this algorithm indicates are likely to be suffering from this illness—but were not yet diagnosed—would have their potential condition revealed to anyone who cared to run the algorithm on the data set. Stuart provided an example of a hypothetical individual, diagnosed with this degenerative mental illness, having to live the remainder of his life with every friend and colleague concerned that his every behavior might be the result of a mental condition predicted by this algorithm.

In the third hypothetical example, Stuart described how researchers might abuse a publicly-available repository of stolen health records from minority groups (e.g., racial minorities or LGBT). He described how researchers at religiously-affiliated anti-homosexual universities might use the data to argue that homosexual youth were more likely to engage in a socially undesirable behavior (e.g., smoking) or how other researchers might use data on racial minorities to associate them with genetically undesirable traits.

Stuart argued that in many of these cases, the general public would find such research objectionable and question any system of ethical regulation that exempted it from review.

In these cases, Stuart proposed that the standard of ethical behavior should rely on whether researchers could reasonably anticipate that the great majority

of those whose data had been stolen would consent to the research taking place, and that the social benefits outweigh the social costs. Stuart's position is thus that public data should only be exempt from ethics reviews if the data were made public with the consent of its subjects.

References

1. Amitay, D.: Most common iphone passcodes (June 13, 2011), http://amitay.us/blog/files/most_common_iphone_passcodes.php
2. Bonneau, J., Just, M., Matthews, G.: What's in a Name? Evaluating Statistical Attacks Against Personal Knowledge Questions. In: Sion, R. (ed.) FC 2010. LNCS, vol. 6052, pp. 98–113. Springer, Heidelberg (2010)
3. Bonneau, J., Preibusch, S., Anderson, R.: A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In: Keromytis, A.D. (ed.) FC 2012. LNCS, vol. 7397, pp. 25–40. Springer, Heidelberg (2012), http://www.cl.cam.ac.uk/~jcb82/doc/BPA12-FC-banking_pin_security.pdf
4. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: WPES 2005: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, pp. 71–80. ACM, New York (2005)
5. Komanduri, S., Shay, R., Kelley, P.G., Mazurek, M.L., Bauer, L., Christin, N., Cranor, L.F., Egelman, S.: Of Passwords and People: Measuring the Effect of Password-Composition Policies. In: CHI 2011: Proceeding of the 29th SIGCHI Conference on Human Factors in Computing Systems. ACM Press, New York (2011) (to appear)
6. Korolova, A., Motwani, R., Nabar, S.U., Xu, Y.: Link privacy in social networks. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 289–298. ACM, New York (2008)
7. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC 2007, pp. 29–42. ACM, New York (2007)
8. Skloot, R.: The Immortal Life of Henrietta Lacks. Broadway (2010)
9. Vance, A.: If your password is 123456, just make it hackme. New York Times, http://www.nytimes.com/2010/01/21/technology/21password.html (January 2010, retrieved September 2010)
10. Weir, M., Aggarwal, S., Collins, M., Stern, H.: Testing metrics for password creation policies by attacking large sets of revealed passwords. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS 2010, pp. 162–175. ACM, New York (2010), http://doi.acm.org/10.1145/1866307.1866327
11. Wikipedia: Lenna, http://en.wikipedia.org/wiki/Lenna (accessed: November 2, 2011)

Ethics Committees and IRBs: Boon, or Bane, or More Research Needed?

Ross Anderson

University of Cambridge Computer Laboratory
JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
ross.anderson@cl.cam.ac.uk

Abstract. A summary of remarks of the keynote talk.

Institutional review boards in the USA, and ethics committees in the UK, have their roots in medical research. In the US Tuskegee scandal, black patients with syphilis were left untreated even after an effective treatment became available in the form of penicillin; in the UK Alder Hey scandal, pathologists retained body parts from deceased children without informing their parents. Yet simply having a committee of doctors review other doctors' research proposals isn't foolproof, as it disregards the differing perspectives and cultural assumptions between doctors and patients. For example, ethics committees were already well established in Britain by the time of Alder Hey, and it's not entirely obvious that a committee of half a dozen randomly-chosen white doctors in the deep south in the 1940s would have acted any differently from the Tuskegee team.

The current tussle in the UK is between a medical research establishment that wants access without consent to medical records that have been "pseudo anonymised" in that the patients' names and addresses have been removed, and a privacy community which points out that most such records can be re-identified easily. Computer scientists know that anonymity is hard, thanks to the work of Denning, Sweeney, Dwork and others; this knowledge is slowly percolating through to the policy community via Ohm's work. Yet we have already had an incident were over eight million "pseudo anonymised" records were lost when a researcher's laptop was stolen; should such a haul end up on wikileaks or pastebin, we might have a scandal like Alder Hey that could damage public confidence in medical research. Could such a dilemma be fixed by ethics committee?

Here is a second example. One UK university has data on the movements of millions of vehicles taken from automatic number-plate recognition cameras. This has been "pseudo-anonymised" by hashing the license plate numbers, yet someone who knew that a target drove on road X at time t could search for all other sightings of that vehicle. Yet the Department for Transport asserts this is no longer personal data. It follows that anyone should be able to obtain a copy using the Freedom of Information Act — and by that I mean anyone, not just any researcher working within the framework of an ethics committee. The

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 133–135, 2012.

comfort that the committee's existence gave to civil servants may have placed the data in a position from which it could escape control altogether.

It may be said that ethics committees give comfort to researchers who work in the many legal grey areas. An example raised by David Erdös of Oxford is that data protection law can easily be interpreted as prohibiting social science research on living individuals where their consent cannot be obtained, a topic case being when you send off job applications to hundreds of professors in order to assess whether there's any racial or gender bias in their hiring practices for postdocs. In fact, a cautious interpretation of the law would prevent even a book review — criticism of the writing of a living author is personal information about him, made available without his consent and with the potential to do real harm. This highlights the wildly different interpretations put on the law by different institutions. At Oxford, ethics committees are starting to give social scientists a hard time over research which the scientists claim is obviously justified; a Cambridge ethics committee chair said that "an academic who asked for ethics clearance to write a book review would be told to go away and stop being annoying".

The diversity causes real friction. My team planned to do some work with another university on how best to tell people that their PC has been recruited to a botnet, so as to persuade them to clean up the machine without causing undue alarm or distress. This is an important problem, as some 5% of PCs worldwide are infected at any one time. But our research project has been stalled. Ethics approval at our end is done at a departmental level and is straightforward; at the other end it goes to a university-wide committee that has "levelled up" to the much more heavyweight procedures expected by researchers in psychology and medicine.

Yet ethics committees don't do much heavy lifting when we face real problems. Colleagues and I do research into payment systems; fraud victims come to us after being fobbed off by their banks or credit card issues, and we often figure out a new modus operandi. In order to test it, we often have to do experiments on live systems. How do we ensure that we don't get arrested for conspiracy to defraud? The answer is: by taking money only from our own accounts; by reading the law carefully and discussing it with specialist lawyers; by telling the police's e-crime unit what were doing; and by having a policy of responsible disclosure. Even so, we've had a bankers' trade association trying to bully us into removing a student's thesis from the web when it documented a vulnerability that was already being actively exploited and which the banks preferred to cover up rather than fix. Our protection in that case came from the support of university colleagues and others who backed us when we told the bankers where to get off.

So is an institutional review board, or an ethics committee, any use at all? It may well be. It can shield an experimenter by documenting intent and thus removing the *mens rea* element from a possible offence. If there is a real issue of law and policy then the experimenter really has to square up to it; but such issues aren't always visible in advance. The boundaries of the law are fuzzy and context-dependent; and context can change overnight. After 9/11, jokes about

terrorism were not so funny for everyone, and attitudes to matters like race and sexuality also change, though at a slower pace.

So how can we maximise the benefit from ethical review, while minimising the harm? It appears that almost all of the benefit from ethical review comes from its very existence, while the harm escalates once it starts to be elaborated into an intricate bureaucratic system. And this may do harm in more ways than one, for example by moral hazard.

In order to push back on the bureaucracy, we should perhaps investigate whether researchers subject to heavyweight ethical review are more reckless than those whose institutions run ethics with a light touch.

Ethical and Secure Data Sharing across Borders

José M. Fernandez[1], Andrew S. Patrick[2], and Lenore D. Zuck[3]

[1] École Polytechnique de Montréal
jose.fernandez@polymtl.ca
[2] Office of the Privacy Commission of Canada and Carleton University
andrew.patrick@priv.gc.ca
[3] University of Illinois at Chicago
lenore@cs.uic.edu

Abstract. This is a report on a panel that was held on March 2^{nd}, 2012, as part of the Third Workshop on Ethics in Computer Security Research (WECSR 2012). The purpose of the panel was to discuss issues pertaining to ethical and secure data sharing across borders. In particular, (1) Are there ethically-driven data-sharing differences between laws of different nations, (2) Are there ethically different norms between nations? (3) How can one satisfy all norms/codes/acts among nations? (4) Can above be enforceable? automatically so? (5) Are there ever circumstances that justify "breaking the glass"? and (6) Assuming data sanitization is involved, how can we (technically) guarantee such?

1 Introduction

ACM's ethical guidelines (as well as IEEE's) are almost two decades old. The most relevant points to data sharing it makes are "Avoid harm to others" (1.2, with the elaboration: "Well-intended actions, including those that accomplish assigned duties, may lead to harm unexpectedly. In such an event the responsible person or persons are obligated to undo or mitigate the negative consequences as much as possible."), and "Respect the privacy of others" (1.7, with the elaboration: "It is the responsibility of professionals to maintain the privacy and integrity of data describing individuals. This includes taking precautions to ensure the accuracy of data, as well as protecting it from unauthorized access or accidental disclosure to inappropriate individuals. Furthermore, procedures must be established to allow individuals to review their records and correct inaccuracies."). The consequences of not complying with the code are "Treat violations of this code as inconsistent with membership in the ACM" (4.1), but the code itself admits that "Adherence of professionals to a code of ethics is largely a voluntary matter."

Ethical and privacy concerns become more prevalent with the rapid progress of data mining, the constant discovery of flaws in data anonymization/sanitization techniques, and the vast amount of electronic data that exists. It is often beyond the ability of a layperson to understand the privacy policy of organizations (e.g., iTunes' new privacy policy for iPhone spans over 17 pages in tiny print) and one cannot obtain many services, including (legally) playing video games, without "volunteering" PII. But then, one may argue these organizations are not bound by above mentioned code of ethics. Perhaps, they should be. Even the computer science research community is at fault,

J. Blythe, S. Dietrich, and L.J. Camp (Eds.): FC 2012 Workshops, LNCS 7398, pp. 136–140, 2012.

when ethics and PII considerations are sometimes cast aside in the chase of being the first and fastest to publish results (ironically enough) about security and privacy flaws that by themselves reveal PII.

The situation is even more dire when we consider data sharing and dissemination among different countries, that naturally have different ethical codes and policies for dealing with privacy issues concerning data sharing (none that we know of seems particularly current.) Data transfer has no borders, hence, neither does data sharing, which renders ethical data sharing all the more challenging.

One may wonder whether ethical guidelines suffice for data sharing. Perhaps policies and regulations are called for. Yet, policies and regulations are difficult to mandate and they would soon be obsolete. Even if we had ideal, always current, data sharing policies, how can they be enforced, or even checked? How can one ensure that they achieve their goals, both in terms of covering all possible scenarios and in not being contradictory?

2 Towards Providing a Framework for Research Data Sharing

When considering the exchange of sensitive data between researchers in computer security, there is presently a partial void in standards and legislation, not only across international borders but also within them.

Most academic institutions that have a research review process (Ethics Review Board in Canada, Internal Review Board in the USA) hold the person who creates the data set responsible for making sure its confidentiality is not violated and that the data is used for the research purposes it was initially intended for. This often takes the form of a signed contractual engagement, where the researcher formally promises to put in place the countermeasures necessary to ensure appropriate use and confidentiality. In addition, the researcher must guarantee that the human subjects that may have participated in the creation of the data set have given their free and informed consent for the use of that data for that particular intended research purpose. But what if the researcher desires to share the data with another researcher? How can the standards imposed upon the originating research be compared with those of the receiving party? How can he guarantee that the data will be used for that same purpose? Furthermore, there are many sensitive data sets in computer security research that do *not* involve human subjects, and might therefore not fall within the purview or the mandate of such a research review process. Examples of such sensitive data include malware collections, information about unpatched system vulnerabilities, system configuration of systems under study, or even information about criminal activity discovered in the context of research, the release of which could have negative consequences for third parties or even the public at large. When sharing this kind of data, how can the originating researcher make sure that the data will be adequately protected by the intended receiver and that it will not be released to unauthorized parties? Or in other words, how can he compare his own (self-imposed) guidelines, if any, with those of the intended receiver?

While there might be an ultimate benefit for the originator to share data with other colleagues (joint projects and publications, scientific validation of own work, exchange of data sets, notoriety, etc.), sharing the data is tantamount to sharing risk. Indeed, if data is misused or made public in an unauthorized fashion, the originator could be held

accountable and, with the current void of legislation and regulations, it would be hard for the researcher to shift that responsibility to the guilty party in a recognizable way.

Therefore researchers that have created or obtained research data sets must do their own due diligence in evaluating risks from data sharing. This, of course, includes verifying the intended use of the data. What is the kind of the research that will be done by the receiving party? What are its potential benefits? While the spirit of academic freedom should be respected, the level of scrutiny on the intended research by the originator of the data should be proportional to the potential damage to the public and other third parties if the data is misused or improperly disclosed.

This is also the case for the examination of the receiving researchers' security policy and security counter-measures. Do the lab facilities of the receiving researcher allow for the proper containment and protection of the data? Do the physical, logical and personnel security policies of the receiving lab/organization adequately reduce the risk that an internal or external party access and potentially release the data in an unauthorized fashion? This verification can be somewhat difficult, mostly because security laboratories should indeed protect the confidentiality of their own security measures, and hence sharing details about them with other researchers for the purposes of gaining access to their data does potentially increase risk. In some sense the receiving researcher should ask himself whether the originator can be trusted not only with information about security policy but also about time-sensitive ideas about research projects that the originator could easily use to his own benefit.

This need for mutual verification of policies and mutual trust may have its own benefits: it forces researchers to exchange information about procedures and tools for securing the data, allowing them to determine best ideas and practices (e.g. protective technology, procedures, etc.) across different environments. Hopefully, this sharing not only of the data but of the knowhow will help the computer security research community converge towards adequate standards that all should adopt. In time, this should become enforceable or certifiable standards, drafted, adopted and recognized by both national and international research funding organizations. Researchers willing to share data could then check new collaborators against these standards, thus reducing the administrative burden and potentially the legal one as well. By providing a framework within which sharing of risk associated with sharing of data would become less problematic and more common, such recognized standards would go a long way in encouraging something that current computer security research desperately needs: sharing and use of common data sets to support scientific repeatability.

3 A Matter of Principle: Ethical Data Sharing across Borders

Disclaimer: All material, views, and opinions in this section are strictly those of its author, Andrew S. Patrick.

Data sharing across borders is a common occurrence, and probably a necessary part of modern commerce, government operations, and law enforcement. Some of Andrew Patrick's work is an attempt to ensure that international data sharing is done in the best possible way. An organization engaging in international data sharing needs to consider four stages: principles, design, execution, and review.

Principles

It can be impossible to satisfy the laws and regulations in all international jurisdictions. The right set of principles, however, can provide importance guidance on how sharing operations should be designed and operated. For example, the Organization for Economic Co-operation and Development (OECD) Guidelines represent an almost universal list of privacy principles: collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability. Organizations adopting and implementing these principles will be well on the way to establishing good privacy protection.

As an example of the importance of establishing data sharing principles, consider an international organization whose purpose is to share personal, sensitive information on an international scale. Can such an organization still act ethically? INTERPOL is the world's largest international police organization with 190 member countries. One of Patrick's many roles is to serve on the Commission for the Control of INTERPOL's Files (CCF), an independent body charged with overseeing ethics and privacy protection.

INTERPOL's 190 member countries having many different legal and social traditions, so it would be impossible for INTERPOL to comply with each and every law and regulation. Instead, strong principles are key for determining what is proper conduct. The key principles that INTERPOL has adopted include:

1. the widest possible mutual assistance within the limits of the spirit of the Universal Declaration of Human Rights;
2. being limited to ordinary law crimes, excluding police actions that are of a political, military, religious or racial character;
3. adhering to the OECD principles for privacy protection, and
4. a strong commitment to independent oversight.

INTERPOL is not perfect, but the adoption of key principles, and associated procedures to ensure they are followed, are necessary steps in the quest to be the most ethical organization possible.

Design for Sharing

An important factor to consider when sharing data is to distinguish between *disclosures* versus *transfer for processing*. Transfer for processing is sharing data with a third party for the purpose for which the data was collected. With transfer, all of the responsibilities and safeguards for privacy protection must be maintained, and in some jurisdictions users have to be told, at the time of collection, about the third parties that will receive the transfers. Assuming the information is being used for the purpose it was originally collected, additional consent for the transfer may not be required.

Disclosure, on the other hand, is sharing data with a third party where the purpose is beyond that established at the time of collection, and/or the responsibility for privacy and data protection is no longer maintained by the first party. Disclosure of personal information without consent is illegal in many jurisdictions.

Organizations cannot control the legal environment in foreign countries, so organizations designing data sharing programs must pay particular attention to the legal requirements of the jurisdictions of their partners, as well the potential political, economic and social conditions that may increase sharing risks. Such an analysis may not necessitate a measure-by-measure comparison of laws, but it does require organizations to conduct an assessment of all the relevant elements that might be important.

Execution

A key issue when considering international data sharing is determining whom to the data share with. When considering privacy protection, Europe has taken a state-by-state approach to determine when sharing is allowed. Organizations are only supposed to share with partners from states that have been determined to be *adequate* in terms of privacy protection. This means that recipients of personal data must be in places that have privacy legislation that is substantially similar to the EU Directives. Getting this adequacy assessment can be important for countries wanting to receive data from the EU through outsourcing arrangements.

Canada, on the other hand, has taken an organization-by-organization approach, where it is the adequacy of the receiving organization that is important. Under Canadian law, an organization collecting and processing personal information is responsible for safe storage, proper use, destruction, etc., and they maintain that obligation even if the data is shared across borders. Organizations and governments are required to setup business arrangements and procedures to ensure that privacy controls are maintained. Organizations cannot outsource their privacy responsibilities, and privacy regulators can audit compliance and respond to complaints.

In contrast, the U.S. has developed a patchwork of state and federal regulations, industry-specific laws, and jurisprudence. Neither the European, Canada, or U.S. approach is necessarily better than another, but it is important that there be a thorough analysis for determining appropriate sharing partners in any environment.

Review

It is not enough to plan and execute. Organizations must also follow-up by conducting reviews, assessments, and audits. There should be a method to accept and deal with complaints and issues, whether they come from within or outside the organization. Also, where possible, independent oversight mechanisms should be put in place, with appropriate visibility and powers.

4 Conclusion

International data sharing, by governments, organizations, and companies, is happening and will continue to happen. Maintaining security, privacy, and ethical conduct can be difficult when disparate parties come together. However, developing strong practices about sharing based on important principles can go a long way towards making sure the right things happen. Such practices will benefit the research community as well as other communities for whom data sharing is vital.

Author Index