Andrew A. Adams
Michael Brenner
Matthew Smith (Eds.)

LNCS 7862

Financial Cryptography and Data Security

FC 2013 Workshops, USEC and WAHC 2013
Okinawa, Japan, April 2013
Revised Selected Papers

Springer

Lecture Notes in Computer Science 7862

Andrew A. Adams Michael Brenner
Matthew Smith (Eds.)

Financial Cryptography and Data Security

FC 2013 Workshops, USEC and WAHC 2013
Okinawa, Japan, April 1, 2013
Revised Selected Papers

Springer

Volume Editors

Andrew A. Adams
Meiji University
Centre for Business Information Ethics
Kanda Surugadai 1-1
Tokyo 101-8301, Japan
E-mail: aaa@meiji.ac.jp

Michael Brenner
Leibniz Universität Hannover
Distributed Computing and Security Group
Schloßwender Straße 5
30159 Hanover, Germany
E-mail: brenner@dcsec.uni-hannover.de

Matthew Smith
Leibniz Universität Hannover
Distributed Computing and Security Group
Schloßwender Straße 5
30159 Hanover, Germany
E-mail: smith@dcsec.uni-hannover.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-41319-3 e-ISBN 978-3-642-41320-9
DOI 10.1007/978-3-642-41320-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013949274

CR Subject Classification (1998): E.3, K.4.4, K.6.5, C.2, 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 Okinawa, Japan on 1st April, 2013.

USEC 2013: Workshop on Usable Security

Networked information systems permeate modern life. From the ATM to the tablet computer, they're ubiquitous, as is increasingly their connectivity to the Internet. Securing these systems is a hard job at the technical level. The socio-technical side adds even more complexity. It is now well-understood that the human side of securing systems is at least as difficult as the technical side. While great strides have been made in making systems usable and technically secure, there is often an inherent contradiction in providing usable security: security is both an emergent property of the system as a whole and for some users/system elements/circumstances the aim of security is explicitly to apply the ultimate opposition of usability: denial of access.

This apparent contradiction underpins this workshop, which brings together researchers from a variety of disciplines including, but not limited to, computer science, psychology, business studies, to present and discuss recent advances in enabling systems to provide more easily usable security and privacy.

Google, Inc. and the Centre for Business Information Ethics at Meiji University sponsored the workshop's keynote speaker, Dr. Alessandro Acquisti of Heinz College, Carnegie Mellon University, who gave a highly engaging talk on his recent work on the behavioral economics of security and privacy:

Confessions of a Privacy Economist

What drives people to disclose or protect their personal information? What are the tangible and intangible consequences of those decisions? In this talk, I will discuss the transition from the economics to the behavioral economics of privacy. In particular, I will present and contrast a series of opposing "frames," or ways to frame and analyze the privacy debate, using the lenses of behavioral economic research. I will start from frames I have analyzed in my research (for instance: is privacy really about "transparency" and "control"?) and progressively move onto less settled, and perhaps more controversial, frames of the debate.

The organisers, Steering Committee and Program Committee of USEC 2013 thank the International Financial Cryptographers Association and the organisers of Financial Crypto and Data Security 2013 for their support in hosting the workshop.

July 2013 Andrew A. Adams

Organization

USEC Steering Committee

Jean Camp	Indiana University, USA
Jim Blythe	University of Southern California, USA
Angela Sasse	University College London, UK

USEC 2013 Organising Committee

Andrew A. Adams	Meiji University, Japan
Kiyoshi Murata	Meiji University, Japan

USEC 2013 Program Committee

Sadia Afroz	Drexel University, USA
Rainer Böhme	University of Münster, Germany
Pam Briggs	Northumbria University, UK
Lorrie Cranor	Carnegie Mellon University, USA
Neil Gandal	University of Tel Aviv, Israel
Seda Gürses	K.U. Leuven, Beligum
Peter Gutmann	University of Auckland, New Zealand
Raquel Hill	Indiana University, USA
Tiffany Hyun-Jin Kim	Carnegie Mellon University, USA
Markus Jakobsson	PayPal, USA
Timothy Kelley	Indiana University, USA
Brian LaMacchia	Microsoft Research, USA
William Lehr	MIT, USA
Hui Kai Lung	Hong Kong University of Science and Technology, China
Hitoshi Okada	National Institute of Informatics, Japan
Frank Stajano	University of Cambridge, UK
Andrew Patrick	Office of the Privacy Commissioner of Canada, Canada
Hovav Schacham	University of California at San Diego, USA
Bruce Schneier	BT, USA
Dan Schutzer	BITS, USA
Sean Smith	Dartmouth College, USA
Douglas Stebila	Queensland University of Technology, USA
David Wagner	University of California, Berkeley, USA
Nicholas Weaver	University of California, Berkeley, USA
Tara Whalen	Carleton University, Canada

WAHC 2013: Workshop on Applied Homomorphic Cryptography

Homomorphic Cryptography has become one of the hottest topics in mathematics and computer science since Gentry presented the first fully homomorphic scheme in 2009. This has also enhanced the interest in secret function evaluation, private information retrieval or searchable encryption in general. Many new cryptographic schemes have been proposed, creating a diverse mathematical basis for further theoretical research. Research on practical applications of homomorphic encryption, secret function evaluation, private information retrieval or searchable encryption is still less advanced due to the poor performance resulting on the complexity assumptions made in current encryption schemes. The goal of the WAHC was to bring together professionals, researchers, and practitioners in the area of computer security and applied cryptography with an interest in practical applications of homomorphic encryption, secure function evaluation, private information retrieval or searchable encryption to present, discuss, and share the latest findings in the field, and to exchange ideas that address real-world problems with practical solutions using homomorphic cryptography and other privacy preserving mechanisms.

The workshop received 12 submissions, each of which was reviewed by at least 3 Program Committee members. While all the papers were of high quality only 6 papers were accepted to the workshop. We want to thank the researchers of all 12 submissions, the members of the Program Committee, the workshop participants, the FC general chair, Kazue Sako, the program chair, Ahmad-Reza Sadeghi and the USEC workshop chair, Andrew A. Adams. Special thanks go to Vinod Vaikuntanathan who traveled all the way to Japan to give the workshop's keynote speech, which was also enjoyed by the attendees of the Financial Crypto and Data Security 2013. The talk surveyed the recent progress in the areas of fully homomorphic encryption and functional encryption – two very powerful methods for computing on encrypted data. It also described the exciting work towards making these technologies practical, and some future directions in this field.

July 2013

Michael Brenner
Matthew Smith

WAHC 2013 Program Committee

Jose Maria Alcaraz Calero	HP Labs, UK
Lynn Batten	Deakin University, Australia
Zvika Brakerski	Stanford University, USA
Dario Fiore	Max Planck Institute for Software Systems, Germany
Kristin Lauter	Microsoft, USA
Aggelos Kiayias	University of Connecticut, USA
Vladimir Kolesnikov	Bell Labs, USA
David Naccache	Ecole Normale Superieure, France
Maire O'Neill	Queen's University Belfast, UK
Elizabeth O'Sullivan	Queen's University Belfast, UK
Henning Perl	Universität Hannover, Germany
Benny Pinkas	Bar Ilan University, Israel
Kurt Rohloff	BBN Technologies, USA
Christoph Sorge	Universität Paderborn, Germany
Osman Ugus	Hochschule HAW Hamburg, Germany
Yevgeniy Vahlis	AT&T Labs, USA
Vinod Vaikuntanathan	University of Toronto, Canada
Marten van Dijk	MIT CSAIL, USA
Fré Vercauteren	K.U. Leuven, Belgium
Adrian Waller	Thales, UK
Dirk Westhoff	Hochschule Furtwangen University, Germany
Xun Yi	Victoria University, Australia

Table of Contents

The Workshop on Usable Security (USEC 13)

The Workshop on Applied Homomorphic Cryptography (WAHC 13)

I Think, Therefore I Am: Usability and Security of Authentication Using Brainwaves*

John Chuang[1], Hamilton Nguyen[2], Charles Wang[2], and Benjamin Johnson[3]

[1] School of Information, UC Berkeley
chuang@ischool.berkeley.edu
[2] Department of EECS, UC Berkeley
{hamiltonnguyen,charleswang}@berkeley.edu
[3] Department of Mathematics, UC Berkeley
benjamin@math.berkeley.edu

Abstract. With the embedding of EEG (electro-encephalography) sensors in wireless headsets and other consumer electronics, authenticating users based on their brainwave signals has become a realistic possibility. We undertake an experimental study of the usability and performance of user authentication using consumer-grade EEG sensor technology. By choosing custom tasks and custom acceptance thresholds for each subject, we can achieve 99% authentication accuracy using single-channel EEG signals, which is on par with previous research employing multi-channel EEG signals using clinical-grade devices. In addition to the usability improvement offered by the single-channel dry-contact EEG sensor, we also study the usability of different classes of mental tasks. We find that subjects have little difficulty recalling chosen "pass-thoughts" (e.g., their previously selected song to sing in their mind). They also have different preferences for tasks based on the perceived difficulty and enjoyability of the tasks. These results can inform the design of authentication systems that guide users in choosing tasks that are both usable and secure.

Keywords: pass-thoughts, EEG, authentication, usability.

1 Introduction

Advances in EEG (electro-encephalography) bio-sensor technologies have opened up brainwave research and application development at an unprecedented level in recent years. Traditionally, EEG data capture has been performed in clinical settings using invasive probes under the skull or wet-gel electrodes arrayed over the scalp. Now, similar data can be collected using consumer-grade non-invasive dry-contact sensors built into audio headsets and other consumer electronics. This opens up immense possibilities for using brainwave signals in different application domains. Originally limited to neuroscience research and clinical treatment

* This research was supported in part by the National Science Foundation under award CCF-0424422 (TRUST).

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 1–16, 2013.

of neurological diseases, EEG technologies are now being deployed for education, training, entertainment, and other ubiquitous computing applications.

Given the growing commercial availability of this technology, an important research agenda is to develop and evaluate different practical methods for regular users to apply their own brainwave data, in everyday (i.e., non-laboratory) settings, for different computer-based applications. In this work, we take a first step by focusing on the problem of user authentication using brainwaves. We propose and evaluate different classes of mental and/or motor tasks that users may perform while wearing a headset with EEG sensors. In addition to collecting EEG data from human subjects as they performed these tasks, we also collected experimental and questionnaire data to measure the usability of the tasks. Taken together, we compare the performance of different mental/motor tasks using metrics for signal similarity, authentication accuracy, task difficulty, task enjoyability, and task repeatability.

We make a significant departure from previous EEG-based authentication studies by studying the efficacy of single-channel as opposed to multi-channel EEG signals. Modern clinical EEG systems employ dense arrays of electrodes to provide 32, 64, 128, and 256 channels of EEG data. In contrast, for our experimental study, we use a consumer-grade headset that provides a single-channel EEG signal. Specifically, the Neurosky MindSet [1] places a single dry-contact sensor over the left frontal lobe region of the brain (Figures 1 and 2). Other than the EEG sensor, the headset is indistinguishable from a conventional Bluetooth headset for use with mobile phones, music players, and other computing devices. The headset can be purchased in the market for approximately $100.

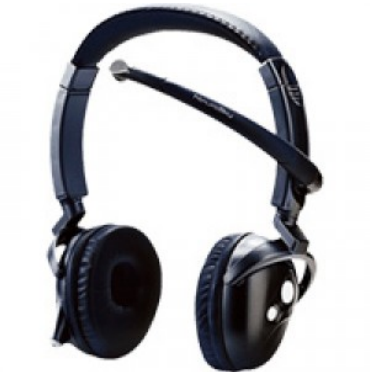


Fig. 1. EEG Headset Used in the Study: Neurosky MindSet

The headset form factor and the non-intrusiveness of the sensor imply a significant lowering of the usability barrier for EEG-based authentication. On the other hand, does the switch from multi-channel to single-channel signals lead to information loss that may render EEG-based authentication infeasible? This is a key motivating question of our study.

Our first key finding is that single-channel EEG signals do exhibit patterns that are subject-specific. Using standard measures of statistical similarity, we find higher signal similarity within subjects than across subjects. This is true across different mental tasks performed by the subjects; and it is true even for the brainwave signals of the same subjects that were collected over different experimental sessions on different days.

Our second key finding is that single-channel EEG authentication can be just as accurate as multi-channel EEG authentication. Leveraging our first finding, we propose and evaluate a suite of threshold-based authentication protocols that makes accept/reject decisions based on statistical similarities of signals. By combining the use of custom tasks and custom thresholds for each user, we can reduce false error rates down to the 1% level, which is comparable to the error rates achieved with multi-channel EEG signals.

Our third key finding is that neither signal similarity nor authentication performance are significantly affected by the categories of mental tasks performed by the subjects. In particular, personalized mental tasks (e.g., sing their favorite song silently, focus on their personal pass-thought) do not produce higher signal similarity or authentication accuracy over mental tasks that are common to all subjects (e.g., close eyes and focus on breathing).

On the other hand, as our fourth key finding, we find that the different categories of mental tasks score very differently in terms of user-perceived difficulty and enjoyability. When asked to choose a mental task that they would be willing to repeat on a daily basis, different subjects assign different weights to difficulty and enjoyability in making their choice. However, recall rates are consistently high for those mental tasks that require the subjects to remember their chosen secrets across sessions.

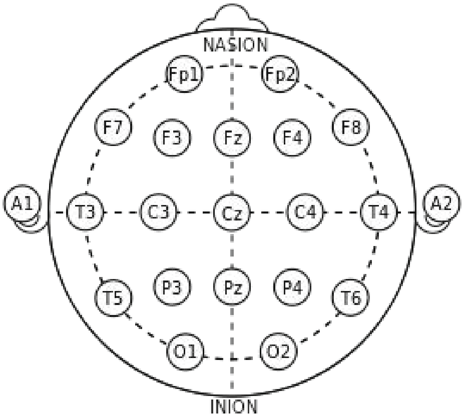


Fig. 2. Electrode placements for the International 10-20 Standard. The placement of the Neurosky Mindset electrode corresponds to the Frontal Polar 1 (Fp1) location.

Taken together, these findings suggest that designers of EEG-based authentication systems do not have to make a hard choice between security and usability. The authentication system should be designed to allow users to experiment with different categories of mental tasks, so that each user repeats a customized task – one that they find easy and enjoyable, but that is also capable of producing high authentication accuracy.

2 Related Work

This research draws upon foundations and recent advances in multiple disciplines, ranging from neuroscience, human-computer interaction, computer security, signal processing, and machine learning. To the best of our knowledge, this work is the first experimental study of the usability design of brainwave-based authentication.

2.1 Brainwave-Based Authentication

The use of brainwave signals for user authentication has received widespread attention in recent years. Thorpe et al. motivate and outline the design of a "pass-thoughts" system [17]. By thinking of a pass-thought rather than typing in a password, this method of authentication promises numerous security advantages, including the resistance to dictionary attacks and shoulder-surfing.

A number of researchers have separately established the feasibility of using EEG signals to classify and/or authenticate users. With a focus on accuracy, they apply a range of statistical, signal processing, and machine learning techniques on multi-channel EEG signals. Poulos et al. use an artificial neural network to classify 4 subjects based on their EEG signals [16]. Marcel and Millan employ gaussian mixture model and maximum a-posteriori model for authentication with 9 subjects [12]. Palaniappan achieved 100% accuracy in classifying 5 subjects using a linear discriminant classifier [14], as well as zero False Acceptance Rate (FAR) and zero False Rejection Rate (FRR) using a two-stage threshold-based authentication process [15]. In each of these studies, the EEG data are captured using clinical-grade multi-channel sensors. More recently, Ashby et al. achieved 100% authentication accuracy with 5 subjects using consumer-grade multi-channel sensors [3]. In each of these studies, all the subjects performed identical tasks, ranging from baseline relaxation to imaginary motor movement, visualization, and solving math problems. None of these studies addresses task personalization or system usability.

2.2 Usability of Novel Authentication Systems

It is well understood that authentication systems must strike a balance between security and usability. Many security solutions fail not because of any flaw in the underlying technical design, but because of difficulties faced by humans in using the system in real-world settings as intended by the system designers. For example, users may find it difficult to remember one different password for each account they own, and resort to writing down the passwords on paper, thereby introducing new vulnerabilities to the system.

Such considerations underpin the development of graphical passwords as usable alternatives to text-based passwords [4]. In systems such as Draw-A-Secret [11], Deja Vu [7] and Passfaces [2], users authenticate themselves via recalling or recognizing images, rather than typing in a sequence of alphanumeric characters as in traditional password-based systems. A key usability metric for these systems is recall, i.e., the ability for users to remember their chosen secrets (e.g., images, faces) over different experimental sessions that are separated by periods of days or weeks. Usability studies demonstrate far higher recall rates for graphical passwords than for text passwords [7, 5]. In our experiment, we also investigate the ability for users to recall their chosen pass-thoughts across different sessions.

More generally, the different approaches to biometrics, including fingerprinting, iris scanning, facial recognition, voice recognition, each introduce different usability challenges and opportunities [6]. With the embedding of EEG and other bio-sensors into mobile phones, headsets, wearable computing devices, and other consumer electronics, the collection of brainwave signals for authentication and other purposes may become more natural and less intrusive than the collection of fingerprints, voice samples, and other biometric signals.

2.3 EEG and HCI

Research in brain-computer interface (BCI) has established the feasibility of using EEG signals to control computers and other devices. BCI systems can reliably evoke and measure event-related potentials (ERP) such as the P300, and use them to spell words and move computer cursors based on a user's intent [8–10, 13]. This proves very valuable in restoring the ability to communicate for patients suffering from the locked-in syndrome and other neurological diseases, and can be generalized to healthy users as well. While our work does not seek to infer user intent from their EEG signals, our choice of user tasks involving external stimuli are informed by the efficacy of eliciting and capturing these event-related potentials.

3 Experiment

3.1 Overview

Our research involved human subjects, and our experimental procedures were approved by an Institutional Review Board. We recruited a total of 15 subjects to participate in our study, all of whom were UC Berkeley undergraduate or graduate students. Each subject met with two investigators in a quiet, closed-room setting for two 40-50 minute sessions on separate days. We briefed subjects on the objective of the study, fitted them with a Neurosky MindSet headset, and provided instructions for completing each of seven tasks. As the subjects performed each task we monitored and recorded their brainwave signals.

3.2 Tasks

The following tasks were repeated five times in each session for each subject.

Breathing Task (breathing). Subjects close their eyes and focus on their breathing for 10 seconds.

Simulated Finger Movement (finger). Subjects imagine in their mind that they are moving their right index finger up and down in sync with breathing, without actually moving their finger, for 10 seconds.

Sports Task (sport). Subjects select a specific repetitive motion from a sport of their choosing. They then imagine moving their body muscles to perform the motion, for 10 seconds.

Song/Passage Recitation Task (song). Subjects imagine that they are singing a song or reciting a passage for 10 seconds without making any noise.

Eye and Audio Tone Task (audio). Subjects close their eyes and listen for an audio tone. After 5 seconds, the tone plays; upon hearing the tone, the subjects open their eyes and stare at a dot on a piece of paper in front of them for an additional 5 seconds.

Object Counting Task (color). Subjects are asked to choose one of four colors – red, green, blue, or yellow. They are then shown on a computer screen a sequence of six images. Each image contains a 5x6 grid of colored boxes. As each grid appears, subjects count, silently in their mind, the number of boxes corresponding to their chosen color. A new grid appears after each 5 seconds. The task continues 6 rounds for a total of 30 seconds.

Pass-Thought Task (pass). Subjects are asked to choose their own pass-thought. A pass-thought is like a password; however, instead of choosing a sequence of letters and numbers, one chooses a mental thought. When subjects are instructed to begin, they focus on their pass-thought for 10 seconds.

3.3 Questionnaire

In addition to the brainwave data, we also asked subjects a series of survey questions. At the end of each session, we asked the subjects to select the one task (out of seven) that they would be most willing to repeat every day. After subjects completed both sessions, we asked them to rate each of the tasks according to the following binary choices: (i) difficult or easy, and (ii) enjoyable or boring.

3.4 Brainwave Data

As subjects completed each task, we recorded their raw EEG data on a computer. The data was transmitted via a bluetooth network connection from the headset to the computer. The raw data includes single-channel EEG signals in both the time and frequency domains. We specifically use the power spectrum data, a two-dimensional matrix which gives the magnitude of the signal for every frequency component at every point in time. With 15 subjects repeating seven tasks, five times per session, and two sessions per subject, we have a total of 1050 brainwave data samples.

3.5 Data Preprocessing

Before performing any analysis on the brainwave data, we first pre-process the power spectrum data to compress the samples. In the temporal dimension, we extract only the middle five seconds out of the total ten seconds of each recorded signal (the exception is the *color* task, for which we chose a five-second section corresponding to a specific image). In the frequency dimension, we extract only the data corresponding to the alpha wave (8-12 Hz) and the beta wave (12-30 Hz) ranges of the signals. We apply our analysis to both ranges.

The second step in our data preparation is to take this two dimensional signal and compress it into a one dimensional signal. Our chosen compression method flattens the signal in the time dimension – specifically, for each frequency component, we compute the median magnitude corresponding to that frequency component over all time. The end result is a one-dimensional column vector with one entry for each measured frequency. This column vector representation is how brainwave samples are stored and manipulated within the authentication system.

4 Data Analysis

After collecting and processing the brainwave data, we begin evaluating the effectiveness of the signals in the context of authentication. This problem requires us to distinguish the signals among different subjects.

We begin by quantifying the similarity between two signals u and v as the cosine similarity of the vector representation of the signals, given by the equation:

$$\text{similarity}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}.$$

Similarity gives a value between 0 and 1, where a similarity of 1 would indicate a perfect match.

We next define two additional notions related to similarity – self-similarity and cross-similarity. Self-similarity refers to the similarity of signals within a single subject, while cross-similarity refers to the similarity of signals between different subjects. Our hypothesis is that self-similarity should be consistently greater than cross-similarity for all subjects, in all tasks. If this is true, we will be able to leverage this difference in our authentication system.

For a fixed task t and given subject s, we define the self-similarity of s in t to be the average of the similarity of every possible pair of samples belonging to s. Likewise, for a fixed task t and given subject s, we define the cross-similarity of s in t to be the average of the similarity of every possible pair for which one sample in the pair belongs to s and the other sample does not belong to s.

Table 1 displays the results of testing our similarity metric. For a given subject, we compute his or her self- and cross-similarity for every task, and then take the average of these values. The final average is the number displayed under the Self and Cross columns. Lastly, we look at the relative difference between self- and cross-similarity for each subject rather than the absolute difference. The last column corresponds to the percent difference between the Self and Cross columns.

From these results, we can make a few observations. First, self-similarity is higher than cross-similarity for all subjects, which is an important pre-requisite in using this metric in our authentication system. Second, there is noticeable variation in percent difference between the 15 subjects. This second result will be used in improving our protocol.

Next, Table 2 gives an alternative visualization of our results. For a given task, we compute the self- and cross-similarity of each subject, and then take the average over all subjects. This gives similarity values associated with tasks rather than subjects. Again, we can see that self-similarity is higher than cross-similarity in all cases. Interestingly, we can observe that the variance in difference in Table 1 is higher than the variance in difference in Table 2. This suggests that the similarity measure has greater variation between subjects than between tasks.

Table 1. Similarity Comparison of Subjects

Subject	Self Similarity	Cross Similarity	Percent Difference
subject 0	0.7207	0.6653	7.99%
subject 1	0.7268	0.6745	7.46%
subject 2	0.7014	0.6602	6.05%
subject 3	0.7577	0.6397	16.89%
subject 4	0.7232	0.6617	8.88%
subject 5	0.6771	0.6702	1.02%
subject 6	0.7147	0.6264	13.17%
subject 7	0.7253	0.6817	6.20%
subject 8	0.7368	0.6828	7.61%
subject 9	0.6941	0.6435	7.57%
subject 10	0.7161	0.6847	4.48%
subject 11	0.7142	0.6816	4.67%
subject 12	0.711	0.6817	4.21%
subject 13	0.7028	0.6106	14.04%
subject 14	0.7099	0.6702	5.75%

Table 2. Similarity Comparison of Tasks

Task	Self Similarity	Cross Similarity	Percent Difference
breathing	0.7304	0.6834	6.65%
finger	0.7282	0.6567	10.33%
sport	0.7144	0.676	5.52%
song	0.7013	0.6498	7.62%
audio	0.7283	0.6637	9.28%
color	0.6664	0.599	10.65%
pass	0.6931	0.632	9.22%

5 Authentication

5.1 Problem Definition

The authentication problem is also referred to as the user verification problem. Given an (identity, sample) pair, the authentication system must determine if the sample provides a legitimate match to the identity.

Authentication systems make two types of errors: False Acceptance (FA) errors occur when the system accepts an impostor, while False Rejection (FR) errors occur when the system rejects an authorized user. The performance of an authentication system can thus be measured in terms of its False Acceptance Rate (FAR) and False Rejection Rate (FRR). The two error measures are often merged to form the Half Total Error Rate (HTER), defined as:

$$HTER = (FAR + FRR)/2.$$

5.2 Testing Schema

Before discussing the implemented authentication protocols themselves, we briefly describe our testing schema used to evaluate the performance of the protocols. Recall that for each task and subject we collected and processed 10 brainwave samples. Our testing schema randomly selects 5 of these samples (for each task and user) to train the authentication protocol. The remaining samples are used to test the protocol.

Evaluating FRR. To assess false rejection we may focus our attention on a single user at a time. Given a specific task, each user has only 5 samples in the testing set for that task, and our testing schema runs the relevant authentication protocol on each of them along with the user's correct identity. If the protocol were to work perfectly it would always accept these (user, sample) pairs. The FRR is computed as the average percentage of such tests that do not accept, taken over all matching pairs of users and samples in the test set.

Evaluating FAR. To assess false acceptance we must focus on many users at a time. Indeed as there is only one legitimate user but many potential impostors, there are many more opportunities for false acceptance than for false rejection. Given a specific task and user, our testing schema randomly selects 5 samples that do not match the user, and runs the relevant authentication protocol on this set of false (user, sample) pairs. If the protocol were to work perfectly it would always reject these pairs, and the FAR is computed as the average percentage of such tests that incorrectly accept.

5.3 Protocols and Results

Baseline Protocol. Our baseline protocol will also be referred to as the Common Task Common Threshold protocol. In this system, all brainwave samples correspond to a single, fixed task. We then choose a common threshold T to be used for all subjects.

The core authentication mechanism is as follows: a user provides as input his claimed identity and brainwave sample. We compute the value $selfSim$ to be the average similarity between the given sample and all 5 samples known to belong to the user. We then randomly select a set of 5 samples such that none of the samples in this set belong to the user. Next, we compute the value $crossSim$ to be the average similarity between the given input sample and the samples in this new set. Finally, if the percent difference between $selfSim$ and $crossSim$ is greater than or equal to T, we accept the authentication attempt. If not, we reject it.

Table 3 shows the result of testing the baseline protocol for each of the tasks. Although the protocol performs better than random guessing, it is still far from practically usable. At best, the HTER is at .322 for the *audio* task. We also observe that FAR is lower than FRR for every task (and for most tasks, many times lower), which implies the current protocol is more effective at determining impostors than confirming legitimate users. In the following sections, we explore improvements over the baseline protocol.

Table 3. Authentication with Common Thresholds

Task	FAR	FRR	HTER
breathing	0.156	0.578	0.367
finger	0.044	0.733	0.389
sport	0.089	0.644	0.367
song	0.155	0.578	0.367
audio	0.244	0.400	0.322
color	0.244	0.622	0.433
pass	0.356	0.400	0.378

Table 4. Authentication with Customized Thresholds

Task	FAR	FRR	HTER
breathing	0.000	0.280	0.140
finger	0.067	0.120	0.093
sport	0.027	0.187	0.107
song	0.000	0.093	0.047
audio	0.027	0.147	0.087
color	0.120	0.440	0.280
pass	0.000	0.120	0.060
customized	0.000	0.022	0.011

Customized Threshold. Our first improvement over the baseline is the Common Task Customized Threshold protocol. This new protocol is nearly the same as the baseline except for one key difference – rather than comparing against a common threshold T, we compare against T_i, a customized threshold optimized specifically for user i.

The first seven rows of Table 4 show the results of testing the performance of this new protocol for each of the tasks. With customized thresholds, we were able to decrease FRR significantly for every task, and in almost every case, we did not sacrifice performance with regards to FAR – the lone exception is the $finger$ task, for which FAR actually increased when customized thresholds were implemented. Overall however, the HTER of the $finger$ task decreased as well.

Further, we were able to achieve a reasonably high success rate for nearly all tasks. Put another way, these results do not suggest that there is one particular kind of task that is definitively most effective for authentication.

Customized Task Customized Threshold. Our final protocol is the Customized Task Customized Threshold protocol. In the previous two protocols, the chosen task was fixed for all subjects. We add an additional step of precomputation in which we determine for each subject, the optimal task to maximize the difference between self and cross similarity for the subject. Then, within that task we determine the optimal threshold specific for that subject, as above.

The last row in Table 4 shows the result of using customized tasks. This version of the protocol outperforms every instance of using common tasks, achieving an HTER of 1.1%. The success of the customized task protocol further reinforces our belief that there does not exist one single task that is the best to use for authentication.

Additionally, we can remove tasks one-by-one from the pool of tasks considered by the protocol and observe how this affects performance. In one instance, we were able to reduce the pool of tasks to only two – specifically *breathing* and *audio* – and still maintain the same HTER of 1.1% as when all seven tasks are used.

6 User Identification

We next consider the more challenging problem of user identification, i.e., given a brainwave signal, can we identify the user to which the signal belongs. This corresponds to the classification problem in machine learning, and we apply standard classification techniques to our data. As in our approach to authentication, we first prepare a truncated signature for each trial by restricting to alpha wave and beta wave frequencies, and averaging across the middle portion of the time domain.

Our testing schema is then as follows: we select one trial signature to be a testing sample. The remainder of the trial signatures are treated as training samples, i.e., we assume the subject identities of these signatures are known. We ask our classifier to classify the testing sample, and record whether the classifier identified the correct subject. This process is repeated for every trial signature.

Our classifier is a basic adaptation the K-Nearest Neighbors (KNN) algorithm for coloring graphs. Given a complete graph with distances between each node and with all but one node colored, the KNN algorithm colors the uncolored node with the most common color among its K nearest neighbors. If there is a tie among colors in the nearest neighbor set, we restrict to nodes having those tied colors, and run the algorithm again with K decremented. Any ties remaining when $K = 1$ are resolved by a fair coin flip. Our adaptation of this algorithm has trial signatures as nodes, subject identities as colors, and Cosine Similarity as the distance metric.

Figure 3 summarizes the classification success rates for $K = 5$. The classifier generally does two to three times better than random guessing. (Since there are 15 colors, random guessing has a classification success probability of $\frac{1}{15} \approx 6.7\%$.) The *audio*, *sport*, and *color* tasks have the best overall classification rates. For example, the classifier can correctly identify a user 22% of the time based on

EEG samples from the *audio* task. This corresponds to a 3.3x improvement over random guessing. Nonetheless, a 22% success rate still falls far below levels acceptable for practical user identification systems.

The reason for the discrepancy in performance between user authentication and user identification is instructional. For user authentication, we can pick custom tasks that provide the highest authentication accuracy for each subject. For user identification, on the other hand, knowledge of which task was performed for a given EEG sample does not help in the classification at all.

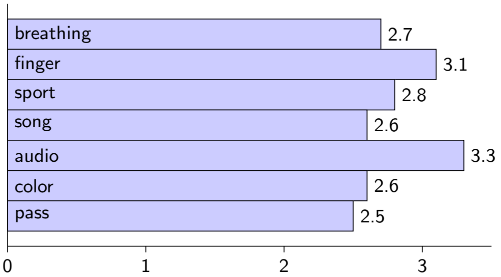


Fig. 3. Classification Performance (random-guessing = 1.0)

7 Usability

There are two dimensions of usability to consider: the usability of the EEG hardware, and the usability of the mental tasks.

In terms of hardware, a single-channel EEG sensor in the form of a dry-contact electrode integrated with a wireless headset is much less intrusive than an array of electrodes that must be carefully placed over the scalp. Having established that single-channel EEG signals collected with consumer-grade EEG sensors over a range of mental tasks can provide the same level of authentication accuracy as multi-channel EEG signals collected with clinical-grade EEG sensors, we can posit that the usability vs. security tradeoff is now tipping in favor of the consumer-grade single-channel approach.

Let us turn to the usability of the mental tasks. At the conclusion of the second experimental session, each of the fifteen subjects was asked in a questionnaire to rate each of the seven tasks as either "difficult" or "easy", and as "boring" or "enjoyable". The responses are summarized in the first three columns of Table 5. For example, seven of fifteen subjects found the *pass* task to be difficult to perform, because their chosen pass-thoughts involve feelings or events that proved hard to repeat on a consistent basis. Similarly, seven of fifteen subjects found the *sport* task to be difficult to perform, because they found it unnatural to imagine the movement of their muscles without actually moving them. On the other hand, all fifteen subjects found the *breathing*, *audio*, and *color* tasks to be easy to perform.

Eight of fifteen subjects rated the *finger* task as boring. Presumably, the task is monotonous just as it is easy. On the other hand, twelve of fifteen subjects rated the *breathing*, *sport*, *song*, and *color* tasks as enjoyable.

At the conclusion of both the first and second experimental sessions, the questionnaire also asked the subjects to choose one task that they would most like to repeat on a daily basis. The responses are summarized in the last column of Table 5.

We can see that the *finger* task, rated boring by more than half of the subjects, was not chosen at all. The *sport* task, rated difficult by almost half of the subjects, received the next fewest votes. On the other hand, the *color* and *breathing* tasks received overall the most repeatability votes. These two tasks are the least boring and least difficult tasks, as evaluated by the subjects.

Table 5. Usability Comparison of Tasks

Task	**Was Difficult**	**Was Boring**	**Would Repeat**
breathing	0/15	3/15	7/30
finger	3/15	8/15	0/30
sport	7/15	3/15	1/30
song	4/15	3/15	5/30
audio	0/15	4/15	4/30
color	0/15	3/15	9/30
pass	7/15	6/15	4/30

Two of the seven tasks require the subjects to respond to external stimuli – an audio tone for the *audio* task, and a sequence of images for the *color* task. Both tasks were perceived to be easy and likely candidates for daily repetition.

Four of the seven tasks provide the subjects an opportunity to choose their own secret: *sport*, *song*, *color*, and *pass*. In contrast, the other three tasks, *breathing*, *finger*, and *audio*, do not involve a personal secret. We do not observe any relationship between the utilization of a secret and the difficulty, enjoyability, or repeatability of a task.

During the second experimental session, we tested each subject on their ability to recall their chosen secrets for the *sport*, *song*, *color*, and *pass* tasks. As seen in Table 6, the subjects had no difficulty in recalling their personalized sport, song, and pass-thought choices. One of the fifteen subjects could not recall the color he chose from the previous session. This suggests the possibility that users are better able to remember secrets that they come up with themselves, than secrets that they select from a menu of discrete choices.

An open question is whether the changing of a chosen secret, as part of a user-initiated password change routine, may affect the authentication performance or even the usability of the task.

Table 6. Recall Rate of Tasks

Task	Recall Rate
song	15/15
sport	15/15
color	14/15
pass	15/15

8 Conclusion and Future Work

In this paper, we study the usability and performance of brainwave-based authentication. Motivated by the trend of low-cost EEG sensors embedded in various consumer electronic devices, we conduct an experimental study to capture brainwave signals from human subjects using consumer-grade EEG headsets in a non-clinical environment. We design a number of different mental tasks for the subjects to perform, and evaluate the usability of the tasks based on their difficulty, enjoyability, and repeatability.

We find that brainwave signals, even those collected using low-cost non-intrusive EEG sensors in everyday settings, can be used to authenticate users with high degrees of accuracy. We show it is possible to compensate for the lower fidelity of single-channel EEG signals by intelligently matching signal similarity thresholds and customized tasks to each user. This means that we can now bypass the usability challenges associated with conventional EEG systems designed for clinical applications.

Different mental tasks also vary in their usability. Subjects will not opt for repeating tasks that are perceived as either difficult or boring. Similar to the experience with graphical passwords, we find that pass-thoughts chosen by the subjects can be recalled by the subjects without much difficulty. In comparing the results of the usability analysis with the results of the authentication testing, we observe that there is no need to sacrifice usability for accuracy. It is possible to achieve accurate authentication with easy and enjoyable tasks.

There are a number of limitations of our study that point to interesting directions for future work.

We are able to maintain a high level of authentication accuracy with a subject pool that is 66% to 275% larger than those from previous studies [3, 12, 14–16], thus demonstrating the feasibility of authentication in a small population, e.g., a work group setting [14]. Nonetheless, it would still be valuable to investigate the scalability of the results to even larger populations.

Going beyond the small set of mental tasks evaluated in this study, a systematic exploration of additional categories of tasks would be of great value. From there, we can seek to gain a more complete understanding of which factors influence the usability and security performance of mental tasks.

While the primary focus of this paper is on user authentication, we also encountered the relative difficulty of accurate classification of users. It would be useful to ascertain whether the classification performance can be improved with other classification algorithms or with other mental tasks.

The robustness of brainwave-based authentication against impersonation attacks is an interesting problem. If an attacker gains knowledge of a target's customized task and chosen secret (e.g., the specific song or passage that a user repeats to herself), how easy or difficult is it for the attacker to fool the authentication system by performing the same customized task?

Finally, if the authentication system works by choosing customized tasks for each subject, the user enrollment process becomes an important design consideration. Today's users may balk at a 45 minute initiation process to set up their password, so the number and choice of mental task categories have to be carefully selected to optimize for both the duration of user enrollment and the accuracy of authentication.

References

1. Neurosky MindSet, `http://www.neurosky.com/`
2. Passfaces, `http://www.passfaces.com/`
3. Ashby, C., Bhatia, A., Tenore, F., Vogelstein, J.: Low-cost electroencephalogram (eeg) based authentication. In: Proceedings of 5th International IEEE EMBS Conference on Neural Engineering (April 2011)
4. Biddle, R., Chiasson, S., van Oorschot, P.: Graphical passwords: Learning from the first twelve years. ACM Computing Surveys 44(4) (2011)
5. Brostoff, S., Sasse, M.A.: Are passfaces more usable than passwords? A field trial investigation. In: Proceedings of HCI (2000)
6. Coventry, L.: Usable biometrics. In: Cranor, L., Garfinkel, S. (eds.) Usability and Security (2005)
7. Dhamija, R., Perrig, A.: Deja vu: a user study using images for authentication. In: Proceedings of the 9th Conference on USENIX Security Symposium (2000)
8. Donchin, E., Spencer, K., Wijesinghe, R.: The mental prosthesis: assessing the speed of a p300-based brain-computer interface. IEEE Transactions on Rehabilitation Engineering 8(2), 174–179 (2000)
9. Farwell, L.A., Donchin, E.: Talking off the top of your head: A mental prosthesis utilizing event-related brain potentials. Electroencephalography and Clinical Neurophysiology 70, 510–523 (1988)
10. Hinterberger, T., Kubler, A., Kaiser, J., Neumann, N., Birbaumer, N.: A brain-computer interface (bci) for the locked-in: comparison of different eeg classifications for the thought translation device. Clinical Neurophysiology 114(3), 416–425 (2003)
11. Jermyn, I., Mayer, A., Monrose, F., Reiter, M., Rubin, A.: The design and analysis of graphical passwords. In: Proceedings of 8th USENIX Security Symposium (August 1999)
12. Marcel, S., del, J., Millan, R.: Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4) (April 2007)
13. Nijboer, F., Sellers, E., Mellinger, J., Jordan, M., Matuz, T., Furdea, A., Halder, S., Mochty, U., Krusienski, D., Vaughan, T., Wolpaw, J., Birbaumer, N., Kubler, A.: A p300-based brain-computer interface for people with amyotrophic lateral sclerosis. Clinical Neurophysiology 119(8), 1909–1916 (2008)

14. Palaniappan, R.: Electroencephalogram signals from imagined activities: A novel biometric identifier for a small population. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 604–611. Springer, Heidelberg (2006)
15. Palaniappan, R.: Two-stage biometric authentication method using thought activity brain waves. International Journal of Neural Systems 18(1), 59–66 (2008)
16. Poulos, M., Rangoussi, M., Alexandris, N., Evangelou, A.: Person identification from the eeg using nonlinear signal classification. Methods of Information in Medicine (2002)
17. Thorpe, J., van Oorschot, P., Somayaji, A.: Pass-thoughts: Authenticating with our minds. In: Proceedings of the New Security Paradigms Workshop, NSPW (2005)

Usability and Security of Gaze-Based Graphical Grid Passwords

Majid Arianezhad[1], Douglas Stebila[2], and Behzad Mozaffari[2]

[1] School of Engineering Science, Simon Fraser University, Burnaby, B.C., Canada
arianezhad@sfu.ca
[2] Science and Engineering Faculty, Queensland University of Technology,
Brisbane, Australia
stebila@qut.edu.au, behzad.mozaffari@connect.qut.edu.au

Abstract. We present and analyze several gaze-based graphical password schemes based on recall and cued-recall of grid points; eye-trackers are used to record user's gazes, which can prevent shoulder-surfing and may be suitable for users with disabilities. Our 22-subject study observes that success rate and entry time for the grid-based schemes we consider are comparable to other gaze-based graphical password schemes. We propose the first password security metrics suitable for analysis of graphical grid passwords and provide an in-depth security analysis of user-generated passwords from our study, observing that, on several metrics, user-generated graphical grid passwords are substantially weaker than uniformly random passwords, despite our attempts at designing schemes to improve quality of user-generated passwords.

Keywords: graphical passwords, eye-tracking, usable security.

1 Introduction

Graphical password schemes have the potential to improve user authentication due to easier memorability and use. Typically, a user indicates various regions of the screen or draws some pattern using mouse, touch, or gaze input methods. Gaze-based password input is promising due to its resistance to shoulder surfing and because it may be easier for people with disabilities to use. The usability of graphical password has been extensively studied, but there has been very little investigation into the quality of user-generated graphical passwords. We investigate several variants of graphical grid passwords to determine if variations can improve the quality of user-generated passwords—in terms of point and stroke distribution and symmetry—while maintaining usability.

An extensive survey of the vast literature on graphical passwords was recently given by Biddle, Chiasson, and van Oorschot [4]. They describe three main categories of schemes: *recall-based schemes*, such as Draw-A-Secret [14], where the user must recall and enter a secret drawing or pattern from memory; *recognition-based schemes*, where a user must recognize a few personal objects from a set of objects, either images (Passfaces [20], Faces [7]) or text [28]; and *cued-recall*

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 17–33, 2013.

schemes, such as PassPoints [24,25] or Cued Gaze-Points [12], where the user is given an image cue and must recall and enter certain points or a pattern.

Recall-based schemes can be divided into two main subcategories. In *free-form drawmetric schemes*, such as Draw-A-Secret [14] or Pass-Go [23], the user draws an arbitrary image on a blank canvas. *Grid schemes* restrict the valid target points to a grid; some, such as PassShapes [26] and the gaze-based Eye-PassShapes [8], restrict moves to adjacent points in the grid or use limited patterns [10], whereas others, such as GridSure [5] and the popular 'pattern lock' 3 × 3 grid screens for Android and other [22] mobile phones allow users to enter arbitrary patterns of grid points. A few schemes [15] have users enter text-based passwords using *on-screen keyboards.*

Shoulder-surfing, where an attacker watches a user enter their password, is a well-known problem for graphical password schemes [11,29]. Grid schemes on mobile phones can be vulnerable to smudge attacks [3], though shoulder-surfing and smudge attacks can be mitigated using biometric characteristics from entering the password [9]. Magnetic entry schemes also resist shoulder-surfing attack [21]. *Gaze-based passwords* may be more resistant to such attacks, since no visual feedback of the user's entry is displayed on-screen, and may also be suitable as an input method for users with disabilities. (Gaze-based entry is not a security panacea, however: video cameras or attackers surreptitiously watching a user's eye movement may still be able to gain enough information to attack passwords with some success [8].)

A well-known weakness of traditional text-based passwords is that human-generated passwords are not truly random. While an eight-character mixed-case alphanumeric password may be chosen from a large theoretical password space ($(26 + 26 + 10)^8 = 62^8 \approx 2^{47.6}$), humans pick passwords from a non-uniform distribution with much lower entropy.

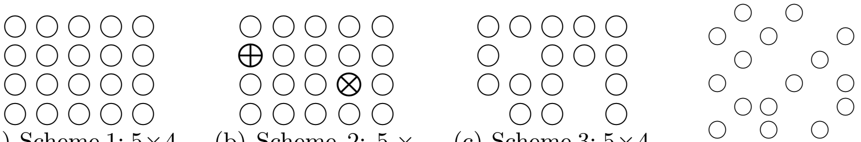
Unfortunately, most papers on graphical password schemes only mention the theoretical password space with no analysis of user-generated passwords, though some research has been done on the password security of some schemes. A line of research by van Oorschot and Thorpe has analysed the space of human-generated passwords in free-form drawmetric schemes [18] as well as the prevalence of image hot spots in cued-recall graphical passwords [17,19], though hot spots can be reduced using masking [6]. User-generated passwords in recognition-based schemes can also have poor entropy and be susceptible to educated guess attacks based on demographic information [7] or personal knowledge [13].

We focus on recall-based graphical grid schemes using eye-tracking for data entry. From the usability perspective, we aim to determine if gaze-based entry of graphical grid passwords, which have no recall cues, can achieve comparable success rates and entry times to cued-recall schemes. On the security side, we aim to provide metrics for the security of human-generated grid passwords, as previous security analyses do not directly carry over to grid schemes. We hypothesize that human-generated passwords will have more symmetry and not use uniformly distributed points and strokes, so we test several variants to see if they improve password quality.

2 Schemes

In a pre-trial phase, we had a handful of users try out a basic grid scheme, and noticed that the passwords they created tended to being symmetric and have poor distribution of the first and last points; in particular, a significant proportion of users chose the top-left point as their first point. Pre-study results were similar to the results for Scheme 1; see Appendix B for distribution of points during the main study. This motivated us to design several variants to see if we could improve the quality of user-generated passwords.

We propose four gaze-based graphical grid password schemes as shown in Figure 1. Scheme 1 is a basic 5 × 4 grid, a generalization of the 'pattern lock' screen popular on Android devices. In Scheme 2, we cued the user to start and end at the specified points, visually displayed with different colours; by picking the first point for the user, we hypothesize that the second point (i.e., the first user-selected point) might have better distribution; this also eliminates perfectly symmetric shapes that use the same start and end point. In Scheme 3, we removed a few random grid points: users may be less likely to pick symmetric shapes since not all the points were available to them. We also wanted to know if a bigger, sparser, less grid-like scheme induced more random passwords: Scheme 4 was a much sparser subset of a larger, 6x6 grid. Note that while we designed schemes 2 through 4 by selecting/removing points at random, we did this randomization once: all users used the exact same fixed grids in Figure 1.



(a) Scheme 1: 5×4 grid

(b) Scheme 2: 5 × 4 grid with cued start ⊕ and end ⊗ points

(c) Scheme 3: 5×4 grid with holes

(d) Scheme 4: 6× 6 sparse grid

Fig. 1. Gaze-based graphical grid password schemes in our study

To enter passwords, users gaze at the first point in the password, press the space bar to tell the system to begin recording, gaze at each subsequent point for at least 0.5 seconds, then press the space bar again to stop recording. Note in particular that users do not have to press the space bar at each point, just gaze at it for at least 0.5 seconds. No visual feedback is displayed to the user while entering their password — no indication of points gazed or even when a gaze is registered; the only visual feedback comes after they press the space bar to stop recording, which results in a dialog box indicating successful or failed entry. Subsequent points have no restriction for adjacency; the same point cannot be gazed at twice in a row, though can be later used again.

The entry grid was displayed on a 19" monitor running at a resolution of 1920 × 1080 pixels. Gaze points were displayed as circles of radius 65 pixels with a 11 × 11 pixel 'cross' (+) displayed in the centre of the circle to help users focus on a target. The user did not need to gaze directly at the circle: we took the closest circle to their gaze fixation.

3 Experiment Design

We conducted a within-subjects lab study. Participants were approached through personal contacts and received no compensation; the study was approved by the university's ethics board.

A standard Windows 7 desktop PC with a 19" monitor was equipped with a Mirametrix S2 Eye Tracker, placed just below the monitor. The device has a data rate of 60 Hz with infrared binocular tracking. The accuracy range is 0.5° to 1° and the drift range is less than 0.3°. Our gaze-based password scheme was a custom-written C# program.

Each participant was assigned to use three of the four schemes: all participants used Schemes 1 and 2, and were randomly assigned to either Scheme 3 or 4. First, participants were introduced to the system and ran the eye-tracker's 9-point calibration routine. We told users to gaze at points for at least 1 second, even though the system would register a gaze after just 0.5 seconds. For each of the three schemes assigned (1, then 2, then either 3 or 4) participants were directed to (a) *create* a new password "of at least 6 points that would be easy for [them] to remember but hard for others to guess"; (b) *confirm* the password; (c) answer three short survey questions[1]; and (d) *login* using the password. After doing this for the three assigned schemes, the participant did (e) a *final login* using the password from Scheme 1. During confirmation and login sub-tasks, participants could keep trying until successful, skip the task, or restart the task (recreate and reconfirm a new password).

The login (d) after sub-task (c) and the final login at the end were designed to test recall after a passage of time. The login (d) after distraction task (c) typically occurred approximately 1 minute after completing steps (a)–(b). This is similar to the 30-second distraction task of Forget et al. [12].

The final login (e) in Scheme 1 typically occurred approximately 10 minutes after completing steps (a)–(d) for Scheme 1. In fact, in our study we also emailed participants two days after their participation, asking them to reply with the scheme 1 password, but not enough participants responded for us to report results.

It should be noted that, by having all subjects proceed sequentially through the tasks, a potential learning effect is introduced in which users find the later schemes easier to use: thus *usability* results may not be fully comparable between schemes. However, studies have found that *security* behaviour can change if the user has been "primed" for security (for example, Whalen and Inkpen [27] found

[1] Survey questions in Appendix A. User password dataset and Java code for metrics available at `http://eprints.qut.edu.au/58524/`

that no users looked for web browser security indicators before being asked to do so). In our context, this means that a user who sees scheme 2 before scheme 1 may choose different start/end points in scheme 1 than had she seen scheme 1 before scheme 2. To compare password quality consistently across schemes and to avoid priming subjects to choose more random or asymmetric passwords in scheme 1, we used a fixed sequence of tasks. This tradeoff between learning effects in usability or in password security seems to be inherent to any study where subjects use multiple variants.

Participants were randomly assigned to either scheme 3 or scheme 4 before they arrived; time constraints prevented us from having participants use both schemes.

Some of our survey questions, regarding security and computer expertise, are a subset of the survey questions of Arianezhad et al. [2].

4 Results

We had 25 participants total, though the eye-tracking equipment only recorded results for 22 of them due to astigmatisms. Participants ranged in age from 19–41 with an average age of 26.1. Most had a high degree of computer expertise; only 1 reported using Android pattern lock.

To help the reader understand what types of passwords are entered by users, we include in Appendix C the points for the passwords entered by our users in Scheme 1.

4.1 Security

Since passwords in our scheme are user-generated, not randomly generated, it is not appropriate to assume that all possible passwords are equally likely. Table 1 reports several measures of password randomness; cells in sections (b)–(d) of the table are of the form $^{a}/_{b}$, where a is the value of the metric for passwords our users created and b is the value for passwords generated uniformly at random, computed either algebraically (for (b)) or on a sample of 100000 passwords of length 7 generated uniformly at random (for (c) and (d)).

Password Length. Users were directed to create a new password "of at least 6 points that would be easy for [them] to remember but hard for others to guess". As reported in Table 1, the average length of passwords in all schemes around $7^{1}/_{3}$ characters. Note that in Scheme 2, users seemed to interpret this instruction for length of at least 6 as including the cued start and end points, hence in the table we report only the number of user-selected—and hence secret—points.

Point Frequency. For all four schemes, the frequency of points selected by users is quite close to random: for example, in Scheme 1, the entropy of user-selected points is 4.11 bits, compared to the maximum 4.32 bits for random points.

Table 1. Security metrics for user-generated passwords versus uniformly random passwords

	Scheme 1 Grid	**Scheme 2** Grid with cued start/end	**Scheme 3** Grid with holes	**Scheme 4** Sparse grid
(a) User-generated password length				
Mean* (SD)	7.59 (2.42)	5.36 (2.01)	7.33 (2.69)	7.23 (1.64)
(b) Binary entropy of points				
All	4.11/4.32	3.87/4.17	3.75/4.00	3.95/4.00
First†	2.18/4.32	2.54/4.25	2.50/4.00	2.78/4.00
Last†	3.54/4.32	2.63/4.25	2.50/4.00	2.14/4.00
(c) Binary entropy of stroke direction & length‡				
	3.47/5.65	3.05/5.54	3.20/5.64	3.73/6.33
(d) Symmetry score‡ (higher = more symmetry)				
Vertical	0.71/0.58	0.70/0.55	0.66/0.57	0.48/0.47
Horizontal	0.66/0.57	0.68/0.59	0.63/0.56	0.43/0.46
(e) Search estimate for 7-point passwords				
Theoretical	$2^{30.2}$	$2^{29.4}$	$2^{28.0}$	$2^{28.0}$
Point entropy	$2^{28.8}$	$2^{27.1}$	$2^{26.3}$	$2^{27.7}$
First+strokes	$2^{23.0}$	$2^{20.8}$	$2^{21.7}$	$2^{25.2}$

* For Scheme 2: excluding cued start/end points.

† First & last *user-selected* points. Thus, for Scheme 2: second & second-last.

‡ Values for uniformly random passwords calculated from 100000 uniformly randomly generated samples of length 7.

However, first and last user-selected points are not very random. In Scheme 1, the entropy of user-selected first points was just 2.18 out of 4.32 bits; in fact 50% of user-generated passwords started in the top-left corner. Scheme 2 was no better: the second and second-last points (i.e., the first and last *user-selected* points) were clustered around the cued points and had low entropy (2.54 and 2.63 out of 4.25 bits). Frequency tables for all, first, and last points of all schemes are given in Appendix B.

Strokes. We next consider the distribution of "strokes", meaning the direction and length between subsequent points. For example, a password where the first point was $(1,1)$ and the second point was $(2,3)$ corresponds to the stroke $(1\downarrow,2\rightarrow)$. We observed that the entropy of strokes in user-generated passwords is quite poor, in all cases between 55% and 62% of the entropy of strokes in randomly generated passwords. Frequency tables for stroke distribution for all schemes are given in Appendix B.

Symmetry. We observed that many users entered passwords that looked to be quite symmetric. For example, consider the fifth password (second row, second column) in Appendix C that was entered by participant #5 entered in Scheme 1.

We devised a metric to measure the symmetry present in a graphical grid password based in part on symmetry analyses of free-form drawmetric schemes [16,18]. The vertical (respectively, horizontal) symmetry score is computed as follows: for each possible vertical (horizontal) axis (axes exist either in between or along columns (rows) of points), fold along the axis, count the number of password points that match on both sides of the fold, and divide by the total number of password points; the vertical (horizontal) symmetry score is the maximum over all possible axes. (Note that both previous works on symmetry analyses of drawmetric schemes include at least some off-centre axes [16] or maximize over all possible axes [18].)

For example, for the firth password in Appendix C, the vertical symmetry score is 1.0, since by folding along the optimal vertical axis (through the third column), we have perfect overlap, whereas the horizontal symmetry score is $7/8 = 0.875$, since by folding along the optimal horizontal axis (through the second row), we have 7 of 8 points overlapping.

Note that although schemes 2–4 are somewhat asymmetric by design, the symmetry score does not become obsolete. Rather, the question becomes: are user-generated passwords more or less symmetric that randomly generated passwords in the same scheme?

As seen in Table 1(d) Schemes 1, 2, and 3 had higher vertical and horizontal symmetry scores than randomly generated passwords, suggesting that Schemes 2 and 3 did not introduce much "asymmetry". However, user-generated passwords in Scheme 4 were as asymmetric as random passwords.

Password Space Estimate. We used the above password metrics to estimate an upper-bound on the amount of work to search for a 7-point password using three different strategies: the *theoretical* search space (computed as $7 \cdot$ (ideal entropy of all points)); based on the *point entropy* of user-generated passwords ($7 \cdot$ (user entropy of all points)); and *first+strokes*, based on the entropy of the first point and subsequent strokes of user-generated passwords ((user entropy of 1^{st} point) $+ 6 \cdot$ (user entropy of strokes)). For Schemes 1–3, these techniques show decreases in search space of at least 7 bits compared to the theoretical space.

Limitations. The symmetry measures we employ do not address rotational symmetry or reflection on non-vertical/horizontal axes. While such symmetries are natural [16] for free-form drawmetric schemes, it is not clear how to correctly define them for grid schemes.

4.2 Usability

Table 2 (following reporting techniques of Forget et al. [12]) reports a wide variety of metrics on the usability of our 4 schemes and compares with 3 other gaze-based schemes. We use non-parametric tests due to small sample size.

Successes and Errors. As described in Table 2, the mean number of password creation operations per participant in Scheme 1 was (mostly) significantly smaller than Schemes 2 (Wilcoxon signed-rank $V = 3$, $p = 0.037$), 3 ($V = 0$, $p = 0.098$), and 4 ($V = 0$, $p = 0.034$). Scheme 1 also required fewer tries for confirmation, but the difference was significant only versus Scheme 4 ($V = 0$, $p = 0.021$).

For number of tries for successful login after the short distraction task (3 survey questions, ∼45 seconds), Schemes 1, 2, and 3 all performed well, and better than Scheme 4, though the difference was not statistically significant. However, the success rate for final logins to Scheme 1, which participants did at the end of the study after doing the Scheme 2 and Scheme 3 or 4 tasks (∼10 minutes later), was quite poor ($\leq$ 3 tries: 55%). This suggests users may forget grid passwords quickly, may become confounded when working with several grid passwords, or did not have enough repetition to ensure memorability. Our recall rate is not far off that of EyePassShapes [8], though theirs was after a much longer period (5 days vs. ∼10 minutes). A Spearman rank correlation test observed no statistically significant correlation between password length and number of confirmation or login errors.

Times. Table 2 reports for creation, confirmation, and login. Note that we report two different types of times:

- *total* time required for creation, confirmation, or login, which includes time elapsed during errors and re-tries, but does not include eye-tracker calibration
- time per point for *successful* creation, confirmation, or login, which includes only the time elapsed during the entry that actually succeeded, and is averaged on a per point basis.

We report both times to allow meaningful comparison with other schemes, some of which (Cued Gaze Points (CGP) T-51) reported total time and some of which (EyePassShapes, EyePassword) reported successful time. Note CGP T-51 [12] times also include time for a 1-point calibration and keyboard-based username entry; all other times do not include calibration or username entry. At the start of the study, we used our device manufacturer's 9-point calibration, which requires ∼20 seconds.

Times required for Schemes 1, 2, and 3 were fairly similar, whereas Scheme 4 had higher creation and confirmation times. Due to high standard deviation, only a few of the differences in means were statistically significant: creation time, Scheme 1 vs. 4 (Wilcoxon signed-rank $V = 16$, $p = 0.043$); confirmation time, Scheme 1 vs. 2 ($V = 63$, $p = 0.041$) and 1 vs. 4 ($V = 6$, $p = 0.006$).

Since we allowed users to choose the length of their password, we separately report times for just the successful creation/confirmation/login operations, averaged over the number of points in the password. Mean time per point is relatively consistent across all schemes and tasks: on average, users require 1.77 seconds per spot. Hence, an experienced user who makes no errors should be able to login with a 7-point password in around 12 seconds or less.

Table 2. Usability metrics of our schemes and other gaze-based schemes

	Scheme 1 Grid	**Scheme 2** Grid with cued start/end	**Scheme 3** Grid with holes	**Scheme 4** Sparse grid	**CGP T-51** Cued-recall [12]	**EyePassShapes** Grid with adjacent movements [8]	**EyePassword** On-screen keyboard (Dwell-QWERTY) [15]
# of participants	22	22	9	13	25	24	18
# of trials	22	22	9	13	169	–	–
Mean creates per participant	1.09	1.55	1.89	2.77	–	–	
Successful confirms on 1st try	91%	64%	67%	38%	67%	–	–
Successful confirms on $\leq$ 3 tries	95%	91%	78%	69%	82%	–	–
Successful logins on 1st try	73%	91%	89%	54%	73%	86%	97%
Successful logins on $\leq$ 3 tries	91%	95%	100%	77%	93%	–	–
Successful final logins on 1st try	45%*	–	–	–	–	57%*	–
Successful final logins on $\leq$ 3 tries	55%*	–	–	–	–	–	–
Mean confirm errors per trial	0.18	0.68	0.89	1.54	1.21	–	–
Mean login errors per trial	0.55	0.45	0.22	0.64	0.51	–	–
Mean (SD) total create time	14.6 (6.4)	19.2 (12.4)	24.2 (18.5)	38.9 (41.1)	44.2 (22.0)†	–	–
Mean (SD) total confirm time	17.9 (22.5)	21.6 (16.5)	26.3 (16.5)	36.4 (42.3)	47.1 (78.5)†	–	–
Mean (SD) total login time	21.4 (21.9)	18.0 (20.2)	19.5 (24.0)	17.4 (12.2)	36.7 (35.9)†	–	–
Mean (SD) succ. create time/point	1.78 (0.50)	1.76 (0.59)	1.94 (0.87)	1.75 (0.58)	–	–	–
Mean (SD) succ. confirm time/point	1.70 (0.51)	1.80 (0.62)	2.04 (0.98)	1.52 (0.65)	–	–	–
Mean (SD) succ. login time/point	1.68 (0.66)	1.60 (0.65)	1.98 (1.29)	1.26 (0.93)	–	1.56	1.08
Ease of use (4-point Likert scale: very easy-easy-hard-very hard)	5-14-3-0	3-17-2-0	1-4-3-1	3-4-5-1	See [12] Fig. 2	mid-scale (2.67/5)	–

* Scheme 1 final login performed 10 minutes after initial use; EyePassShapes final logins performed 5 days after initial study.

† CGP T-51 times included username entry and calibration time; other results do not.

Our times are generally comparable with other schemes. In particular, per-point time during successful logins (ranging from 1.26–1.98 seconds per point) is on par with that for EyePassShapes (1.56 seconds), although a bit higher than EyePassword (1.08 seconds per point).

User Perception. For each scheme, participants rated “how difficult it was to complete the task” on a 4-point Likert scale (very easy, easy, hard, very hard). Nearly all participants rated Schemes 1 and 2 easy or very easy (slightly lower than Cued Gaze-Points [12]; higher than EyePassShapes [8]), but only about half did for Schemes 3 and 4.

5 Conclusion

We have studied the usability and security of various recall-based graphical grid password schemes when used with gaze-based user interfaces. Though it can be difficult to precisely compare usability results across studies, in general our success rates and entry times are comparable with existing gaze-based cued-recall schemes. We give the first thorough treatment of the quality of passwords generated by users in graphical grid password schemes.

Assessing the strength of user-generated passwords on a variety of metrics is essential. User-generated graphical passwords may perform well on some metrics but poorly on others. Thus, for user-generated passwords, a simple password-space calculation in which all potential passwords are considered equally likely is overly optimistic. We have proposed the first metrics for assessing randomness of grid password schemes, which can be applied to all grid schemes, including for example Android pattern lock. In all of our schemes, the distribution of the first and last user- points was quite poor. The distribution of strokes between subsequent points in a password was also quite poor. Our attempt in Scheme 4 at increasing asymmetry in user-generated passwords worked, but at the cost of significantly longer creation and confirmation time and significantly lower confirmation and login success rates. Of the four schemes we proposed, the basic grid scheme, Scheme 1, seems to provide the best ease-of-use (high success rates, small time), with password distribution quality comparable to the other schemes.

Larger-scale real-world studies testing gaze-based graphical grid password scheme would provide insight into several open questions, such as the usability of gaze-based authentication in a non-laboratory setting, generalization to other user populations, suitability for users with disabilities, long-term recall rates, whether use of multiple grid passwords has a confounding effect, and the relative security of human-generated grid passwords in settings with more realistic risks.

Acknowledgements. The authors gratefully acknowledge helpful discussions with Margot Brereton and Daniel Johnson.

References

1. Proc. 30th International Conference on Human Factors in Computing Systems, CHI 2012. ACM (2012)
2. Arianezhad, M., Camp, L.J., Kelley, T., Stebila, D.: Comparative eye tracking of experts and novices in web single sign-on. In: Proc. 3rd ACM Conference on Data and Application Security and Privacy (CODASPY 2013) (to appear, 2013), http://www.douglas.stebila.ca/research/papers/acks13/
3. Aviv, A.J., Gibson, K., Mossop, E., Blaze, M., Smith, J.M.: Smudge attacks on smartphone touch screens. In: USENIX WOOT 2010 (2010), https://www.usenix.org/conference/woot10/smudge-attacks-smartphone-touch-screens
4. Biddle, R., Chiasson, S., van Oorschot, P.C.: Graphical passwords: Learning from the first twelve years. ACM Computing Surveys 44(4), 19:1–19:41 (2012)
5. Bond, M.: Comments on Gridsure authentication (March 2008), http://www.cl.cam.ac.uk/~mkb23/research/GridsureComments.pdf
6. Bulling, A., Alt, F., Schmidt, A.: Increasing the security of gaze-based cued-recall graphical passwords using saliency masks. In: Proc. 30th International Conference on Human Factors in Computing Systems (CHI 2012) [1], pp. 3011–3020 (2012)
7. Davis, D., Monrose, F., Reiter, M.K.: On user choice in graphical password schemes. In: Proc. 13th USENIX Security Symposium, pp. 151–164 (2004), http://static.usenix.org/event/sec04/tech/davis.html
8. De Luca, A., Denzel, M., Hussmann, H.: Look into my eyes!: can you guess my password? In: Proc. 5th Symposium on Usable Privacy and Security (SOUPS 2009), pp. 7:1–7:12. ACM (2009)
9. De Luca, A., Hang, A., Brudy, F., Lindner, C., Hussmann, H.: Touch me once and I know it's you!: implicit authentication based on touch screen patterns. In: Proc. 30th International Conference on Human Factors in Computing Systems (CHI 2012) [1], pp. 987–996 (2012)
10. De Luca, A., Weiss, R., Drewes, H.: Evaluation of eye-gaze interaction methods for security enhanced PIN-entry. In: Proc. 19th Australasian Conf. on Computer-Human Interaction (OZCHI 2007), pp. 199–202. ACM (2007)
11. Dunphy, P., Heiner, A.P., Asokan, N.: A closer look at recognition-based graphical passwords on mobile devices. In: Proc. 6th Symposium on Usable Privacy and Security (SOUPS 2010), pp. 3:1–3:12. ACM (2010)
12. Forget, A., Chiasson, S., Biddle, R.: Shoulder-surfing resistance with eye-gaze entry in cued-recall graphical passwords. In: Proc. 28th International Conference on Human Factors in Computing Systems (CHI 2010), pp. 1107–1110. ACM (2010)
13. Hayashi, E., Hong, J., Christin, N.: Security through a different kind of obscurity: evaluating distortion in graphical authentication schemes. In: Proc. of the 29th International Conference on Human Factors in Computing Systems (CHI 2011), pp. 2055–2064. ACM (2011)
14. Jermyn, I., Mayer, A., Monrose, F., Reiter, M.K., Rubin, A.D.: The design and analysis of graphical passwords. In: Proc. 8th USENIX Security Symposium (1999), http://static.usenix.org/events/sec99/full_papers/jermyn/jermyn.pdf
15. Kumar, M., Garfinkel, T., Boneh, D., Winograd, T.: Reducing shoulder-surfing by using gaze-based password entry. In: Proc. 3rd Symposium on Usable Privacy and Security (SOUPS 2007), pp. 13–19. ACM Press (2007)
16. Nali, D., Thorpe, J.: Analyzing user choice in graphical passwords. Technical Report TR-04-01, School of Computer Science, Carleton University (May 2004), http://www.cs.carleton.ca/research/tech_reports/2004/TR-04-01.pdf

17. van Oorschot, P.C., Salehi-Abari, A., Thorpe, J.: Purely automated attacks on passpoints-style graphical passwords. IEEE Transactions on Information Forensics and Security 5(3), 393–405 (2010)
18. van Oorschot, P.C., Thorpe, J.: On predictive models and user-drawn graphical passwords. ACM Transactions on Information and System Security 10(4), 5:1–5:33 (2008)
19. van Oorschot, P.C., Thorpe, J.: Exploiting predictability in click-based graphical passwords. Journal of Computer Security 19(4), 669–702 (2011)
20. Passfaces: The science behind Passfaces (September 2001), `http://www.passfaces.com/enterprise/resources/white_papers.htm`
21. Sahami Shirazi, A., Moghadam, P., Ketabdar, H., Schmidt, A.: Assessing the vulnerability of magnetic gestural authentication to video-based shoulder surfing attacks. In: Proc. 30th International Conference on Human Factors in Computing Systems (CHI 2012) [1], pp. 2045–2048 (2012)
22. Tafasa: PatternLock, `http://www.tafasa.com/patternlock.html`
23. Tao, H.: Pass-Go, a new graphical password scheme. Master's thesis, University of Ottawa (2006), `http://site.uottawa.ca/~cadams/papers/HaiTaoThesis.pdf`
24. Weidenbeck, S., Waters, J., Birget, J.C., Brodskiy, A., Memon, N.: Authentication using graphical passwords: Basic results. In: Proc. Human-Computer Interaction International (HCII 2005) (July 2005), `http://clam.rutgers.edu/~birget/grPssw/susan3.pdf`
25. Weidenbeck, S., Waters, J., Birget, J.C., Brodskiy, A., Memon, N.: Authentication using graphical passwords: effects of tolerance and image choice. In: Cranor, L.F., Zurko, M.E. (eds.) Proc. Symposium on Usable Privacy and Security (SOUPS 2005), pp. 1–12. ACM (2005)
26. Weiss, R., De Luca, A.: Passshapes: utilizing stroke based authentication to increase password memorability. In: Proceedings of the 5th Nordic Conference on Human-Computer Interaction (NordiCHI 2008), pp. 383–392. ACM (2008)
27. Whalen, T., Inkpen, K.M.: Gathering evidence: use of visual security cues in web browsers. In: Inkpen, K.M., van de Panne, M. (eds.) Proceedings of Graphics Interface 2005, vol. 112, pp. 137–144. Canadian Human-Computer Communications Society (2005), `http://portal.acm.org/citation.cfm?id=1089532`
28. Wright, N., Patrick, A.S., Biddle, R.: Do you see your password?: applying recognition to textual passwords. In: Cranor, L.F. (ed.) Proc. 8th Symposium on Usable Privacy and Security (SOUPS 2012), pp. 8:1–8:14. ACM (2012)
29. Zakaria, N.H., Griffiths, D., Brostoff, S., Yan, J.: Shoulder surfing defence for recall-based graphical passwords. In: Cranor, L.F. (ed.) Proc. 7th Symposium on Usable Privacy and Security (SOUPS 2011), pp. 6:1–6:12. ACM (2011)

A Survey

[In the survey questions reproduced below, we use ______ to indicate that the question allowed a free-form answer, ○ to indicate that a single choice could be made, and □ to indicate that multiple choices could be made. Participants completed questions 1–4 during the distraction during Scheme 1, questions 5–8 during the distractions during Scheme 2, questions 9–12 during the distractions during Scheme 3 or 4, and questions 13–16 at the end of the study.]

You can skip any questions you prefer not to answer.

1. What is your participant number? ______
2. What is your age? ______
3. What is your gender?
   ○ Male ○ Female ○ Prefer not to say
4. What is the highest level of education you have completed?
   ○ Some high school
   ○ High school diploma
   ○ TAFE diploma[2]
   ○ Some university education
   ○ Bachelor's degree
   ○ Master's degree
   ○ Doctoral degree
   ○ Other
5. Are you currently a student?
   ○ Yes ○ No
   If yes, what are your year and major? ______
6. Are you currently employed?
   ○ Yes ○ No
   If yes, what is your occupation? ______
7. Do you use a computer daily for work?
   ○ Yes ○ No
8. Do you have a degree in OR are currently studying toward a degree in an IT-related field (e.g., information technology, computer science, electrical engineering, etc.)?
   ○ Yes ○ No
9. Have you ever (select all that apply)
   □ Designed a website
   □ Registered a domain name
   □ Used SSH
   □ Configured a firewall
   □ Created a database
   □ Installed a computer program
   □ Written a computer program
   □ None of the above
10. Have you ever taken or taught a course on computer security?
    ○ Yes ○ No
11. Please check all of the following statements that describe your password habits.
    □ I use the same password for every website.
    □ I have a few passwords that I use interchangeably.
    □ I have one password that I use for important sites and another password I use for less important sites.
    □ I use different passwords for each site.
    □ I use my web browser's password manager to store my passwords.
    □ I write my passwords down on a piece of paper.
    □ I use a separate program to store my passwords.

[2] *[In Australia, TAFE stands for Technical and Further Education, and such institutions typically offer vocational tertiary education courses.]*

12. Please specify the brand and model of your mobile phone. ______
13. If you have an iPhone, which of the following options best describes your passcode lock habits?

    □ I have set a numerical passcode to lock/unlock my iPhone
    □ I have installed a third-party application to simulate Android grid lock screen on my iPhone
    □ I have no lock screen setting on my iPhone
    □ I don't have an iPhone

14. If your mobile supports Android, which of the following options best describes your lock screen habits?

    □ I have set a numerical passcode to lock/unlock my mobile
    □ I use grid lock screen on my mobile phone
    □ I have no lock screen setting on my mobile phone
    □ I don't use an Android mobile phone

15. Please rate each task in the study based on how difficult it was to complete the task (1=very easy, 2=easy, 3=hard, 4=very hard).
    (a) T1: Creating password in a grid.
    (b) T2: Creating password in a grid with start and end points.
    (c) T3: Creating password in a grid with holes, if you did this task.
    (d) T4: Creating password in an asymmetric screen, if you did this task.
16. After completing these tasks, would you use this password scheme if your computer was equipped with an eye-tracking device?

    ○ Yes ○ No

    Why or why not? ______

B Frequency Tables

B.1 Scheme 1: Grid

Scheme 1: All points

0.0719	0.0479	0.0838	0.0599	0.0599
0.0539	0.0778	0.0958	0.0599	0.0479
0.0060	0.0479	0.0838	0.0539	0.0299
0.0060	0.0240	0.0419	0.0299	0.0180

Scheme 1: First point

0.5000	0.1818	0.0455	0.0000	0.0455
0.0909	0.0455	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0909	0.0000	0.0000

Scheme 1: Last point

0.0455	0.0000	0.0000	0.1364	0.1364
0.0455	0.0455	0.0000	0.0455	0.0455
0.0000	0.0000	0.0909	0.0909	0.0000
0.0000	0.0455	0.0909	0.0455	0.1364

Scheme 1: Stroke frequency

	4 ←	3 ←	2 ←	1 ←		1 →	2 →	3 →	4 →
3 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2 ↑	0.0000	0.0000	0.0000	0.0000	0.0207	0.0000	0.0069	0.0069	0.0000
1 ↑	0.0000	0.0000	0.0069	0.0207	0.0690	0.0345	0.0000	0.0069	0.0000
	0.0000	0.0000	0.0138	0.1310	0.0000	0.2276	0.0069	0.0000	0.0000
1 ↓	0.0000	0.0138	0.0276	0.0276	0.2414	0.0621	0.0138	0.0138	0.0000
2 ↓	0.0000	0.0000	0.0000	0.0069	0.0069	0.0138	0.0069	0.0000	0.0000
3 ↓	0.0000	0.0000	0.0000	0.0000	0.0138	0.0000	0.0000	0.0000	0.0000

B.2 Scheme 2: Grid with cued start/end

Scheme 2: All points

0.0085	0.0932	0.0508	0.0593	0.0169
0.0000	0.1017	0.1186	0.1017	0.0424
0.0085	0.1017	0.1102	0.0254	0.0424
0.0085	0.0254	0.0254	0.0254	0.0339

Scheme 2: Second point

0.0455	0.2273	0.0000	0.0000	0.0455
0.0000	0.4091	0.0455	0.0000	0.0455
0.0455	0.0909	0.0000	0.0000	0.0000
0.0000	0.0455	0.0000	0.0000	0.0000

Scheme 2: Second last point

0.0000	0.0000	0.0455	0.0000	0.0000
0.0000	0.0000	0.0455	0.3636	0.0000
0.0000	0.0000	0.2273	0.0455	0.0455
0.0000	0.0455	0.0455	0.1364	0.0000

Scheme 2: Stroke frequency

	4 ←	3 ←	2 ←	1 ←		1 →	2 →	3 →	4 →
3 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1 ↑	0.0000	0.0071	0.0071	0.0143	0.0357	0.0286	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0071	0.0929	0.0000	0.2357	0.0000	0.0214	0.0000
1 ↓	0.0000	0.0214	0.0071	0.0643	0.3357	0.0643	0.0286	0.0000	0.0000
2 ↓	0.0000	0.0000	0.0000	0.0071	0.0000	0.0000	0.0000	0.0000	0.0000
3 ↓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

B.3 Scheme 3: Grid with holes

Scheme 3: All points

0.0455	0.0303	0.0758	0.0303	0.0455
0.0455		0.0152	0.0455	0.0303
0.1061	0.1515	0.1061		0.0303
	0.1212	0.0758		0.0455

Scheme 3: First point

0.2222	0.1111	0.0000	0.0000	0.0000
0.1111		0.0000	0.0000	0.0000
0.2222	0.0000	0.0000		0.0000
	0.1111	0.0000		0.2222

Scheme 3: Last point

0.0000	0.0000	0.1111	0.0000	0.0000
0.0000		0.0000	0.0000	0.2222
0.0000	0.2222	0.1111		0.0000
	0.2222	0.1111		0.0000

Scheme 3: Stroke frequency

	4 ←	3 ←	2 ←	1 ←		1 →	2 →	3 →	4 →
3 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0175	0.0000
2 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1 ↑	0.0000	0.0000	0.0000	0.0877	0.0877	0.0526	0.0175	0.0000	0.0000
	0.0000	0.0000	0.0000	0.1579	0.0000	0.1754	0.0175	0.0000	0.0000
1 ↓	0.0000	0.0000	0.0175	0.0000	0.2632	0.0175	0.0000	0.0000	0.0000
2 ↓	0.0000	0.0000	0.0000	0.0351	0.0000	0.0000	0.0000	0.0000	0.0000
3 ↓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

B.4 Scheme 4: Sparse grid

Scheme 4: All points

	0.0532		0.1064		0.0000
0.0532		0.1064		0.0000	
	0.0745			0.0957	0.0000
0.0426			0.0851		
		0.0213			0.0213
	0.0000		0.0000	0.0426	

Scheme 4: First point

	0.0769		0.0000		0.0000
0.3077		0.0769		0.0000	
	0.0000			0.0000	0.0000
0.0000			0.0000		
		0.0000			0.0000
	0.0000		0.0000	0.0000	

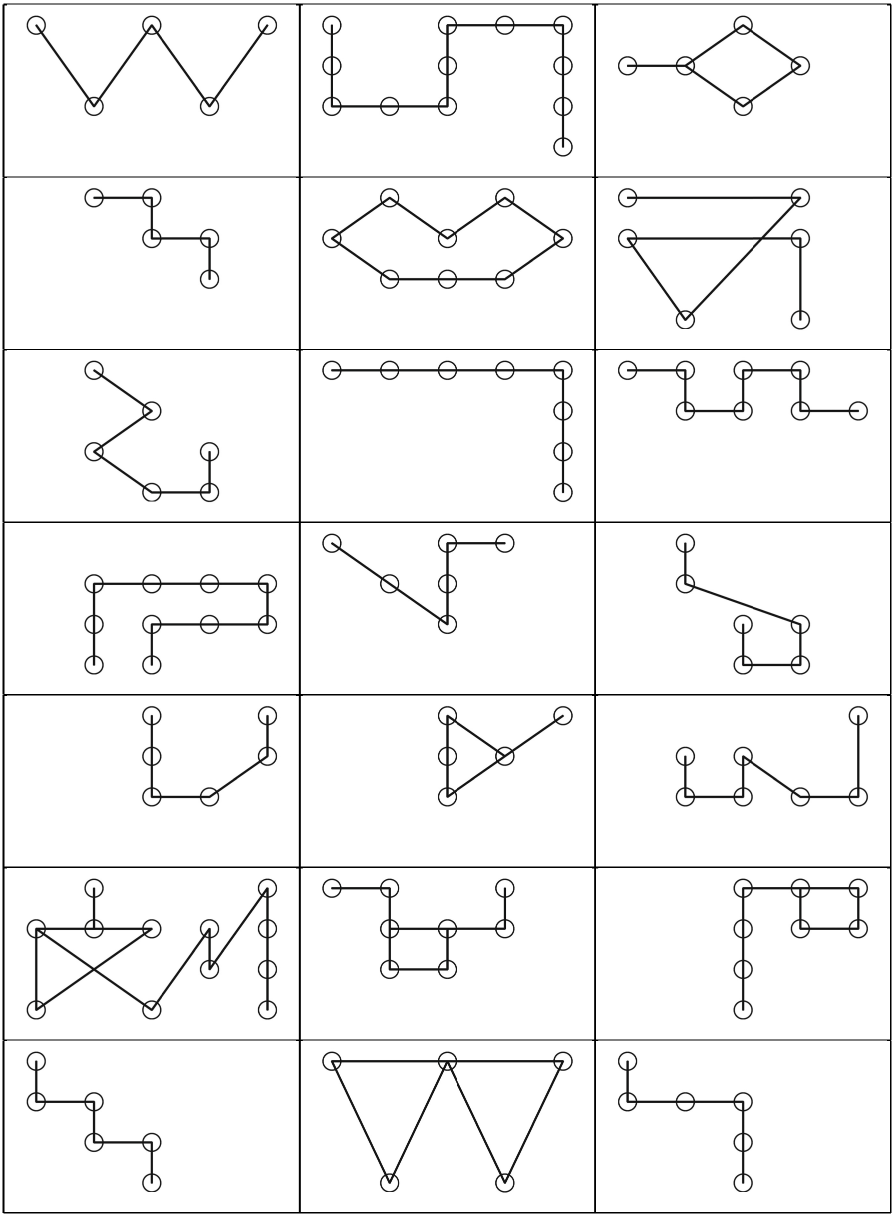
Scheme 4: Last point

	0.0000		0.0769		0.0000
0.0000		0.0000		0.0000	
	0.0000			0.0000	0.0000
0.0769			0.0769		
		0.0000			0.0769
	0.0000		0.0000	0.2308	

Scheme 4: Stroke frequency

	5 ←	4 ←	3 ←	2 ←	1 ←		1 →	2 →	3 →	4 →	5 →
5 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0123
2 ↑	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1 ↑	0.0000	0.0000	0.0000	0.0370	0.0370	0.0000	0.0864	0.0000	0.0000	0.0000	0.0000
	0.0000	0.0000	0.0123	0.0370	0.0247	0.0000	0.0617	0.0370	0.0123	0.0247	0.0000
1 ↓	0.0123	0.0000	0.0123	0.0123	0.2840	0.0370	0.0741	0.1111	0.0000	0.0000	0.0000
2 ↓	0.0000	0.0000	0.0000	0.0247	0.0000	0.0123	0.0247	0.0000	0.0000	0.0000	0.0000
3 ↓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4 ↓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5 ↓	0.0000	0.0123	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

C Sample User-Generated Passwords — Scheme 1



The Impact of Length and Mathematical Operators on the Usability and Security of System-Assigned One-Time PINs

Patrick Gage Kelley*, Saranga Komanduri, Michelle L. Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor

University of New Mexico and Carnegie Mellon University

Abstract. Over the last decade, several proposals have been made to replace the common personal identification number, or PIN, with often-complicated but theoretically more secure systems. We present a case study of one such system, a specific implementation of system-assigned one-time PINs called PassGrids. We apply various modifications to the basic scheme, allowing us to review usability vs. security trade-offs as a function of the complexity of the authentication scheme. Our results show that most variations of this one-time PIN system are more enjoyable and no more difficult than PINs, although accuracy suffers for the more complicated variants. Some variants increase resilience against observation attacks, but the number of users who write down or otherwise store their password increases with the complexity of the scheme. Our results shed light on the extent to which users are able and willing to tolerate complications to authentication schemes, and provides useful insights for designers of new password schemes.

1 Introduction

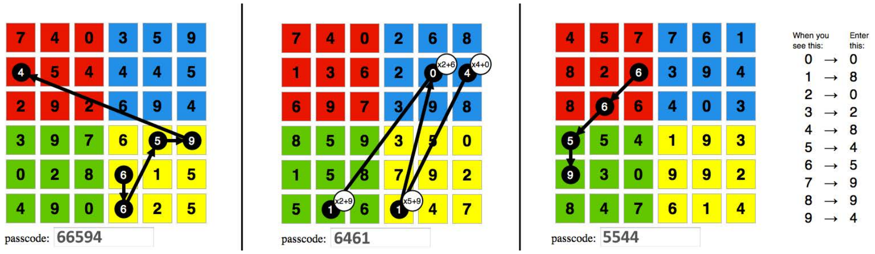
Personal identification numbers (PINs), or short numeric passwords, are commonly used at automated teller machines and to restrict entry into secure physical spaces. Both scenarios are potentially vulnerable to observation attacks, in which an attacker observes a user entering her password in order to learn about it. Attackers may be physically present and witness the password by looking over a person's shoulder (shoulder surfing) [21], or through recording devices (e.g., keyloggers or cameras) [17].

One solution to the problem of shoulder surfing is a *one-time PIN*, which is valid for only a single authentication. An attacker who observes a one-time PIN cannot replay it to gain access. Large numbers of one-time PINs may be computed in advance and shared between the system and the user. However, the user must have the next PIN on the list with her every time she wishes to authenticate, which requires carrying the list or having the PINs delivered on demand.[1] Alternatively, the system may display a challenge to the user and prompt her to use a shared secret to compute a response that

* Corresponding author.

[1] `http://gmailblog.blogspot.com/2011/02/advanced-sign-in-security-for-your.html`

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 34–51, 2013.



7	4	0	3	5	9
4	5	4	4	4	5
2	9	2	6	9	4
3	9	7	6	5	9
0	2	8	6	1	5
4	9	0	6	2	5

passcode: 66594

7	4	0	2	6	8
1	3	6	2	0	4
6	9	7	3	9	8
8	5	9	3	5	0
1	5	8	7	9	2
5	1	6	1	4	7

passcode: 6461

4	5	7	7	6	1
8	2	6	3	9	4
8	6	6	4	0	3
5	5	4	1	9	3
9	3	0	9	9	2
8	4	7	6	1	4

passcode: 5544

When you see this:		Enter this:
0	→	0
1	→	8
2	→	0
3	→	2
4	→	8
5	→	4
6	→	5
7	→	9
8	→	9
9	→	4

Fig. 1. The first image shows PGlength5, a standard length-five PassGrid. The second shows PGx+4, a length-four PassGrid with four different multiplication-addition rules. The third shows PGcodecard, a length-four PassGrid where each element must be translated with a single function, always shown to the right of the grid (The first grid element is a 6, so a user would enter a 5).

demonstrates that she knows the secret. Such challenge-response systems typically involve cryptographic functions that require the use of a computational device. However, it may be more convenient for a user if she can compute a response in her head.

Some graphical-password schemes allow the user to derive the correct one-time response from a combination of the screen display and their knowledge of the secret [7, 26]. These schemes offer some convenience, but only modest advantages over PINs in terms of resistance to observation attacks. These schemes can often gain observation attack resistance by requiring the user to remember a longer secret or perform simple mathematical operations. However, the tradeoffs between usability and security that such schemes may present have not been studied previously.

In this paper we explore the usability and security benefits of enhancing system-assigned one-time PIN systems with longer secrets or mathematical operators. We present a usability case study in which we analyze *PassGrids*, an implementation of a one-time PIN authentication mechanism in which users memorize a secret pattern on a 6x6 grid. Each time a user attempts to authenticate, she is presented with a grid filled with random digits, and she enters the digits that correspond with the elements of her pattern. While our study is limited to a specific authentication system, our approach allows us to examine security and usability tradeoffs that are generally applicable to a range of system-assigned one-time PIN systems as well as other authentication systems. Rather than testing the PassGrid scheme per se, we primarily use it to assess the relative security gains and usability impacts associated with adding various complications to its base design.

Surprisingly, we find that neither increased length nor mathematical operators greatly impacts usability. Although added complexity generally reduces user enjoyment, it does so far less than could be expected. However, added complexity does increase the tendency of users to write down or otherwise store their passwords. Using mathematical operators provides larger security gains than lengthening the pattern, while achieving similar usability. We also find that users are able to perform basic modular arithmetic operations as part of the authentication process, but dislike having to remember and perform multiple operations. More generally, our results shed some light on the extent to which users are willing and able to tolerate complications to authentication schemes, which in turn could be useful to designers of new schemes.

The remainder of this paper is organized as follows: In Section 2 we review related work and present the PassGrids case study in detail. Section 3 explains our methodology and reports on our study participants. In Section 4 we present our security analysis and main usability results. Section 5 concludes with a discussion of the implications of our findings.

2 Background and Related Work

In this section we review related work on graphical one-time PINs and graphical passwords, and introduce PassGrids, the scheme whose variants we focus on in this paper.

Graphical One-Time PINs. Graphical one-time PIN systems are a subset of one-time PIN systems, in which the authentication challenge is presented graphically. One example is the GrIDsure system,[2] studied by Brostoff et al. in an 83-participant user study [7]. GrIDsure is similar to PassGrids, with a five-by-five grid and user-selected patterns. Since users tend to select somewhat predictable patterns, this reduces the effective password space [22]. This erosion of practical entropy, along with other security issues related to graphical one-time PINs, is detailed by Bond [6].

Other examples of graphical one-time PINs are PassFaces and the commercial GridPIN system [8, 20]. In one variation of PassFaces, the user enters a one-time PIN calculated by locating previously selected pictures of faces within a grid. GridPIN displays a keypad in which each digit is surrounded by eight smaller digits; the user selects a direction (e.g., bottom left), and enters a one-time PIN calculated by locating her original PIN digits and then selecting the associated smaller numbers based on her direction.

We expand on previous studies of graphical one-time PINs by conducting a large online user study that examines a larger six-by-six cell grid while varying pattern length as well as the use of mathematical operators. Our findings provide insight into the tradeoffs between usability and resistance to observation attacks for this class of systems.

Graphical Password Schemes. For more than a decade, various graphical password schemes have been proposed to combat weaknesses of text passwords. Surveys by Biddle et al. [5] and earlier by Suo et al. [19] provide a comprehensive discussion of the breadth and history of graphical passwords. We focus here on a subset of graphical password schemes that Biddle et al. refer to as *recall-based systems*, in which users reproduce a secret.

Recall-based systems, or authentication through "what you know," are a general class that also includes text passwords. Many types of recall-based, single-factor authentication are subject to observation attacks (e.g. shoulder-surfing). When a user provides input to the authenticator, an attacker can observe the secret, effectively allowing the attacker to impersonate the user in the future. The quintessential example of this attack is during PIN entry at an ATM [2]. Sophisticated attacks may infer passwords from keypad acoustics or electromagnetic emanations from computer displays.

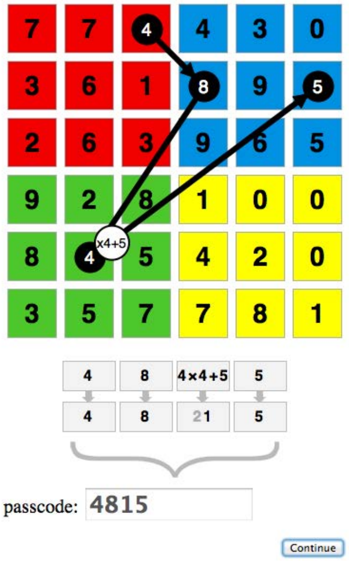
Many graphical recall-based systems require users to draw a sketch or pattern of their own creation, normally on a grid [5, 14]. A simple version of this is the current

[2] http://www.gridsure.com/

Android phone-unlock screen, where users trace a pattern of their own choosing on a three-by-three grid. The Android system is susceptible to shoulder-surfing attacks and to “smudge” analysis [4].

Weiss and DeLuca introduce a highly memorable graphical-password system using shapes [24]. Unfortunately, this scheme provides no more security against observation attacks than do traditional passwords [10]. This is not atypical, as many recall-based graphical passwords are vulnerable to single-observation attacks, and others are resistant to attacks after a small number of observations [12].

One example of a graphical password system designed to resist observation attacks is the convex hull click system, in which users must click a point within the convex hull created by locating the correct icons within a field of other icons [25]. This system has been shown to be vulnerable to repeated observation attacks based on the frequency with which the secret icons appear as compared to other icons [3]. To address this, previous work has looked at obscuring part of the challenge-response from an attacker [10, 18]. When successful, these systems are resistant to any number of observation attacks, but become vulnerable when an attacker can observe all parts of an authentication.



7	7	4	4	3	0
3	6	1	8	9	5
2	6	3	9	6	5
9	2	8	1	0	0
8	4	5	4	2	0
3	5	7	7	8	1

4	8	4x4+5	5
4	8	21	5

passcode: 4815

Fig. 2. An excerpt from the tutorial we developed to explain patterns with operators. Here, a participant assigned the PGx+1 condition must modular multiply the third digit in their pattern by the constant ‘4’ and then add ‘5.’ If the result is greater than 10, they should enter only the digit in the ones place, as shown in the example.

PassGrids. For this study, we implemented the user interface for a graphical one-time PIN system based on designs provided by PassRules US Security LLLP, creators of the “It’s Me!” graphical one-time PIN system. We call our implementation “PassGrids,” a name we made up for the study. We also created video-based tutorials and a JavaScript animation to teach study participants how to use PassGrids. PassGrids examples can be seen in Figures 1 and 2.

The PassGrids user interface contains a six-by-six grid and a password text-entry field. The grid displays the challenge: 36 colored squares, each containing a single randomly generated digit. Each quadrant of squares is a different color: red, blue, yellow, or green. The user’s secret is a *pattern*, formally an n-tuple, of locations on the grid, that users memorize. The grid shape and color are intended to aid the user in remembering the pattern and are the same for every user.

To authenticate, a user identifies the digits that correspond to the grid locations in her pattern and enters them in the text field using a keyboard or number pad. We will refer to the resulting one-time PIN entered by the user as a *passcode*. For example, a pattern of length five would require a user to memorize five locations in order, as shown on the left grid in Figure 1. On the left grid, the user would enter 66594 to authenticate.

The random digit generation is constrained so that each digit appears either three or four times in each six-by-six grid.[3] This means any user input matches multiple patterns in the grid. For example, in the center grid in Figure 1, "8440" matches 144 permutations of points. If grids were randomly generated without this constraint, some digits might appear only once in some grids greatly reducing the number of permutations.

In this study, we tested several variations on system-generated PassGrid passwords, selected to represent a range of resistance to observation attacks. These variations, which include varying the length of the pattern and requiring users to apply mathematical operators to elements of the pattern, are described in detail in Section 3.1.

3 Methodology

To test multiple variations of PassGrids and PINs, we conducted an online study using Amazon's Mechanical Turk service.[4] Mechanical Turk facilitates the recruitment of workers to complete short online tasks for small payments. Despite concerns about blindly relying on Mechanical Turk [1], several studies have found that properly-designed MTurk tasks provide high-quality user-study data, with much more diverse participants than are typically available in lab-based studies [9, 11, 13, 15, 23].

We conducted a two-part study with 1600 participants, using an experimental protocol similar to that used previously to study password-composition policies for text passwords [16].

In part one, we assigned each participant a four-digit numeric PIN or a PassGrids variant, described in the Conditions subsection below. Throughout this paper, and in all communications with participants, we refer to a participant's assigned PIN or PassGrids pattern as her password. We told users to imagine their password was assigned to them for use with their main email account after their previous password was compromised, and we asked them to behave as if this were their real password. While this hypothetical scenario may not produce the same results as a situation in which users actually use their passwords to protect high-value accounts, any behavioral bias introduced by this hypothetical scenario would likely impact all experimental conditions similarly, and thus the impact on our comparative analysis should be small.

3.1 Conditions

Participants were assigned to one of the following eight conditions, selected to represent a range of resistance to observation attacks.

- **PGbasic.** Participants were assigned a randomly generated four-element PassGrids pattern. In all conditions no location in the grid appeared more than once in a pattern.

[3] Six digits appear four times, four digits appear three times.

[4] http://mturk.amazon.com

- **PGlength5.** Participants were assigned a randomly generated five-element PassGrids pattern.
- **PG+1.** Participants were assigned a random length-4 PassGrids pattern. A randomly generated addition operator (for example, add 4) was applied to a randomly selected element of the pattern.
- **PG+4.** Participants were assigned a random length-4 PassGrids pattern. Separate randomly-generated addition operators (for example, add 4, add 1, add 4, and add 9) were applied to each of the four pattern elements.
- **PGx+1.** Participants were assigned a random length-4 PassGrids pattern. A randomly-generated multiplication/addition operator (for example: multiply by 3, then add 4) was applied to one of the pattern elements, also randomly selected.
- **PGx+4.** Participants were assigned a random length-4 PassGrids pattern. Separate randomly-generated multiplication/addition operators (for example: multiply by 3, then add 4; multiply by 2, then add 8; multiply by 3, then add 5; multiply by 5, then add 2) were applied to each of the four pattern elements.
- **PGcodecard.** Participants were assigned a random length-4 PassGrids pattern. Participants were told that a "swap" function would be applied to the numbers they typed in. For example, whenever they saw a 3, they must enter a 0. They were also told the entire translation would always be shown to the right of the grid, in a table we call a *codecard*. This condition roughly simulates providing each user with a paper codecard, which could be carried in her wallet and used for each login. The protocol design assumes a best-case scenario, one where the participant cannot lose the card or leave it at home. Note that if the codecard is also observed, this condition has security properties similar to PGbasic.
- **PIN.** Participants were assigned a randomly generated 4-digit PIN.

In each of the above conditions, the user of the system must memorize the pattern (the location of cells in the grid) as well as any operators (including type and quantity) that they must apply. The exception is PGcodecard which always displays the function box beside the grid.

In any condition involving mathematical operators, the math is modulo 10: once the result has been calculated, only the one's place digit is retained, so that the final passcode has the same number of digits that the pattern has cells. From left to right, Figure 1 illustrates conditions PGlength5, PGx+4, and PGcodecard. Figure 2 illustrates modular math for PGx+1.

3.2 Protocol Details

Participants in each condition were first shown an introductory video. The video welcomed them to the study and showed two examples of a basic password (a length-4 PassGrid or a length-4 PIN). Participants in the non-basic conditions were shown a third example, in the style they would be assigned, to demonstrate how the operator(s) were used or how the codecard function worked. The videos themselves ranged in length from 28 seconds for PIN to 117 seconds to PGx+4.

Immediately after the video, participants were assigned their password. For PassGrids participants, the password was animated on the screen, and math, if present, was detailed below the grid. A screenshot of this is shown in Figure 2. We then told participants they would need to successfully authenticate three times using the password they had just been assigned. After each attempt, we prompted them to enter the password again, until three successful authentications were achieved. After any three consecutive incorrect attempts, we displayed the password again. Throughout the process, a counter at the top of the screen reminded participants how many more authentications were needed. This set of authentications is considered the "practice" period.

After three successful authentications, participants were presented with 24 randomly selected, single-digit arithmetic problems: eight addition problems, eight multiplication problems, and eight hybrid multiplication/addition problems. These problems served as a distractor task between password entry attempts, as well as to measure the speed and accuracy with which participants could perform simple arithmetic problems like those used in some PassGrids variants. To mirror passcode entry, we instructed participants to perform modular arithmetic, saying: "For example, if you were to see the following addition problem: 4 + 9 = 13, Enter only '3' into the box."

Next, participants completed an online survey about demographics, password habits, and opinions of the password system used in the study.

Finally, participants were required to authenticate successfully one more time to complete the first part of the study. (Again, we displayed the password after three unsuccessful attempts.) We told participants we would contact them for follow-up surveys and displayed a completion code that they entered into Mechanical Turk to receive a 55-cent payment.

Two days after a participant completed part one, we sent an email asking her to return for part two of the study for a 70-cent bonus payment. URL customized for each participant. Participants who returned were asked to recall their passwords. Those who failed to recall their password after three tries were shown their password. Participants were then presented with a second survey, which included additional questions about password creation, storage, and usage.

3.3 Participants

Over a five-week period in August and September 2011, 4731 participants began our study. Of those, 3250 (68.7%) completed part one; the other 1481 (31.3%) are discussed in Section 4.6. Of the 3250 who completed day one, 2000 returned and successfully completed day two. From each condition, we selected the first 200 participants that successfully completed both days; unless stated otherwise, our analysis focuses on those 1600 participants.

The mean age of these participants was 30; 843 (52.7%) reported being male and 739 (46.2%) female. 449 (28.1%) reported studying or working in a technical field, and 195 (12.2%) in art or design. With Kruskal-Wallis and χ^2 tests, there were no significant differences between conditions for any of these characteristics.

4 Results

Across the PIN condition and PassGrids variations we tested, we first evaluate the security properties of each variation and then explore how successfully participants authenticated (accuracy), whether they memorized or stored their passwords (memorability), how they felt about the system (perception), the rate at which potential participants dropped out, and how much time was required to successfully log in. Our results show that all PassGrids conditions are more resistant to a single observation attack than PINs, but that with multiple observations PassGrids variants can also be compromised. We found that most variations of PassGrids are entered less accurately on first use by users than PINs, though users quickly comprehend how to authenticate with the system. Users report PassGrids to be a little bit more difficult but considerably more fun than PINs. We found that although users can generally authenticate surprisingly accurately even with arithmetic operations, adding such operations to PassGrids increases the rate of dropout from the study, decreases enjoyment, and greatly motivates people to write down information about their PassGrid password.

4.1 Security

We examined several security metrics for evaluating PassGrids: password space, passcode strength, and resistance to observation attacks. We focus most on observation attacks, and consider a particular threat model in which the attacker is given three chances to authenticate after n observations.

Password Space and Passcode Strength. One simple measure of security is *password space*, or the set of all possible passwords that could be assigned. In PassGrids, the password space can be increased by increasing the length of the pattern, increasing the size of the grid, or introducing mathematical operators. Conversely, the password space could be reduced by pruning patterns from the password space, for example patterns with points far apart from each other might be removed in an attempt to improve usability. Table 1 quantifies the password space for each of our conditions.

Table 1. Experimental conditions, shown in order of resistance to observation attacks, with number of possible passwords. Note that if the codecard is also observed, PGcodecard behaves similarly to PGbasic.

	possible	**% guessed after n observations**						
condition	passwords	1	2	3	4	5	6	7
PIN	1.0E+4	100						
PGbasic	1.4E+6	6.0	96.3	100				
PGlength5	4.5E+7	2.2	93.9	100				
PG+1	5.7E+7	2.7	64.5	99.2	100			
PG+4	1.4E+10	0.3	7.6	90.3	99.9	100		
PGx+1	5.7E+8	0.2	11.2	51.5	82.3	93.6	97.4	98.9
PGx+4	1.4E+14	1.2	3.32	18.6	67.9	91.2	97.0	99.0
PGcodecard	1.4E+16	0.4	0.8	1.0	1.4	2.1	5.0	21.6

Another security metric is *passcode strength*. With a randomly assigned PIN, all passcodes are equally likely. If an attacker with no knowledge of the user's password guesses a PIN at random, he has a 1 in 10,000 chance of gaining access. As a result, the password space and passcode strength are the same for randomly assigned PINs. With PassGrids, however, the probability of success from guessing a random passcode (with no knowledge of the user's password) can increase if the attacker analyzes the grid presented at login time. Since some digits are repeated more than others in the grid, some passcodes might be more likely to grant access than others. We can measure this effect by examining the distribution of passcodes produced by each PassGrid scheme.

Our analysis finds that the attacker's benefit from this kind of grid analysis is negligible. The weakest of the conditions we considered was PGbasic, in which an attacker has a 1.8 in 10,000 chance of gaining access with this kind of educated guess. Therefore, an attacker would still require a large number of guesses to gain access. Some of the PassGrids conditions we tested — PG+4, PGx+4, and PGcodecard — have the same passcode strength as a randomly assigned PIN. The other three PassGrids conditions have more passcode strength than PGbasic but less than a corresponding PIN. As a result, we don't consider passcode strength a very useful security metric for comparing these systems.

Estimating Observation Resistance. With traditional PINs, only a single observation is required for an attacker to learn the password, because a user's PIN is always the same. With PassGrids, the passcode is a function of the randomly generated grid and the user's password (pattern and operators), where each passcode maps to multiple unique passwords. Nevertheless, PassGrids are not immune to observation attacks.

With each observation, the attacker can reduce the space of possible passwords (ignoring degenerate cases where the same grid is observed multiple times). To provide an intuition of how this works, imagine that the attacker observes a victim in the PGbasic condition enter a passcode of "1234." If the digit "1" appears in four different cells in the grid, then the attacker knows that the first element of the victim's pattern must be in one of those four cells. If the attacker observes the victim again, he can eliminate any cells that don't correspond to the first digit the victim enters on the second observation. After a sufficient number of observations, the space can be reduced to a single password.

In our threat model, though, we allow the attacker to make three guesses before being locked out. Because passcodes map to multiple passwords, this gives the attacker more power than one might expect — for example, each incorrect guess can eliminate multiple passwords. After developing an algorithm which makes optimal guesses in this way, we used simulations to estimate the strength of PassGrids against observation attacks. This allowed us to quickly test many PassGrid variants. Our threat model assumes the attacker can see the complete grid and the victim's passcode in each observation, as might be available in an ATM skimming attack. We also assume the attacker knows the victim's password policy, i.e. the space of possible patterns and operators from which the password was assigned.

In each simulated observation attack, we randomly select a password from the password space S and generate n observations, i.e. random grids and corresponding passcodes. Our algorithm receives this data and removes any passwords from S which could never have produced the given data. The algorithm is then given the chance to authenticate with a new grid. It selects the most likely passcode to guess based on the remaining passwords in S. If this guess fails, the algorithm uses this failure to prune the space before guessing again on a new grid. The simulation is counted as successful if the algorithm's passcode is accepted within three guesses.

Observation Attack Results. Table 1 presents the results from 980,000 simulations run on the PassGrids conditions. For each condition and number of observations, we report the percentage of the 20,000 simulated attacks that were successful.

Overall, we see that all of the PassGrids conditions are better than PINs against a single observation, as might occur in an opportunistic shoulder-surfing attack. However, even PGcodecard is compromised after six observations. This is realistic if, for example, an attacker uses a hidden camera to record the same victim multiple times.

To compare between PassGrid conditions, we can select a threshold for success probability. For example, if we consider a condition compromised when 5% or more of the attacks succeed, then both PIN and PGbasic are compromised after a single observation, although PGbasic poses a greater challenge to the attacker. Conditions PGlength5, PG+1, PG+4, and PGx+1 are compromised after two observations, PGx+4 is compromised after three observations, and PGcodecard is not compromised until six. Choosing different cutoff points would result in different equivalence classes among the conditions.

A possible variation on PassGrids restricts the space of possible patterns by choosing only cells that are relatively close to one another, in an effort to increase usability. A similar approach is to allow users to select their patterns, which we expect them to do non-uniformly. This leads to a reduction in password space and increased vulnerability to observation attacks. To evaluate this, we simulated observation attacks on a variant of PGbasic in which patterns were selected such that the Euclidean distance between cells did not exceed a given threshold.[5] Our analysis indicates that this scheme provides little to no benefit over traditional PINs against observation attacks, which were successful more than 50% of the time after just one observation. We did not test this variant further.

It is important to note that these attacks require a relatively powerful attacker who can record both the grid and the victim's passcode, and calculate the optimal passcode to try on each attempt. Many realistic attackers, for example, an opportunistic shoulder-surfer, may be weaker.

4.2 Math Results

We analyzed participants' responses to the 24 random modular arithmetic problems to determine whether any of the operators we tested would be particularly problematic. Overall, participants completed these problems accurately, getting 96%, 94%, and 92% of addition, multiplication, and combined multiplication addition problems correct, respectively. Each problem type differs significantly in the proportion participants got right (Holm-corrected FET, $p < 0.05$). The mean number of incorrect problems per participant was 1.4; only 83 participants (5%) got nine or more problems wrong (two standard deviations above the mean).

[5] The threshold was set to $(3 * \sqrt{2}) = 4.24$. This number allows for a pattern that has all four points on a diagonal and reduces the total number of possible patterns to $10,060$.

The mean completion time was 5.0 seconds per addition problem, 5.6 seconds for multiplication, and 8.0 seconds for combined problems.[6] These results suggest that including simple modular arithmetic does not pose a significant barrier to authentication for our population (Mechanical Turk workers); it could be more problematic for others.

4.3 Accuracy

Participants authenticated with our system five times: three times on day one immediately after being assigned a PIN or pattern, once more on day one after the survey and math questions, and a fifth time when they returned for day two. We consider a participant to have successfully logged in if she can authenticate within three tries (before she is shown her password again). Many participants performed poorly in the first trial, as they first used the system, but by the end of the third trial most participants seemed to grasp password entry. Figure 6 (Appendix A) shows the percentage of participants who successfully logged in per condition, per trial.

Surprisingly, by the final authentication all conditions show similar accuracy, despite the large differences in success in the previous trials. This indicates that after some practice, users can authenticate just as successfully with the complicated conditions as with the simpler ones. In this final trial, no differences between conditions were found using Fisher's exact test at the $\alpha = 0.05$ significance level. (Seven pairs of conditions were selected a priori and tested without correction; all other pairs were Holm-corrected.)

We also examined the types of mistakes participants made, finding that many mistaken entries were very close to correct. Across the PassGrids conditions, 308 of 1400 participants (22%) made errors in the day two recall. In 7% of these cases, the participant entered a passcode with too many digits; 16% used too few. 65% percent of participants who made errors used only one wrong digit in their passcode; 33% of these got the last digit wrong. 13% of mistaken participants entered the right digits in the wrong order, and 6% entered a passcode that appears to result from transposing their pattern to an incorrect starting cell. Forty-three of 1000 participants in conditions with operators (4%) entered a passcode for the correct pattern with no operators. Note that individual participants may exhibit more than one type of error.

4.4 Storage and Memorability

We use storing behavior as a rough proxy for perceived memorability; that is, conditions in which participants wrote down their passwords more often can be considered harder to remember. During the survey at the end of day two, we asked, "Did you write down or store any information to help you remember your password pattern? (please be honest, you get paid regardless, this will help our research)." The results of this question are shown in Figure 3.

Unsurprisingly, patterns in conditions that are intuitively more difficult, specifically those where participants needed to memorize more information, were more frequently written down. PGx+4 (83%) and PG+4 (75%) were each stored significantly more often

[6] Measured from page load to submission of answer, which includes the time needed to load the page, navigate the mouse/cursor to the answer field, type the answer, and submit.

than all the other conditions.[7] PGcodecard, while much more resistant to observation attack, was not significantly different in storage frequency from a standard length-4 PassGrid (44% and 43% respectively). This is not surprising, because we always showed the codecard on the screen; in practice, the user would need to store it.

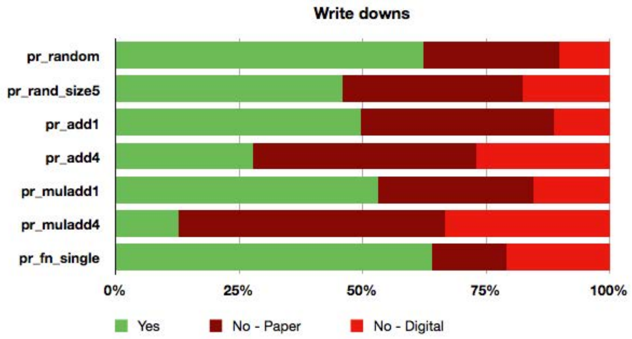


Fig. 3. Percentage of participants who stated they did or didn't store their password, by condition

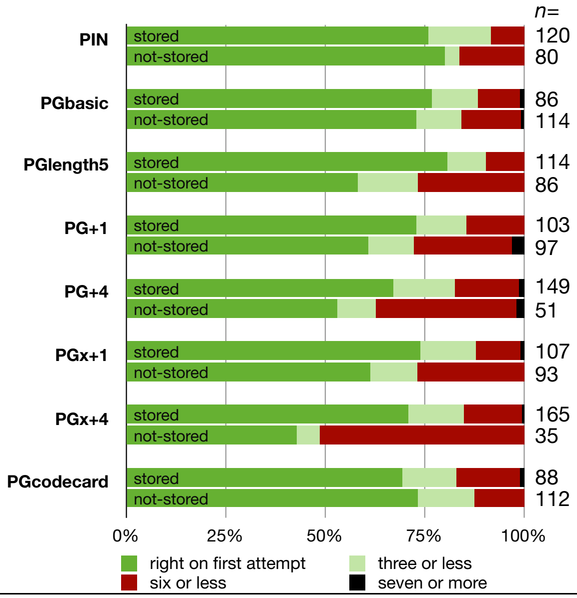


Fig. 4. For this analysis, we split the participants in each condition into two groups, those who stored their password (top) and those who did not (bottom), and compared their day 2 login success rates. Note sample sizes are different, and noted on the right. We a priori selected PIN, PGbasic, PG+4, and PGx+4 for significance testing, only the latter two showed significant differences (FET, $p < .006$).

Surprisingly, some PassGrids conditions, such as PGbasic, were stored significantly less frequently than PINs (43% and 60% respectively). However, that difference may be caused, at least in part, by the fact that writing down a pattern is less straightforward and familiar than writing down a PIN, rather than by differences in perceived memorability.

We also compared accuracy between participants who said they did not write down their password and those who said they did. We show the results in Figure 4. Across all conditions, 87% of participants who wrote down their passwords authenticated successfully on day two. The success rate for those who did not write down was only 76%, a significant difference (FET, $p < .001$). Within conditions, we saw no significant difference in accuracy between writers and

[7] All comparisons in this paragraph HFET, $\alpha = 0.05$.

non-writers for PIN and PGbasic. In PG+4 and PGx+4, writers were significantly more accurate than non-writers (FET, $p < .006$). (These conditions were selected a priori for significance testing.)

As a rough estimate of memorability, we also consider how many participants did not write down their password, yet authenticated successfully on day two. Just 33.5% of our 200 PIN participants didn't write down their password and still logged in; this is similar to PGx+1, at 34.0%. Excepting PGcodecard, where participants were told they did not need to memorize or store the translation, the condition that performed best was PGbasic, where 48% of participants successfully authenticated on day 2 without password storage. In the worst case, PGx+4, only 8.5% of participants were able to successfully log in without password storage, suggesting it may not be tractable by memory alone.

One further note is that some participants indicated they were not storing the entire PassGrid password, but only the operators they needed to apply.

4.5 User Perception

To explore user perception of PassGrids, we asked a series of Likert questions on day two (1-5 from strongly agree to strongly disagree). Three of these were: "Using PassGrids was ..." annoying, difficult, and fun. A fourth asked the participant if remembering their password was difficult. We graph the responses in Figure 7 in Appendix A.

Our results indicate that most PassGrids variations are more fun than PINs. Some are also more difficult, but in some cases additional complexity can be achieved without decreasing usability.[8] PGbasic and PGlength5 performed best, being rated significantly more fun than PIN but not significantly different in annoyance or difficulty. The other PassGrids conditions were significantly worse than PIN in annoyance and difficulty.

Comparing PassGrids conditions to each other, PGx+1 was not significantly different in user perception than PG+1, but users rated it significantly easier to remember and use than PG+4. PGx+4 was the worst overall.

4.6 Dropouts

While there are many reasons for participants to leave a study, it is likely the dropout rate can be abstracted as a rough metric for difficulty and user frustration. We examined the number of participants who accepted the task from Mechanical Turk, but then left the study before completing day one, and found that the rate at which users dropped out of our study increased roughly in line with the number of operators added in the PassGrids conditions. The results are shown in Figure 5.

The dropout rate among participants assigned to PassGrids was significantly higher than among PIN participants (33% and 16% respectively).[9] Within PassGrids conditions, those in PGlength5 were not significantly different from those in PGbasic, but were significantly less likely to drop out than those in PGcodecard. Similarly, PGx+1 was not significantly different from PG+4, but had significantly more dropouts than

[8] All comparisons in this paragraph HFET, $\alpha = 0.05$.

[9] All comparisons in this section HFET, $p < 0.001$.

PG+1. Lack of a significant difference does not mean that dropout rates were the same, but it does indicate that the size of any difference is small. On the other end of the spectrum, PGx+4 participants were nearly as likely to drop out as they were to complete day one (44.8%).

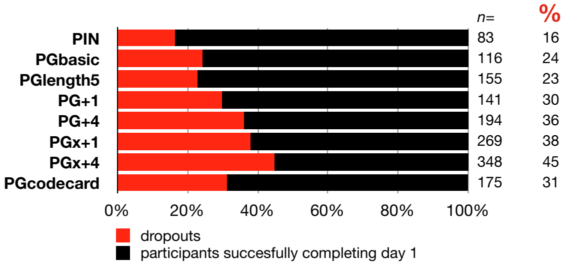


Fig. 5. Percentage of participants who upon accepting the MTurk task successfully completed day one. Ranges from 16.4% for PIN to 44.8% for PGx+4.

We did find that in conditions with the highest dropout rates (PG+4, PGx+1. PGx+4) accuracy on the math problems was highest (94-96%, compared to 91% in PIN). This may mean that those participants who dropped out were not as confident at mathematics. While we might expect demographic differences between conditions due to the differences in dropout rates, we did not see that.

4.7 Timing Information

Password authentication includes many subtasks, such as reading the web page, remembering the password, entering the password, and retrieving written notes. Identifying and timing these subtasks in an online study is infeasible. Here, we present instead two measures of timing: *entry* time and *login* time. Entry time estimates the amount of time spent entering the password and is taken from the first successful attempt of the third authentication. We used the third practice entry here, to attempt to reduce the impact of memory-based recall, focusing just on time to actually authenticate, as the participants had just done this twice prior. Login time encompasses an entire authentication, including unsuccessful attempts, and is taken from the final authentication. All times were measured from server-side events, which do not account for client-side delays like page loading. Therefore, these times might be overestimates. The timing data is shown in Table 2.

In entry time and login time, PIN is the clear winner, as expected. Among PassGrids conditions, the cost of additional complexity in authentication time is clearly illustrated. Median entry times range from 12 sec for PGbasic to 38.5 sec for PGx+4. Login times are almost three times longer, even though the mean number of attempts required on day two was 1.8. This suggests that authentication takes longer when users haven't used their password in several days.

4.8 Tutorial

A potential problem with comparing PassGrids to PINs is pre-existing user familiarity with PINs. People have seen and used PINs many times, but are likely completely unfamiliar with one-time graphical PIN systems in general, as well as with our specific implementation. To attempt to address this, we created a series of video tutorials, based on PassRules system specifications but with some modifications for our conditions. Since this was an online study, we have no way of knowing if a video was actually watched. Users may have covered the video with another application, muted the audio, or otherwise underutilized the tutorial.

We recorded the number of times participants returned to the video tutorial after being shown their password. Of the 1400 participants across the seven PassGrids conditions, 363 participants (26%) returned to watch the tutorial at least once, with most returning only once (299 participants), and one participant returning 5 times.

Table 2. Median login times in seconds per condition

	entry time (s)	login time (s)
PIN	6.0	20.0
PGbasic	12.0	35.0
PGlength5	15.0	51.0
PG+1	15.0	51.5
PG+4	23.0	74.5
PGcodecard	20.0	56.0
PGx+1	17.0	50.0
PGx+4	38.5	96.5

Despite the tutorials, we found there was still some confusion about PassGrids. In the free-response portion of the day 1 survey, many participants described the concept as both “interesting” and “new.” Some found it confusing, and it is unclear whether they understood that the passcode would be different each time. When asked how they remembered their pattern, participants gave responses such as, “just keep retrying the combination” or “I had a very hard time with remembering due to the fact that you changed the numbers around on the side and I had to put different numbers for each number.” From such comments, it seems that improving the tutorial so that more participants truly understand how PassGrids work could prove beneficial.

5 Discussion

Our results show that a system-assigned one-time PIN system such as PassGrids is a viable PIN replacement for systems where observation attack prevention is a priority. While not invulnerable against observation attacks, attackers must be technologically assisted, with complete knowledge of the grid and a non-trivial algorithm for determining passcodes with a high likelihood of success.

We found that several methods can be used to increase the security of PassGrids, including increased length, mathematical operators, and codecards. Modular arithmetic is not difficult for our participants when explained in straightforward terms.

While mathematical operators do provide additional security without suffering a loss in ability to authenticate, there are substantial usability drawbacks as the complexity increases. Participants considered our most difficult condition, PGx+4, substantially more difficult to remember and use. We also saw greatly increased rates of password storage, and we lost nearly half of our incoming participants in that condition, more than any other. All this together suggests math should be used in moderation, with as few constants as possible and a minimal number of pattern elements affected. Additionally, we must keep in mind that increased complexity here will lead to more password storage, which depending on the threat model may be more harmful than the benefits gained.

Increasing length did result in slight observation resistance gains, with little difference in accuracy, reported enjoyment, or timing, and only a slight increase in storage;

however, it is likely that length cannot continuously be increased without more substantial usability losses.

Finally, the codecard functionality allowed us to examine a much greater observation resistance, by increasing the space of the translation in a more diverse way than simple operators will ever allow. Overall, PGcodecard performed well, but the requirement of a written source that must be kept secret in order to achieve the observation resistance benefits may not in practice be usable. Additionally, this requirement may not align with the motivation behind one-time PIN systems, especially if the goal is to allow participants to log in, simply from memory, as with a standard PIN.

Applying similar types of modifications to other systems, such as the closely related GrIDsure system is straightforward; users could select longer patterns, or select operators in combination with their patterns. GridPIN could be extended in a similar way, after using the displayed keypad and selected direction to map the original PIN digit to a one-time digit, an additional mathematical operator or codecard could be applied. A similar modification to PassFaces might require users to find a starting number associated with the correct face, then modify that number using mathematical operators or a codecard.

We tested modifications on a basic one-time PIN system, increasing the length, adding mathematical operators or a digit translation, each designed to increase the observation resistance of the system. We measured how these modifications affected the usability of a system in various ways including memorability, storage, enjoyability, and accuracy. We believe that these modifications and the results we described here are not unique to this system, but give password system designers an understanding of how these techniques can enhance the security of other one-time PIN systems.

Acknowledgments. This research was supported by NSF grants DGE-0903659 and CNS-1116776, by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office, by Air Force Research Lab Award No. FA87501220139, and by gifts from PassRules US Security LLLP and Microsoft Research.

References

1. Adar, E.: Why i hate mechanical turk research (and workshops). In: Proc. CHI Workshop on Crowdsourcing and Human Computation (2011)
2. Anderson, R.: Why cryptosystems fail. In: ACM CCS 1993, pp. 215–227 (1993)
3. Asghar, H.J., Li, S., Pieprzyk, J., Wang, H.: Cryptanalysis of the convex hull click human identification protocol. In: Burmester, M., Tsudik, G., Magliveras, S., Ilić, I. (eds.) ISC 2010. LNCS, vol. 6531, pp. 24–30. Springer, Heidelberg (2011)
4. Aviv, A.J., Gibson, K., Mossop, E., Blaze, M., Smith, J.M.: Smudge attacks on smartphone touch screens. In: WOOT 2010, pp. 1–7 (2010)
5. Biddle, R., Chiasson, S., van Ookrschot, P.: Graphical passwords: Learning from the first twelve years. ACM Computing Surveys (2011) (to appear)
6. Bond, M.: Comments on gridsure authentication (2008), `http://www.cl.cam.ac.uk/~mkb23/`

7. Brostoff, S., Inglesant, P., Sasse, M.A.: Evaluating the usability and security of a graphical one-time PIN system. In: BCS Conference on HCI (2010)
8. Brostoff, S., Sasse, A.: Are passfaces more usable than passwords? a field trial investigation. In: HCI 2000, pp. 405–424 (2000)
9. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6(1), 3–5 (2011)
10. De Luca, A., Denzel, M., Hussmann, H.: Look into my eyes!: Can you guess my password? In: SOUPS 2009, pp. 1–12. ACM (2009)
11. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system? Screening Mechanical Turk workers. In: Proc. CHI (2010)
12. Golle, P., Wagner, D.: Cryptanalysis of a cognitive authentication scheme. In: IEEE SP 2007 (2007)
13. Jakobsson, M.: Experimenting on Mechanical Turk: 5 How Tos (July 2009), `http://blogs.parc.com/blog/2009/07/experimenting-on-mechanical-turk-5-how-tos/`
14. Jermyn, I., Mayer, A., Monrose, F., Reiter, M.K., Rubin, A.D.: The design and analysis of graphical passwords. In: USENIX Security Symposium, p. 1 (1999)
15. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with Mechanical Turk. In: Proc. CHI (2008)
16. 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 (2011)
17. Krebs, B.: ATM skimmers: Hacking the cash machine (2011), `http://krebsonsecurity.com/2011/04/atm-skimmers-hacking-the-cash-machine/`
18. Sasamoto, H., Christin, N., Hayashi, E.: Undercover: authentication usable in front of prying eyes. In: SIGCHI 2008, pp. 183–192. ACM (2008)
19. Suo, X., Zhu, Y., Owen, G.S.: Graphical passwords: A survey. In: ACSAC 2005, pp. 463–472 (2005)
20. SyferLock. Syferlock technology, `http://www.syferlock.com/day1/demovidpin.htm`
21. Tari, F., Ozok, A.A., Holden, S.H.: A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In: SOUPS 2006, pp. 56–66 (2006)
22. Thorpe, J., van Oorschot, P.C.: Human-seeded attacks and exploiting hot-spots in graphical passwords. In: USENIX Security Symposium, pp. 8:1–8:16 (2007)
23. Toomim, M., Kriplean, T., Pörtner, C., Landay, J.: Utility of human-computer interactions: toward a science of preference measurement. In: Proc. CHI (2011)
24. Weiss, R., De Luca, A.: Passshapes: Utilizing stroke based authentication to increase password memorability. In: 5th Nordic Conference on HCI (2008)
25. Wiedenbeck, S., Waters, J., Birget, J.-C., Brodskiy, A., Memon, N.: Authentication using graphical passwords: effects of tolerance and image choice. In: SOUPS 2005, pp. 1–12 (2005)
26. Wiedenbeck, S., Waters, J., Sobrado, L., Birget, J.-C.: Design and evaluation of a shoulder-surfing resistant graphical password scheme. In: AVI 2006, pp. 177–184 (2006)

A Additional User Study Results

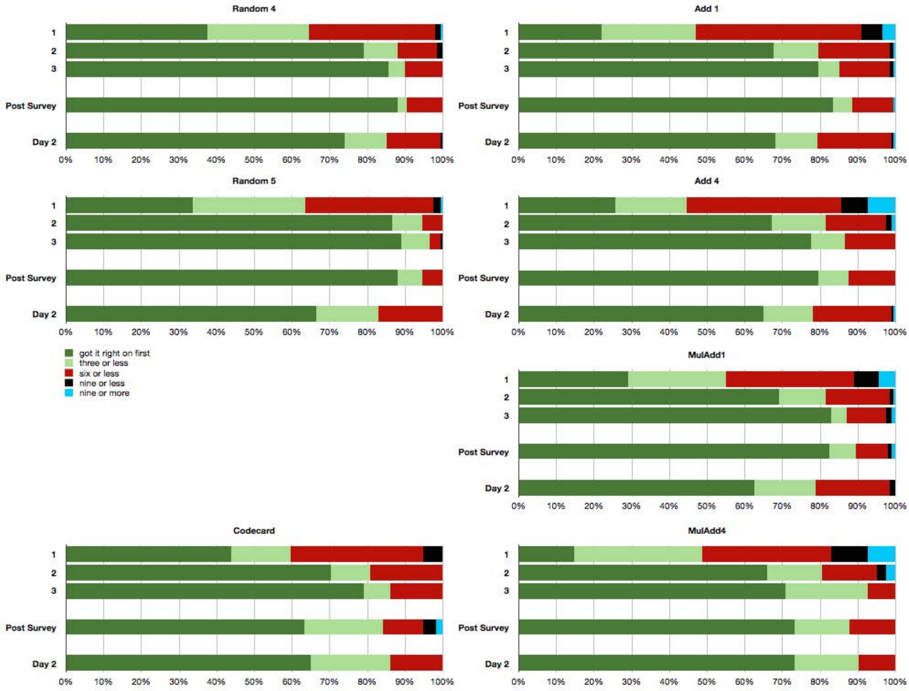


Fig. 6. Percentage of participants who logged in successfully (either on their first attempt or within three), across three practice authentications, day 1 recall, and day 2 entry, by condition

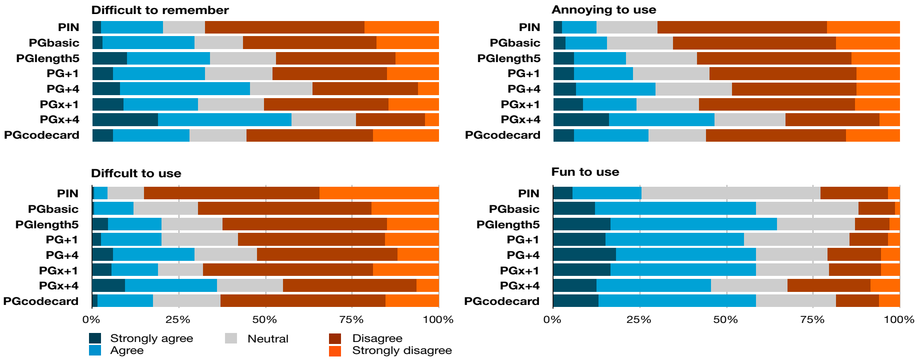


Fig. 7. Likert responses graphed by response, by condition. All participants answered four standard questions on day two about difficulty to learn, difficulty to use, annoyance, and fun to use, for each password system.

QRishing: The Susceptibility of Smartphone Users to QR Code Phishing Attacks

Timothy Vidas, Emmanuel Owusu, Shuai Wang, Cheng Zeng,
Lorrie Faith Cranor, and Nicolas Christin

Carnegie Mellon University
Pittsburgh, PA, USA
{tvidas,eowusu,shuaiwang,chengzeng,lorrie,nicolasc}@cmu.edu

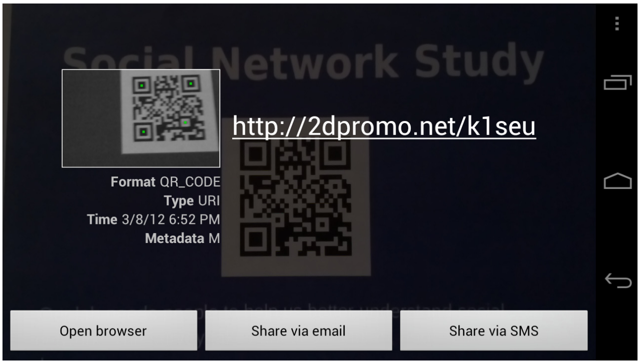
Abstract. The matrix barcodes known as Quick Response (QR) codes are rapidly becoming pervasive in urban environments around the world. QR codes are used to represent data, such as a web address, in a compact form that can be scanned readily and parsed by consumer mobile devices. They are popular with marketers because of their ease in deployment and use. However, this technology encourages mobile users to scan unauthenticated data from posters, billboards, stickers, and more, providing a new attack vector for miscreants. By positioning QR codes under false pretenses, attackers can entice users to scan the codes and subsequently visit malicious websites, install programs, or any other action the mobile device supports. We investigated the viability of QR-code-initiated phishing attacks, or QRishing, by conducting two experiments. In one experiment we visually monitored user interactions with QR codes; primarily to observe the proportion of users who scan a QR code but elect not to visit the associated website. In a second experiment, we distributed posters containing QR codes across 139 different locations to observe the broader application of QR codes for phishing. Over our four-week study, our disingenuous flyers were scanned by 225 individuals who subsequently visited the associated websites. Our survey results suggest that curiosity is the largest motivating factor for scanning QR codes. In our small surveillance experiment, we observed that 85% of those who scanned a QR code subsequently visited the associated URL.

Keywords: Phishing, Mobile, Security, QR Code, Smartphone.

1 Introduction

A Quick Response code (QR code) is a two-dimensional matrix of black and white pixels [20] that can be used to store information in a compact and optically-scannable form. QR codes have gained popularity due to their higher information density and improved readability compared to one-dimensional barcodes. As the number of smartphone users grows rapidly [8], businesses are turning to QR codes en masse to provide a fun and simple way to direct smartphone users to their websites and products. QR codes are designed to be readable regardless of orientation and in cases where a code is partially damaged or masked. These properties facilitate the use of QR codes in consumer applications to convey information to users.

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 52–69, 2013.



http://2dpromo.net/k1seu

Format QR_CODE
Type URI
Time 3/8/12 6:52 PM
Metadata M

Fig. 1. Screen capture of the most popular "barcode scanner" on Android: ZXing. With default settings, the URL is prominently shown to the user after scanning.

QR codes are typically "scanned" by photographing the QR code using a mobile device, such as a smartphone. The image is then interpreted by a QR code reader that users may install as an application on their mobile device. The reader decodes the message and performs an operation based on the message. For example, if the encoded data contains a link to a mobile application download, the reader may launch a marketplace application such as Google Play or Apple App Store. The content represented by a QR code is often a hyperlink, and the associated action is to launch the device's web browser and visit the website specified by the code.

QR codes can be found on store-front windows, magazines, newspapers, websites, posters, mass mailings, and billboards. Businesses display QR codes on advertisements to direct people to their websites. One study found over 14 million U.S. mobile users scanning QR codes during June 2011 [26].

The ease with which one can create and distribute QR codes has not only attracted businesses, but also scammers seeking to direct people to phishing websites. Phishing is a semantic attack that cons individuals, under the guise of a legitimate organization or individual, into visiting a malicious website or providing sensitive information [21]. With the increased usage of QR codes, QR code phishing, or QRishing (phonetically: "*krihsh*-ing"), presents a threat to this new, convenient technology. Concerns for the safety of QR codes are increasing [11, 24, 26, 28]. An attacker might place a sticker of a QR code containing malicious content over a legitimate QR code or create an entirely new QR code advertisement masquerading as a legitimate entity.

Some QR code reader applications may perform actions without first presenting the human-readable QR code content to the user. For example, an application may automatically open a hyperlink in the device's web browser without permitting the user to first verify the hyperlink. In this case, it is easier for attackers to deceive users into divulging private information or, even worse, installing malicious software on their phones. On the other hand, if the barcode application displays the URL to the user, an astute user may notice a suspicious-looking URL. However, use of "URL shorteners" can make it more difficult for users to evaluate a URL. Figure 1 depicts an application displaying the URL to the user and awaiting further action by the user.

To frame the scope of the problem, we tested the most popular reader applications from the Android and Apple marketplaces in January 2012. We downloaded and tested the top ten free applications for "barcode scanner" from Google Play and the Apple App Store. Thirty percent of these top ten free scanning applications in the Google Play Market and 50% in the Apple App Store immediately visit a scanned URL in the default configuration. Tables showing the particular applications and results can be found in the appendix.

The purpose of this study is to measure the threat QR codes pose as a phishing attack vector and to identify ways to improve the safety of QR code interaction. We are interested in the behaviors of smartphone users when they see QR codes posted in public places, including whether or not they look for context around the QR code, scan the QR code, and visit the website from the QR code.

We further motivate the problem with related work in Section 2. The user study consisted of two experiments: (1) A QRishing experiment and (2) a baseline surveillance study of user interaction with QR codes, which we describe in Section 3 and Section 4, respectively. Security implications of the study are presented in Section 5 and conclusions in Section 6.

2 Related Work

Phishing is a type of semantic attack where the malicious party attempts to gain sensitive information (e.g., account credentials or credit card numbers) by baiting victims with communications and content that appears to be from a legitimate party (e.g., a counterfeit password change website). Existing research has repeatedly shown that typical computer users have difficulty distinguishing between legitimate content and phishing content [14–16, 32]. QRishing is an extension of phishing that utilizes QR codes.

Downs et al. interviewed 20 non-expert computer users about their decision-making process when they encountered suspicious looking emails [16]. Their study suggests that simply being aware of Phishing-style scams is insufficient. Furthermore, their findings suggest that providing message-specific contextual cues (e.g., "this website is requesting a password") may be more effective than sender-specific cues, as scammers exploit the fact that many victims have a real account with the entity that is being faked.

The economic viability of *typosquatting* demonstrates the usefulness of misspelled and misleading domain names [22]. The only technological controls currently available to counter QRishing rely on the user to identify questionable URLs or involve cues from external tools, such as domain blacklisting services. Such cues have previously been shown to be misunderstood by users [17] who may not even understand the difference between positive and negative cues [14].

Dhamija et al. found that nearly a quarter of their 22 participants did not use browser-based cues (e.g., the address bar and status bar) leading to incorrect identification of fraudulent websites [15]. We find similar results, in the case where QR codes are used as a medium for phishing – specifically, 36% of our phishing participants indicated that they did not or could not recall checking the link.

The ease with which one can access web content via QR codes may induce more users to ignore browser-based cues as compared to entering or following a link. A typical QR code use case involves a scan and a click. Perceptive and security-conscious users may pause to examine the hyperlink but, in general, there is very little explicit interaction with the encoded data. The use of shortened URLs and the limited screen space of smartphones further obscures browser-based cues.

Similar to other phishing work, non-technical controls, such as increasing user education [27], may produce similar effects when applied to QRishing. A combination of automated detection systems along with user education may prove to be the best approach [21].

The security and usability research communities have explored various proposals for combating phishing attacks. Zhang et al. leveraged message-specific contextual cues for automated phishing detection in their implementation of *CANTINA* [32]. Dhamija et al. presented a defense against phishing attacks that made use of trusted paths to prevent window spoofing and independently computed images that allowed users to authenticate the remote party by visually verifying that the expected image was received [14].

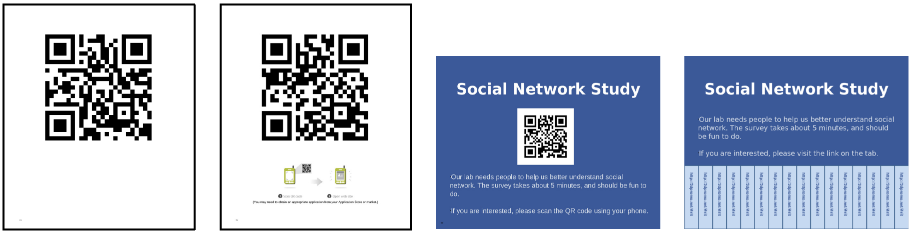
Phishing and the effects of malware are perhaps more threatening on mobile devices than on traditional computers. The management model, long patching cycle, limited screen space, and myriad of input types and sensors found on mobile devices make mobile oriented malware particularly distressing [19, 30].

Concurrent to our research, an Internet vigilante claimed to have conducted a QRishing attack by changing his Twitter icon to a QR code. This QR code represented a shortened URL that ultimately led victims to a webpage that reportedly hosted a WebKit browser exploit and secondary exploits for iOS and Android devices. This five-day attack is claimed to have garnered 1200 victims of which 500 successfully executed the secondary OS-specific payload [31].

In addition to the claimed QRishing attack via Twitter, industry researcher Eric Mikulas has recently presented work on QR code phishing attacks [29]. Following our study, Mikulas conducted a similar attack predominately using small stickers to place QR codes around Pittsburgh, PA. Similar to our study, these QR codes lead smartphone users to an informative website about QR code risks. After posting 80 stickers and receiving fewer website visits than expected, Mikulas said he plans to "attach them to fliers offering a false incentive or even place his stickers on top of existing advertisements and QR codes."

3 QRishing Experiment

The ease with which one can create and distribute QR codes may make them attractive to scammers seeking to direct people to phishing websites. The purpose of this study is to understand how users interact with QR codes in public spaces and to assess the susceptibility of smartphone users to QRishing attacks. In this experiment, we posted flyers around the city of Pittsburgh, PA. Each passerby who scanned one of the flyers was directed to a a brief online survey.


Social Network Study
Our lab needs people to help us better understand social network. The survey takes about 5 minutes, and should be fun to do.
If you are interested, please scan the QR code using your phone.
Social Network Study
Our lab needs people to help us better understand social network. The survey takes about 5 minutes, and should be fun to do.
If you are interested, please visit the link on the tab.

(a) *qrcode_only* (b) *qrcode_inst.* (c) *qrcode_SNS* (d) *ripoff_SNS*

Fig. 2. An example flyer for each of the four conditions deployed in the QRishing experiment. (a) shows *qrcode_only* –flyers with a QR code. (b) shows *qrcode_instructions* –flyers with a QR code and usage instructions. (c) shows *qrcode_SNS* –flyers advertising a mock SNS study with QR code. (d) shows *ripoff_SNS* –flyers advertising a mock SNS study with rip-off tabs.

3.1 Methodology

We posted flyers with QR codes both on the Carnegie Mellon University campus and in public locations around Pittsburgh (e.g., at bus stops, public bulletin boards at restaurants, coffee shops, etc). All flyers were posted in public locations where flyers are routinely placed. Each QR code on a poster represented a unique URL to our webserver, allowing us to unambiguously know in which location the participant observed our flyer. We used random, unique URLs similar to popular "URL shortening" services for each flyer. Such URLs are commonly used in QR code advertising. Further, the use of random URLs minimizes the risk that after scanning one flyer, curious participants could easily determine and visit URLs associated with other flyers. In the last week of January and first week of February 2012, we posted flyers at 139 different locations: 104 campus locations, 35 off-campus locations. Each flyer was checked weekly and, if needed, replaced. This experiment had four conditions (pictured in Figure 2):

- ***qrcode_only.*** A flyer with only a QR code.
- ***qrcode_instructions.*** In addition to the QR code graphic, includes instructions on how to use a QR code.
- ***qrcode_SNS.*** Innocuous flyer utilizing a QR code (a "social networking" user study advertisement).
- ***ripoff_SNS.*** A user study flyer similar to 3, but with traditional rip-off tabs instead of a QR code.

All conditions were randomly distributed across the locations and ran simultaneously for four weeks. When a person scanned the QR codes (or visited the URL on the rip-off tab), they were taken to our website where they were informed about the experiment and prompted to participate in an optional survey.

Conditions *qrcode_only* and *qrcode_instructions* did not have any advertised function, thus any participant in these conditions is likely to have scanned the QR code out of curiosity, compulsion, fun, etc. Conditions *qrcode_only*, *qrcode_instructions* and *qrcode_SNS* all involve the use of QR codes and thus provided insight into the frequency with which QR codes on flyers are scanned. Without

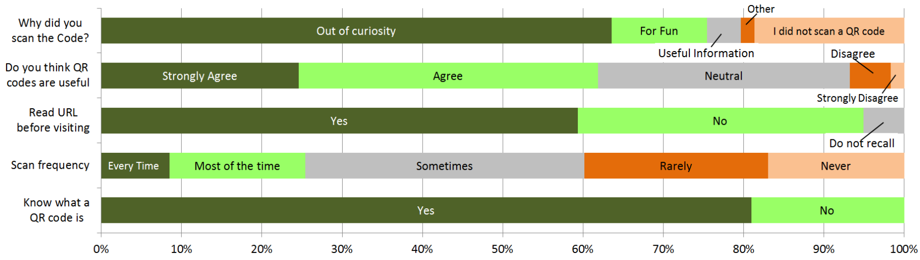


Fig. 3. Survey responses. Most participants scanned QR codes out of curiosity, agree than QR codes are useful, read the URL prior to visiting the website, and know the term "QR code."

a QR code, *ripoff_SNS* served as a performance baseline to compare with the other three conditions.

Regardless of condition, upon visiting the URL, participants were notified of the study via webpage and given the choice to follow a link to take an optional survey. We also recorded the access time, IP address, and user-agent from the server web log. Upon completion of the survey (or electing not to participate in the survey) the participant was automatically taken to a debrief webpage for the experiment. Participants who reported being under 18 years old were informed that their data would not be used in research and we discarded associated data.

3.2 Results

Of the 139 posted flyers, 85 (61%) were utilized by participants to visit the study website at least once, totaling 225 hits across all conditions. Examination of source address, access time and poster location (URL) indicated that only once did the same device scan a QR code twice. One hundred twenty-two participants (54%) completed the optional survey. Seventeen participants started, but did not complete the survey, and five participants self-reported to be under 18, and were removed from the study.

In the survey, participants were asked "Do you know what a QR code is?" The majority (83%) of survey takers responded "Yes," indicating some familiarity with the technology. Even 51% of participants in *ripoff_SNS*, which did not use a QR code, answered "Yes," further indicating that participants were aware of the technology. We posit that although some smartphone users may not know the term "QR code," the majority of users know the function of a QR code when presented with one.

We also asked participants about the primary reason they chose to scan the QR code (including an option for "I did not scan a QR code"). We observed far more participants scanning the QR code out curiosity than for related information. Figure 3 shows the distribution of survey responses from participants. More than 75% of the survey respondents scanned the flyer out of curiosity (64%) or for fun (14%). Less than 4% claim to have scanned the QR code because the related information seemed useful. Twenty percent of the respondents indicate that they did not scan a QR code, and all of these participants were in *ripoff_SNS*.

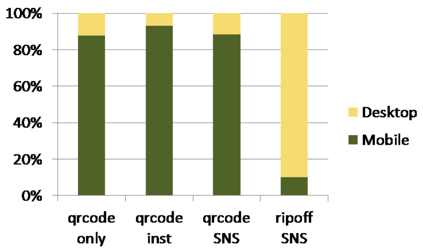


Fig. 4. Mobile vs desktop users by condition.

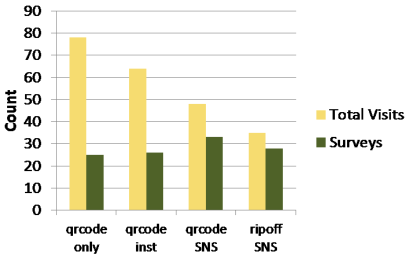


Fig. 5. Visited URLs and Survey Completion by condition

As expected, participants not using a mobile device were also predominately in the condition without a QR code, *ripoff_SNS*, though not exclusively (Figure 4).

Among the four conditions, *qrcode_only* had the most participants while *ripoff_SNS* had the least number of participants. Figure 5 shows the distribution of participants who both visited the URL and completed the survey across the four conditions. While curiosity was reported to be the main reason for initially scanning a QR code, participants were significantly more likely ($\chi^2 = 8.7344$, df = 1, p = 0.003) to complete the survey in conditions that explicitly advertised a study than those that had no advertised functionality.

Fifty-eight percent of respondents report reading the URL prior to visiting the link. While this behavior is likely safer than that of the 36% who did not read the URL, they still visited an obscure URL to an unrecognized domain (we registered the domain just prior to the study).

Across all four conditions we found that men were at least 2.5 times more likely to participate, especially in *qrcode_only* where we observed more than 7.6 times as many male participants. While we are uncertain of exactly how many individuals passed our flyers, nor how many of them possessed mobile devices, we can approximate percentages based on demographic data for the respective areas. For the on-campus flyers we can compare to CMU's general population [25], and for off-campus flyers we can compare to Pittsburgh census [12] data for the area we posted flyers. Further, we can use market penetration data [4] [5] to approximate percentages of smartphone owners. The gender distribution on-campus is 63% male, 37% female [25], and off-campus in the Pittsburgh area is 52% male, 48% female [12]. Among U.S. smartphone users, the gender is distributed 47% male, 53% female [5]. The incumbent population suggests that approximately 50–60% of our participants should be male, yet we observed 75%.

The only condition that fell within the expected gender ratio was *qrcode_instructions*. As shown in Figure 6, *qrcode_instructions* has fewer male respondents and more "Prefer not to answer" than the other three conditions. There is no way to tell the gender of those who selected "Prefer not to answer." While the result is not statistically significant, it is clear that in our experiment *qrcode_instructions* had more respondents who wished not to reveal their gender.

Our two most observed age ranges (on and off campus) were 18–24 and 25–34, together accounting for 78% of our participants. This closely aligns with the two age groups that have most adopted smartphones [4].

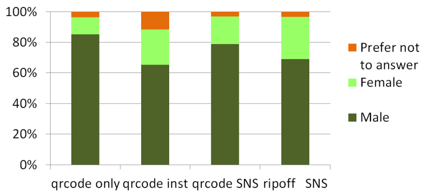


Fig. 6. Self-reported gender by condition. *qrcode_instructions* has fewer male and more "Prefer not to answer" participants than any other condition.

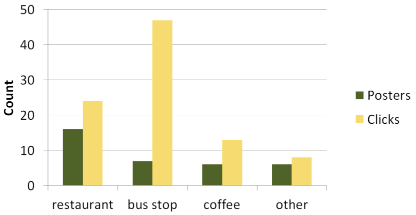


Fig. 7. Poster performance by location type. For each location type, green bars show the number of fliers posted, and yellow bars show accumulated clicks.

We found that flyers at bus stops far outperformed other locations. On average, flyers posted at bus stops solicited nearly seven URL visits per flyer. Figure 7 shows the distribution of off-campus flyers as grouped by bus stop, restaurant, coffee shop and other. Flyers posted at bus stops may receive more attention simply due to behavior at such a location. For example, those waiting for the bus to arrive are likely bored and are forced to wait idly at the location for a non-trivial duration.

We examined other metrics such as day-of-week, time-of-day, and user perception of QR code usefulness, all of which did not prove useful as a predictor of behavior. Additionally, we examined the networks from which devices were connecting, and the results were as expected in the U.S. Of the cellular network users 54% used Verizon Wireless, 31% AT&T, and 15% Sprint. The non-cellular users primarily (63%) used campus networks while the primary home Internet providers where Comcast (18%) and Verizon (6%).

3.3 Limitations

Unlike the envisioned attack scenario, we are bound by ethical, legal, and Institutional Review Board limitations in the presentation and placement of QR codes. A would-be attacker may have little consideration for vandalism, covering existing QR codes with his own, or any number of other less scrupulous activities.

Like many on-campus studies, our observed population for on-campus flyers is biased to the local population of CMU. Similarly, our off-campus flyer locations were subject to the respective populations in Pittsburgh and may not be representative of other areas.

In *qrcode_SNS* and *ripoff_SNS*, the "social networking user study," the flyers will have only attracted individuals interested in such a study. Other false pretenses could have been employed, such as a local band or work-from-home opportunities, but would have similarly limited the set of individuals attracted to the flyer.

In our experiments we used "shortened URLs," which have their own security implications [23]. It is possible that users may be more likely to follow a typical URL, but we felt that using a shortened URL exhibited more realistic conditions as shortened URLs are often used in QR code advertisements. The short property of shortened URLs also fits nicely with mobile devices as the limited screen space will cause many URLs to be truncated for display, resulting in the user only having the ability to observe part of the URL. Further, we wanted seemingly random URLs so that users could not easily predict the URL of a poster they had not physically encountered.

Particular to *ripoff_SNS*, participants may have been less likely to correctly type or may have had less desire to participate in the study due to the URL format. The URLs used in all conditions were similar to those found in popular URL shortening services (e.g., `http://bit.ly`, `http://goo.gl`). Such a random pattern (e.g., skx0r132) may be perceived differently by participants than a link consisting of a domain name and a common webpage naming convention (e.g., study.php).

4 Surveillance Experiment

Since QR codes are abundant in urban areas, we wanted to observe how people interact with them in public spaces. Specifically, we wanted to identify how many participants would scan the QR code and also visit the associated website versus the number of participants who would scan the QR code but elect not to visit the website. This observation provides insight about the potential for QR codes as a phishing attack vector because examining the URL is a practical and effective defense against many phishing threats. The use of QR codes minimize the person's effort in obtaining a URL; the person does not have to manually transcribe the URL from the source material. Such reduced interaction may encourage the unsafe behavior of visiting a questionable website without seeing the URL, sacrificing security in favor of usability.

This section describes the methodology, experimental design, and analysis of the surveillance experiment. We refer to this experiment as *surveillance_exp* in the subsequent text.

4.1 Methodology

We posted a flyer containing a QR code on a bulletin board at Carnegie Mellon University and placed it under video surveillance. By comparing captured video footage of people scanning the QR code with server logs, we were able to identify the number of participants who scanned the code as well as the number of participants who actually visited the URL encoded in the QR code. If a corresponding entry did not exist on the server we assumed that the participant scanned the QR code, but chose not to visit the website. The experiment had two conditions: an incentive condition and a no-incentive condition.

- ***surv_qrcode_only.*** In the no-incentive case, we collected two weeks of footage using a flyer containing only a QR code, similar to Figure 2(a).
- ***surv_incentive.*** Following the incentive case, we collected two additional weeks of footage using a flyer offering the chance to win a $50 Amazon gift card.

The sequence of events a participant followed in both conditions is as follows. First, a person who walks by the bulletin board noticed our flyer. They became a participant in our study when they entered the field of view of the camera and scanned the QR code. If the participant chose to visit the website (or the reader application automatically opened the link), they were presented with a simple web page that thanked the person for their interest in the study and asked them to take a survey. The person may have selected "continue" to further participate by taking an online survey, selected "cancel" to continue directly to the debrief material, or simply elected to close the browser. Similar to the *qrishing_exp* experiment, participants who reported that they were under the age of 18 received an additional debrief message stating that their data would not be used in the study and that they were not eligible to receive the incentive.

Every time a participant accessed our secure server, we recorded the time of access, the IP address, and user-agent in the web server log. The IP address was used to assess the connection type (e.g., campus Wi-Fi). Participants in *surv_incentive* were asked to provide their email address in order to be notified in the event they won the gift card. Providing an email address was at the sole discretion of participant. Furthermore, we ensured that a participant's email was not correlated with her survey responses.

4.2 Data Capture

We posted a flyer containing a QR code in the Gates Center for Computer Science on CMU's Pittsburgh campus. The flyer was posted on an announcements board located on the main floor of the building, an area which is open to the public and access-controlled only at night. A camera and netbook were mounted above the board to capture the activity of people around the poster. We checked the flyer daily to ensure it was unobstructed and that there were no other QR codes nearby. After some field trials at the site using both Android and iPhone smartphones, we concluded that a 3-by-3 inch QR code would best ensure the participant was within the field of view of the camera.

The netbook was configured to capture data only when motion was detected. The camera recorded four frames per second for as long as there was motion, and for 60 seconds thereafter. Each time a picture was captured, it was immediately processed with an edge-detection algorithm in order to minimize storing more data than required by the experiment and to protect the privacy of the participant. The experiment configuration is depicted in Figure 8.

An example of data captured from one participant's interaction is shown in Figure 9. This figure shows a participant approaching the flyer, photographing the flyer, and leaving the experiment area.

4.3 Removing False Positives

Due to the sensitivity of the software motion detection and the communal nature of the experiment site, the vast majority of collected images are not imminently useful. In many cases, passersby will briefly trigger data capture, people will

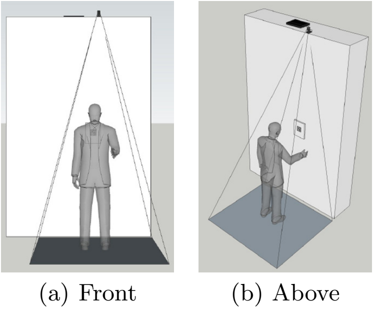

(a) Front (b) Above

Fig. 8. Surveillance experiment equipment configuration. The camera and netbook are mounted on the area above the announcements board. The box around the person represents the field of view of the camera. (a) shows the setup from the front, facing the board. (b) shows an isometric view from above.

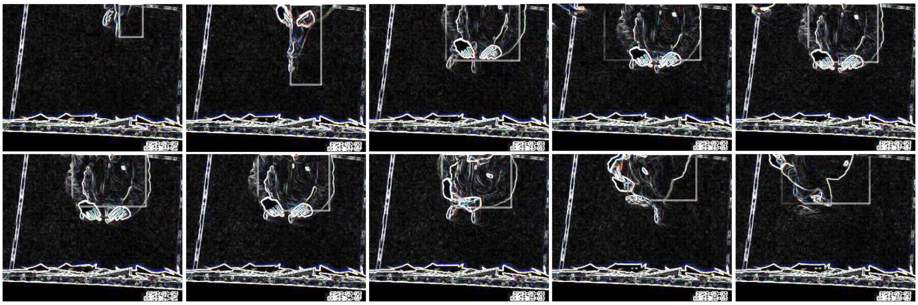
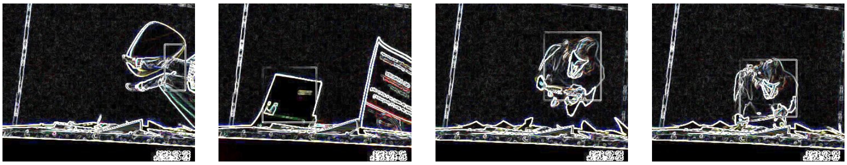


Fig. 9. A participant photographing the flyer posted on the bulletin board at the bottom of each frame. The sequential progression is from left-to-right and top-to-bottom. The rectangles that overlay the participant indicate software motion detection. Each frame is processed using an edge-detection algorithm in order to minimize capturing more data than required by the experiment and to protect the privacy of the participant.

move chairs into the field-of-view or otherwise congregate or loiter. Our flyer was secured at each corner with thumb-tacks, however other flyers may have been secured only at the top leading to some circumstances where activity outside of camera view caused flyers to move. The situation most apt to provide a false positive is when a subject appears to be facing the flyer, but it is not clear if the subject is actually photographing the flyer. By examining such situations in context with time-adjacent images, we are able to identify unrelated activities, such as posting or retrieving a flyer unrelated to our study. Examples of each of these false positives are provided in Figure 10. The captured images provide enough fidelity to accurately determine which should be discarded from analysis.

4.4 Results

We collected data for four weeks beginning February 7, 2012, two weeks using the *surv_qrcode_only* display (10 participants) followed by two weeks using the



(a) Foreign object (b) External effects (wind) (c) Possible participant (d) Posting an unrelated flyer

Fig. 10. The communal nature of the experiment site encourages inhabitants to loiter, as seen in (a) where an individual has moved a chair into the experiment area and relaxed. (b) depicts nearby "wind" which occasionally caused nearby flyers to trigger motion detection. From the frame shown in (c), it is difficult to discern if the person in is scanning the flyer or not. However with the additional context of time-adjacent frames, it is obvious that the person is searching for the best location to post a flyer unrelated to the study. The posting, (d), is performed several seconds after (c).

surv_incentive display (two participants). We conducted a follow-on experiment by re-posting *surv_qrcode_only* for two more weeks (six participants) at which point we could no longer use the location. From video analysis we determined that three individuals likely scanned the QR code, but elected not to visit the URL. Of these three individuals, one was from the *surv_incentive*, one was from the *surv_qrcode_only* and one was from the follow-on no incentive condition. In our study 85% (15/18) of people that scanned a QR code proceeded to visit the website, however our results may not be representative of a larger population. Nine participants visited the URL in the *surv_qrcode_only* (plus five more in the follow-on), and only one in *surv_incentive*. This ratio suggests that the incentive may not have actually enticed the participants to scan the QR code. Moreover, more people scanned the poster in the follow-on than in *surv_incentive*, further re-enforcing the tendency to scan the no incentive condition.

Five participants, all in the *surv_qrcode_only*, started the survey. Of these five, one was under 18 and discarded, another selected "Prefer not to answer" and "Neutral" for every question. The remaining three participants were all students of age 18-24, two male and one female. Of these three, two completed the survey. Interestingly, one answered the question "How often do you scan a QR code?" with "Every time I see one" while the other answered "Rarely." The devices of the nine participants include four iPhones, four Android devices, and one BlackBerry.

4.5 Limitations

This experiment was conducted in a single, on-campus location, limiting results to a single population. The location was in the computer science building, leading to a relatively technologically-sophisticated population. The location was near a primary walking path and near a coffee shop, both of which contribute to a wider demographic, but the single location certainly has population bias. Further, the participant pool may have degraded between conditions, since the same location was used for both conditions. For example, participants may not scan a new poster, anticipating that the new poster is part of the same study.

Another technical limitation was the subjectivity in determining whether a person scanned the QR code. If a correlated entry appeared in the server logs, the person certainly scanned the QR code. However, without the server log entry, we are forced to decide whether the images indicate that a person scanned the QR code. As shown in the results, nearly all (85%) of people who scanned the code also visited the website leaving 15% subject to scrutiny. None-the-less, the analysis is subjective and it is possible that some instances may have been misclassified.

5 Security Implications

Unsurprisingly, of the 229 participants we observed, more than 80% (184) used a mobile device in our studies. Contrary to national metrics from around the time of our study showing an Android majority [9] [4], we observed 57% (105) using an Apple iOS device and 38% (69) using an Android based device (the remaining five percent used Blackberry and Windows mobile devices). Given that the majority used either an iOS or Android device, it is also no surprise that of the mobile clients, 96% of browsers are WebKit based. Figure 11 shows user-agent distribution for the measured devices.

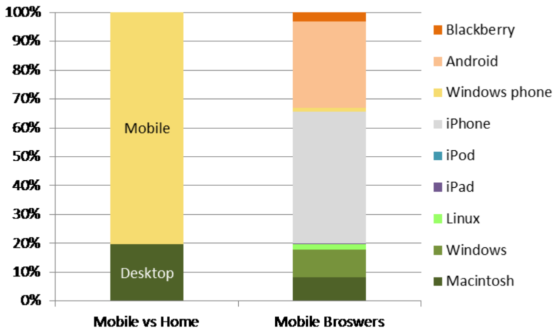


Fig. 11. Observed user-agent and mobile vs desktop browser distribution

At the time of this research, known vulnerabilities and public exploits existed that target the WebKit browser directly [6] [1] or a content handler (such as PDF files [2]) for all Android and iOS devices observed in this study. In order to take advantage of such vulnerabilities, an attacker must persuade the user to visit a web site serving malicious content, possibly using a QR code. A successful attack will result in the attacker having remote access to the same resources as the browser.

In many cases, executing in the same context as the browser may be enough to achieve attack objectives such as reading browser cookies or stealing website passwords. Other attack objectives require more privileged access. This elevated access is often referred to as "rooting" or "jailbreaking" the mobile device. Rooting exploits are often very dependent on the operating system version.

We observed the operating system distribution for iOS and Android devices. The fragmentation present for Android is consistent with what has been reported in the literature and by Google [7,18,30].[1] We observed nine different iOS versions, but greater than 80% (89) of iOS devices were running the most recent, 5.0.1. In Android we observed 11 different iOS versions, and no single version was present on more than 30% of devices. The two versions with the highest percentage of devices were 2.3.3(18%) and 2.3.4 (27%), neither of which were the most current software within an Android branch at the time of the study (in this case 2.3.7). Rooting exploits not requiring physical device access were publicly available at the time [13] [3] for 83% (59) of Android and 17% (19) of iOS devices we observed.

Compared to the actions required from an unscrupulous attacker, conducting our study demanded significantly more effort. For example, we spent considerable time ensuring that flyers at all locations were available. Since each location required a unique flyer, we tracked which specific flyers were posted so that the distribution of conditions and specific URLs were maintained throughout the experiment. Similarly, our exposure was limited to locations where we were ethically and legally permitted to post QR codes. A would-be attacker does not have such problems, permitting the attacker to expend significantly less time and receive significantly more exposure when compared to this study. For this reason, our findings should be considered an extreme lower bound on the susceptibility of QRishing.

In many cases QRishing would be conducted *physically*, meaning that an attacker would have to find some way to post QR codes where a user might approach and scan the code. The physical format of QRishing could be realized in many forms: full posters, sticker overlays, etc. In our study we simply posted disingenuous flyers. When compared to digital forms of phishing, such as email, the cost of performing a QRishing attack is likely comparatively high. The cost of printing QR codes is negligible, but it takes time to post them and other risks, such as being physically caught placing a malicious QR code, represent considerable potential cost. On the other hand, an attacker, unbound by legal and ethical issues, can place QR codes in a wider range of places than we were permitted.

Mobile browsers are largely not employing technical controls that have been available in desktop browsers for some time. For instance, technical controls may be used to assist the user in making security-conscious decisions. Some reader applications already display the QR code content prior to performing an action, such as visiting a website. While this simple action requires the user to "click one more button," the opportunity to at least assess the potential for a questionable domain is beneficial. This could be augmented with security-specific controls that are already ubiquitous in desktop browsers such as comparing the scanned URL to a blacklist or some other "safe browsing" technology.

[1] Google's self-reported numbers [7] are grouped less precisely than ours. But when we group ours accordingly, they align closely.

Security indicators for valid certificates and SSL/TLS connections are widely adopted by desktop-browser vendors and allow consumers to assess the security of their communications over the web. However, the deployment of technical controls and security indicators to mobile browsers is complicated by the relatively small screen real estate for handheld devices. In an empirical study of ten mobile browsers and two tablet browsers, Amrutkar et al. find that many of the World Wide Web Consortium's (W3C) guidelines for security indicators in web user interfaces are not implemented on mobile browsers and that there is little consistency among mobile browsers that do implement security indicators [10].

Another technological control specific to smartphones is to enable timely application of security updates to mobile browsers and core device software. Feature updates could be separated from security updates. The separation would allow security updates to be applied quickly and independently of feature updates, allowing economic motivations to drive the release (or not) of feature updates. This control does not specifically address the threat of QR codes, but can mitigate the subsequent threats posted by the malicious websites. Similarly, if the browser was a self-contained component, similar to other mobile applications, it could be updated independent of the core software of the device. In this light, alternative browsers such as Firefox mobile, provide a method to use an updated browser on old devices where system software is no longer updated.

6 Conclusion

We presented two experiments demonstrating that QR codes are a viable method for conducting phishing attacks. We posted QR code posters across 139 different locations and found that 225 individuals scanned at least one poster over a four-week period. Overall, 61% of the disingenuous posters were scanned by at least one person.

Most users (75%) scanned the QR code out of curiosity or for fun. Comparatively, very few scanned in order to solicit more information about the context surrounding the QR code. The results of our surveillance experiment indicate that most users who scan a QR code will subsequently visit the related URL, even if the domain is unfamiliar and uses "URL shortener" style URLs. Providing security controls that already exist in desktop browsers to mobile browsers may foster safer behavior than what we observed in this study.

While a QRishing attack likely requires more resources than a typical email oriented phishing attack, the cost of conducting a QRishing attack is negligible. However, indirect costs, such as physically being caught, present considerable additional risk over traditional phishing mechanisms. None-the-less, if the attacker wishes to target a particular audience, such as smartphone users, QRishing may be a viable option. The ease with which such an attack can be mounted against current smartphones is particularly concerning given the long patching cycle and potential for an attacker to gain elevated privileges on the device. With or without the security-specific controls, user awareness of new threats like QRishing will be critical as mobile devices become increasingly popular.

Acknowledgments. This work is supported in part by National Science Foundation under ITR award CCF-0424422 (Team for Research in Ubiquitous Secure Technology) and IGERT award DGE-0903659 (Usable Privacy and Security). We would like to thank Jim Skees for his assistance in facilitating the physical location for the surveillance experiment.

References

1. About the security content of iOS 4.3 (March 2011), http://support.apple.com/kb/HT4564
2. About the security content of iOS 5.0.1 (November 2011), http://support.apple.com/kb/HT5052
3. CVE-2011-3874 - libsysutils rooting vulnerability (zergRush) (November 2011), http://code.google.com/p/android/issues/detail?id=21681
4. Generation app: 62% of mobile users 25-34 own smartphones (November 2011), http://blog.nielsen.com/
5. The Male vs. Female Debate Goes Mobile (November 2011), http://blog.compete.com
6. Android bug opens devices to outside control: experts (February 2012), http://www.reuters.com/article/2012/02/24/us-google-android-security-idUSTRE81N1T120120224
7. Android Developer Guide: Platform Versions (February 1, 2012), http://developer.android.com
8. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2011–2016 (February 2012), http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html
9. comScore Reports December 2011 U.S. Mobile Subscriber Market Share (February 2012), http://www.comscore.com/Press_Events/Press_Releases/2012/2/comScore_Reports_December_2011_U.S._Mobile_Subscriber_Market_Share
10. Amrutkar, C., Traynor, P., van Oorschot, P.C.: An Empirical Evaluation of Security Indicators in Mobile Web Browsers. Technical Report GT-CS-11-10, Georgia Institute of Technology (2011)
11. Borrett, L.: Beware of Malicious QR Codes (June 2011), http://www.abc.net.au/technology/articles/2011/06/08/3238443.htm
12. U. C. Bureau. Pittsburgh census map (2000), http://www.city.pittsburgh.pa.us/cp/html/census_map.html
13. chpwn, MuscleNerd, and chronicdevteam. iOS Jailbreaking Website, http://jailbrea.kr/
14. Dhamija, R., Tygar, J.: The battle against phishing: Dynamic security skins. In: Proceedings of SOUPS 2005, pp. 77–88. ACM (2005)
15. Dhamija, R., Tygar, J., Hearst, M.: Why phishing works. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 581–590. ACM (2006)
16. Downs, J., Holbrook, M., Cranor, L.: Decision Strategies and Susceptibility to Phishing. In: Proceedings of SOUPS 2006, pp. 79–90. ACM (2006)
17. Egelman, S., Cranor, L., Hong, J.: You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 1065–1074. ACM (2008)
18. Gahran, A.: Why 'Android fragmentation' isn't so bad (February 2012), http://www.cnn.com/2012/02/17/tech/mobile/android-fragmentation-gahran/

19. Han, J., Owusu, E., Nguyen, T.-L., Perrig, A., Zhang, J.: ACComplice: Location Inference using Accelerometers on Smartphones. In: Proceedings of the 4th COMSNETS (January 2012)
20. Hara, M., Watabe, M., Nojiri, T., Nagaya, T., Uchiyama, Y.: Optically readable two-dimensional code and method and apparatus using the same (March 10, 1998) US Patent 5,726,435
21. Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L., Hong, J.: Teaching Johnny not to fall for phish. ACM Transactions on Internet Technology 10(2), 7 (2010)
22. Moore, T., Edelman, B.: Measuring the perpetrators and funders of typosquatting. In: Sion, R. (ed.) FC 2010. LNCS, vol. 6052, pp. 175–191. Springer, Heidelberg (2010)
23. Neumann, A., Barnickel, J., Meyer, U.: Security and privacy implications of url shortening services. In: Proceedings of the Workshop on Web 2.0 Security and Privacy (2010)
24. Newman, R.: Consumer Alert: QR Code Safety. Better Business Bureau (June 2011), http://sandiego.bbb.org/article/consumer-alert-qr-code-safety-28037
25. Office of Institutional Research and Analysis. Carnegie mellon factbook (February 2012), http://www.cmu.edu/ira/factbook/pdf/facts2012/entire-fb-for-web-as-of-3-1-121.pdf
26. Radwanick, S.: 14 Million Americans Scanned QR Codes on their Mobile Phones in june 2011 (August 2011), http://www.comscore.com
27. Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L., Hong, J., Nunge, E.: Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In: Proceedings of SOUPS 2007. ACM (2007)
28. Tamir, C.: AVG (AU/NZ) Cautions: Beware of Malicious QR Codes. PCWorld (June 2011), https://appsec-labs.com/blog/tag/qrcode
29. Todd, D.M.: Security expert warns smartphone users of the risks in scanning cybercoding, http://www.post-gazette.com (accessed June 2012)
30. Vidas, T., Votipka, D., Christin, N.: All your droid are belong to us: A survey of current android attacks. In: Proceedings of the 5th USENIX WOOT, p. 10. USENIX Association (2011)
31. Wagenseil, P.: Anti-anonymous hacker threatens to expose them, http://www.msnbc.msn.com (accessed March 2012)
32. Zhang, Y., Hong, J., Cranor, L.: Cantina: a content-based approach to detecting phishing web sites. In: Proceedings of the 16th International Conference on World Wide Web, pp. 639–648. ACM (2007)

Appendix

Table 1. QR code reader applications tested. Five of the top ten free iOS applications and three of the top ten free Android applications automatically visit URLs scanned from QR codes.

No.	Application [Vendor]	Auto Visit
1	Barcode Scanner [Versolab]	no
2	ShopSavvy (Barcode and QR Scanner) [ShopSavvy, Inc.]	yes
3	RedLaser Barcode and QR Scanner [eBay, Inc.]	no
4	ScanLife Barcode and QR Reader [Scanbuy, Inc.]	yes
5	AT&T Code Scanner [AT&T Inc]	no
6	pic2shop - Barcode Scanner [Vision Smarts]	no
7	Bakodo - Barcode Scanner [Dedoware, Inc]	no
8	NeoReader - QR reader [NeoMedia Technologies, Inc]	yes
9	i-nigma QR Code scanner [3GVision]	yes
10	MOBILETAG - Barcode Scanner [Mobile Tag]	yes

(a) iOS Applications

No.	Application [Vendor]	Auto Visit
1	Barcode Scanner [ZXing]	no
2	ShopSavvy Barcode Scanner [ShopSavvy, Inc.]	yes
3	QuickMark Barcode Scanner [SimpleAct, Inc.]	no
4	RedLaser Barcode and QR Reader [eBay Mobile]	no
5	ScanLife Barcode and QR Reader [Scanbuy, Inc.]	yes
6	Barcode scanner [george android]	no
7	i-nigma Barcode Scanner [3G Vision]	yes
8	AT&T Code Scanner [AT&T Service, Inc.]	no
9	ixMAT Barcode Scanner [ixellence.com]	no
10	BARCODE SCANNER [Jet Ho]	no

(b) Android Applications

Many mobile devices do not have any QR code reading software pre-installed. Tables 1(a) and 1(b) show the specific applications tested, whether the application automatically visits a URL retrieved from a barcode, and the order (top to bottom) of popularity on March 8, 2012. Several of the most popular iOS applications were either not free, or did not scan QR codes. We did not test any applications that were not free. Thirty percent of these top ten free scanning applications in the Google Play Market and 50% in the Apple App Store immediately visit a scanned URL in the default configuration. When applications employ this feature, the user has no opportunity to visually inspect the URL prior to visiting that URL.

"Comply or Die" Is Dead: Long Live Security-Aware Principal Agents

Iacovos Kirlappos, Adam Beautement, and M. Angela Sasse

University College London, Department of Computer Science,
London, United Kingdom
{i.kirlappos,a.beautement,a.sasse}@cs.ucl.ac.uk

Abstract. Information security has adapted to the modern collaborative organisational nature, and abandoned "command-and-control" approaches of the past. But when it comes to managing employee's information security behaviour, many organisations still use policies proscribing behaviour and sanctioning non-compliance. Whilst many organisations are aware that this "comply or die" approach does not work for modern enterprises where employees collaborate, share, and show initiative, they do not have an alternative approach to fostering secure behaviour. We present an interview analysis of 126 employees' reasons for not complying with organisational policies, identifying the perceived conflict of security with productive activities as the key driver for non-compliance and confirm the results using a survey of 1256 employees. We conclude that effective problem detection and security measure adaptation needs to be de-centralised - employees are the principal agents who must decide how to implement security in specific contexts. But this requires a higher level of security awareness and skills than most employees currently have. Any campaign aimed at security behaviour needs to transform employee's perception of their role in security, transforming them to security-aware principal agents.

Keywords: Information security management, compliance, decision-making.

1 The Need for Information Security

Organisations today face an ever-increasing number of information security threats: intellectual property theft can severely impact competitiveness, loss of customer information can damage corporate profiles and loss of access to corporate systems can impact the organisation's productivity [1]. Despite the significant amount of time being invested in producing effective security solutions by researchers and industry experts, the challenges and potential threats organisations face today are higher than ever [1].

After implementing technical controls strong enough to minimise an organisation's exposure to all but the most sophisticated (and costly) attacks, security researchers and practitioners today focus on humans as the "weakest link" in the security chain [2]. Information security turned to the disciplines of Human-Computer Interaction

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 70–82, 2013.

(HCI) and Behavioural Economics for security solutions that their employees can, and will, comply with [3-5]. Research in usable security and economics of security has yielded some valuable insights, but the problem of non-compliance is still rife. There have been steps to re-designing security solutions to fit human capabilities and limitations [5-7], and to base on them on people's real security needs, rather than what experts think [8], but we are still lacking an understanding of drivers of security behaviour 'beyond the interface'.

In this paper, we examine real-world non-compliance examples to understand drivers for non-compliant actions in information security. We present a study designed to identify the drivers of deliberate non-compliance, and then consider how this understanding can be used to transform Information Security Management. We begin by summarising existing literature on managing security behaviour.

2 Organisational Approaches to Information Security

Information Security management currently attempts to reduce an organisation's exposure to security risks primarily by formulating policies of how they should behave to avoid those risks, and communicating those policies to employees. Policies are usually in the form of documents, which define the security objectives of the organisation, the responsibilities of employees, and sanctions for non-compliance. Policies are vital for organisations - without them, specific security implementations can be developed without a clear understanding of the organisation's wider security objectives and employee responsibilities [9][10]. But current security policies do not address the security challenges organisations face for two reasons:

1. Employees have no insights on policy design [11]: policies are designed to reflect the way the policymakers *believe* employees should behave, usually adding elements required to comply with regulations, audit checks and international standards.
2. The formulation of both policies and standards is largely based on lessons learnt from past failures, and is rarely grounded in scientific principles [12]. Security is currently a craft, that is only useful for securing organisations against breaches that closely resemble past events. It also makes assumptions about the context and the environment in which the interaction of employees with information-handling systems takes place, ignoring factors like employee workload, and treating all compliance scenarios as the same [13]. This results in policies ending up as long lists of *dos* and *don'ts* located on web pages most employees only access when they have to complete their mandatory annual "security training" and which has little to no effect on their security behaviour.

So employees don't comply with security policies. Most organisations respond by trying to reduce the possibility for non-compliance through technical mechanisms – such as making downloading of information impossible. Enforcement usually takes the form of access control, restricting which employees can have access to which files. Compliance with the policies may be monitored. (In the case of access control, though, what tends to be monitored is whether access entitlements are still appropriate – rather than if an employee is in possession of a document they should not have). Security training and risk communication are used to influence employee behaviour towards compliance and reduce security risks. There are indications, though, that this set of current measures is not effective:

1. Compliant behaviours are being associated with specific threat scenarios or working practises, but there is little understanding of principles, or culture of secure behaviour. This means most employees are unable to take the initiative and make local decisions when new security problems arise [14][15].
2. Enforced compliance with cumbersome mechanisms consumes valuable employee resources, reducing the organisation's productivity [3][4]. In reality, large parts of the organisation (consider line managers, for instance) are complicit in employees' non-compliance, because – whatever the policies say – they value productivity more.
3. Compliance enforcement creates tension and deepens the *value gap* between security enforcers and the rest of the organisation [16]. Frustration with security is attributed back to the enforcers, which can result to any information coming from them being treated with scepticism or ignored and breeds a negative attitude towards information security in general [5] which can discourage compliance with security mechanisms - even sensible and well-designed ones [17].

Recent industry reports state that information security risks are increasing [1][18-20], so Information security research needs to develop more effective and sustainable approaches to managing non-compliant employee behaviour. Our contribution, presented in this paper, is a detailed, empirically-based understanding of reasons for non-compliance. This provides decision makers with a framework for identifying plausible ways of managing employee behaviour more effectively, and evaluating their effectiveness in a systematic fashion.

3 Understanding Non-compliance

To obtain more detailed insights into employee compliance, researchers need access to employees who are willing and able to honestly speak about their security behaviour within the work environment. We have built relationships with a number of partner organisations that were prepared to grant us access to their employees, encourage participation, and publicly assure them there would be no reprisals. Over the past two years we have conducted studies in two partner organisations as part of a process to identify areas of friction between the business and security processes, and to design and deploy appropriate interventions.

The first stage in this process is to conduct a series of interviews. This stage has been completed in two organisations. 126 interviews were conducted with the US and the UK parts of a major energy company, and 86 interviews with the UK employees of a telecommunications company. The interviews were semi-structured and probed aspects of security awareness and compliance, including:

1. The employee's awareness of the sensitivity of information they handle, and why they need to protect it.
2. Their knowledge of existing security policies, and what mechanisms they should or could use to reduce security risks.
3. Their experiences when interacting with the existing security policies and mechanisms.
4. Examples of, and reasons for non-compliance: how they circumvent policies and mechanisms, and their understanding of risks associated with these.

The majority of employees reported non-compliance in the organisation's day-to-day operations; interviewers then asked follow up questions to identify the conditions that led to the use of workarounds, the factors they used to decide whether to comply or not comply, and their understanding of the risks involved in their actions.

The insights we present here are based on a subset of the 126 interviews conducted in the first company, and a complete analysis of all interviews with respect to one mechanism: access control [21]. These were analysed using a thematic coding analysis based on the three Grounded Theory stages [22]: *open, axial* and *selective* coding. This led to the identification of three different non-compliance situations: *high compliance cost, lack of understanding, unavailable compliance mechanisms*.

The second stage of the process, completed in the utility company and underway in the telecommunications company, is the deployment of a scenario-based survey that presents participants with an example of a conflict situation drawn from an analysis of the interviews. Participants are offered 4 non-compliant courses of action that would allow them to resolve the conflict and were asked to rank the options in order of how likely they would be to use them and also to rate how severe a breach of policy the course of action is. A statistical analysis of the 1256 results from the survey (utilising MANOVA, Spearman's Rho and Chi-Squared tests) revealed several key "hotspots" where options rated as insecure were still being highly ranked as viable options Additionally, we were able to identify a US/UK cultural difference through the analysis of the results, which allowed us to further refine our understanding of the problem, and potential effective solutions. We also analysed 874 voluntary free-text comments left by participants using a Grounded Theory coding approach.

Using findings from these studies, summarised in the following sections and grouped according to the non-compliance situation they relate to, we aim to devise tangible suggestions to reduce the friction between the existing security implementation and business processes, provide guidelines for the design and deployment of future security mechanisms, and also aid in the development and maintenance of a more mature and resilient security culture.

3.1 Could Comply, But Cost too High

The first reason we identified as a driver for non-compliance is the high individual resource investment (such as time, or cognitive or physical effort) that certain security mechanisms demand. The main focus of the majority of employees is not to be secure, but to efficiently complete a *primary production task* – such as manufacturing goods, financial investment, or delivering CNI services. This results in employees being willing to spend a limited amount of both time and effort on secondary tasks, such as security (the *Compliance Budget*, [3]). Security mechanisms that impose high workload overheads make non-compliance an attractive option for quick primary task completion [3][23]. Most organisations are unaware of, or ignore, the impact of security mechanisms on users. Cormac Herley [4] has pointed out, that in the consumer context, "security people value customers' time at zero". Our studies show that in the work context, organisations work on the assumption that employees can simply absorb the effort associated with security compliance. But because most security mechanisms are difficult and cumbersome to use, employees literally feel their time/effort being drained. This experience drives non-compliance: the perceived risk mitigation achieved by complying does not seem worth the perceived cost of effort and disruption to the primary task [4]. The greater the perceived urgency and importance of the primary task, the more attractive or acceptable non-compliant options become - even when employees are aware of potential risk. Employees re-organise their primary tasks to avoid or minimise their exposure to security mechanisms that slow them down significantly [24]. Our interviews yielded several examples of this around file sharing [21]. In our subsequent survey, we included a file sharing scenario, in which a group of employees had to share a large volume of files, but incorrect permissions prevented some of them from accessing those. The pressure of an upcoming deadline, combined with employees knowing that setting up access takes about a week, led to the most frequently chosen response (selected by 32.6% of employees) being *"to email the restricted document archive directly to all recipients on his work group mailing list"*. The same respondents rated this as the second most risky option, giving it a severity rating of 4/5). In the (voluntary) free-text comments for this scenario, most respondents described the consequences of not completing the primary task as definite and severe, whereas the risk associated with breaching the security policy was only a potential one.

In our interview analysis, we identified the following frequent non-compliance instances driven by the primary-task focus:

1. 50% of employees shared their passwords for quick access to systems if colleagues needed access for work purposes, but did not have the necessary permissions, because it *"would take ages"* to get the permissions changed. Password and account sharing is a common workaround. Our interviewees also expected their colleagues to do the same for them. Even some managers reported this as common and acceptable practice: "*employees newly-involved in a project access the system using someone else's credentials until their access is sorted out*". This is an example of organisations becoming complicit in circumvention of policies and mechanisms which do not fit with the primary task.
2. 53% of employees reported having used personal unencrypted USB drives to share data perceived to be sensitive with colleagues because it is faster and easier than company-issued encrypted ones. The effort involved in using the latter did was perceived to be *"not worth it for simple file transfers around the office"*. Some interviewees said "*they immediately wiped the drives afterwards*" to prevent data falling to the wrong hands.

In both cases, the delay to completing the primary task is perceived as "not worth the effort" of guarding against a potential, unclear risk; implicit in these statements is *"we've done it many times and nothing bad happened, so surely it cannot be that bad?"* Employees knew they were not complying with policies but felt this was justified by getting their job done or helping a colleague. The survey also supports our conclusions: in the scenario where an employee does not have an encrypted USB stick, the use of an unencrypted one was second most popular choice, scoring less than 1% behind the most popular option of borrowing an encrypted drive from a colleague. It was also rated as the second least severe risk; only uploading the files to public data storage received a higher severity rating.

3.2 Could Comply, But Why Should I?

Inaccurate perceptions of risk and technology underlie many insecure behaviours [25][26]. In particular employees under-estimate the risk mitigation that can be achieved by compliance with some policies – and this, in turn, makes non-compliance appear a more attractive option. Examples of this include:

1. Employees rarely considered the possibility that their actions might lead to malware being introduced to their organisation's systems – hence the perception that using a personal USB stick would cause no harm.
2. Employees did not consider that deleted data can be easily recovered from drives if those are lost; they believed that deleting all the data from a drive after finishing with a file transfer provides adequate protection.
3. Employees considered any data stored on their company laptops to be secure because a Windows password was required to access them - but the Windows password was only used for access control purposes. This resulted in unsafe practices, like storing sensitive files locally on the laptops, assuming they are adequately protected when travelling on public transport [27].

We also found most employees did not have a good understanding of what information security is, and what it tries to protect. Security risks were described as *"just to confidentiality not security"* when confidentiality is a key goal of information security. There were also varying and inaccurate statements of what particular security policies permitted or prohibited - creating many security myths.

The survey results indicate that even when employees are aware of a policy and interpret it correctly, this is not a strong motivator for individual behaviour. We linked each of the options in the scenarios to a behaviour and attitude type. When asked what to do when observing a clear breach of policy by a colleague or visitor, the most frequently chosen option was *"report suspicions but take no direct action"* Employees took a passive approach – they did not think they had any responsibility to promote compliance with security policies. It is not sufficient for organisations to just correct employee misconceptions about policies and risks of their actions. They should also make adherence to security policies, and actively promoting adherence, part of the psychological contract they have with employees [28] – but this will not work if security interferes with individual and organisational tasks and processes to the point that compliance is perceived as *"not worth it"*.

3.3 Something's Awry, Just Can't Comply

In some cases, compliance may not even be an option, regardless of how much time or effort employees are willing to invest. Employees reported being unable to comply because the implementation the corresponding security mechanisms did not match basic requirements:

1. Employees justified copying files to laptops because there was insufficient space on their network drive, or because they had experienced problems accessing files they needed from home or while travelling.
2. Employees found the encrypted USB drives provided by the organisation were too small, so alternative file-sharing methods such as using unencrypted drives or emailing files had to be used.
3. The large number of passwords required in order to ensure access to the various corporate systems resulted in employees being unable to recall those from memory. This led to writing their passwords down, either in electronic form on their laptop or in a document they carry with them all the time.

In the above cases, most employees were aware of the increased risks associated with their behaviour, but felt that the organisation's failure to provide a "*properly working technical implementation*" forced them into workarounds so they could keep working and complete their primary task. The employees' perception was that the organisation would prefer security transgressions to "*letting everything grind to halt*" – and this was confirmed by similar responses from respondents with managerial responsibility in the survey. This is another example of how the organisation is complicit in employees' non-compliance.

4 Rethinking Information Security Management

Organisations looking to have effective information security need balance between the productivity and risk management goals. Our observations suggest that currently, organisations do not manage this balancing act: they set high targets for both productivity and security, and leave it to employees to resolve any conflicts between them. Most of the time, employees will chose productivity because 1) their behaviour is focussed on the primary task, and 2) they are principal agents who are trying to maximise their own benefit [29]. Based on our results here and those of other studies [5, 24] we suggest that most organisations are complicit in security non-compliance. They enable and reinforce their employees' non-compliance choices because they

1. Reward employees for productivity not security,
2. Fail to identify and fix security policies and mechanisms that create friction, and
3. Rarely enact the sanctions they threaten in case of non-compliance - very few organisations that threaten 'comply or die' on paper act on it[1].

[1] One of the authors has been involved in a (as yet unpublished) study of a company that publicly declares that non-compliance with any of its 'principal security rules' is grounds for instant dismissal. It would have to dismiss half of its workforce every month if it acted on this declaration; it would not be able to continue operating if it did.

Pallas [29] has applied the economic concept of *Principal-Agent* relationship to managing information security; we found his approach extremely helpful both in explaining the behaviours we identified, and to identify changes that organisations can make to break the non-compliance cycle. Employees are rational actors and to motivate them to comply with security policies, they have to perceive compliance as serving their own best interest [4]. The traditional "*command and control*" approach – where policies are set centrally by security experts, who select mechanisms and specify behaviours that must be complied with, without considering individual tasks or business processes – does not work in modern, flat, geographically distributed organisations who want to be agile, and want productive employees with ideas and initiative. Most organisations and policy makers have moved from compliance to risk-based information security standards (such as ISO27001), but have failed to make the same shift when it comes to managing employees' security behaviour; in that case organisations are 'unwittingly complicit' as they do not realise they are acting in a schizophrenic and uncoordinated way, negatively influencing employee compliance. Central policies and mechanisms cannot fit the variety of local and situational contexts in which individual employee decisions take place. Greater flexibility is needed to adapt to local circumstances, and solve conflict with tasks and business processes as they arise. Employees need to understand the risks surrounding their roles and the benefits of compliance to both themselves and the organisation, and then be trusted to make their own risk decisions in a way that mitigates organisational risks [15]. To aid the effective implementation of this security management approach the implemented security mechanisms need to be better *aligned* with the primary task, aiming to improve the identified employee misconceptions and misunderstandings that lead to non-compliance.

4.1 Align Security Policies with Main Productivity Objectives

As we previously mentioned, security implementations need to act as enablers to the primary tasks not blocking those. Teo and King [30] introduce the term *Information Systems Alignment* to describe "*The degree to which the information systems plan reflects the business plan*". We argue that the same needs to apply to information security: The more a security policy and its implementation accommodate employee priorities and values, the more it improves the alignment of incentives in the enforcer-employee principal-agent relationship [29]. Thus, the security policy is less likely to be resisted [31].

To achieve this *Information Security Alignment*, employee attitudes and beliefs need to be considered when formulating security policies [30]. As shown in Section 3.2 high-level, abstract information security goals are not a strong motivator for employees – they cannot compete with concrete demands of business processes that employees know well, and for which they understand the consequences of failure to deliver.

Failure to take into account the beliefs and attitudes of employees results in the target group (end users) not adequately participating in the design of security mechanisms, or the creation and maintenance of a strong security culture, which inevitably are going to affect their day-to-day jobs. Participatory design [33] has been at the core of most successful human factors and usability engineering processes, and security designers cannot afford to ignore it. The reasons for non-compliance identified in our findings provide a good starting point for incorporating similar procedures into security design. Those need to be communicated to policymakers and security designers, so that information security solutions more suited to employee daily routines can be created. This can re-adjust employees' cost-benefit decisions, increasing compliance rates and creating a positive attitude towards security, which can also render employees more susceptible to attempts to instigate and maintain a stronger security culture within the organisation.

4.2 Adjusting the Cost-Benefit Perception

To improve employee compliance decisions we also need to target their individual *cost-benefit analysis*. After creating policies and security implementations that accommodate for employee needs and priorities, we need to target the cost-benefit balance to shift it towards compliance by making it an economically attractive option for employees [34]. Beautement et al. [3] identify four factors through which this balance can be influenced (*Design, Culture, Monitoring, Sanctions*). In the remainder of this section we discuss how these four factors relate to our current findings, explaining how each one of those can be targeted to encourage compliance by changing the employees' perceived *cost-benefit* balance.

Design. Even for the most risk-aware and knowledgeable employees, the cost-benefit balance will favour non-compliance when implemented systems impose high overheads on their primary tasks [24]. Reduced compliance costs can eliminate the identified "*cost too high*" and "*can't comply*" non-compliance instances. To improve on the security design an organisation needs to:

1. Check that security mechanisms work in a given context. A network drive on which employees are encouraged to store their documents should be adequately sized so that they do not run out of space, combined with auto-archiving systems to prevent employees travelling around with confidential data on their laptops. In addition, encrypted laptop drives could reduce the risks when employees need to have some files stored locally and VPN access should be improved to reduce the need to transfer data through other channels. Single sign-on systems can eliminate the need to write down passwords, while providing every employee with an encrypted USB drive can reduce the need to use unencrypted ones. In all cases the secure option should also be the easiest one to use.
2. Provide flexibility to make local and situational adjustments. Employees who need access to systems to proceed with their primary tasks cannot wait for a few working days for that to be granted, otherwise they will find another way to get access (usually through their trusted colleagues). Many interviewees reported that outsourcing of IT services had removed previously available routes to getting local and temporary adjustments made. The ability to make such adjustments would reduce password sharing and information sharing through unauthorised channels that is driven by the focus on productivity. The processes required for security, as

well as the necessary mechanisms and technology, should mesh cleanly with the needs of the primary task.

Communicating the Value of Security. Once compliance-enabling systems are implemented, the organisation can consider raising employee awareness of risks and principles for managing them. Blanket 'security education campaigns' are not effective – messages need to be targeted at the perceptions held by specific groups of employees. The question "*why should I care?*" needs to be answered – what are the benefits? Organisations have to move away from the 'fear' sell of breaches and sanctions, and emphasise information security's contribution to achieving organisational objectives, and personal values, such as professionalism, instead [35]. This can be achieved through improved understanding of:

- *Everyone contributes to security*. Employee perception of security needs to be changed from "*getting in the way of achieving organisational goals*" to "*important for the organisation achieving its goals*" [36]. Employees need to realise that by following recommended security practices they are contributing to the smooth and efficient operation of business processes, as security ensures the availability of the resources required for the primary task to be successfully completed.
- *My specific contribution to protecting the organisation.* All employees can damage the organisation when not complying, even in relatively small ways. Thus, they all bear some responsibility for organisational security. Employees need to know what precautions they should be taking to reduce the organisation's exposure to security risks.

The two points above need to be communicated to employees through well-designed Security Awareness, Education and Training (SAET) campaigns. Those need to be formulated on a role-specific basis based on the identified employee misconceptions and non-compliance drivers, rather than flooding them with generic, organisation-wide advice that ends up doing more harm than the attacks they seek to prevent [37]. This approach also allows for increased flexibility, as organisations whose employees are adequately aware about the need for security, can tailor their behavioural change campaigns to start from the education stage. When employees are adequately knowledgeable on threats and vulnerabilities surrounding their role, organisations only need to implement an effective training scheme, testing their knowledge and only reverting back to education when misunderstanding is identified. Once the 3 steps have been effectively implemented, role-specific reminders of the key messages are needed to reinforce awareness and keep the employees informed on new risks. Also, education material should always be available for employees that need to refer back to it.

Monitoring, Sanctions – maybe. Trust, Definitely. When the security systems of an organisation are designed in a way that favours compliance and employees are well-aware of the information security risks related to their roles, expensive *architectural means* (physical and technical mechanisms to prevent unwanted behaviours [29]) become obsolete: compliance now comes from employees motivated to behave securely [38], based on norms developed by the existence of both *formal* and *informal* rules that are significantly cheaper to enforce [29]. This can also create a positive environment where employees feel well-trusted by the organisation, inducing further compliance. The definition of trust as "*willingness to be vulnerable based on positive expectations about the actions of others*" [39] may sound like an oxymoron to old-school *command and control* security managers, but organisations where employees have increased responsibilities are more likely to establish a high-level of security awareness and improved understanding of the need for security [15][28][40]. On the other hand, employees that abuse trust should be visibly punished; clever monitoring implementations can detect employee trust abuse [41] and employees that observe sanctions enforced, are less likely to attempt to knowingly abuse trust.

5 Conclusions

Our results show that a better understanding of real-world employee compliance decisions creates a new perspective for information security management. Many organisations know that 'comply or die' is dead – but some still keep conjuring up its ghost, while others struggle to find an alternative paradigm for managing their employees' security behaviour. We suggest that the first necessary step is to recognise employees' primary task focus, and design security that fits into individual tasks and business processes. Only when this can been achieved should organisations focus on communication. Identifying misconceptions and myths that justify insecure behaviour helps to design targeted campaigns to bust or transform these. A clear set of information security principles needs to be identified and communicated to create employees who are risk-aware and know how to manage the risks that apply to them.

5.1 Future Research

We are currently expanding our research to include other organisations, aiming for a better multi-organisational understanding of employee security perceptions and compliance-affecting factors. This will allow the generalisation of our research findings to provide an industry-wide view of current problematic information security mechanisms and practices, together with suggestions on how those practices can be improved to increase compliance rates. The focusing of our research on the analysis of empirical data, gathered by investigating real-world problems from active operational environments, can result in improved effectiveness of security decision making and wider adoption of the underlying principles by organisations when designing their security solutions.

References

1. GRT Corporation, `http://www.grtcorp.com/content/british-intelligence-speaks-out-cyber-threats`
2. Schneier, B.: Secrets and lies: digital security in a networked world. Wiley (2000)

3. Beautement, A., Sasse, M.A., Wonham, M.: The compliance budget: managing security behaviour in organisations. In: NSPW 2008: Proceedings of the 2008 Workshop on New Security Paradigms, pp. 47–58 (2008)
4. Herley, C.: So long, and no thanks for the externalities: the rational rejection of security advice by users. In: Proceedings of the 2009 Workshop on New Security Paradigms Workshop (NSPW 2009), pp. 133–144. ACM, New York (2009)
5. Adams, A., Sasse, M.A.: Users Are Not The Enemy: Why users compromise security mechanisms and how to take remedial measures. Communications of the ACM 42(12), 40–46 (1999)
6. 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(3), 122–131 (2001)
7. Weirich: Persuasive password Security. PhD thesis, University College London (2005)
8. Friedman, B., Howe, D.C., Felten, E.: Informed consent in the Mozilla browser: Implementing value-sensitive design. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, HICSS. IEEE (2002)
9. Fulford, H., Doherty, N.F.: The application of information security policies in large UK-based organizations: an exploratory investigation. Information Management & Computer Security 11(3), 106–114 (2003)
10. Higgins, H.N.: Corporate system security: towards an integrated management approach. Information Management and Computer Security 7(5), 217–222 (1999)
11. Bartsch, S., Sasse, M.A.: Guiding Decisions on Authorization Policies: A Participatory Approach to Decision Support. In: ACM SAC 2012, Trento, Italy (2012)
12. Björck, F.: Security Scandinavian style. PhD diss., Stockholm University (2001)
13. Fléchais, I.: Designing Secure and Usable Systems. PhD diss., University College London (2005)
14. Wood, C.C.: An unappreciated reason why information security policies fail. Computer Fraud & Security (10), 13–14 (2000)
15. Flechais, I., Riegelsberger, J., Sasse, M.A.: Divide and conquer: the role of trust and assurance in the design of secure socio-technical systems. In: Proceedings of the 2005 Workshop on New Security Paradigms (NSPW 20005), pp. 33–41. ACM, New York (2005)
16. Albrechtsen, E., Hovden, J.: The information security digital divide between information security managers and users. Computers & Security 28(6), 476–490 (2009)
17. Karyda, M., Kiountouzis, E., Kokolakis, S.: Information systems security policies: a contextual perspective. Computers & Security 24(3), 246–260 (2005)
18. PWC (2012), `http://www.pwc.co.uk/audit-assurance/publications/uk-information-security-breaches-survey-results-2012.jhtml`
19. Ashford, W.: (2012), `http://www.computerweekly.com/news/2240148942/Infosec-2012-Record-security-breaches-cost-UK-firms-billions`
20. Deloitte (2009), `http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/UK_ERS_2009_CB_Security_Survey.pdf`
21. Bartsch, S., Sasse, M.A.: How Users Bypass Access Control and Why: The Impact of Authorization Problems on Individuals and the Organization. In: ECIS 2013: The 21st European Conference in Information Systems (in press, 2013)
22. Strauss, A., Corbin, J.: Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage Publications, Incorporated (2007)

23. Bulgurcu, B., Cavusoglu, H., Benbasat, I.: Information Security Policy Compliance: An Empirical Study of Rationality-Based Beliefs and Information Security Awareness. MIS Quarterly 34(3), 523–548 (2010)
24. 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, pp. 383–392. ACM, Atlanta (2010)
25. Adams, J.: Risk. University College London Press (1995)
26. Wash, R.: Folk models of home computer security. In: Proceedings of the Sixth Symposium on Usable Privacy and Security. ACM (2010)
27. http://www.pcworld.com/article/261754/does_the_windows_logon_password_protect_your_data_.html
28. Sasse, M.A., Ashenden, D., Lawrence, D., Coles-Kemp, L., Fléchais, I., Kearney, P.: Human vulnerabilities in security systems. Human Factors Working Group, Cyber Security KTN Human Factors White Paper (2007)
29. Pallas, F.: Information Security inside organisations. PhD Thesis, Technical University of Berlin (2009)
30. Teo, T.S.H., King, W.R.: Integration between business planning and information systems planning: an evolutionary-contingency perspective. Journal of Management Information Systems, 185–214 (1997)
31. Trompeter, C.M., Eloff, J.H.P.: A framework for the implementation of socio-ethical controls in information security. Computers & Security 20(5), 384–391 (2001)
32. Dhillon, G., Backhouse, J.: Current directions in IS security research: towards socio-organizational perspectives. Information Systems Journal 11(2), 127–153 (2001)
33. Checkland, P.B., Poulter, J.: Learning for Action: A short definitive account of Soft Systems Methodology and its use for Practitioners, Teachers and Students (2006)
34. Furnell, S.M., Jusoh, A., Katsabas, D.: The challenges of understanding and using security: A survey of end-users. Computers & Security 25(1), 27–35 (2006)
35. James, H.L.: Managing information systems security: A soft approach. In: Proceedings of the 1996 Information Systems Conference of New Zealand (ISCNZ 1996). IEEE Computer Society, Washington, DC (1996)
36. Von Solms, B., von Solms, R.: From information security to business security. Computers & Security 24(4), 271–273 (2005)
37. Kirlappos, I., Sasse, M.A.: Security Education against Phishing: A Modest Proposal for a Major Rethink. IEEE Security & Privacy 10(2), 24–32 (2012)
38. Vroom, C., Von Solms, R.: Towards information security behavioural compliance. Computers & Security 23(3), 191–198 (2004)
39. Riegelsberger, J., Sasse, M.A., McCarthy, J.D.: The mechanics of trust: a framework for research and design. International Journal of Human-Computer Studies 62(3), 381–422 (2005)
40. Schlienger, T., Teufel, S.: Analyzing information security culture: increased trust by an appropriate information security culture. In: Proceedings of the14th International Workshop on Database and Expert Systems Applications, pp. 405–409. IEEE (2003)
41. Caputo, D., Maloof, M., Stephens, G.: Detecting insider theft of trade secrets. IEEE Security & Privacy 7(6), 14–21 (2009)

Information Security as a Credence Good

Ping Fan Ke[1], Kai-Lung Hui[1], and Wei T. Yue[2]

[1] Hong Kong University of Science and Technology
pfke@ust.hk, klhui@ust.hk
[2] City University of Hong Kong
wei.t.yue@cityu.edu.hk

Abstract. With increasing use of information systems, many organizations are outsourcing information security protection to a managed security service provider (MSSP). However, diagnosing the risk of an information system requires special expertise, which could be costly and difficult to acquire. The MSSP may exploit their professional advantage and provide fraudulent diagnosis of clients' vulnerabilities. Such an incentive to mis-represent clients' risks is often called the *credence goods* problem in the economics literature[3]. Although different mechanisms have been introduced to tackle the credence goods problem, in the information security outsourcing context, such mechanisms may not work well with the presence of *system interdependency risks*[6], which are introduced by inter-connecting multiple clients' systems by the MSSP. In particular, we find that allowing clients to seek alternative diagnosis of their vulnerabilities may not remove the MSSP's fraudulent behaviors. We shall explore alternative ways to solve the credence goods problem in the information security outsourcing context.

Keywords: Information security outsourcing, credence good, interdependency risks.

1 Introduction

Enhancing the security of information systems has become an important task for organizations. An accurate risk assessment is often important in implementing a cost-efficient security protection. By knowing the actual risk level, an organization can procure the appropriate level of security protection.[1] However, it is not easy to accurately diagnose the risk of an information system, especially for organizations without proper security expertise. Therefore, many organizations would prefer to outsource their security protection to a managed security service provider (MSSP). Yet, the information asymmetry between the MSSP and his clients introduces an incentive for the MSSP to cheat his clients, which could

[1] An excessively high security protection could lead to wastage of resources and poor usability. Similarly, sub-standard security protection could expose the organization to excessive risks and losses.

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 83–93, 2013.

subsequently lead to fraudulent behaviors.[4][5][2] In this study, we investigate such an incentive and discuss the implications for practices of security protection. We also study the challenges brought by system interdependency, which is a key feature of information security outsourcing that introduces new risks to the clients.[2][6]

Our model is founded on contract theory in economics, which studies how the MSSP and his clients behave based on their incentives.[1] While the prior literature in credence goods studies mechanisms to prevent *inefficient treatment*[3][4][5], our study focuses on *fraudulent diagnosis* and how clients can obtain their true risk level from the MSSP's diagnosis. In particular, we shall discuss how the MSSP decides his pricing mechanism and how it variously relates to his incentive to provide honest/dishonest diagnosis.

Section 2 presents our basic model. We start by showing that the MSSP will always charge one price to all clients of different risks, and hence his diagnosis is un-informative. Then, we introduce the self-diagnosis option to the clients and show that it incentivizes the MSSP to provide truthful diagnosis. Section 3 discusses the impact of introducing system interdependency risk. In particular, we show that in the presence of system interdependency risks self-diagnosis is insufficient to rectify the MSSP's incentive to hide the clients' risks.

2 Basic Model

We make the following assumptions in the basic model: [A1] There are n clients and one managed security service provider (MSSP). Each client values her system at v. [A2] Each client's system face a particular risk $\omega \in \{h, l\}$ decided by the nature. The probability of being high risk is r, and the probability of being low risk is $1 - r$. [A3] A high risk system will be attacked by a hacker with probability $a_h \in (0, 1)$. A low risk system will be attacked with probability $a_l \in (0, 1)$, $a_h > a_l$. [A4] The clients do not know their risk levels. The MSSP can accurately diagnose clients' risk levels. [A5] The unit cost of security protection quality q, which represents the probability of deterring an attack, is c_k for clients and c_s for MSSP, $c_k > c_s$. [A6] v, r, a_h, a_l, c_k, c_s are public information. [A7] $a_h v < c_s$. [A8] The clients cannot verify the MSSP's effort in security protection (i.e., there is no verifiability).

Assumption A7 implies complete security protection is cost-inefficient, and so it avoids a corner solution with $q^* = 1$. Assumption A8 implies that the MSSP will choose the security quality independent of the protection fee he charges.

The game begins with nature chooses clients' risk level, and the MSSP chooses capacity m and publishes contract information such as price p. After that, a client will decide whether to consult the MSSP. If not, she will directly choose protection quality in-house. Otherwise, she will visit the MSSP and receive a diagnosis. Based on the diagnosis result and offered price, the client will decide whether to accept the service. The MSSP will choose protection quality if she

[2] For example, the MSSP may exaggerate his clients' risks and over-charge them without working hard to protect them.

accept, or she will just choose protection quality in-house otherwise. After the protection quality is decided, the hacker launches attacks and outcomes are realized.

A client who does nothing in security protection will have expected utility $u_0 = (1 - \bar{a})\, v$, where $\bar{a} = ra_h + (1 - r)\, a_l$ is the expected attack rate. Suppose that a client has decided to develop security protection in-house. Her expected utility would be

$$u_k = [1 - \bar{a}\,(1 - q_k)]\, v - \frac{1}{2} c_k q_k^2, \tag{1}$$

where q_k is the security quality from in-house development. Differentiating u_k with respect to q_k, the optimal quality is $q_k^* = \frac{\bar{a}v}{c_k}$. Therefore, the expected utility of the client with in-house development is

$$u_k^* = (1 - \bar{a})\, v + \frac{1}{2} \frac{(\bar{a}v)^2}{c_k}, \tag{2}$$

which is greater than the expected utility of not protecting the system, i.e. $u_k^* > u_0$. Hence, u_k^* is the client's reservation utility.

To attract clients to use his service, the MSSP has to introduce a compensation term ("Liability") $\beta \in (0, 1]$ in the contract. If a client is attacked under the MSSP's protection and loses v, then the MSSP has to compensate her by βv. Without such compensation, by assumption A8, the MSSP can always minimize his cost by providing $q_s = 0$, which is undesirable to the clients.

2.1 One Price Solves All

We first consider the case where the MSSP charges a single price p on security protection to all clients. A client's expected utility of outsourcing to the MSSP would be

$$u_s = r\,[1 - a_h\,(1 - \beta)\,(1 - q_{s,h})]\, v + (1 - r)\,[1 - a_l\,(1 - \beta)\,(1 - q_{s,l})]\, v - p, \tag{3}$$

and the MSSP's expected profit would be

$$\pi = r\left[p - a_h\,(1 - q_{s,h})\,\beta v - \frac{1}{2} c_s q_{s,h}^2\right] + (1 - r)\left[p - a_l\,(1 - q_{s,l})\,\beta v - \frac{1}{2} c_s q_{s,l}^2\right]. \tag{4}$$

Differentiating π with respect to $q_{s,h}$ and $q_{s,l}$, the optimal quality is $q_{s,h}^* = \frac{a_h \beta v}{c_s}$ and $q_{s,l}^* = \frac{a_l \beta v}{c_s}$. By backward induction, the client's expected utility of outsourcing becomes

$$u_s = (1 - \bar{a})\, v + \bar{a}\beta v + \frac{\left(\bar{a}^2 + \sigma_a^2\right) v^2}{c_s} \beta\,(1 - \beta) - p, \tag{5}$$

where $\sigma_a^2 = r\,(1 - r)\,(a_h - a_l)^2$ is the variance of the attack rate. Substituting $q_{s,h}^*$ and $q_{s,l}^*$ into (4), the MSSP's profit maximization problem becomes:

$$\max_{p,\beta} \left[p - \bar{a}\beta v + \frac{1}{2} \frac{\left(\bar{a}^2 + \sigma_a^2\right) (\beta v)^2}{c_s}\right]$$

$$\text{s.t. } p \leq \bar{a}\beta v + v^2 \left[\frac{\beta(1-\beta)\left(\bar{a}^2+\sigma_a^2\right)}{c_s} - \frac{1}{2}\frac{\bar{a}^2}{c_k} \right].$$

The price constraint ensures that the clients are not worse off after using the MSSP's service, i.e. $u_s \geq u_k^*$. The optimal solution is $\beta^* = 1$, $p^* = \bar{a}v - \frac{1}{2}\frac{(\bar{a}v)^2}{c_k}$, and $\pi^* = \frac{v^2}{2}\left(\frac{\bar{a}^2+\sigma_a^2}{c_s} - \frac{\bar{a}^2}{c_k}\right) > 0$.

Will the MSSP price discriminate, i.e., offer p_h to high risk clients and p_l to low risk clients? It turns out that he will not. If he sets the prices honestly, the clients will learn their own risk levels from the MSSP's diagnosis and pricing. This will help the clients select the proper q_k with respect to their risk levels, which would increase their reservation utility and so decrease the MSSP's profit. On the other hand, if the MSSP "cheats" the clients on pricing, then they could always maximize their utility by only accepting a low price, $p_l \leq p^*$.

From the above reasoning, we propose that the MSSP will prefer to offer a single price contract in the information security outsourcing market:

Proposition 1. In information security outsourcing, setting a single price contract with liability term, which does not reveal any risk information of the clients, is optimal for the MSSP.

Note that the low risk clients are worse off because they will be subsidizing the high risk clients. Therefore, the MSSP will tend to exaggerate clients' risk to encourage them to use his service. Once the clients recognize this fact, they will probably ignore the MSSP's recommendation and protect their own system using the average quality. This results in either over-protected for low risk clients or under-protected for high risk clients, which makes the system less usable.

We next consider the case when the clients can seek alternative diagnosis (we call this "self-diagnosis").

2.2 Self-diagnosis

With self-diagnosis, we assume that the clients can pay d_k to a third-party consultant to reveal her risk. After self-diagnosis, the clients can treat themselves using the corresponding quality, i.e. $q_{k,h}^* = \frac{a_h v}{c_k}$ and $q_{k,l}^* = \frac{a_l v}{c_k}$. The reservation utility of a client with risk level ω after self-diagnosis and self-treatment will be

$$u_{k,\omega}^{d*} = (1-a_\omega)v + \frac{1}{2}\frac{(a_\omega v)^2}{c_k} - d_k. \tag{6}$$

Further, the client will choose between in-house protection and outsourcing, depending on which option gives more utility, after self-diagnosis. Therefore, the minimum expected utility of a client after self-diagnosis would be

$$u_k^{d*} = (1-\bar{a})v + \frac{1}{2}\frac{(\bar{a}v)^2}{c_k} + \frac{1}{2}\frac{(\sigma_a v)^2}{c_k} - d_k. \tag{7}$$

If $d_k \leq \frac{1}{2}\frac{(\sigma_a v)^2}{c_k}$, then $u_k^{d*} \geq u_k^*$, which means that it is efficient for clients to seek self-diagnosis. We will assume that such a condition holds in the following analysis.

We first consider the case where the MSSP charges different prices for different types of clients. Suppose that the MSSP charges honestly, i.e., offering p_h to high risk clients and p_l to low risk clients. This situation is similar to serving two market segments with $r = 1$ and $r = 0$, which is both profitable. Therefore, he will serve both types of clients with the following profit maximization problem:

$$\max_{p_h, p_l, \beta} r \left[p_h - a_h \beta v + \frac{1}{2} \frac{(a_h \beta v)^2}{c_s} \right] + (1-r) \left[p_l - a_l \beta v + \frac{1}{2} \frac{(a_l \beta v)^2}{c_s} \right],$$

$$\text{s.t. } u_{s,h} \geq u^*_{k,h},\ u_{s,l} \geq u^*_{k,l}.$$

The constraints show that the MSSP charges the clients honestly so that a client with a particular type of risk will not be worse off. The solution is $\beta^* = 1$, $p^*_h = a_h v - \frac{1}{2} \frac{(a_h v)^2}{c_k} > p^*_l = a_l v - \frac{1}{2} \frac{(a_l v)^2}{c_k}$, and $\pi^* = \frac{(\bar{a}^2 + \sigma^2_a) v^2}{2} \left(\frac{1}{c_s} - \frac{1}{c_k} \right)$.

However, the MSSP has incentive to overcharge the low risk clients with p^*_h, which is the main reason that price discrimination is unsustainable in the case without self-diagnosis. A client will always accept p^*_l since it is beneficial for either type, and self-diagnose only when p^*_h is offered. By doing so, the clients can punish a dishonest MSSP by turning down the p^*_h offer and do in-house protection instead. Therefore, the MSSP will earn nothing if he overcharges the clients, and the clients know it.

We next consider the possibility of a mixed self-diagnosis strategy. To construct such a strategy, consider the profit of serving a low risk client with low price:

$$\pi_{l,p^*_l} = \frac{(a_l v)^2}{2} \left(\frac{1}{c_s} - \frac{1}{c_k} \right), \tag{8}$$

and the profit of serving a low risk client with a high price:

$$\pi_{l,p^*_h} = (1-\rho) \left[\frac{(a_l v)^2}{2} \left(\frac{1}{c_s} - \frac{1}{c_k} \right) + (a_h - a_l) v \left(1 - \frac{1}{2} \frac{a_h v}{c_k} - \frac{1}{2} \frac{a_l v}{c_k} \right) \right], \tag{9}$$

where ρ is the probability of self-diagnosis when a client was offered a high price ("Re-diagnosis Rate"). An effective mixed strategy should result in $\pi_{l,p^*_h} \leq \pi_{l,p^*_l}$, which gives rise to the re-diagnosis rate:

$$\rho \geq \frac{(a_h - a_l) \left(1 - \frac{1}{2} \frac{a_h v}{c_k} - \frac{1}{2} \frac{a_l v}{c_k} \right)}{(a_h - a_l) \left(1 - \frac{1}{2} \frac{a_h v}{c_k} - \frac{1}{2} \frac{a_l v}{c_k} \right) + a^2_l v \left(\frac{1}{c_s} - \frac{1}{c_k} \right)}. \tag{10}$$

The client would maximize her utility by minimizing the re-diagnosis rate, and so the equality holds for (10) in equilibrium. This re-diagnosis rate removes the MSSP's incentive to cheat and supports the price discrimination equilibrium.

We now consider the case where the MSSP charges a single price p to all clients. If all client prefer to self-diagnose, then at least one type of clients would benefit from using the revealed risk information for in-house treatment. So, the

MSSP can earn more by serving both types of clients with price discrimination. On the other hand, the clients would prefer to use the MSSP's service directly without self-diagnosis if and only if the MSSP charges them a low price. But, by doing so he will get sub-optimal profit because he is practically giving out surplus to high risk clients.

From the above discussion, the MSSP could earn more profit by price discrimination. Hence, the self-diagnosis option removes the MSSP's incentive to conceal risk information.

Proposition 2. With a cheap self-diagnosis option, the MSSP will truthfully reveal the clients' risk information.

When the MSSP's diagnosis result is verifiable at a low cost, clients can actually learn from the MSSP's recommendation. As a result, they can protect their own system according to their own risk, so that the systems are secured without losing usability. However, in reality, different systems are often interconnected to address users' need, which introduce new challenges. In the next section, we will examine how system interdependency risks affect the current situation.

3 System Interdependency Model

We add the following assumption to extend the basic model with system interdependency risks: [A9] A client who joined the MSSP's network will lose εv if at least one other system in the MSSP's network is compromised. The MSSP needs to compensate $\beta\varepsilon v$ to all affected clients who are not directly attacked.

Consider the MSSP's network with m clients. The probability of at least one system being attacked is

$$P_{X>0} = 1 - \prod_{i=1}^{m_h} \left[1 - a_h \left(1 - q_{s,h,i}\right)\right] \prod_{i=1}^{m_l} \left[1 - a_l \left(1 - q_{s,l,i}\right)\right], \tag{11}$$

where m_h is the number of high risk clients in the network, m_l is the number of low risk clients in the network, $m_h + m_l = m$. The loss of a client j with risk level ω will be $L_{\omega,j}v = a_\omega \left(1 - q_{s,\omega,j}\right)(1 - \varepsilon) v + \varepsilon v P_{X>0}$. Since the loss involves m-th order terms, to simplify the analysis, we approximate it by only retaining the first order terms:

$$\tilde{L_{\omega,j}} = a_\omega \left(1 - q_{s,\omega,j}\right)(1 - \varepsilon) + \varepsilon \left[\sum_{i=1}^{m_h} a_h \left(1 - q_{s,h,i}\right) + \sum_{i=1}^{m_l} a_l \left(1 - q_{s,l,i}\right)\right]. \tag{12}$$

3.1 Without Self-diagnosis

Suppose that the MSSP charges p to all clients. The expected utility of client j who uses the MSSP's service would be

$$u_{s,j} = r \left[\left(1 - L_{h,j}\right) v + L_{h,j}\beta v - p\right] + (1 - r) \left[\left(1 - L_{l,j}\right) v + L_{l,j}\beta v - p\right], \tag{13}$$

and the MSSP's expected total profit would be

$$\pi = \sum_{i=1}^{m_h} \left(p - L_{h,i}\beta v - \frac{1}{2} c_s q_{s,h,i}^2 \right) + \sum_{i=1}^{m_l} \left(p - L_{l,i}\beta v - \frac{1}{2} c_s q_{s,l,i}^2 \right). \tag{14}$$

Differentiating π with respect to $q_{s,h,i}$ and $q_{s,l,i}$, the optimal quality is $q_{s,h,i}^* = \frac{Ta_h\beta v}{c_s}$ and $q_{s,l,i}^* = \frac{Ta_l\beta v}{c_s}$, where $T = 1 + \varepsilon\,(m-1)$ is the (amplified) risk factor due to system interdependency.

The expected utility of outsourcing the protection would then become

$$u_s = \left[1 - \bar{L}\,(1-\beta)\right] v - p, \tag{15}$$

where $\bar{L} = T\bar{a} - \frac{T^2\left(\bar{a}^2+\sigma_a^2\right)\beta v}{c_s}$ is the expected loss after outsourcing. Now, suppose that the MSSP is committed to serve a client after diagnosis, which means that he cannot freely choose m_h and m_l. In a network with m clients, the expected number of high risk clients would be $E\left[m_h\right] = rm$, and the expected number of low risk clients would be $E\left[m_l\right] = (1-r)\,m$. Therefore, the MSSP's profit maximization problem becomes

$$\max_{p,\beta,m} m \left[p - \bar{L}\beta v - \frac{1}{2} \frac{\left(\bar{a}^2+\sigma_a^2\right)\left(T\beta v\right)^2}{c_s} \right]$$

$$\text{s.t. } p \le \left(\bar{a} - \bar{L}\right) v + \bar{L}\beta v - \frac{1}{2}\frac{\left(\bar{a}v\right)^2}{c_k}.$$

The solution is $\beta^* = 1$, $p^* = \bar{a}v - \frac{1}{2}\frac{\left(\bar{a}v\right)^2}{c_k}$, and the number of clients served by the MSSP satisfies the following equation: $m^* = \frac{1}{2} + \frac{E[aq_s^*]v - \frac{1}{2}c_s E\left[q_s^{*2}\right] - \frac{1}{2}\frac{\left(\bar{a}v\right)^2}{c_k}}{2\varepsilon v\left(\bar{a} - E[aq_s^*]\right)}$, where $E\left[aq_s^*\right] = ra_h q_{s,h}^* + (1-r)\,a_l q_{s,l}^*$ and $E\left[q_s^{*2}\right] = rq_{s,h}^{*2} + (1-r)\,q_{s,l}^{*2}$.

If the MSSP can freely choose m_h and m_l, when he will charge p^*, a low risk client who uses the MSSP's service will be subsidizing the high risk clients. Therefore, the optimal decision for the MSSP is to serve only the low risk clients in equilibrium, and get the subsidies as profit.

However, once the clients realize this, they will demand for a lower price since p^* is not a desirable price for low risk clients. Therefore, the MSSP can no longer charge p^* if he does not commit to serve the clients, which results in sub-optimal profits.

What if the MSSP sets different prices for different clients? Since the interdependency risk limits the MSSP's capacity, and serving a high risk client with p_h^* is more profitable compared with serving a low risk client with p_l^*, the MSSP will prefer to serve only the high risk clients. Specifically, the optimal decision for capacity satisfies $m_l^* = 0$ and $m_h^* = \frac{1}{2} + \frac{a_h v q_{s,h}^* - \frac{1}{2}c_s q_{s,h}^{*2} - \frac{1}{2}\frac{\left(a_h v\right)^2}{c_k}}{2\varepsilon a_h v\left(1 - q_{s,h}^*\right)}$. Therefore, the MSSP has great incentive to overcharge the low risk clients, since kicking out a low risk client is not a problem.[3] Hence, the MSSP always prefers to offer

[3] If a low risk client accepts p_h, the MSSP can earn even more since the required protection level and the interdependency risk brought by this client is lower.

a high price p_h, which means that posting two prices cannot be an equilibrium strategy.

Yet, clients will only accept a low price p_l, and they will suspect that they get overcharged when p_h is offered. These competing strategies cause the market to breakdown.

If the MSSP uses a mixed strategy and offers p_h and p_l sometimes, clients will only accept when p_l is offered, which result in sub-optimal profit compared with the case of using single price with service commitment.

From the above discussion, if the MSSP does not commit to serve every clients, then he will end up serving only one type of clients. This reveal the clients' risk information, and hence result in sub-optimal profit, or even market breakdown. Therefore, the MSSP will prefer to charge a single price and commit to serve every client, which leads to a similar outcome as Proposition 1.

3.2 With Self-diagnosis

Continue from the above discussion, when the MSSP posts two prices, he is committed to serve any clients with p_h. However, with self-diagnosis, the clients can verify whether they really get overcharged. Hence, the clients who learn that they have a high risk from self-diagnosis will continue to use the MSSP's service.

The market will not breakdown and the MSSP's aggressive pricing strategy is actually "resurrected" by self-diagnosis. The MSSP has no incentive to deviate from this strategy, since offering $p_l < p_h$ will result in sub-optimal profit.

Therefore, even when self-diagnosis is feasible, the credence goods problem still remains when system interdependency is present.

Proposition 3. In the presence of system interdependency, when there are sufficient high risk clients in the market and the clients can cheaply self-diagnose, then the MSSP will always charge a single price p_h, and only high risk clients will use the MSSP's service. In other words, self-diagnosis will not dissuade the MSSP's from concealing the clients' risk information.

In this situation, low risk clients are rejected by the MSSP, so that they cannot enjoy a better protection. Even worse, every clients need to verify the MSSP's diagnosis, which results in duplication of diagnosis cost.

4 Final Remarks

The typical credence goods problem is often solved by introducing verifiability of the service provider's efforts. Here, we show that by introducing verifiability in the MSSP's diagnosis (which is done by self-diagnosis), the MSSP will truthfully reveal the clients' risks in the basic setting. However, when we introduce system interdependency risks into the model, the MSSP will have incentives to exaggerate clients' risks and offer a high price, which seems common in reality. This brings challenges to organizations that want to learn their risk level and avoid constantly over-paying for security protections. In future work we shall study alternative mechanisms that can tackle this challenge.

References

1. Akerlof, G.A.: The market for “lemons”: Quality uncertainty and the market mechanism. The Quarterly Journal of Economics 84(3), 488–500 (1970)
2. Anderson, R., Moore, T.: The economics of information security. Science 314(5799), 610–613 (2006)
3. Dulleck, U., Kerschbamer, R.: On doctors, mechanics, and computer specialists: The economics of credence goods. Journal of Economic Literature 44(1), 5–42 (2006)
4. Emons, W.: Credence goods and fraudulent experts. The Rand Journal of Economics 28(1), 107–119 (1997)
5. Fong, Y.: When do experts cheat and whom do they target? RAND Journal of Economics 36(1), 113–130 (2005)
6. Kunreuther, H., Heal, G.: Interdependent security. Journal of Risk and Uncertainty 26(2), 231–249 (2003)

Appendix: Proof of Propositions

Proof of Proposition 1

We first derive the equilibrium profit under single price contract. The Lagrange function of the profit maximization problem is

$$\varLambda = p - \bar{a}\beta v + \frac{\left(\bar{a}^2 + \sigma_a^2\right)(\beta v)^2}{2c_s} - \lambda\left\{p - \bar{a}\beta v - v^2\left[\frac{\beta(1-\beta)\left(\bar{a}^2+\sigma_a^2\right)}{c_s} - \frac{\bar{a}^2}{2c_k}\right]\right\} \tag{16}$$

where $\lambda \geq 0$. The first order conditions are:

$$\frac{\partial \varLambda}{\partial p} = 1 - \lambda = 0 \tag{17}$$

$$\frac{\partial \varLambda}{\partial \beta} = \frac{\left(\bar{a}^2 + \sigma_a^2\right)v^2}{c_s}(\beta - 2\beta\lambda + \lambda) - \bar{a}v(1-\lambda) \tag{18}$$

and the Kuhn-Tucker condition is:

$$-\lambda\left\{p - \bar{a}\beta v - v^2\left[\frac{\beta(1-\beta)\left(\bar{a}^2+\sigma_a^2\right)}{c_s} - \frac{1}{2}\frac{\bar{a}^2}{c_k}\right]\right\} = 0 \tag{19}$$

Solving the above equations, we have $\lambda = 1$, $\beta^* = 1$, and $p^* = \bar{a}v - \frac{1}{2}\frac{(\bar{a}v)^2}{c_k}$. Substitute them back to (4) and (5) yields the client’s expected utility and the MSSP’s expected profit:

$$u_s^* = (1-\bar{a})v + \frac{1}{2}\frac{(\bar{a}v)^2}{c_k} = u_k^* \tag{20}$$

$$\pi^* = \frac{v^2}{2}\left(\frac{\bar{a}^2+\sigma_a^2}{c_s} - \frac{\bar{a}^2}{c_k}\right) \tag{21}$$

We then prove that price discrimination is sub-optimal. Suppose the MSSP charges two price honestly, then clients can infer their risk information and use it to decide in-house protection quality. The reservation utility of a client with risk level ω is

$$u_{k,\omega}^{**}=\left(1-a_{\omega}\right)v+\frac{1}{2}\frac{\left(a_{\omega}v\right)^{2}}{c_{k}} \tag{22}$$

which can be obtained by considering a degenerated market with $r=1$ and $r=1$ on (2). Hence, the overall expected reservation utility of a client will be

$$u_{k}^{**}=\left(1-\bar{a}\right)v+\frac{1}{2}\frac{\left(\bar{a}^{2}+\sigma_{a}^{2}\right)v^{2}}{c_{k}} \tag{23}$$

Since the overall reservation utility u_k^{**} is increased, and the total welfare between a client and the MSSP does not change, the MSSP's profit is decreased. Specifically, it becomes:

$$\pi^{**}=\frac{\left(\bar{a}^{2}+\sigma_{a}^{2}\right)v^{2}}{2}\left(\frac{1}{c_{s}}-\frac{1}{c_{k}}\right) \tag{24}$$

By comparing (21), (22), (24) and (25), part of the MSSP's surplus $\frac{1}{2}\frac{(\sigma_{a}v)^{2}}{c_{k}}$ moves towards the client. Therefore, offering a single price is optimal for the MSSP.

Proof of Proposition 2

We first discuss the way to obtain the equilibrium profit, which is basically applying the result in Proposition 1. Consider the MSSP serves two different market with $r=1$ and $r=0$, and substitute them into the equilibrium profit from Proposition 1, i.e. (22). By taking the weighted average, we can obtain the equilibrium profit:

$$\pi^{*}=\frac{\left(\bar{a}^{2}+\sigma_{a}^{2}\right)v^{2}}{2}\left(\frac{1}{c_{s}}-\frac{1}{c_{k}}\right) \tag{25}$$

We then show that the MSSP will not stick on offering a single price in the equilibrium. Firstly, it is trivial to see that if not all clients uses the MSSP's service, his profit will be sub-optimal and he can increase it by price discrimination. Secondly, If every clients uses the MSSP's service, the total welfare the MSSP and a client will be

$$W=r\left\{\left[1-a_{h}\left(1-q_{s,h}^{*}\right)\right]v-\frac{c_{s}q_{s,h}^{*2}}{2}\right\}+\left(1-r\right)\left\{\left[1-a_{l}\left(1-q_{s,l}^{*}\right)\right]v-\frac{c_{s}q_{s,l}^{*2}}{2}\right\} \tag{26}$$

which is obtained by applying the MSSP's cost c_s into clients' problem. Note that the total welfare $W=u_s+\pi$ in this case. From previous analysis, the optimal quality for the MSSP will always be $q_{s,h}^{*}=\frac{a_{h}\beta v}{c_{s}}$ and $q_{s,l}^{*}=\frac{a_{l}\beta v}{c_{s}}$. Hence, (27) could be re-written as:

$$W=\left(1-\bar{a}\right)v+\left(\bar{a}^{2}+\sigma_{a}^{2}\right)v^{2}\left[\frac{\beta\left(2-\beta\right)}{2c_{s}}\right] \tag{27}$$

A client will only use the MSSP's service when $u_s \geq u_k^{d*}$. Hence, the MSSP's profit will be

$$\pi \leq \left(\bar{a}^2 + \sigma_a^2\right) v^2 \left[\frac{\beta\left(2-\beta\right)}{2c_s} - \frac{1}{2c_k}\right] + d_k \tag{28}$$

The equality holds when $u_s = u_k^{d*}$, which means the MSSP extracts all surplus from clients. And the right hand side of (29) reaches the maximum when $\beta = 1$. However, $u_s = u_k^{d*}$ and $\beta = 1$ are contradicting. In order to have $u_s = u_k^{d*}$, the price p must satisfy the following:

$$p = a_h \beta v - \frac{1}{2}\frac{\left(a_h v\right)^2}{c_k} + \frac{\left(a_h v\right)^2}{c_s}\beta\left(1-\beta\right) \tag{29}$$

$$p = a_l \beta v - \frac{1}{2}\frac{\left(a_l v\right)^2}{c_k} + \frac{\left(a_l v\right)^2}{c_s}\beta\left(1-\beta\right) \tag{30}$$

Since $a_h > a_l$ and $a_h v < c_k$, $\beta = 1$ cannot solve both (30) and (31) together. Therefore, a sufficient small d_k would guarantee that the profit of offering single price is smaller than that of offering two different prices. Hence, the MSSP will offer two prices honestly and it solves the credence good problem.

Sorry, I Don't Get It: An Analysis of Warning Message Texts

Marian Harbach, Sascha Fahl, Polina Yakovleva, and Matthew Smith

Distributed Computing and Security Group, Leibniz University Hannover
Schlosswender Str. 5, 30159 Hannover, Germany
{harbach,fahl,yakovleva,smith}@dcsec.uni-hannover.de

Abstract. Security systems frequently rely on warning messages to convey important information, especially when a machine is not able to assess a situation automatically. There is a significant body of work studying the effects of warning message design on users with numerous suggestions on how to optimise their effectiveness. Design guidelines and best practises help the developer to display urgent information. In this paper, we present the first empirical analysis on the extent of the influence of linguistic properties on the perceived difficulty of the descriptive text in warning messages. We evaluate warning messages extracted from current browsers and present linguistic properties that can improve a warning message text's perceived difficulty. Our results confirm that, while effects of attention, attitude and beliefs are at least as important as the linguistic complexity of the text, several steps can be taken to improve the text's difficulty perceived by the user.

Keywords: Usable Security, Comprehension, Warning Messages, Readability.

1 Introduction

Designing and writing warning messages can be considered a form of art. In the past, users and IT professionals alike were confused by complicated warning and error messages that seemed to consist of only hex numbers and stack traces, such as the famous "blue screen of death". A considerable amount of work has continuously improved the quality of warning messages for many different applications and proposed guidelines on how to compose useful and understandable dialogues (e. g. [3,6,13]). However, users still seem to struggle with warnings on a regular basis, suggesting that there are still open problems in creating understandable and helpful warning messages.

The reception of warning messages by a user is often explained using Wogalter's Communication-Human Information Processing (C-HIP) model [15] or Cranor's extension of C-HIP: the human-in-the-loop (HITL) framework [4]. In both models, information is conveyed from a source through a channel to a human receiver. At the receiving end, the information first needs to gain sufficient attention before the information enters the comprehension stage. Afterwards, attitudes and beliefs as well as motivation further influence the information before

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 94–111, 2013.

the processing results in behaviour. A lot of work has been put into optimising colours, fonts, symbols and icons to attract attention and facilitate reception.

In this paper, we investigate the comprehension stage: does the structural composition (syntax and vocabulary) of a warning message's text influence the user's overall perception and support comprehension? Or, in other words: if a user chose to read a warning message, would he or she be able to extract the necessary information and find the text easy to parse and understand?

It has been recognised that the descriptive text provided in warning messages needs to convey important information about the problem and be understandable by most computer users at the same time. In 2011, Bravo-Lillo et al. [3] compiled a set of design guidelines and present rules for descriptive text, including:

- "describe the risk; describe consequences of not complying; provide instructions on how to avoid the risk;"
- "be brief; avoid technical jargon."

However, these guidelines are hard to quantify, especially since there is no example of a perfect warning message to date. Thus, judging whether or not the requirements and advice of the guidelines are sufficiently met usually needs an expert's opinion or dedicated testing through user studies. Consequently, there is considerable effort and knowledge involved in analysing and optimising warning messages. Small development efforts, such as start ups or app developers, often do not have the resources to thoroughly analyse the warning messages used in their products. They could benefit from more concrete and possibly objectively testable instructions on how to create useful warning messages in particular.

This paper investigates several methods to automatically assess warning message texts and analyses to which extent linguistic properties in general influence the user's perceptions of a warning message text. We will present an evaluation of existing readability measures on current browser security warnings as well as four empirical studies to assess the user's perceptions. Our results indicate that existing warning messages are too hard to read for the average user and that particular sentence structures as well as technical terms, which can be found in indexes of computer security textbooks, significantly correlate with the perceived difficulty of warning messages.

To the best of our knowledge, there has not been any work that empirically investigates the role of text comprehension and readability for computer warning messages to date. This work expands on preliminary results that have provided an overview of warning message readability using existing measures [8].

We offer three main contributions:

- We validate whether or not existing readability measures are suitable to judge warning message texts and determine the linguistic difficulty of existing warning messages.
- We investigate the effect of linguistic properties of warning message texts on the users' perceptions and provide empirical evidence for the influence of grammatical structures and vocabulary on warning message comprehension.

- We present quantifiable properties of text that influence warning message readability and comprehension.

The remainder of this paper is structured as follows: First, we introduce related work, before summarising readability of browser messages using a set of existing readability measures and an analysis thereof. Section 4 reports on the results of a user study that assess the applicability as well as the results of these readability measures. Sections 5 and 6 describe two online studies, collecting users' ratings of warning messages and comparing them to several linguistic properties. Additionally, effects of translation and a comparison between different software products are presented. Section 7 presents the results of interviews that discussed particular problems on a word and sentence level with users. Section 8 discusses limitations before Section 9 finally summarises the implications of our results and concludes this paper.

2 Related Work

A considerable amount of research has investigated warnings in the digital realm. Cranor's Human-In-The-Loop (HITL) framework [4] is a specialisation of Wogalter's C-HIP model [15] and describes how interactions between computers and humans can cause security problems.

Egelman et al. [6] presented a first study on warning efficacy for phishing prevention in 2008. They found that a large part of their test subjects chose to heed warnings that required interaction from the user and offer guidelines to improve warnings. According to their results, effective warnings need to interrupt the primary task, provide clear choices, fail safely and prevent habituation.

In a similar fashion, Sunshine et al. [13] tested the efficacy of certificate warnings presented by browsers and tried to improve the state of the art by modifying colours based on context and providing more detailed and interactive information on risks. While their changes improved efficacy, they concluded that the warnings still leave users vulnerable to man-in-the-middle attacks. Maurer et al. [9] also showed that warnings based on user input data types can help to prevent phishing and decrease habituation by increasing the context of a warning.

Bravo-Lillo et al. [3] provided another perspective on improving warning messages. They found that design changes can improve understanding and motivation but also realised that warning messages were not able to help users to differentiate between low and high-risk situations. Understanding and motivation were also found to be strongly connected and important factors in safely responding to warnings. Additionally, Bravo-Lillo et al. [2] offer qualitative insight into warning assessment by users of different skill levels and conclude that all aspects of warning design need to be considered in order to improve warnings. They also explicitly mention that the process of reading a warning is a central concern for warning message reception.

In another line of work, previous research has empirically investigated readability issues of end-user license agreements [7] and found shortcomings in informing the user before demanding consent.

The related work has conferred many valuable insights into the effectiveness and design of warning messages as well as problems with readability. We hope that the analyses presented in this paper complement the existing results by investigating the role of linguistic properties for the comprehension of warning message texts.

3 Readability Measures

In a previous publication [8], we explored the application of readability measures from the domain of educational psychology for computer warning messages. These measures take a piece of text and predict a level of reading skill necessary for comprehending the contents. For example, obtaining a value of 11 from a readability measure, such as SMOG [10], for a piece of text implies that an average reader needs to have the reading level of a student in 11th grade to be able to process the linguistic structure of this text. It is important to note that readability measures do not address the semantic difficulty of a text, but focus on linguistic difficulty, which is related to complicated sentence construction, long or polysyllabic words and similar properties. However, a text can be deemed to be "readable" using a certain measure but still confuse a reader. Yet, the linguistic difficulty is an important precursor for the overall comprehension of a text and therefore a useful indicator. If readability, as obtained from a suitable readability measure, is bad, the semantic information is harder to extract. *In the remainder of this paper, we generally address linguistic difficulty as described above, as opposed to semantic difficulty or other aspects of text layout, such as typesetting.*

Previously, we presented an analysis of security warnings based on warning messages from the two most common open-source browsers, Google Chrome and Mozilla Firefox. We extracted 24 English warning texts (15 for Chrome, 9 for Firefox) and added another four certificate warnings (hostname verification or unknown root CA warnings) from Internet Explorer 8, Safari, Outlook and iTunes to our sample to offer a broader cross-product comparison for a particularly common warning message. Warnings include certificate and phishing warnings, as well as messages indicating connectivity problems or unreachable servers. We also collected the same warnings in German. The selected warnings have at least about 50 words, because the readability measures we used are not validated for shorter samples of text. An abbreviated list of the warnings can be found in Appendix A.

We found that the predicted reading skills for this set of warnings differ depending on which measure is applied. However, all measures suggested at least an average reading level of an eighth grade student, while the SMOG measure, which is most suitable for warning messages due to its construction, even predicted the reading level of a first year college student for the average warning message. Details can be found in [8]. The extent to which these values are appropriate and useful is discussed in the following section.

4 Exploratory Study

To validate the readability results described above, we conducted an exploratory study of readability and linguistic comprehension. In order to minimise the effects of differences in language skills, we decided to test only native speakers. Since the study was conducted in Germany, we used the German versions of the set of 28 warning messages introduced above.

4.1 Design

Participants took a standard reading ability test to judge their individual reading level (Metze's "Stolperwoerter" test [1]). Next, they were presented with a cloze test (a piece of text where every fifth word is removed and has to be filled in by the participant) on six selected warning messages and scored based on their success rate. Cloze tests are commonly used as comprehension tests for the construction of the existing readability measures [5]. We selected four German warnings from Chrome and two from Firefox, since their readability scores (Amstad's measure for German texts) were distributed across the range we found in the tests described above. We stripped the warnings of all identifying and distracting features, using the same font and background for all messages. We introduced a fictitious browser named *InterBrowse*, as well as a fictitious banking website *mybank.com*, and replaced all references to the original software and websites with these names. Participants were given a simple working scenario stating that they were trying to surf to www.mybank.com using InterBrowse and then encountered a warning. We also reminded them that we intended to test the messages and not the participants' performance. After completing the cloze tests, participants re-read the full messages and sorted the texts by their feeling of comprehension. We pre-tested our protocol in a laboratory setting, discussed in previous work [8].

4.2 Participants

Based on this study protocol, we invited 1,486 students on a university-wide mailing list to participate in an online study. We advertised a study on browsing behaviour that would take 20 to 25 minutes and offered participants the chance to win a lottery of two 100 € Amazon vouchers as motivation. We received 311 complete responses, after removing non-native speakers and respondents with IT-related majors. The participants' average age was 22.8 and 130 came from the faculty of arts (cf. Table 2 in the Appendix). Technical experience among our participants was rather high, with an average of 2.29 on a scale from 1 (high expertise) to 5. Upon completion of the tasks, 216 participants (69.5 %) reported that they had seen one of the six warnings before and 49 (15.8 %) were unsure.

4.3 Results

For each participant, we collected the Stolper score, i. e. the individual's reading level, the cloze performance, i. e. how many of the gaps in the text were filled in correctly, the time taken per cloze text, and each participant's ranking in terms of subjective readability of the six presented warning messages, i. e. which messages did the participant find harder or easier to read and understand. Cloze performance was automatically assessed using a Levenshtein distance of 3 on the provided answers. Therefore, a word in a gap was counted as correct if the edit distance was equal or less than 3 compared to the original word, accounting for typos. This approach was chosen over an individual assessment of the semantics of the provided solution, since manually assessing each solution would have been too time consuming and could have biased results due to subjective scoring. To compensate for this strict assessment of performance, we chose a lower criterion score (see below).

We found significant differences in the cloze test performances between participants with high or low technical expertise. Since the cloze performances were found to be non-normally distributed (Kolmogorov-Smirnov Z between 1.579 and 2.862, $p < .031$ in all cases), we applied the Mann-Whitney U test and found significant differences in all messages (U between $5,762$ and $6,344$, Z between -2.301 and -3.144 with $p < 0.05$) except one (Message 6, $U = 7,595.5$, $Z = -.493$ and $p = .622$). While all other messages received higher scores from high-expertise participants, this particular message took the longest time to complete on average and received similar scores from both groups. The seldom seen message was about the use of a weak signature algorithm in a certificate and might therefore have been perceived as equally complicated by high- and low-expertise participants. Interestingly, this message also received the best average performance across all warnings, which suggests that complicated messages can be understood if enough time is spent.

In our reading ability test (Stolper-Test), the 311 respondents achieved an average score of 77.85 % ($sd = 17.95$), which is above average for their age group. The average score for participants between 21 and 25 years is 70.7 % and for people of 26 years and older is even lower (66 %), according to [1]. This effect can be explained by the above-average education of students.

Readability Results. Using the participants' reading abilities, we calculated readability scores for each of the six tested warnings to compare with existing measures. This procedure was adopted from the original construction of several other readability measures which use cloze tests on passages of selected texts to derive the readability formula through regression [5]. The scores are based on a criterion score or threshold of correct answers on the corresponding cloze test. A criterion score of 90 % or higher is necessary for important information that needs to be well understood by readers [5,8]. However, since cloze performance was automatically assessed, we chose a criterion score of 70 % to account for synonyms. Using this criterion, we calculated readability scores for the six warnings as the average reading level (Stolper score) of participants that performed better than the criterion score on a particular warning message. Therefore, lower values for the readability score indicate higher readability.

According to the results (cf. Table 3 in the Appendix), our score correlates highly with the number of words in a message ($\rho = .943$, $p = .005$). While there are no other significant correlations due to the small sample size, we found indications of potential correlations with Amstad ($\rho = .714$, $p = .111$) and LIX ($\rho = -.600$, $p = .208$) scores. However, the implied direction of correlation is conflictive: These numbers suggest that better readability according to our Stolper-score-based measure is connected with worse readability according to Amstad and LIX. We could not find a significant correlation with the participants' rankings of messages either.

Because of the small number of warnings in this exploration, we cannot generally reject the applicability of readability measures for warning messages. However, the results suggest that the existing measures for German texts (i. e. the Amstad and LIX scores) do not fit the scores we collected directly from participants.

Another important trend is that for those students achieving 70 % or more correct answers in cloze testing, the mean reading ability is considerably higher ($> 79\,\%$) than the average score in their age group and older age groups ($66 - 70\,\%$). This implies that the average person would find these warnings hard to read.

The results also suggest that the readability scores we derived from Stolper scores somewhat mirror the participants' perceptions: scores are higher for messages rated as having the best subjective readability and lower scores for those perceived as worst. Another interesting implication of our results is that we did not find any correlation at all between the existing readability measures for German texts and the participants' subjective ratings of warning comprehension. The next section investigates this further.

5 Rating Study

The study described in the previous section focused on gaining direct measurements of text readability to evaluate the applicability of readability measures. The results suggest that the readability scores obtained from existing measures may not mirror the participants' perceptions of warning messages.

With the study presented in this section, we aimed to gather how easy people perceive understanding a warning message text to be. If a text is easier to read, the problem of users not reading or skimming warning messages might be alleviated. Therefore, we collected user ratings for the 28 warning messages introduced in Section 3. Again, we used the German versions of the texts and tested native speakers, to minimise effects of language skill levels.

5.1 Design

We prepared an online survey that presented each participant with six out of our set of 28 warning messages. Participants were primed with the same scenario as in the previous study. The order and selection of the messages was randomised for each participant. For each warning message, participants were asked to read the message, to summarise the contents roughly in one sentence and then rate their perception of the warning message with four items on a 7-point scale from "I completely agree" to "I completely disagree". The items addressed comprehension of the entire message, the words used in the message, previous exposure and understanding of why the message appears. We also added two additional items, which were semantically inverse to two items in the original set. Before starting the rating exercise, we asked participants an attention question, that required participants to answer "No" even though the correct answer was obviously "Yes". At the end of the survey, we collected demographics.

5.2 Participants

We invited 1,522 students of the same mailing list[1] to participate in the survey. The study was advertised as a follow-up of the previous study that would take 8 to 12 minutes to complete, welcoming new and returning participants. Once again, we offered participation in a lottery for two 50 € Amazon vouchers as compensation. 250 participants successfully completed the survey. First, we removed participants that wrongly answered the attention question with "Yes" instead of the required "No". We also removed records of participants that study IT or a related subject, whose native language was not German and whose browser language was not German, to remove effects stemming from the level of language skill as well as daily exposure to warnings in different languages. Furthermore, responses that had a mean difference of three or more between the two inverse items and the corresponding original items were removed. Lastly, we filtered respondents that always chose the same answers on the rating items and those who either entered nonsensical summaries or copy-and-pasted parts of the warning message.

After filtering, 119 complete and validated responses remained. 40.3 % of our participants were female, 51.3 % had participated in the study described above and 60.5 % reported to have seen one of the warnings they were shown before (cf. Table 4 in the appendix). On average, it took the participants about 16 minutes to complete the survey, which is considerably longer than anticipated by pretesting in a laboratory setting.

5.3 Results

Initially, we checked for demographical imbalances in our rating results, using the non-parametric Mann-Whitney U test, since normality testing indicated significant deviations from the normal distribution in many of the rating variables. We found a few imbalances on the item for message comprehension: Messages 5, 21 and 27 were rated significantly better by participants that had previously participated in the first study. Message 12 received better ratings from men and

[1] The number of subscribers increased between studies.

messages 18 and 22 received significantly different ratings by participants that stated they had seen some of the warnings before. Since there was no obvious pattern in these differences, we accept them for further analysis.

We used Spearman's rho as a robust measure to test the monotonic relationship between rating ranks. The average ratings for comprehension and understanding the cause are strongly correlated ($\rho = .937$, $p < .001$) as is comprehension and difficulty of vocabulary ($\rho = -.797$, $p < .001$). Additionally, there is a relationship between previous exposure and the three other items ($\rho = -.65$, $\rho = .76$ and $.80$, $p < .001$): having more experience with a warning may support comprehension and understanding the cause.

Linguistic Properties. To see if particular linguistic properties of a warning message influence the users' perceptions, we used the Stanford Parser [11] and Part-of-Speech (POS) tagger [14] for German texts to analyse the structure of the warning texts. We gathered frequencies for 54 types of tags from the "Stuttgart-Tübingen-Tagset", as well as parse-tree parameters, including average number of nominal and verb phrases per sentence, as well as maximum and average parse-tree depth.

Several POS tag types showed medium to strong correlations with the ratings: Articles (ART, $\rho = .593$, $p = .001$) and the participle perfect (VVPP, prefix or infix "ge", $\rho = .564$, $p = .002$) appear to positively correlate with ratings, while the occurrence of the particle "zu" (english: "to") in front of an infinitive (PTKZU, $\rho = -.63$, $p < .001$) showed a negative correlation. Linear regression showed that VVPP and PTKZU can explain 54.7 % of the total variance in the participants' comprehension rating. Additionally, we did not find any meaningful correlation with the existing readability measures Amstad and LIX.

We also found correlations between the readability score we calculated based on cloze testing in the previous study with the maximum parse-tree depth ($\rho = -.872$, $p = .054$) and the number of attributive adjectives per sentence (ADJA, $\rho = -.90$, $p = .037$), but not with the ratings collected in this study. However, these correlations lack power, since the previous study only investigated six warning messages.

6 English Rating Study

In order to explore if similar effects exist for English warnings, we ran an additional rating study with the same setup on Amazon's Mechanical Turk (MTurk). Furthermore, warning messages for international software projects, such as Firefox and Google Chrome, are usually written in English and then translated into the different languages for localisation. It is possible that translation may cause the resulting warning messages to have a different linguistic structure compared to one written directly in the target language. Thus, we also used this study to compare the results of the translated warning texts with their original counterparts to see if translation has any effects on the ratings.

6.1 Design

We used the English versions of the set of 28 warning messages and created a HIT that advertised a task to rate ten browser warning messages on MTurk. We offered 1.50 $ as compensation for each successful completion and stated that only non-random and honest answers would receive the compensation. The study included the same validation questions as before and presented ten randomly selected warnings to each participant after introducing the InterBrowse and mybank.com scenario.

6.2 Participants

Our HIT was completed by 120 workers and took an average time of 20 minutes and 13 seconds (sd = 12 minutes and 29 seconds). We applied the same filtering methods as described in the previous study and hence retained 68 valid responses. Each message received an average of 24.3 ratings, ranging from 15 to 32. The average age of participants was 37 years ($sd = 12.7$), exactly half were female, and the overall self-reported technical experience was 2.44 ($sd = 1.01$). Respondents stated their occupation as student (8.8 %), full-time employee (14.7 %), part-time employee (47.1 %), self-employed (20.6 %) and other (8.8 %), including unemployed and homemakers.

6.3 Results

Similar to the results above, many of the rating variables showed significant deviations from a normal distribution (Kolmogorov-Smirnoff Test). We therefore ran the remaining analysis using non-parametric tests. First of all, the data was checked for demographical imbalances. For the comprehension rating, we found that messages 25 and 26 were perceived to be more difficult by younger participants. Interestingly, as in the results for the German versions of the messages, message 12 was perceived as being significantly easier to comprehend by men (Mann-Whitney $U = 48$, $Z = -2.297$, $p = .026$). Similar to above, the different ratings show significant correlations, although the strength is slightly weaker.

To identify structural features that influence ratings in English messages, we again applied the Stanford Parser and POS tagger for English texts to the English warnings. We used the 36 POS tags of the Penn Treebank Tagset[2], as well as the number of nominal and verb phrases, number of words per sentence, maximum number of words in a sentence, and (maximum) parse-tree depth. In contrast to before, we found only two correlations: the number of determiners (DT, similar to articles, $\rho = -.60$, $p < .001$) negatively influenced the ratings on difficulties with the vocabulary and the comprehension rating ($\rho = .491$, $p = .008$). In this case, linear regression was able to explain 46.2 % of the variance in the comprehension rating, using the number of words in the longest sentence as well as the number of wh-determiners (WDT, e. g. "which") and co-ordinating conjunctions (CC, e. g. "and"). There also was no meaningful correlation with the existing readability measures for English texts.

[2] http://www.cis.upenn.edu/~treebank/home.html

Comparison with German Results. We found a medium to strong correlation between the ranks for the German messages from the previous study to the English pendants (ρ between .68 and .78 for the four rating items, $p < .001$), indicating that messages perceived as complicated in German were also perceived as such in English and vice versa. Therefore, we conclude that the effects observed in the German messages do not purely stem from translation.

Next, we ranked all messages according to the three rating categories comprehension, understanding the cause and difficulty of vocabulary in the respective language. Based on the top and bottom five messages in each category, we found that three messages performed very well and four messages performed very poorly in both languages. Messages 18 and 19, (Firefox: "Reported Attack Page" and "Suspected Web Forgery"), and 28 (Safari, "Invalid Certificate") were consistently among the highest ratings. These warnings use easy, non-technical vocabulary and give direct recommendations on possible actions for the user.

The four messages receiving consistently bad ratings comprise three messages from Chrome ("Weak Signature Algorithm", "Unlisted Server Certificate", and "No Revocation Mechanism"), as well as one from Firefox ("SSL Disabled"). These messages address very technical issues and have probably never been seen by any of our participants: they also received very low previous exposure ratings.

Comparison between Products. Between the six certificate warning messages of different products that we included in the set of warnings, results showed that the Safari message was consistently found to be the easiest to comprehend and to use the easiest words. Likewise, we found that the message from Internet Explorer 8 was consistently rated worst. While the messages have comparable length (42 and 59 words respectively), the Internet Explorer message repeatedly uses the word "certificate" and other technical terms. The Safari message, in contrast, uses simple language, states a cause, the involved risk and asks the user to decide on a course of action.

Two Chrome warnings in our set differed only by their headline. One read: "This is probably not the site you are looking for!" and the other said "The site's security certificate is not trusted!". The message that did not mention certificate in the headline received consistently better ratings in both languages. Even though the difference is not statistically significant, this trend may imply that technical terms at the very beginning of a warning message can negatively influence the users' perceptions. To further investigate which factors influence users' perceptions of a warning message text in particular, we conducted interviews.

7 Interview

The previous studies have shown that there can be particular linguistic properties that may influence a user's perception of a warning message. The use and placement of technical terms as well as specific grammatical constructs showed correlations with the user ratings. We conducted interviews to directly analyse the participants' perceptions of technical terms and linguistic features, such as sentence composition.

7.1 Design

The interview was introduced to the participants as an investigation of readability in Internet browser warning message texts. We reminded them that this test was not about their abilities to comprehend the warnings but that their insights as to why a certain message might be hard to understand was of interest. Participants were presented with six warning messages as well as our InterBrowse scenario and would then be asked to carefully read the message. Next, we queried which sentences or parts of sentences were hard to read and their explanation. Afterwards, participants ranked all 6 warnings according to the perceived level of complexity. In a last task, they were provided with three highlighters and the same set of warning messages once more: we asked them to use a green highlighter to mark easy and clear words, a yellow one for words of medium difficulty that they still knew the meaning of and a red one for unclear and hard words. While they were working, we asked participants to offer their reasoning and collected their comments.

7.2 Participants

The participants were randomly recruited by phone from the database of more than 1,500 students also used above. Non-native speakers, students of German and Literature or Computer Science were excluded. We offered a compensation of 10€ and interviewed eight students (three female, 19 to 24 years old, four from the faculty of arts and four from the sciences) before our results reached saturation. Two participants had taken part in one of our previous studies, seven stated that they had seen one of the warning messages before or were unsure, four mainly use Firefox while two use Safari, one Chrome and one IE. The mean self-reported technical experience was 2.87 ($sd = .64$).

7.3 Results

Participants' comments can be divided into three main categories, detailed below. Participants are referred to as $P1, \ldots, P8$.

Headlines. Seven respondents stated that a warning's title should be short and precise. Additionally, five claimed that technical terms should not be in a headline. Four participants offered that "if I only looked at the heading, I wouldn't have had any clue what the error message is about" (P7). Participants agreed that an ill-conceived headline would deter them from continuing to read.

Positive Properties of Sentences. Short, precise sentences with an easy structure were appreciated by all respondents. Four of them explicitly requested that a simple sentence structure should be used: "[This] makes the message more colloquial and perfect for people who aren't experts" (P8). All participants offered that technical terms used in error messages hamper the understanding and awareness of the potential problem. The text marking tasks also showed that short sentences are preferred, yet, according to the comments, longer and more complex structures do not necessarily lead to readability problems.

Negative Properties. All participants agreed that the use of technical terms (see below) discourages them from reading (on) and trying to understand the scenario. P2 added: “One has to be really desperate to read this passage thoroughly”. In a similar fashion, half of the participants stated that in daily life, they would simply ignore paragraphs with many technical details. Six participants attempted to decode the meaning and the possible impact of the information in some of the warning messages, but failed. They felt “insufficiently informed” (P6) by the messages. P1 stated: “You simply want to get to the desired website and I don’t understand the problem itself nor when or how it will get solved”. These findings generally confirm the general preconceptions and the results of previous work.

Word-level Observations. During the word marking exercise, participants often indicated words as hard that had a technical background or referred to unclear concepts. The list included words such as “certificate” or “entity”, but also simple adjectives, including “attacking” and “weak”. Table 5 in the Appendix provides an overview of all words mentioned by participants.

Using this list of words, we counted occurrences in our set of 28 German warning messages. Again using Spearman’s Rho, the counts of hard words showed a correlation of .559 ($p = .002$) with the ratings of comprehension obtained in the studies described above, even though the list of words was only obtained on 6 of the 28 messages. Expanding on the implications of these results, we used the index terms of a computer security textbook [12][3] as an extended word list. The count of words from this list found in the 28 warning messages provided a slightly stronger correlation with ratings ($\rho = .646$, $p < .001$). The three best-rated and four worst-rated messages identified in section 6.3 also consistently received corresponding index-word counts of one match or less and three matches or more respectively. The same holds for headlines: the best-rated messages only used “website” in their headings while the worst-rated messages used technical terms (e. g. “certificate” or “revocation”).

8 Limitations

There are several limitations which need to be taken into account: First, our participants were either students or Mechanical Turk workers, which both represent a special group of people. Especially the students may present a best-case scenario for text comprehension, due to the exposure to difficult reading assignments in many subjects. However, the groups are quite different in terms of age and education, as well as professional background. Yet, we still found similar results in both studies.

Second, collecting self-reported measures likely causes a certain amount of bias. However, we implemented measures to try and mitigate these effects, by randomising messages and their order, as well as using only relative comparisons.

[3] We chose this textbook because it was the most recent security textbook digitally available at our library with an index.

Finally, we did not address the efficacy of warning messages explicitly, but used user ratings. While "pleasant" readability is a goal within itself, the correlation between readability and efficacy needs to be explicitly studied in future work. As noted above, related work suggests that facilitating the understanding of warning messages can predict user behaviour [13].

9 Discussion

During the course of our investigations, we found several aspects of warning message texts that influence their reception by users. First, cloze testing indicated that the required average reading level for warning messages is higher than the average reading level of most adults, mirroring the common image of warning messages often being too complicated. Results also hinted at the possibility that complicated messages can be understood by many readers if they spend enough time. However, these tests also indicated that the set of existing readability measures does not predict warning message difficulty accurately.

We then conducted the rating studies to collect users' ratings of warning messages and analyse if there are linguistic properties that can explain the rating differences. In both English and German warning texts, linguistic properties were able to explain about half of the variance in the ratings. Grammatical constructs that increase the information content of a sentence, for example coordinating conjunctions in English texts and German infinitive constructions, as well as grammatical tenses, such as the participle perfect in German texts, cause texts to be perceived as harder to understand. Additionally, we found that in both German and English versions of the warnings, messages with easy and non-technical vocabulary consistently received positive ratings while those that addressed specific technical problems consistently received negative ratings. A comparison between warnings from different products showed similar results.

Finally, we interviewed users and gathered aspects of warning message texts that may influence comprehension: headlines, non-technical vocabulary and short sentences were among the most frequently stated issues influencing the users' perceptions of warning message texts. Interestingly, the stated need for precise statements can cause conflicts: technical vocabulary is commonly used to make statements precise and short.

We were able to show that the linguistic properties identified in our studies can also be found in the best and worst message texts, according to the collected ratings. The set of words extracted from our interviews as well as a computer security textbook's index showed significant correlations with the ratings.

As stated above, our findings were able to explain about half of the variance in ratings using linguistic properties. We thus conclude that the linguistic properties of warning message texts and consequently issues that users might have with complicated sentence structures or difficult compounded words are one part of the larger puzzle, which entirely needs to be taken into account when designing new warning messages. Additional factors, such as missing context, previous exposure, unclear semantics, and effects of attitudes and beliefs can also strongly influence the users' perceptions of warning messages and their text.

Altogether, we found quantitative empirical evidence that linguistic properties can help to improve warnings: keeping headlines simple, using as few technical words as possible and creating short sentences without complicated grammatical constructions makes warning messages more pleasant for the user. A final take-away is that warning messages should not contain words that can be found in IT security textbook indexes. It is of course a challenge to describe the warning without such terms, however our results suggest it is a challenge worth working on.

References

1. Backhaus, A., Brügelmann, H., Knorre, S., Metze, W.: Forschungsmanual zum Stolperwörter-Lesetest (2004), `http://www.agprim.uni-siegen.de/lust/stolpermanual.pdf`
2. Bravo-Lillo, C., Cranor, L.F., Downs, J., Komanduri, S.: Bridging the Gap in Computer Security Warnings: A Mental Model Approach. IEEE Security & Privacy Magazine 9(2), 18–26 (2011)
3. Bravo-Lillo, C., Cranor, L.F., Downs, J., Komanduri, S., Sleeper, M.: Improving Computer Security Dialogs. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part IV. LNCS, vol. 6949, pp. 18–35. Springer, Heidelberg (2011)
4. Cranor, L.F.: A Framework for Reasoning About the Human in the Loop. In: Proc. UPSEC. USENIX (2008)
5. DuBay, W.H.: The Principles of Readability, `http://www.impact-information.com/impactinfo/readability02.pdf`
6. Egelman, S., Cranor, L.F., Hong, J.: You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In: Proc. CHI. ACM (2008)
7. Grossklags, J., Good, N.: Empirical Studies on Software Notices to Inform Policy Makers and Usability Designers. In: Dietrich, S., Dhamija, R. (eds.) FC 2007 and USEC 2007. LNCS, vol. 4886, pp. 341–355. Springer, Heidelberg (2007)
8. Harbach, M., Fahl, S., Muders, T., Smith, M.: Poster: Towards Measuring Warning Readability. In: Proc. CCS. ACM (2012)
9. Maurer, M.-E., De Luca, A., Hussmann, H.: Data Type Based Security Alert Dialogs. In: Proc. CHI Extended Abstracts. ACM (2011)
10. McLaughlin, G.H.: SMOG Grading – A New Readability Formula. Journal of Reading 12(8), 639–646 (1969)
11. Rafferty, A., Manning, C.D.: Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In: ACL Workshop on Parsing German (2008)
12. Spitz, S., Pramateftakis, M., Swoboda, J.: Kryptographie und IT-Sicherheit. Springer (2011)
13. Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., Cranor, L.F.: Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In: USENIX 2009 (August 2009)
14. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL (2003)
15. Wogalter, M.S., Conzola, V.C., Smith-Jackson, T.L.: Research-based Guidelines for Warning Design and Evaluation. Applied Ergonomics 33(3), 219–230 (2002)

A Warning Messages

Due to space constraints, the full set of warnings used in the study cannot be shown in the paper. In the following, we present an overview of messages while the full set can be downloaded from `http://benutzerstudie.dcsec.uni-hannover.de/warnings/`.

Table 1. Overview of warning messages used in the study

#	Browser	Beginning of Message
1	Chrome	The site's security certificate has expired! For a certificate which has not expired, the issuer of that certificate is ...
2	Chrome	The site's security certificate is not trusted! You attempted to reach mybank.com, but the server presented ...
3	Firefox	This Connection is Untrusted. You have asked InterBrowse to connect securely to mybank.com, but ...
4	Chrome	The site's security certificate is not trusted! You attempted to reach mybank.com but instead you actually reached ...
5	Chrome	Invalid Server Certificate. You attempted to reach mybank.com, but the server presented an invalid certificate. ...
6	Chrome	The server's security certificate is not yet valid! You attempted to reach mybank.com, but the server presented ...
7	Chrome	This is probably not the site you are looking for! You attempted to reach mybank.com but instead you actually ...
8	Chrome	The site's security certificate is signed using a weak signature algorithm! You attempted to reach mybank.com, but ...
9	Chrome	The server certificate contains a weak cryptographic key! You attempted to reach mybank.com, but the server presented ...
10	Chrome	The server's security certificate is revoked! You attempted to reach mybank.com, but the certificate that the server ...
11	Chrome	Unlisted Server Certificate. This site lists all its valid certificates in DNS. However the server used one which isn't listed. ...
12	Chrome	The server's security certificate has errors! When you connect to a secure website, the server hosting that site presents ...
13	Firefox	This Connection is Untrusted. You have asked InterBrowse to connect securely to mybank.com, but we can't confirm ...
14	Chrome	This webpage is not available. InterBrowse's connection attempt to mybank.com was rejected. The website may be down, or ...
15	Chrome	No revocation mechanism found. No revocation mechanism found in the server's certificate. When you connect to ...
16	Chrome	Unable to check whether the server's certificate was revoked. When you connect to a secure website, the server hosting ...
17	Chrome	Unknown server certificate error. An unknown error has occurred. When you connect to a secure website, the server ...
18	Firefox	Suspected Web Forgery. This page has been reported as a web forgery designed to trick users into sharing personal or ...

19 Firefox Reported Attack Page! This web page at mybank.com has been reported as an attack page and has been blocked based on ...

20 Firefox The certificate is not trusted because it is self signed. mybank.com uses an invalid security certificate. ...

21 Firefox Certificate will not be valid until date. mybank.com uses an invalid security certificate. The certificate will not be valid ...

22 Firefox The certificate expired on date. mybank.com uses an invalid security certificate. The certificate expired on ...

23 Firefox SSL protocol has been disabled. An error occurred during a connection to mybank.com. Can't connect securely because ...

24 Firefox Untrusted Connection Error. You have asked InterBrowse to connect securely to mybank.com, but we can't confirm that ...

25 MS IE 8 Security Certificate Problem. There is a problem with this website's security certificate. The security certificate ...

26 iTunes InterBrowse cant verify the identity of the server mybank.com. The certificate for this server was signed by ...

27 MS Outlook Problem with the site's security certificate. The information you exchange with this site cannot be viewed or changed ...

28 Safari InterBrowse can't verify the identity of the website mybank.com. The certificate for this website is invalid. You might ...

B Tables

Table 2. Demographics for the exploratory online study. Self-reported technical expertise was measured on a scale of agreement to the statement "I have a very detailed understanding of computer technology and the Internet" with 1 being complete agreement and 5 complete disagreement. The Stolper score indicates reading ability on a scale from 0-100% of successful completion of 35 reading tasks in five minutes.

N:	311
Age:	$22.8, sd = 4.1$
Tech. Expertise:	$2.29, sd = .92$
Area of Studies:	130 Arts (41.8%)
	181 Sciences and Other (58.2%)
Browser:	195 Firefox (62.7 %)
	56 Chrome (18.0 %)
	14 Internet Explorer (4.5 %)
	17 Opera (5.5 %)
	28 Safari (9.0 %)
	1 Other (.3 %)

Table 3. Results of cloze testing. Higher Amstad and lower LIX scores suggest better readability. The average rank indicates the position within the participants' subjective ordering of warnings (ranks closer to 1 indicate better subjective readability). Lower values of our readability score (70 % criterion score) indicate better readability. The last column shows the number of participants that were above the 70 % criterion score.

Message	Words	Amstad	LIX	Avg. Rank	Score 70	# Respondents
1	61	62.84	39.67	3.05	80.16	110
2	45	43.48	59.44	2.78	79.17	69
3	85	54.99	54.19	3.87	80.73	43
4	114	68.02	38.59	3.25	81.14	70
5	99	71.44	45.79	3.49	80.25	81
6	59	49.64	48.65	4.55	79.54	112

Table 4. Demographics for the Rating Study. Self-reported technical expertise was measured on a scale of agreement to the statement "I have a very detailed understanding of computer technology and the Internet" with 1 being complete agreement and 5 being complete disagreement.

N:	119
Age:	22.7, $sd = 4.02$
Tech. Expertise:	2.34, $sd = .98$
Area of Studies:	51 Arts (42.9 %)
	68 Sciences (57.0 %)
Browser:	82 Firefox (70.1 %)
	16 Chrome (13.7 %)
	8 Internet Explorer (6.7 %)
	3 Opera (2.5 %)
	8 Safari (6.7 %)
	2 N/A (1.7 %)

Table 5. Words mentioned by interview participants, arranged by difficulty and number of participants they were mentioned by. The category "high-one" was omitted because it was empty.

Medium Difficulty		High Difficulty
one	more than one	more than one
to confirm	weak	signature algorithm
to issue	attacking	security certificate
to forge	security settings	certificate
expiry	to expire	entity
to adapt	server	network administrator
to check	to present (a certificate)	proxy server
to contact	manipulation	proxy settings
operating system	security credentials	
to block	identity information	
private information	identification	
communicate	secure connection	

Soulmate or Acquaintance?
Visualizing Tie Strength for Trust Inference

Tiffany Hyun-Jin Kim[1], Virgil Gligor[1], Jorge Guajardo[2]
Jason Hong[1], and Adrian Perrig[1]

[1] Carnegie Mellon University
{hyunjin,gligor,jasonh,perrig}@cmu.edu
[2] Bosch Research and Technology Center
jorge.guajardomerchan@us.bosch.com

Abstract. Prior social science research has shown that tie strength is a useful indicator of context-dependent trust in many real-world relationships. Yet, it is often challenging to gauge trust in online environments. Given a multitude of variables that represent social relationships, we explore how to visualize interpersonal tie strength to empower people to make informed, context-dependent online trust decisions. Our goal is to develop visualizations that are meaningful, expressive, and comprehensible. In this paper, we describe the design of four visualizations. We also report on the results of two user studies, where users commented that our visualizations are highly comprehensive, meaningful, and easy to understand.

1 Introduction

Social interactions are increasingly moving into the online world. For example, traditional physical-world interactions, such as finding a babysitter, a partner, or a renter, used to work through word-of-mouth; however, people find it convenient to perform the same interactions online nowadays. Unfortunately, the online realm suffers from a lack of cues that can help people make informed trust decisions. As Steiner's famous cartoon depicts, "[o]n the Internet, nobody knows you're a dog," referring to the difficulty of verifying one's identity on the Internet [26].

For example, many people receive friend invitations in online social networks (OSNs) from casual acquaintances, friends of a friend, and even total strangers. A major problem here is that little information exists to help differentiate between people one has actually met, and scammers who impersonate an individual; indeed, prior studies have shown that such attackers fooled many OSN users, including security-conscious individuals [1,4,17,24].

One potential approach for trust establishment is to automate trust decisions such that computers make trust decisions for people. However, two major drawbacks render such automation infeasible: context-dependent nature of trust and differences in individuals.

Context-Dependent Nature of Trust. Trust varies depending on different contexts; different types of trust are needed for identifying an appropriate person for a babysitter for your child, for carpooling, or for new renters for your home. An automated system, however, is not clairvoyant and cannot make accurate decisions about which social

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 112–130, 2013.

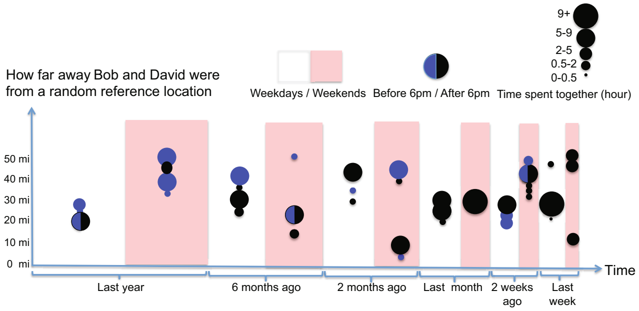


Fig. 1. An example visualization of tie strength between Bob and David. This diagram visualizes how far away from a random reference point two people have been interacting over a period of a year, how much time they have been spending at each location, whether the interactions were before or after 6:00 PM, and whether the interactions were on weekdays or weekends. These data can be feasibly acquired by smartphones (e.g., collocation can be acquired using GPS or Wi-Fi geo-location and duration/time of day/day of the week can be recorded on smartphones).

context the trust decision needs to be made in. For example, OSNs today cannot automatically distinguish between a social friend, a co-worker, an acquaintance of a friend, or a stranger whom you have never met.

Differences in Individuals. Every individual has distinct characteristics which are hard for automated techniques to capture; some people may choose to trust everyone while others may not. For example, extroverts have been found to be more willing to trust other people than introverts [12]. Given such differences in individuals, making trust decisions are difficult to account for in an automated manner.

Our goal in this paper is to understand what kind of information and how to offer it to people so that they can make informed trust decisions. We leverage prior research results which have shown that interpersonal tie strength is a good indicator of large classes of trust relations [16, 19], and social science researchers have established a plethora of parameters that correlate with tie strength [11, 14, 16, 19, 21, 25, 27].

Using parameters that we believe could be feasibly acquired by smartphones or online interactions, we explore the design space of visualizing tie strength to empower users to make informed, context-dependent trust decisions. Past work has shown how collocation data using smartphones and laptops [2], activity data on Facebook [14] and Twitter [13], and sensors and smartphone data [7] can be used to infer a range of characteristics about social relationships between people. Given the past work, one might ask why visualizations are needed, rather than just having, for example, a simple number that summarizes tie strength as, for example, 4 out of 5. A single number, however, is inadequate for at least two reasons: 1) numerical representation of tie strength may not be able to capture the details that are crucial for making informed trust decisions, and 2) deliberate attackers may be able to maliciously enhance numerical tie-strength values. Instead, we suggest visualizing tie strength with a rich set of features, which can be provided to users solely or as a supplement to numerical values. For example, one of the tie-strength visualizations that we propose is Figure 1, which depicts a summary of proximity information over a period of a year such that users can infer tie strength between Bob and David.

Contributions. This paper makes the following research contributions:

1. We explore the design space of interpersonal tie-strength visualizations that empower users to make informed, context-dependent trust decisions.
2. We present the design of four different visualizations illustrating aspects of tie strength (selected from a first-round user study).
3. We analyze usability in terms of meaningfulness, intuitiveness, and applicability to various use cases based on a second round of user study results.

Our user study results show that our visualizations are highly understandable; over 90% of study participants correctly interpreted the tie strength information on our visualizations. Also, study participants reported that our visualizations are intuitive while accurately portraying tie strength, and they provided diverse applications where they can use the visualizations to make informed, context-dependent trust decisions.

2 Background: Interpersonal Tie Strength

Pioneering research by Granovetter explored the strength of ties that exist between individuals [16]. Following his work, researchers studied the theoretical parameters for tie strength: amount of time [16,19], intimacy [16], affection [19], emotional intensity [16], reciprocal interaction [16, 19], structural factors [6], emotional support [27], and social distance [21]. Among multiple dimensions, Gilbert and Karahalios argue that relatively simple proxies can be substituted for determining tie strength in practice [14]: communication reciprocity [11, 16, 19], existence of at least one mutual friend [25], recency of communication [20], and interaction frequency [15, 16]. In our work, we embrace many of these insights. In particular, we designed many of our visualizations to convey communication reciprocity, recency, and frequency.

An extensive amount of literature has demonstrated that the frequency of interaction among people increases their likelihood of forming a friendship or romantic relationship [5]. Some studies have used physical proximity as a proxy for the amount of social interaction between pairs [10,23], for example, showing that communication frequency drops exponentially with the distance between a pair [3, 28]. Cranshaw et al. provide a model for predicting friendship based on the contextual features of users' location trails [2], using collocation and where collocations happened as a primary feature. This past work suggests that physical proximity may be a useful proxy for tie strength, an observation that we rely on in many of our visualizations.

Overall, our work builds on a great deal of past work in social science investigating relationships and strength of ties. Our primary contributions here are in the design and evaluation of new visualizations for conveying aspects of tie strength.

3 Problem Definition

Our interest is to explore visualizations that are based on data that have been shown to be feasibly acquired by smartphones or online interactions. Hence, based on these proxies for the variables in Section 2, we specifically consider the following 11 parameters:

1. **Collocation.** As suggested by prior work [2, 9], this parameter represents the placement when multiple users are physically present at the same location.
2. **Number of collocations.** This parameter represents the number of distinct locations where users physically interact [2, 9].
3. **Duration of interaction.** This parameter represents the time duration when users interact [2].
4. **Time of day.** This parameter represents when the interaction takes place [2, 9].
5. **Day of the week.** This parameter represents whether the interaction occurs during weekdays or weekends [2, 9].
6. **Length of relationships.** This parameter represents how long two users have known each other [15, 16].
7. **Interaction frequency.** This parameter represents how frequently users communicate through online (e.g., emails, chatting) and offline (e.g., face-to-face meeting, phone conversation) interactions [15, 16].
8. **Friendship level.** We propose friendship level to represent the social proximity between two users. For example, Alice may be one of Bob's top 10 best friends based on the quality and the quantity of their interactions.
9. **Interaction reciprocity.** This parameter represents whether the interaction was one-way (e.g., Alice attempts to call Bob who never responds) or reciprocal (e.g., When Bob misses Alice's call, he calls her back) [11, 16, 19].
10. **Recency of interaction.** This parameter represents how recent the previous interaction is [20].
11. **Number of mutual friends.** This parameter represents how many common friends two users share [25].

3.1 Assumptions

In this paper, we explore parameters whose values could be feasibly collected using smartphones or online interactions, and we assume that data acquired by smartphones or online interactions is correct. We also assume that visualizing the combination of parameters can be performed on a smartphone, and that a public-key cryptosystem is used for signing the visualization as follows: Bob, who creates a tie-strength visualization with David, has a private key to digitally sign the visualization, and Alice can validate Bob's signature with Bob's public key. Hence, digital signatures enable verification of the diagram and prevent forgeries.

For privacy, we assume that Bob can, at his discretion, decide to whom or whether at all to release information about his relation with David by signing (or not signing) the visualization. Analogously, David can release visualizations at his discretion.

3.2 Design Goals

Our goal is to accurately capture and visualize tie strength such that users can make informed, context-dependent trust decisions. Our desired properties are as follows:

- **Meaningful.** Visual diagrams should be designed using relevant parameters to convey semantically meaningful and useful tie-strength information to users. That is, presented diagrams should not mislead viewers to draw inaccurate conclusions.

- **Intuitive.** Visual diagrams should be intuitive such that users can interpret and understand the diagrams without difficulty. Ordinary users should understand the diagrams without rigorous training or explanations.

Note that the design goals are in tension with each other. For example, satisfying meaningfulness requires accurately portraying parameters of tie strength, and satisfying intuitiveness is in direct conflict with meaningfulness as accurate information can easily be incomprehensible.

3.3 Mapping Visualization Parameters

A multitude of design options exist to visualize tie strength parameters, including for example position on x- and y-axis, shape, size, color, and connection between objects [22]. Based on various mappings for the visual parameters to tie-strength indication values, we designed 12 different diagrams conveying tie strength as a formative exercise to help us explore the design space and solicit early feedback from participants. In particular, we explore visualizing the combination of multiple, relevant parameters in the same plot to accurately convey tie strength. Due to space limitations, however, we only focus on the top four visualizations that were found to be most useful and meaningful by our participants, as shown in Figures 2–5. Low-ranked visualizations are shown in Figure 1 and Figures 10–16 in Appendix.

We evaluated these diagrams through two rounds of user studies. The goal of the first study was to help us qualitatively understand the pros and cons of each of these visualizations, and filter out less useful visualizations. The goal of the second study was to measure the meaningfulness, intuitiveness, and applicability of these visualizations to a range of use cases. Towards this end, we took the top four visualizations from the first study and conducted a series of tests using Mechanical Turk.

4 Study 1: Formative Study

The objective of this first study was to choose a subset of the 12 diagrams that people find intuitive and helpful in evaluating social tie strength. Note that our goal was not to directly compare the diagrams against each other, but rather to understand what kind of information was useful and desirable to users.

In this study, we recruited 19 volunteers (9 females and 10 males) from diverse locations, including universities, a professional/office building, and a coffee shop. Participants were in the age range of 21 – 54, with various educational backgrounds (from high school graduates to doctoral degrees), and the interview took 20 minutes. In terms of the technical background, all participants were computer-savvy, using computers for at least 10 hours per week.

Procedure. We invited each participant to a room and described 12 diagrams in randomized order. After describing each diagram, we asked the participant to provide feedback on the diagram. Throughout the study, we asked the participant to speak out loud. After seeing all 12 diagrams, we asked the participant to group them in 3 categories: like, dislike, and unsure. We asked reasons behind the decision and asked the participant to pick the best 3 diagrams that (s)he would use to infer tie strength.

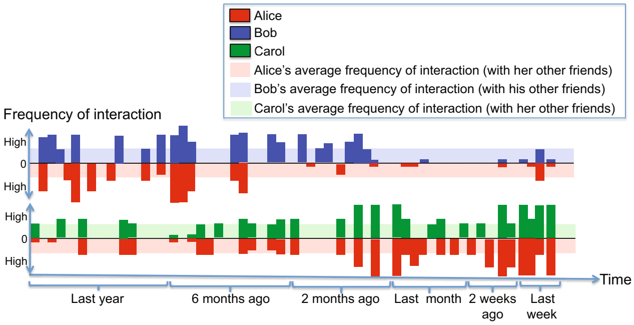


Fig. 2. Diagram A. This diagram presents frequency and recency of interaction, length of relationship, reciprocity, and the number of mutual friends using colored bars. 10 participants selected this diagram as one of their top 3 choices.

Results. In general, participants selected diagrams that they identified as simple, intuitive, and/or fun to examine. Figures 2–5 had the highest rankings overall from the formative study. Below, we describe the design rationale behind each of these 4 diagrams and the feedback that the study participants provided.

4.1 Diagram A: Bar Graph Visualization of Interaction Frequency

Diagram A focuses on displaying how frequently a user has interacted with his friend(s) using bars over the length of their relationships (Fig. 2). In particular, Diagram A illustrates the following parameters:

– *Length of relationships* is displayed on the x-axis in logarithmic scale. We chose this design to let people easily see older information about interactions, as well as more recent interactions.
– *Interaction frequency* is displayed on the y-axis using colored bars. For example, users see that the interaction frequency between Bob and Carol has been decreasing as the sizes of the bars on both Bob's and Carol's sides are decreasing; on the other hand, the interaction frequency between Bob and David has been increasing.
– *Interaction reciprocity* is shown based on the proportion of the bar sizes. For example, equal-sized bars on a graph implies that two users interact reciprocally; however, if one side's bar is significantly longer/shorter than the other side's bar, the interaction has been one-way.
– *Recency of interaction* is portrayed based on the existence of the most recent bars on the graph. In Fig. 2, Bob and Carol's most recent interaction was last week.
– *Number of mutual friends* is represented by the number of distinct graphs on a single plot. In Fig. 2, the viewer and Bob have two mutual friends: Carol and David.

We plot average interaction frequencies with colored background for those who are on a diagram. From Fig. 2, pink background represents the average interaction frequency that Alice has with all her other friends. Hence, this diagram enables users to approximate "friendship level" in comparison to average friend: users can compare if Alice interacts more or less frequently with the given mutual friend, and perceive better tie strengths between Alice and the mutual friend.

Analysis of Diagram A. Diagram A emphasizes the interaction frequency over time. Objectively presenting the actual interaction frequency may be challenging; for

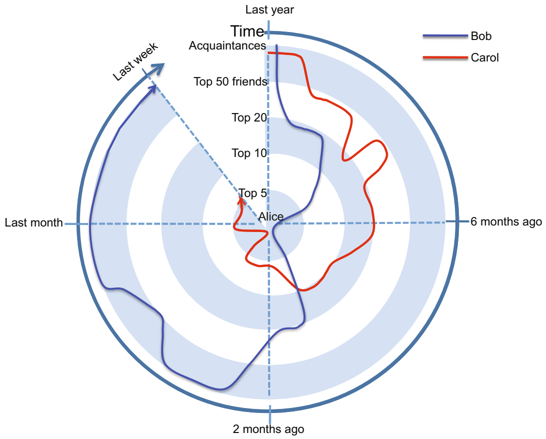


Fig. 3. Diagram B. This diagram presents friendship level, length of relationships, recency of interaction, and number of mutual friends on a Polar coordinate system. 8 participants selected Diagram B as one of their top 3 choices.

example, an introvert user may have low frequency values compared to an extrovert user. On the other hand, by providing an average value on all diagrams and by normalizing the average value to be consistent, Diagram A enables users to remove such biases and evaluate the *relative* frequency values in an intuitive manner.

Extra information regarding the interaction frequency can be encapsulated in Diagram A. For example, users can place the mouse pointer over a bar to get the percentage of online versus offline communications.

A potential limitation of Diagram A may be the scale issue when multiple graphs are shown on a single plot. Fig. 2 displays two graphs, and people may feel overloaded when multiple graphs, with distinct colors, are displayed.

Feedback on Diagram A

Pros. 10 out of 19 participants picked Diagram A as one of their top 3 choices. In particular, participants expressed their preference of this diagram in terms of their familiarity with the bar graphs and the simplicity for understanding its implication. One participant expressed enthusiasm since this diagram can preserve privacy with ambiguity: “[h]aving reciprocal interaction means good relationships, but having no interaction does not necessarily mean negative relationships.”

Cons. Although 10 participants picked Diagram A as one of their top 3 choices, they were cautions of sharing their own Diagram A with others. Two participants mentioned that this diagram seemed to reveal information in detail, and 3 people raised the possibility of misinterpretation: given 2 interaction frequency diagrams – one with the participant’s significant other and the other with the participant’s close friend – on a single plot, the significant other may get upset that the friend is a stronger tie to the participant.

4.2 Diagram B: Polar Coordinate Visualization of Friendship Level

While Diagram A portrays the variations of interaction frequencies over time on the Cartesian coordinate system, Diagram B emphasizes the changes in friendship level on the Polar coordinate system using line graphs (Fig. 3). By placing a user (Alice) on the center, a curve (of Bob) approaching the center can be intuitively interpreted as they are

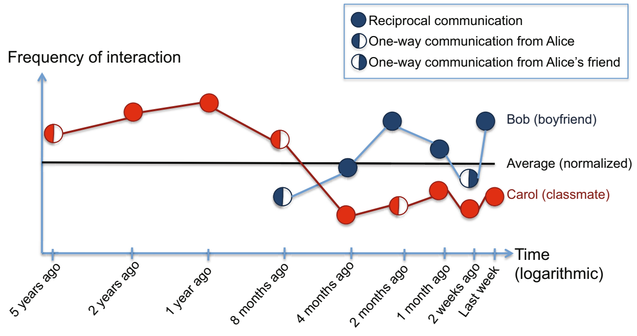


Fig. 4. Diagram C. This diagram maps the frequency and recency of interaction over the length of relationships. 7 participants selected Diagram C as one of their top 3 choices.

getting closer to each other in terms of friendship; on the other hand, a curve moving away from the center may indicate that their friend relationships are not as good as before. Diagram B illustrates the following parameters:

- *Length of relationship* is displayed over the angle in logarithmic scale.
- *Friendship levels* are displayed at uniformly distributed distances away from the origin with the scales of top 5, 10, 20, 50 best friends, and acquaintances.
- *Number of mutual friends* is represented by the number of distinct graphs on a single plot. In Fig. 3, the viewer and Bob have two mutual friends: Carol and David.

Analysis of Diagram B. *Friendship level* is another way of indicating tie strength, and Diagram B illustrates friendship levels using the Polar coordinate system. We assume that a system can automatically deduce friendship ranking among all friends. We conjecture that placing a targeted user on the center of the diagram and showing the changes in friendship level with lines over time is one of the natural ways of visualizing tie strength. Hence, people may find Diagram B attractive and intuitive.

Feedback on Diagram B
Pros. Most people provided positive feedback on Diagram B. For instance, one participant commented that "the information is composed organically." Three people admitted the the circular shape made this graph harder to understand, but they were still attracted to this design. Eight participants selected Diagram B as one of top 3 choices because the information was displayed in a clear manner and they could easily infer relationship changes by examining the flow of the lines.
Cons. Those participants who put Diagram B into the "dislike" category indicated that the circular orientation made this diagram hard to read. One participant also mentioned that this diagram took time to understand how the tie strength was portrayed. Another participant commented that Diagram B did not display too much information.

4.3 Diagram C: Line Graph Visualization of Interaction Frequency

Line graphs are useful in displaying increases and decreases in values over time. We apply line graphs in Diagram C where they depict the variation in interaction frequency over the length of relationships (Fig. 4). Diagram C illustrates the following parameters:

- *Length of relationships* is displayed on the x-axis in logarithmic scale.
- *Interaction frequency* is displayed on the y-axis without a detailed scale.
- *Interaction reciprocity* is shown using the amount of shade on each plotted dot. For example, a fully-colored dot implies that the interaction is reciprocal, and a half-colored dot implies that the interaction is one-way where the originator is based on the side of the color as shown in Fig. 4.
- *Recency of interaction* is conveyed based on the most recent point on the graph. In Fig. 4, Bob and Carol's most recent communication was last week.
- *Number of mutual friends* is represented by the number of distinct graphs on a single plot. In Fig. 4, the viewer and Bob have two mutual friends: Carol and David.

Similar to Diagram A, we introduce an average interaction frequency line on Diagram C, which represents the average interaction frequency that Alice has with all her other friends. This average line enables users to infer approximate "friendship level" relative to average friends. With this average, users can compare if Alice interacts more or less frequently with the given mutual friend, and can perceive better tie strengths between Alice and the mutual friend.

Analysis of Diagram C. Diagram C maps the same set of parameters as Diagram A. However, the reduced reciprocity information on Diagram C enables overlaying the lines which results in a more compact representation as well as the ability to more easily compare the different friendship levels. Instead of a bar graph, Diagram C is a connected line graph. Along with the average line, users may find Diagram C simple to read and easy to interpret. Furthermore, Diagram C can encapsulate extra information (e.g., percentage of online and offline communication and the reciprocity ratio) by placing a mouse pointer over each dot.

Feedback on Diagram C
Pros. Participants enjoyed the representation of reciprocity on this diagram. Seven participants who picked Diagram C as one of their top 3 indicated that this diagram was easy to read and understand. They also mentioned that comparing multiple graphs was straightforward.
Cons. Two participants mentioned that the symbols to represent reciprocity versus one-wayness were confusing. Instead of using the same color within a circle to represent reciprocity as shown in Fig. 4, they suggested using different colors or textures to represent reciprocity on each graph.

4.4 Diagram D: Dot Graph Visualization of Distinct Collocation

People tend to spend a lot of time together with their strong ties. However, the amount of time spent together by itself may not be a robust parameter to infer tie strength due to high false positive rate. For example, co-workers spend a lot of time together while they may not necessarily be close friends. On the other hand, people do not tend to visit many distinct places with casual acquaintances; people only interact with casual co-workers at their work place. Based on this observation, we conjecture that strong ties

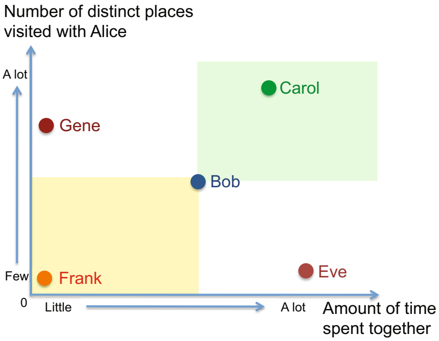


Fig. 5. Diagram D. This diagram displays the magnitude of the number of distinct places Alice visited together with her other friends and how much time Alice has spent interacting with them. 8 participants selected Diagram D as one of their top 3 choices.

can be distinguishable based on the number of collocation and duration of interaction. Diagram D (Fig. 5) maps the following parameters:

- *Number of distinct collocations* is mapped on the y-axis, ranging from a few to a lot of locations.
- *Duration of interaction* is mapped on the x-axis, ranging from little interaction to a lot of interactions, expressed in terms of time. In this diagram, the time duration includes not only physical but also other offline and online interactions.
- *Number of mutual friends* is displayed using dots over the plot.
- *Reciprocal interaction* is implied in this diagram since physical interactions can only occur when two people are near each other's vicinity.

For example, Fig. 5 shows that Alice and Carol have been spending a lot of time together while visiting many distinct places, possibly implying their strong-tie friend relationship. On the other hand, Alice and Eve have been spending a lot of time together but in few places, possibly implying a weak-tie co-worker or classmate relationship.

Analysis of Diagram D. Diagram D incorporates fewer parameters than others. We presumed that such simplicity would be better in preserving users' privacy, and that people would find this simpler diagram easy to understand and suitable for a number of use cases.

Feedback on Diagram D

Pros. All participants emphasized that Diagram D was straightforward and simple to understand. They also enjoyed to see a large number of mutual friends on the same plot. *Cons.* Although participants enjoyed the simplicity, three of them raised the issue that it might be hard to determine the relationship since they might not get as much information from this diagram versus the other diagrams. Also, two participants were confused by the yellow and green quadrant representations and suggested better use of colors.

4.5 Summary of Study 1

The formative study enabled us to pick the top 4 diagrams that people expressed suitability and usefulness in inferring tie strength.

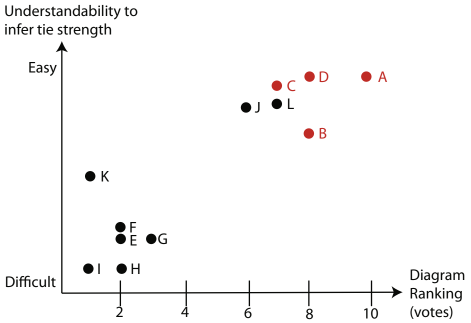


Fig. 6. Scatterplot showing diagram popularity vs. understandability in inferring tie strength. This diagram shows that the popular diagrams (e.g., Diagrams A–D, L) were easier to understand and infer tie strength compared to those low-ranked diagrams (Diagrams E–K). Note that we selected Diagram C over Diagram L as one of the top 4 diagrams because participants indicated that Diagram C carries more information and is easier to infer tie strength than Diagram L.

Fig. 6 summarizes the relationship between the participants' understandability in inferring tie strength and the popularity of 12 diagrams (Diagrams F – L are in Appendix). In summary, the participants favored diagrams that are easy to understand and infer tie strength. Note that Diagram C and Diagram L both received 7 votes. However, we selected Diagram C as one of the top 4 choices based on two reasons: 1) the participants indicated that Diagram C carries more information that would be useful to infer tie strength compared to Diagram L, and 2) Diagram L visualizes the same parameters as Diagram B for which a lot of participants expressed their fondness.

5 Study 2: Evaluation of Visualizations

Using the top 4 diagrams from Study 1 (as rated by participants), we conducted an online user study to analyze if the top 4 diagrams convey semantically meaningful and useful tie-strength information to users, and if these diagrams are easy to interpret and understand. we also studied the applicability of these diagrams to other use cases.

5.1 User Study Background

We conducted an online survey using Amazon Mechanical Turk (MTurk). We followed common methodologies for running MTurk studies [8, 18]. We wanted to focus on U.S. participants first; hence, we set the location restriction flag on MTurk to invite only users located within the U.S.

Our online survey had two rounds: the first round was to analyze meaningfulness and intuitiveness of Diagrams A–D, and to solicit other use cases for the diagrams. Based on the use cases that participants provided, we designed a follow-up survey to evaluate the applicability of Diagrams A–D to various use cases.

From 201 total participants, we analyzed the responses from 96 participants who completed both rounds after eliminating careless users as follows: 1) we eliminated anyone who provided contradicting answers to simple questions that we purposefully asked multiple times with different wording, and 2) we eliminated anyone who provided the same answers (both multiple-choice and fill-in answers) for at least 3 diagrams. The demographics of the 96 participants are as follows: 73% female and 27% male within the age range of 16 – 41 (μ = 36.4, σ = 9.4), all living in the U.S. All participants, even those we eliminated, were paid at least $1.00. Participants who provided accurate

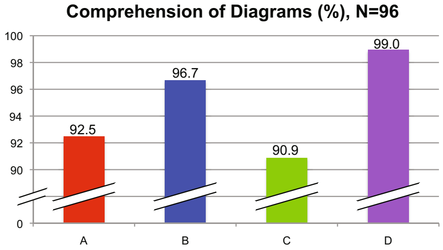


Fig. 7. Comprehension of Diagrams A–D. This bar graph presents the percentage of correctly answered questions on each diagram. Note that the y-axis is skipped over to 90%.

answers in the comprehension section of our study were paid $2.00. Finally, participants who returned for our follow-up survey on use cases were paid an additional $0.30. Thus, the 96 participants whose data we report on were paid $2.30 each.

5.2 Study Scenario

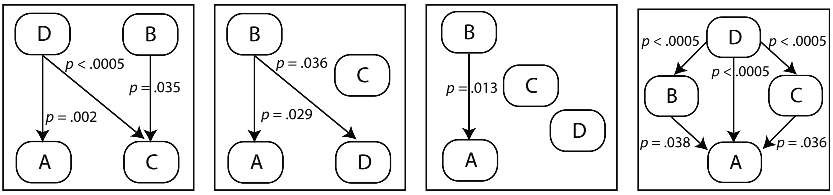
We used a within-subjects design and asked study participants to play a role as follows: "You are coordinating a surprise party for your best friend Alex. You would like to invite Alex's best friends whom you don't know, but you don't want to ask Alex directly. You found an application which analyzes how close people are to Alex. This application can show you 4 different diagrams, each of which draws different features to represent how close of a friend a person is to Alex. You are now ready to explore all 4 options and find out whom to invite to Alex's surprise party. Please explore each diagram carefully and answer the questions."

To minimize biases, we randomized the order of Diagrams A–D, and for each diagram, we described in detail what parameters the diagram visualizes and how users can interpret them along with some examples. We then asked questions on each diagram to test comprehension, meaningfulness, and intuitiveness. At the end of the study, we asked the participants to provide other use cases for each diagram in their own words.

5.3 Study Results

For all analyses reported in this section, we conducted repeated measures ANOVA tests using Greenhouse-Geisser correction (if the sphericity assumption was violated) and post-hoc pairwise comparison tests using the Bonferroni adjustment.

Comprehension. To measure how well participants comprehended Diagrams A–D, we asked 5 questions about each diagram; we asked 3 questions pertaining to the individual parameters that each diagram illustrates (e.g., when was the most recent interaction that Alex and Bailey had?, how many distinct places did Alex and Casey visit?), 1 question for interpreting the graphs in general (e.g., how did the interaction frequency change between Alex and Bailey over the last year?), and 1 question for comparing two different graphs/points on each diagram (e.g., between Casey and Drew, who did Alex interact more frequently with last week?). Note that not all diagrams carry the same parameters and the same information; hence, we modified some questions on the diagrams while maintaining the same relative level of difficulty.



(a) (b) (c) (d)

Fig. 8. Partial order graph from the Bonferroni post-hoc pairwise test based on (a) comprehension, (b) accuracy, (c) appropriateness, and (d) intuitiveness of Diagrams A–D. An arrow from X to Y means that Diagram X is statistically significant than Diagram Y in each property with 95% confidence rate. A *p-value* is shown next to the corresponding arrow.

Figure 7 shows the percentage of correctly answered questions on each diagram. As Figure 7 shows, all diagrams achieved high comprehension rate (over 90%). In particular, some diagrams resulted in significantly better comprehension than others according to the ANOVA test (see Table 1). Post-hoc pairwise comparison tests reported that Diagram D resulted in significantly higher comprehension rate compared to Diagrams A and C, and so did Diagram B compared to Diagram C. Figure 8(a) is the partial order graph based on this pairwise comparison test results.

Meaningfulness. To evaluate meaningfulness, we asked participants to evaluate how accurately each diagram portrays tie strength and how appropriate each diagram is for surprise party invitation, both using the 7-point Likert scales (1: not meaningful at all – 7: very meaningful). We used subjective measures to capture people's perceptions of how accurately each diagram portrays tie strength and how appropriate each diagram is for the use case of surprise party invitation.

An ANOVA test ($\chi^2 = 11.84, p = 0.037$) reported that the mean accuracy ratings of 4 diagrams were statistically significant as shown in Table 1. The partial order graph based on the pairwise test results is shown in Figure 8(b). Based on the results, we can conclude that Diagram B, depicting the level of friendship over length of relationship, is the visualization that participants rated as portraying tie strength most accurately.

In terms of how appropriate each diagram is for the use case of surprise party invitation, an ANOVA with the sphericity assumption satisfaction ($\chi^2 = 16.83, p = 0.005$) reported that mean appropriateness differed with statistical significance among 4 diagrams (see Table 1). The partial order graph based on the pairwise test results is shown

Table 1. Means and repeated measure ANOVA results for design goals (*N* = 96). The highest means that are statistically significant from others are highlighted in bold.

	Comprehension min:0 max:5	Accuracy min:1 max:7	Appropriateness min:1 max:7	Intuitiveness min:1 max:7
A	$4.623 \pm .079$	$5.146 \pm .147$	$5.104 \pm .183$	$4.521 \pm .200$
B	$\mathbf{4.823} \pm \mathbf{.071}$	$\mathbf{5.667} \pm \mathbf{.152}$	$\mathbf{5.708} \pm \mathbf{.159}$	$5.198 \pm .168$
C	$4.544 \pm .072$	$5.229 \pm .139$	$5.188 \pm .160$	$5.167 \pm .146$
D	$\mathbf{4.948} \pm \mathbf{.027}$	$5.031 \pm .154$	$5.115 \pm .175$	$\mathbf{6.135} \pm \mathbf{.100}$
ANOVA	$F(2.365, 285) = 8.30$ $p < 0.0005$	$F(2.748, 285) = 4.29$ $p = .007$	$F(3, 285) = 4.06$ $p = .008$	$F(2.657, 285) = 20.00$ $p < 0.0005$

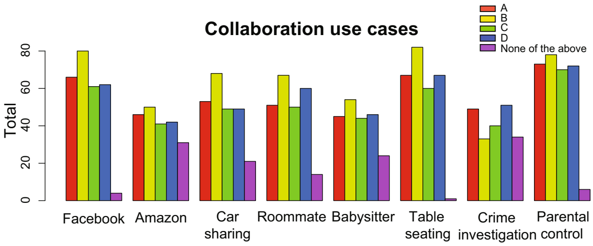


Fig. 9. Applicability of Diagrams A–D to various use cases ($N = 96$)

in Figure 8(c). Based on these results, we can conclude that Diagram B is the most appropriate visualization to infer tie strength for the use case of surprise party invitation.

Intuitiveness. We asked participants to rate how intuitive each diagram was to understand given a Likert scale from 1 (not intuitive at all) to 7 (very intuitive). An ANOVA test ($\chi^2 = 21.17, p = 0.001$) reported that mean intuitiveness differed statistically significantly among 4 diagrams as shown in Table 1. Figure 8(d) is the partial order graph based on the pairwise-test results. Hence, participants found Diagram D as the most intuitive visualization.

Use Cases. We wanted to understand what participants thought about using 4 diagrams for other use cases. To evaluate use cases, we asked participants to provide their own (if possible). Based on participants' feedback, we created a follow-up survey and invited them back to select the diagram(s) they deemed suitable for each use case.

Figure 9 is the bar graph summarizing the result for the use cases. We provided 4 examples based on our conjectures: 1) validating Facebook friend inviter, 2) validating product recommenders on Amazon, 3) verifying the renter of the participants' vehicles, and 4) finding a roommate. Among many examples that the participants provided, we show the following on Figure 9: 1) finding a babysitter, 2) finding close people for determining table seatings, 3) analyzing crime investigation, and 4) learning whom the participants' children hang out with.

Overall, Diagram B had the highest scores on all use cases except crime investigation; for this case, the participants reported that Diagram A, depicting interaction frequency, and Diagram D, depicting collocation, are suitable.

6 Discussion

Table 2 summarizes how Diagrams A–D satisfy the design goals based on the participants' feedback. Based on the study results, we can conclude that Diagram B, depicting the changes in the friendship level over the length of time period using simple lines, is the best tie-strength visualization among 4 designs since it was ranked to be the most meaningful diagram and had a high comprehension rate. For implementation, further study may be needed to study how to represent such concrete friendship levels using online and offline communications.

One major complaint about Diagram B was that the Polar coordinate system was challenging to read (although it graphically depicts distance from the center point); indeed, Diagram C plots the interaction frequency over length of time using simple

Property \ Diagram	A	B	C	D
Comprehension		●		●
Meaningfulness		●		
Intuitiveness				●

Table 2. Summary of design goal satisfactions for Diagrams A–D. A dot is placed on the diagram with the highest mean value that was statistically significant from other diagrams for each property.

lines on the Cartesian coordinate system. However, a recurring downside of Diagram C was the representation of the reciprocity: users expressed the difficulty of understanding the definition of reciprocity. Hence, we leave it as a future study to verify the criticality of reciprocity for inferring tie strength for context-dependent trust decisions.

Privacy. In this paper, our main focus was to study the utility of the visualizations. Although these diagrams show sensitive information, it is also abstracted to minimize specific details, such as when calls are made or what the content of the communications are. Furthermore, privacy-sensitive data is aggregated and normalized, without revealing exact values, and release of tie strength visualization information is entirely voluntary. Thus, a user can suppress releasing information that s/he does not feel comfortable about. For example, one participant in Study 1 mentioned, "I like [Diagram A] since it doesn't look trivial to figure out the exact interaction frequency. To me, this diagram greatly preserves privacy." Another participant also mentioned, "although [Diagram D] shows less information than other diagrams that I've seen so far, I think this diagram can still be useful. But I'm not quite sure how helpful this diagram would be to check how close people are." We plan to study privacy aspects in our future work once the utility is recognized.

7 Conclusion

We explored the design space of visualizing interpersonal tie strength to empower users to make their own informed, context-dependent trust decisions for various collaborative activities. We designed 12 different diagrams for visualizing tie strength, based on data that have been shown to be feasibly gathered from smartphones and online interactions. In our first user study, we solicited qualitative feedback from our participants regarding our designs, and based on this feedback, we narrowed our visualizations down to four. In a second user study, we were able to analyze how comprehensive, meaningful, and easy to understand our visualizations were. Although we found that participants appreciated the applicability of our visualization to a wide range of collaboration use cases, future research still needs to determine the extent of its suitability.

References

1. Sophos Facebook ID Probe, `http://www.sophos.com/pressoffice/news/articles/2007/08/facebook.html`
2. Bridging the Gap Between Physical Location and Online Social Networks (2010)
3. Allen, T.J.: Managing the flow of scientific and technological information. Massachusetts Institute of Technology (1966)

4. Bilge, L., Strufe, T., Balzarotti, D., Kirda, E.: All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In: Proceedings of WWW (2009)
5. Bossard, J.H.S.: Residential propinquity as a factor in marriage selection. American Journal of Sociology 38(2), 219–224 (1932)
6. Burt, R.S.: Structural Holes and Good Ideas. American Journal of Sociology 110(2), 349–399 (2004)
7. Dong, W., Lepri, B., Pentland, A.S.: Modeling the Co-evolution of Behaviors and Social Relationships Using Mobile Phone Data. In: Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (2011)
8. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In: Proceedings of CHI (2010)
9. Eagle, N., Pentland, A., Lazer, D.: Inferring friendship network by using mobile phone data. PNAS 106(36), 15274–15278 (2009)
10. Festinger, L.: Informal social communication. Psychological Review 57(5), 271–282 (1950)
11. Friedkin, N.E.: A Test of Structural Features of Granovetter's Strength of Weak Ties Theory. Social Networks 2, 411–422 (1980)
12. Gaines, S.O., Panter, A.T., Lyde, M.D., Steers, W.N., Rusbult, C.E., Cox, C.L., Wexler, M.O.: Evaluating the Circumplexity of Interpersonal Traits and the Manifestation of Interpersonal Traits in Interpersonal Trust. Journal of Personality and Social Psychology 73(3), 610–623 (1997)
13. Gilbert, E.: Predicting Tie Strength in a New Medium. In: Proceedings of CSCW 2012 (2012)
14. Gilbert, E., Karahalios, K.: Predicting Tie Strength With Social Media. In: Proceedings of CHI (2009)
15. Gilbert, E., Karahalios, K., Sandvig, C.: The Network in the Garden: An Empirical Analysis of Social Media in Rural Life. In: Proceedings of CHI (2008)
16. Granovetter, M.S.: The Strength of Weak Ties. The American Journal of Socialogy 78(6), 1360–1380 (1973)
17. Hamiel, N., Moyer, S.: Satan Is On My Friends List: Attacking Social Networks. In: Black Hat Conference (2008)
18. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing User Studies with Mechanical Turk. In: Proceedings of CHI (2008)
19. Krackhardt, D.: The Strength of Strong Ties: The Importance of *Philos* in Organizations. In: Nohria, N., Eccles, R. (eds.) Networks and Organizations: Structure, Form, and Action, pp. 216–239 (1992)
20. Lin, N., Dayton, P.W., Greenwald, P.: Analyizing the Instrumental Use of Relations in the Context of Social Structure. Sociological Methods Research 7(2), 149–166
21. Lin, N., Ensel, W.M., Vaughn, J.C.: Social Resources and Strength of Ties: Structural Factors in Occupational Status Attainment. American Sociological Review 46(4), 393–405 (1981)
22. Mullet, K., Sano, D.: Designing Visual Interfaces: Communication Oriented Techniques. Prentice Hall (1994)
23. Newcomb, T.M.: The acquaintance process. Holt, Rinehart, and Winston (1961)
24. Ryan, T.: Getting in Bed with Robin Sage. In: Black Hat Conference (2010)
25. Shi, X., Adamic, L.A., Strauss, M.J.: Networks of Strong Ties. Physica A: Statistical Mechanics and its Applications 378(1), 33–47 (2007)
26. Steiner, P.: On the Internet, nobody knows you're a dog. The New Yorker (July 1993)
27. Wellman, B., Wortley, S.: Different Strokes from Different Folks: Community Ties and Social Support. The American Journal of Sociology 96(3), 5538–5588 (1990)
28. Zipf, G.K.: Human behavior and the principle of least effort. Addison-Wesley Press (1949)

A Low-Ranked Diagrams

The following diagrams, along with Diagram E in Figure 1 are the ones that were not selected as top 4 diagrams from Study 1.

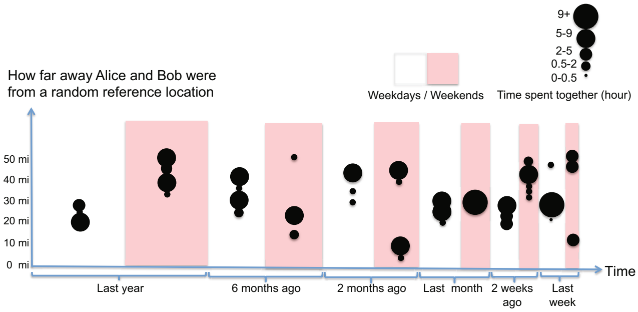


Fig. 10. Diagram F. This diagram shows how far away from a random reference location two people have been interacting over a period of a year, how much time they have been spending at each location, and whether the interactions happened on weekdays or weekends.

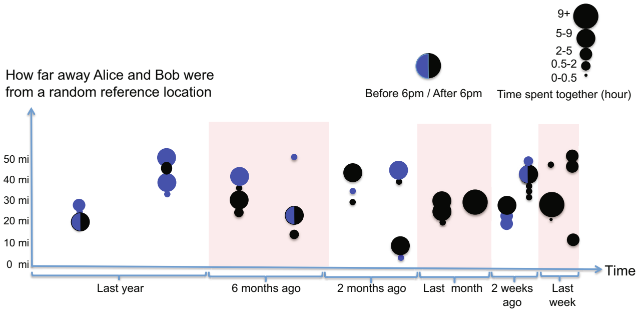


Fig. 11. Diagram G. This diagram shows how far away from a random reference location two people have been interacting over a period of a year, how much time they have been spending at each location, and whether the interactions were before 6:00 PM or after 6:00 PM.

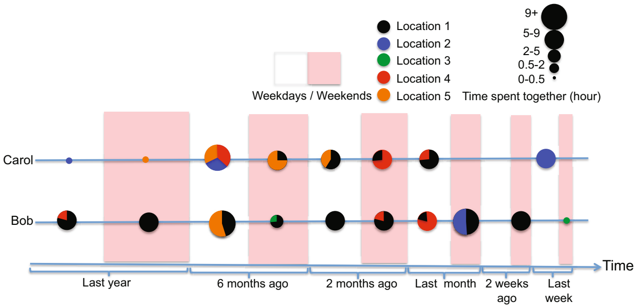


Fig. 12. Diagram H. This diagram shows the number of distinct locations that Alice has physically been collocated with her friends Bob and Carol over a period of a year, and how much time they have been spending at each location, and whether the interactions took place on weekdays or weekends.

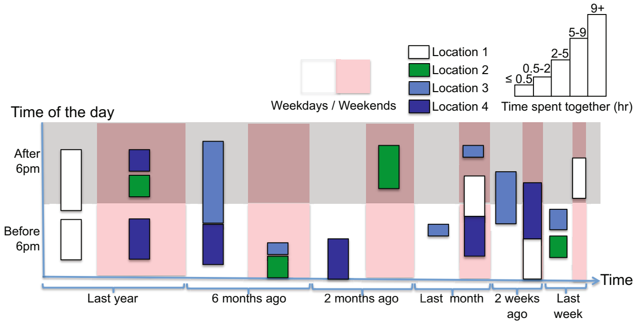


Fig. 13. Diagram I. This diagram is a variation of Diagram H, emphasizing distinct collocations. Unlike Diagram H, this diagram visualizes collocations of two people using blocks. Different colored blocks represent distinct locations and the size of the blocks indicates the amount of time two people have spent together.

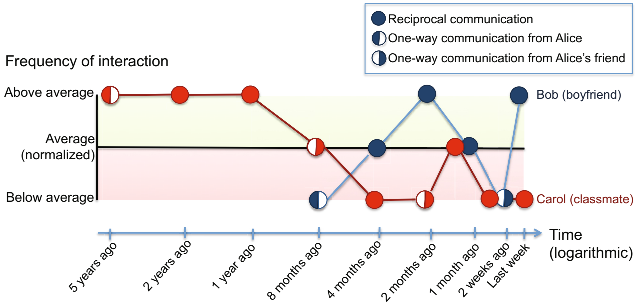


Fig. 14. Diagram J. This diagram displays the same parameters as Diagram C. The distinction is that Diagram J categorizes the interaction frequency into three groups: above average, average, or below average.

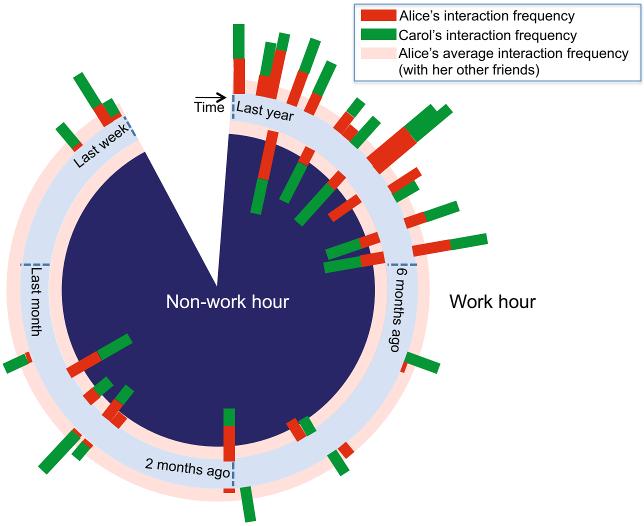


Fig. 15. Diagram K. This diagram plots two user's interaction frequency over time, as in Diagram A. In contrast to Diagram A, Diagram K displays an additional parameter: time of day.

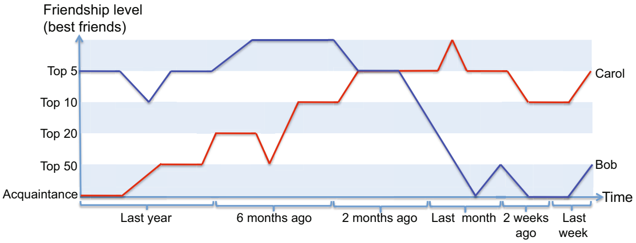


Fig. 16. Diagram L. This diagram is a variation of Diagram B. Instead of using the Polar coordinate system, Diagram L displays changes in friendship levels using the Cartesian coordinate system.

Awareness about Photos on the Web and How Privacy-Privacy-Tradeoffs Could Help

Benjamin Henne and Matthew Smith

Distributed Computing & Security Group, Leibniz Universität Hannover
{henne,smith}@dcsec.uni-hannover.de

Abstract. Many privacy issues concerning photos on the Web and particularly the social Web have been discussed in the past. However, much of this discussion is based on anecdotal evidence and has focused on media uploaded by users themselves. We present the results of a survey conducted with 414 participants that studies user awareness of privacy issues concerning the sharing of media including media shared by others. We additionally investigate the current perception of metadata privacy, since metadata can amplify threats posed by photos on the Web, for instance by tagging people or linking photos to locations. Furthermore, we present how this metadata can be used to help to protect private information and discuss the concept of a *privacy-privacy-tradeoff* and how this can be used to enable people to discover photos relevant to them and therefore regain control of their media privacy.

Keywords: privacy, awareness, social media, photo sharing, metadata, privacy-privacy-tradeoff.

1 Introduction

A multitude of privacy issues with online photos have been discussed in the past years [1,3,2,6,7,10,12], with photography in general being a point of contention for privacy issues for over a century [13]. The thought of being depicted in a photo somewhere on the Web is already a privacy concern for some people: Even a picture of someone at a perfectly harmless location may raise objections. People feel even more threatened by pictures showing them in embarrassing situations, doing socially questionable things, or at a place or with someone they would rather deny having been with. Research has shown that people feel their privacy threatened by photos taken by nearly any other person, no matter if they are from people outside [1] or inside their social circle, including friends and family [3]. Furthermore, media content may not just harm personal privacy, but can also cause immediate effects, since employers, insurance companies and banks use such information to gather knowledge about employees or clients. An increasing number of people have become cautious about sharing personal data in social network services (SNS). Yet, SNS users still create threats to their own privacy by accidentally disclosing compromising pictures of themselves to the public. Access control facilities offered by SNS help people keeping their

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 131–148, 2013.

media private up to a certain degree, though usability or comprehension issues often complicate the effective deployment of privacy settings [6,10]. Aside from these relatively obvious problems, other threats have not yet received sufficient attention: Shared photos not only affect the uploaders' privacy, but the privacy of all persons visible in the photo. Threats posed by such photos are particular insidious, since the potential victims are not involved in the uploading process and thus cannot take any preemptive measures against being depicted online. While for instance tagging people in photos can be prevented in current SNS, there currently are no countermeasures to the upload itself except legal actions or demanding that the media be taken offline again. Since online sharing of media cannot be simply prohibited, raising awareness about shared media on the Web is the key issue to address privacy concerns arising from the increased use of the social Web.

For privacy threats of shared media to take effect, two requirements have to be fulfilled: To cause harm, media needs to be able to be associated to a person. In addition, the media in question must contain objectionable content for that person. The association and the content can both be either non-technical – i. e. only recognizable by humans – or technical – i. e. content actively linking to a personal profile, or metadata containing a compromising time or location. In this context, the metadata plays an integral role: It stores additional information besides the picture itself and is easily machine-readable. The use of contemporary cameras and especially smartphones amplifies the privacy threat posed by shared media: Current cameras are capable of gathering location information via GPS or Wi-Fi-tracking and automatically embed it into photos. Latest applications additionally integrate facial recognition functions that aim to automatically tag individuals in photos. Modern devices ease the annotation of shared media with information that may give rise to privacy concerns.

In this paper, we focus on threats posed by photos shared by others and analyze their relevance by presenting results from an online survey. We discuss *awareness* of media sharing as a key issue in Sect. 2 and examine the current importance of metadata privacy in Sect. 3. Our analysis is based on the results of an online survey with 414 participants, showing that while most participants are aware of possible privacy threats, they also see a need for a better chance to effectively control which photos depicting them are shared. Using a prototypical system, designed to raise awareness about media sharing, we discuss *privacy-privacy-tradeoffs* that disclose certain private information to a social network privacy service to regain control over more important private information in Sect. 4. Finally, we conclude this paper in Sect. 5

1.1 Survey Design and Participants

The remainder of this paper describes the results of an online survey. We will introduce the individual parts of our survey in combination with the respective results in separate sections.

1,418 members of a university-related mailing list were invited to participate in the survey. The invitation asked for participation in a survey on privacy issues of media sharing. While explicitly mentioning privacy has possibly caused selection bias, we intended to recruit users interested in this topic to investigate a best-case scenario. As an incentive for participation, we offered participants an option to enter a raffle for two $ 60 vouchers from Amazon.

We received 414 complete and valid answers. 53.9 % of our participants were male and 46.1% female. About 25 % of the participants already had at least one university degree. The average age of participants was 23±4 years. 22.2 % indicated a high or very high technical expertise. According to Westin's privacy segmentation index [9], 91.8 % of the participants were classified as privacy pragmatists, 6.0 % as fundamentalists and 2.2 % as unconcerned. Thus most of our participants handle their online privacy pragmatically depending on the situation, indicating that most of them would therefore not simply be uninterested in privacy controls nor demand them regardless of the real threat, but present differentiated opinions on the topic at hand.

Normality testing indicated significant deviations from the normal distribution (Kolmogorov-Smirnov) for most rating variables as expected, which is why we employ non-parametric test measures to discuss our results.

2 Online Photo Awareness

When reports about employers and banks using social media to gain knowledge about their employees and customers increased, privacy problems of shared media began to catch the public's i.e. the media's attention. However, the extent to which this attention actually translates into user actions or awareness is unknown. Thus, one goal of our survey was to learn about the extent of the awareness users currently have concerning online photos they might be depicted in. Are people really aware of the threat posed by pictures shared by others and their possible impact? Do users realize that pictures they are not tagged in also cause privacy issues?

2.1 Linking Media to People

Most of the popular SNS like Facebook or Google+ and media-sharing sites like Flickr allow users to tag objects and people in the media they upload. Media can be commented on, annotated with keywords, or directly linked to a person. The direct link between profiles and photos thereby was initially met with a great outcry of privacy concerns. Such links simplify finding pictures of people beyond the content they consciously share in their profiles. For this reason, current SNS allow their users to either completely forbid others to link them in shared media or to approve links before they become visible to the public. However, such links also have a positive side: When tagged in a photo, users usually receive notifications about the link and consequently about the photo that might raise privacy concerns. Based on this notification, users can check the picture and possibly have unwanted content removed or access restricted [2,12]. One goal of our survey was to find out to what extent users are aware of the positive effect of such tags. To gather reasons for tagging others in photos, we asked our participants how frequently they tag someone for specific reasons, using a 7-point scale from *not at all* to *very often* (cf. Fig. 1). 30 % of the participants stated that they never tag people in their photos just to notify the tagged user. The remaining 70 % rated this item with a mean rating of 5.34 (sd = 1.42), indicating that this is a valid reason for tagging for most users. Likewise, 54.8 % of all participants stated that they never tag someone in a photo to make other people aware of this photo. For the remaining participants, this also appears to be a less important reason with a mean rating of 3.73 ($sd = 1.55$). The numbers indicate that our participants rather tag their friends to notify them about their presence in pictures than to distribute their photos to others.

To assess the perception of being tagged, we asked participants to rate their feelings on the effects of being tagged in photos, on a 7-point scale from (1) *like it very much* to (4) *neutral* and (7) *dislike it very much* (cf. Fig. 2). The results indicate that becoming aware of photos of oneself is not the most important effect of tagging for our participants. This mirrors the Web 2.0 spirit: Most participants state they significantly prefer (Wilcoxon test, $Z = -3.41$, $p = .001$) finding photos of others with a mean of 3.51 ($sd = 1.37$) to finding photos of themselves with a mean value of 3.79 ($sd = 1.8$). However, participants also stated that they rather dislike that others can find their photos because of tags with a mean of 4.77, $sd = 1.55$. These results confirm typical assumptions about social sharing: SNS users like to be able to easily find photos of others while they dislike others being able to easily find pictures of themselves. Feelings about being informed about pictures of oneself tend to be more neutral which indicates that they see only little to no awareness benefits in being tagged. Therefore, people rather tag to follow the Web 2.0 spirit than for privacy reasons.

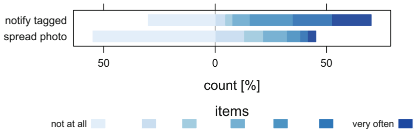


Fig. 1. (q13) How frequently do you tag for these reasons?

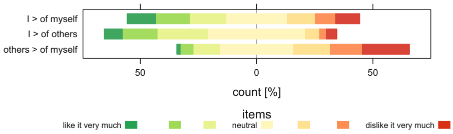


Fig. 2. (q14) Rate the effect of people tags: Who finds photos of whom

Limits of Tagging. The positive side of tagging people with profile links should not be underestimated. Indeed, such links are the only solution available in current SNS to notify people of photos of themselves besides any out-of-band communication between photographers and depicted people. The links offer a certain level of awareness, but one has to keep in mind that they are limited to photos of friends or indirect friends, because outside of these circles, access control, missing social connections and the lack of interest prevent notification.

To judge the seriousness of this deficit, we tried to assess the origin of privacy issues from the users' viewpoint. We asked our participants to rate the extent of a possible privacy violation by photos shared by different groups of people on a 7-point scale from *very low* to *very high* (cf. Fig. 3). Most respondents rated any violation higher than *very low*: Only 1.4 % of the participants rate a possible violation to be *very low* regardless of who shared the photo. The participants rate the violation level of photos shared by friends to be the lowest with a mean of 3.64 (sd = 1.85). Photos shared by friends of friends were rated to caused a medium level of violation on average (4.69, sd = 1.66) and the media shared by strangers was rated highest with an average rating of 5.23 (sd = 1.95). The differences in mean ratings are significant (Friedman test, $\chi^2_2 = 185.41$, $p < .001$)). Additionally, 47 % of participants rated privacy violations by strangers' photos consistently higher than those caused by direct and indirect friends. We conclude that participants perceive threats caused by strangers' photos to be worse than other privacy violations. In contrast to photos posted by direct or indirect friends, photos uploaded by strangers are neither tagged with entailing links, nor do they result in any notification. Therefore, profile links as a privacy feature have serious deficits because they do not cover this scenario.

The results on the extent of a possible privacy violation suggest that participants seem to believe that others do not comply with a "moral obligation", as described in [2], even though most people declare they think about other users' privacy when sharing media: We asked the participants to rate the influence of threats to others and threats to themselves as decision-making criteria for sharing a photo using a 7-point scale from *not at all* to *very much.* Only 2 % of our participants answered that they do not think about threats to others at all when sharing photos on the Web. Within the remaining participants, about 61 % rate threats to others and threats to themselves with the same value. Interestingly, 6.6 % of participants rated threats to others as a sharing criterion higher than threats to themselves.

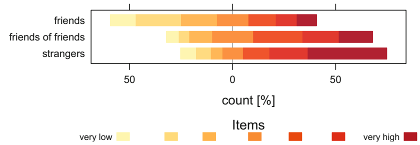


Fig. 3. (q17) Estimate a possible privacy violation of photos shared by ...

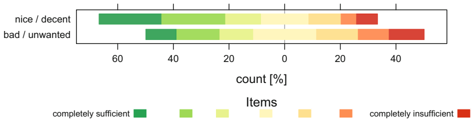


Fig. 4. (q19) How well do you feel informed about all photos of yourself?

2.2 Awareness Today

In the context of shared photo awareness, we also need to consider photos that contain identifying information but are not linked to profiles. Compared to photos directly linked to a person's profile and therefore immediately discoverable photos, unlinked photos are more critical: A tag that contains identifying information is attached to a photo, but no link to a person's profile is made. This can technically be implemented in a multitude of ways, ranging from mentioning a name in the headline or a comment in a SNS, to metadata that describes depicted people stored in the image file. While the potential damage of course is smaller, the threat can remain hidden far longer, because no automated mechanism helps to find this image. Currently, the only way to combat this threat is for the concerned person to pro-actively crawl the Web in search of such photos. We asked the participants of our survey to estimate the risk of someone finding a photo of them anytime in the future that this someone should not have seen. They assessed the likelihood of three scenarios of how they could be associated to a picture using a 7-point scale from *very low* to *very high*. While 24 % of the participants rated the risk of someone finding a photo that was previously linked to a SNS profile to be *very low*, only 11 % rated that risk to be *very low* if the photo contained personal references in the metadata or if they are only visible in a photo. This is an obvious result, since the tagged person is notified about photos linked to his or her SNS profile and can therefore be removed if necessary. Users see more future threats in unknown photos with personal references than in those they are only visible in: In the former case, 45 % of the participants rated the risk to be in the worst three elements of the scale, while only 35 % did so in the latter case. This difference is statistically significant (McNemar test, $\chi^2 = 10.32$, $p = .001$). This indicates that participants believed photos with actual personal references in the metadata to be more easily discoverable, for instance using a search engine, than those they are only visible in.

Finally, our study addressed to which extent users are satisfied with currently available options to become aware of photos of themselves. Thus we first queried respondents how they are currently becoming aware of photos of themselves, using a multiple-choice question. 75 % stated that they automatically get notifications by email when tagged in a photo (94 % of these were Facebook users); 52 % of the participants stated that they get to know about photos of themselves by chance; 39 % of them hear about photos of themselves in conversations and 30 % in friends' messages; 18 % actively look for photos; 4.6% get informed by messages from non-friends; and 3.4 % stated that they do not become aware of photos of themselves at all. Automated notifications are only possible in the case of profile-linked tags in current SNS. It is important to note that all the means of becoming aware of photos presented to the participants are not applicable in the case of non-linked tagging or missing tags.

Furthermore, we asked our participants to rate how well they feel informed about several types of photos of themselves on the Web, on a 7-point scale from *completely sufficient* to *completely insufficient* (cf. Fig. 4). Concerning decent photos, their perceived level of available information was a little better than neutral (3.2, *sd* = 1.85) and concerning objectionable photos, their average perception was exactly neutral (4.0, *sd* = 1.85). In detail, 22 % stated that their level of information is *completely sufficient* concerning decent photos of themselves while 25 % chose a level from worse than neutral to *completely insufficient*. In contrast, only 11 % state a level of *completely sufficient* concerning objectionable photos, while 39 % of the participants assert that their level of information about bad photos of themselves was worse than neutral to *completely insufficient*. Again, the differences between these values are statistically significant (McNemar test, $\chi^2 = 50.77, p < .001$). We finally asked the survey participants whether they would like to use a service that helps them finding relevant photos, which requires its users to manually screen potential photos. 53.1 % of them answered with a clear yes and 41.8 % were interested in using such a service. Only 3.6 % argued that the effort of screening would overbalance benefits. Others called on the uploaders' moral obligation or denied being depicted online at all.

2.3 Summary

Becoming aware of uploaded photos that a user is visible in is the key issue for combating privacy threats created by online media. Popular services allow their users to tag people in shared media. Mostly, tagging creates a link to the profile of that person. The tagged person is notified and can take action. Respondents did not see very much awareness benefits in such linked tags. Even if these features were fully appreciated, the privacy benefit is limited to photos of direct and indirect friends within their circles of friends. Photos shared by other people and outside of service boundaries cannot benefit from such mechanisms. Yet, users rate exactly those photos to pose the biggest threat for a possible privacy violation. In order to become aware of all relevant photos, photos with non-linked personal references as well as photos without any reference to a person have to be considered. For these types of photos, there currently are no effective possibilities to increase awareness besides manually crawling the web. When asked in which way and how well they are informed about photos they may be depicted in, participants' answers confirm that improvements are needed in the area of online media awareness and privacy. Although prior research has shown that users tend to spend little effort in privacy settings, nearly all of our participants are willing to invest at least some time in screening potential photos, if this offers a chance of being informed about potential privacy violations. A participant even offered to pay a one-time fee for such a service. The challenge is to implement a service that caters for the users' privacy needs and does not create new threats to the users' privacy at the same time.

3 Photo Metadata

Metadata is used to add valuable context information to images and helps to order, categorize and even find images in huge media libraries or by search engines. Metadata handling is integrated in nearly every image processing software and digital camera today. Modern devices automatically save several pieces of metadata with each photo, including the current date, time and GPS coordinates and even the camera owner's name. Additionally, an increasing number of applications support semi-automatic tagging of photos with textual location information based on reverse geocoding or tagging people within images. Besides the image itself, metadata of that image can also harm the privacy of a person.

Metadata can link people to images, for instance by storing names of photographers or depicted people. It can also contain information about the time or location of taking a photo that can create or amplify privacy threats.

3.1 Knowledge and Nescience of Users

Regarding privacy concerns of metadata, it is important to differentiate between data that is loosely attached to media for instance in the UI of a website and data stored directly in an image file. While the former is typically only accessible within the service and protected by access control, the latter is spread with the image and is generally as persistent as the image itself. Since only a part of all users (61 % of our survey participants) knows the term metadata, we can assume that even less know the difference between these two kinds of metadata storage and their respective implications. In our survey and consequently in this paper, we therefore only use the abstract term *metadata* to refer to additional information of photos, such as time, headline, or tagged people, regardless of how it is stored. During our survey, however, one participant commented: *"No difference was made between embedded metadata and metadata stored externally, that makes a world of differences when spreading a photo"*.

To estimate how users handle metadata, we asked the 253 participants that indicated to know what metadata is to agree or disagree to a set of statements. About 25 % stated that they do not add additional metadata to photos. However, some users might nonetheless do so in SNS, without knowing the term. About 6 % of the 253 participants stated that they remove all metadata from images before they share them on the Web and an additional 35 % stated that they remove parts of the metadata. 2 % said that the online services they use remove metadata on upload. Our participants also admitted to nescience: 58 % answered that they do not know what their SNS or media-sharing sites do with photo metadata. 29 % state that they do not know which additional information is contained in the photos they share. About 27 % of the 253 participants state that they do not think about metadata at all when sharing images on the Web. In contrast, 9 % of the 253 state that metadata is an important part of sharing.

3.2 Private Metadata

Most research and online services consider only few pieces of metadata of photos as confidential or related to privacy. We already discussed the practice of tagging people in current SNS in Sect. 2.1. Beyond these kind of tags, the location of a person or the location a photo was taken is most discussed in other papers and one of the few that is also specifically addressed in current services on the Web. The general term *location* mostly refers to GPS-based WGS-84 coordinates. Other location information, such as the name of a city or a point of interest, or an address where a photo has been shot is often not considered. But, since geocoding has become cheap and easy, coordinates and textual location information have to be dealt with equally. However, this is generally not the case: For instance, at the web-based photo sharing feature of Apple's iCloud Photo Stream, WGS-84 coordinates are removed, but any other location information is retained. This shows that we have to extend the current notion of privacy-related information in media metadata. Additional meta-information will raise privacy concerns in the future: The number of cameras that write a camera identifier into photos rises. These ids may not be as unique as a smartphone's IMEI, but still can be used to re-identify a camera owner. Additional concerns may arise from new metadata standards that allow tagging people with names and bounding boxes directly within image files. Up to now, this was only possible and known in the context of online SNS, but applications like Google Picasa or Windows Live Photo Gallery as well as libraries like exiv2 implement these standards today.

With our survey, we aimed to asses the users' view of the privacy implications of different pieces of metadata and how severe they estimate a possible privacy violation caused by the disclosure of such data to be. We asked our participants to rate the possible privacy impact of adding such metadata to media depicting others on a 7-point scale from *very low* to *very high*. Additionally, we asked them to rate the privacy impact of metadata added by others to media depicting themselves using the same scale. Table 1 in the appendix shows details of both.

Comparing the different kinds of metadata, headline, description, and tags are perceived to have the least impact with a mean rating of 3 ($sd = 1.7$) across both questions on the 7-point scale from *very low* to *very high*. The creation date and time of a photo (3.6, $sd = 1.7$), the photographers' name (3.4, $sd = 1.8$), and also broad location information, such as the city or region where a photo was taken, (3.9, $sd = 1.7$) are considered to have slightly less than medium impact. In contrast, the names of depicted people (4.9, $sd = 1.8$) and exact location information, such as GPS-based coordinates or a postal address, (5.2, $sd = 1.7$) are perceived as having a higher impact.

People. It is interesting to note the difference between names of depicted people and the photographer's name, since both indicate persons related to a photo. Finding the name of camera owners in photos also implicates their presence at that time and place, as long as the camera or smartphone was not lent to others.

Location. Our participants rated location as the kind of metadata with the highest privacy impact. However, recent related work voiced doubts that location still raises much concerns with today's smartphone users, compared to the beginning of the mobile era. For instance, in the "very-upset-ranking" of Porter Felt et al. [11], the participants ranked location-related risks in the bottom half and the actual location was ranked second-lowest out of eleven data types. Fisher et al. [4] show that iOS users seem to pay attention to which apps they allow to use location and do not disable the feature in general. Krumm [8] summarizes different results, showing that people do not seem to care about location privacy. So why does our data differ?

The prior work mainly deals with location in the context of location-based services, the pro-active publication of locations, or the misuse of location permissions by smartphone applications. In all these cases, location and where that information is stored may be less tangible to people. Our survey has been conducted in the context of photo sharing on the Web. In this case, location is at least connected to a picture and eventually to additional meta-information. A photo may be seen to last longer in the public. Photos are indeed not touchable, but much more concrete in the participants' mind than a single location recorded by an abstract service. Caused by the higher familiarity with photos, location data in pictures may raise more privacy concerns than in other contexts. To the best of our knowledge, no previous work compared users' feelings about location data in different contexts. We believe that there are different aspects that may explain the differences of results, which we will investigate in future research.

To examine the influence of the audience when disclosing location data, we asked our participants to rate how they felt if people get to see a photo of them that includes location information, using a 7-point scale from *very unconcerned* to *very concerned* with 4 as *neutral.* When sharing a photo with location data with friends (2.24, $sd = 1.5$) or friends of friends (3.51, $sd = 1.7$), participants state to be more or less unconcerned and more concerned in the case of other people (5.16, $sd = 1.8$). However, when it comes to servers, for instance the service that hosts the photo (5.23, $sd = 1.8$) or a privacy service that searches for depictions (5.28, $sd = 1.9$), people state to be even more concerned, which is contrary to the results of Felt et al [11]. The scenario of a privacy service will be discussed in the next section.

3.3 Summary

In this section, we discussed the role of metadata on the respondents' perception of privacy. While users of SNS know that they can add comments, locations or people tags to images on the Web, the general idea of metadata seems still to be less known to users. Only few people know about metadata that is stored directly in photos. Consequently, few people know about privacy-related data that might already be contained in images before they are uploaded to the Web. Even if they do know about the data, we have to ensure that people are aware of the contained information: For instance, the photographer (and therefore also the likely owner of the camera) was also present when a photo was taken. There also is little difference between GPS-based coordinates and postal addresses due to geocoding. Additionally, we presented results that are in conflict with previous investigations on sharing location data. Further research is needed to examine if and why there is a difference in perception.

In general, the potentially important role of metadata has to be made clear to users who are concerned about their privacy. Additionally, many processes that handle metadata are not forthcoming about which kind of information they handle in which way. For instance, it needs to be clear that if location information is removed, all kinds of location information are removed, including coarse locations or geocoded information. Moreover, there is little awareness of which information is stored in images by software and cameras: A single option in Google Picasa decides if people tags are stored in its database or are written into the files. Most users are not aware of the consequences of this choice. Canon cameras can also write the camera owner's name into the metadata, which also has possible privacy implications. Regarding photos and metadata, transparency and usable privacy mechanisms are needed to lower privacy threats as well as the danger of nescience.

4 A Privacy-Privacy-Tradeoff

Traditional privacy research aims to preserve users' privacy at all cost. We propose that this is not necessary and desirable in real world systems, especially in the social Web that is built around contributing and sharing. Users decide which aspects of their personal data they disclose to others. The Web 2.0 spirit shows that many people are happy about sharing things as long as they benefit from it or appear in a positive light.

Photo metadata can contain various information from technical details about the camera used to context information about the who, when, where and what of a photo. It can be used to preserve the non-visual context of a photo or it can be used to order a huge collection of images. In addition to these traditional use cases, we propose to also use some pieces of metadata for security and privacy purposes. A somewhat related intention can be found in the work of Klemperer et al. [7]: they derive access control rules for images from their keywords. In contrast, we propose to leverage image metadata to protect the privacy of the people affected by an image by allowing people to become aware of it [5].

The following scenario illustrates how metadata can be used to this end: *The service S assists users in finding media that might be relevant to them. S may be implemented as a value-added service within a SNS. Users of S can define private locations on a map or update their current location at the service through "checking in" or similar approaches. Based on co-location checks of users' private areas and the location information of photos uploaded to the SNS via S, the service notifies users who may be depicted in a photo based on respective locations. Additionally, SNS profile pictures can be used as training data for face recognition to improve results.* In this example, the service S leverages location metadata and profile pictures of users to make them aware of photos, so that they can protect themselves against unwanted publication.

Most of the necessary metadata is private to the affected people. If we want to use this information for privacy protection, we face some fundamental questions about the privacy of information that potential users have to decide for themselves: Firstly, is all information that at least some people regard as private also private to the user? Secondly, is all information that the user regards as private equally private in the way that the number or groups of people or services, which he allows to get to know the information, are identical? Otherwise, what information would the user share with which people or services? This creates *privacy levels* containing information that is similarly relevant to users' privacy. While most privacy fundamentalists and privacy unconcerned might have exactly one level of privacy, the number of privacy pragmatists in our study was found to be considerably higher. We therefore suggest building privacy mechanisms based on privacy levels.

To confirm the usefulness of this suggestion, we asked the participants of our survey to what extent they agree to the existence of privacy levels as defined above (cf. Fig. 5). On a 7-point scale from (1) *strongly agree* to (7) *strongly disagree* with 4 as *neutral*, the participants provided a mean agreement of 2.63 (*sd* = 1.7). 13.5 % of the participants indicated disagreement (5.6 % strongly disagree), 12.3 % were neutral, and 74.2 % indicated agreement (33.1 % strongly agree). We found no relation between the answers and the Westin segmentation. According to these results, participants generally feel that there are different levels of privacy, while about one third strongly supported this notion.

If privacy levels exist, we can take advantage of them: A privacy service may leverage some information that is less private to a person to secure other information that is more private to that person. We call this a *privacy-privacy-tradeoff*: If privacy levels exist in a system that builds on (public but also private) information – like current SNS and other social sites – users can choose to disclose less private information to secure other, more private information.

In our survey, we validated this idea by asking if our participants agree to this kind of tradeoff. We accompanied this question with a short description of the above scenario where they could choose to "reveal their location to a service to get notified about photos in which they might be depicted". We asked the participants to rate their agreement using a 7-point scale from (1) *strongly agree* to (7) *strongly disagree* with 4 as *neutral*, concerning if they, in general, would disclose some private information to secure other more private information (cf. Fig. 5). Participants gave a mean agreement of 3.49 (*sd* = 1.7). While 22.7 % of the participants indicated disagreement to this privacy-privacy-tradeoff and 24.5 % of them answered neutrally, 52.7 % of the 414 participants voiced their agreement to the tradeoff.

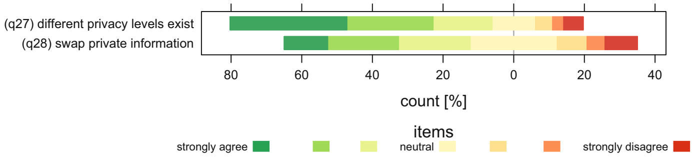


Fig. 5. (q27) Do privacy levels exist? (q28) Would you in general share some private information to secure other more private information?

We also asked participants which information they would trade for being notified about photos they might be depicted in that would otherwise be hard to find or even not accessible. We used the same 7-point scale as above. As shown in Fig. 6, participants mostly agreed using their existing profile pictures (2.97, *sd* = 1.8) to get notifications about photos. These, for instance, could be used to train face recognition. The second information the participants would be willing to disclose to some extent is pre-defined locations (4.0, *sd* = 1.9) that could be used for co-location checks to find photos at static places like home. On average, participants were reluctant to provide additional profile photos that comply with guidelines, like for a passport (4.5, *sd* = 1.9), which would be more suitable to train face recognition. They also indicated slight disagreement on providing their SNS list of friends (4.5, sd = 1.8) that could be used to specifically monitor friends' photos. Participants disagreed most to the use of a location-based service to constantly disclose their current location to the photo-service (5.4, sd = 1.9). This kind of data would obviously allow for the most effective co-location checks with photos.

Altogether, besides the use of existing profile photos, our participants on average disagree to trade private information when it comes to implementing a real tradeoff. However, if we consider respondents that indicated agreement (including those that would not mind) on trading private information as potential users, we get the following percentages: 67.5 % (82.6 %) might allow the use of existing profile photos, while 35 % (51.7 %) would provide extra photos complying to guidelines. 30.9 % (50.5 %) would allow to use their friends list. 39.6 % (61.8 %) would define private locations on a map and 19.1 % (31.2 %) would use the location-based service to update their current location.

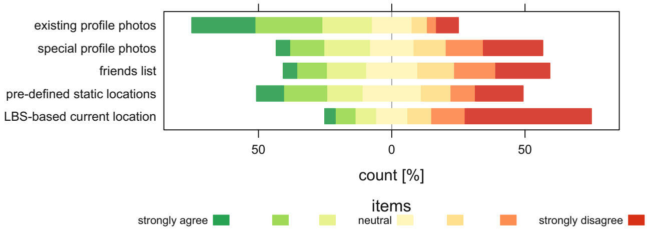


Fig. 6. (q30) What information would you disclose to a photo-sharing service to find photos of yourself that you otherwise would not be able to find or access?

Additionally, we added three questions to our survey that describe specific tradeoff situations to investigate agreement using the same 7-point scale:

q31: *"I am less upset if someone finds out where I have been than if that person gets to see private photos of myself."* — Participants somewhat agreed on average (3.0, sd = 1.7); 66.2 % indicated agreement and 15.7 % answered neutrally.

q32: *"I am less upset if my SNS knows where I have been than if my friends and strangers gets to see unwanted photos of myself."*— Again, Participants somewhat agreed on average (3.3, sd = 1.8); 60.4 % indicated agreement and 16.2 % answered neutrally.

q33: *"If there is a privacy service that notifies me about unwanted photos in which I am depicted but needs to know where I have been, I would use it. I would tell it where I have been to get to see possible photos of myself."*— Participants provided an average agreement of 3.7 (sd = 1.8); 53.2 % indicated agreement and 16.1% answered neutrally.

While all answers differ significantly (Friedman test, $\chi^2_2 = 44.46, p < .001$), answers to the first two questions appear to be weakly correlated (Spearman's $\rho_{1+2} = 0.596, p < .001$) and answers to the third appear to be independent ($\rho_{2+3} = 0.236, \rho_{1+3} = 0.226, p < .001$). Hence, respondents generally indicated agreement to scenarios stating a direct privacy tradeoff, but were more reluctant about disclosing information to a service to get notified of possible photos of them. This may imply that participants do see a privacy tradeoff but are not quite willing to trust another service to keep even less sensitive data private.

4.1 Summary

Our hypothesis that not all private information is equally private to people but is structured into several privacy levels was confirmed by our results; only 5.6 % of participants strongly disagreed. Given that privacy levels exist, we suggested leveraging this circumstance: We proposed to use less private information to secure information that is more private to users. We asked participants to what extent they would agree to a privacy-privacy-tradeoff. In general, 77.2 % agreed or were neutral towards this proposal. However, when participants were asked about a real implementation instead of a general idea, less people agreed to trade private information. While participants agreed to disclose SNS profile pictures for notifications about photos, they were generally more reluctant towards other information, especially location. However, a considerable amount of participants was ready to trade private information and may therefore be considered to be potential users of tradeoff-based privacy mechanisms.

The results of the explicit tradeoff situations confirms this impression: 60.4 % of the participants agreed that they prefer their SNS knowing where they were rather than other people, from inside or outside of their social circle, seeing unwanted pictures. Furthermore, 53.2 % of participants directly agreed to using a service offering this privacy-privacy-tradeoff.

5 Conclusion and Future Work

The results of our survey give a detailed account of the privacy preferences users have concerning the sharing of photos and their perceptions about linking photos to people. The assessment of the users' current degree of awareness shows that improvements are needed in the area of online media awareness and that users are willing to accept additional effort to gain improved awareness.

We investigated the role of metadata and differences in the perceived privacy impact of the unwanted disclosure of specific metadata: Personal references and location data raise most concerns for the users. These findings partly contradict the current views in related work, which state that users are not particularly concerned about location information. We therefore suggest that location privacy needs to be reconsidered in general and especially in the context of shared media, since our survey indicates that there are strong concerns about disclosing this kind of location information.

We also discussed the general idea of a privacy-privacy-tradeoff. Our survey shows that such a tradeoff would be appreciated by a fair number of users. The willingness to use a tradeoff-based service depends on the offered benefits: When participants were asked if they wanted to become more aware of photos of themselves, most agreed. However, the disclosure of meta-information and private data was also considered an issue. Finding the right balance in this tradeoff is an interesting topic of future research. We hope the results presented in this paper can serve as a basis for designing privacy-privacy-tradeoff-based services that take the users' perceptions into account.

References

1. Ahern, S., Eckles, D., Good, N., King, S., Naaman, M., Nair, R.: Over-exposed? Privacy patterns and considerations in online and mobile photo sharing. In: Proc. SIGCHI Conference on Human Factors in Computing Systems, CHI 2007 (2007)
2. Besmer, A., Lipford, H.R.: Moving beyond untagging: photo privacy in a tagged world. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010 (2010)
3. Besmer, A., Lipford, H.R.: Privacy Perceptions of Photo Sharing in Facebook. In: Proc. of the Fourth Symposium on Usable Privacy and Security, SOUPS 2008 (2008)
4. Fisher, D., Dorner, L., Wagner, D.: Short paper: location privacy: user behavior in the field. In: Proc. of the 2nd ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, SPSM 2012 (2012)
5. Henne, B., Szongott, C., Smith, M.: Snapme if you can: Privacy threats of other peoples' geo-tagged media and what we can do about it. In: Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec) (April 2013)
6. Johnson, M., Egelman, S., Bellovin, S.M.: Facebook and privacy: it's complicated. In: Proc. of the 8th Symposium on Usable Privacy and Security, SOUPS 2012 (2012)
7. Klemperer, P., Liang, Y., Mazurek, M., Sleeper, M., Ur, B., Bauer, L., Cranor, L.F., Gupta, N., Reiter, M.: Tag, you can see it!: using tags for access control in photo sharing. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012 (2012)
8. Krumm, J.: A survey of computational location privacy. Personal and Ubiquitous Computing 13(6) (August 2009)
9. Kumaraguru, P., Cranor, L.F.: Privacy Indexes: A Survey of Westin's Studies. Tech. Rep. CMU-ISRI-05-138, Carnegie Mellon University (2005)
10. Liu, Y., Gummadi, K.P., Krishnamurthy, B., Mislove, A.: Analyzing Facebook privacy settings: User expectations vs. reality. In: Proc. of the 2011 ACM SIGCOMM Internet Measurement Conference, pp. 61–70. ACM (2011)
11. Porter Felt, A., Egelman, S., Wagner, D.: I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In: Proc. of the 2nd ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, SPSM 2012 (2012)
12. Squicciarini, A., Xu, H., Zhang, X.: CoPE: Enabling collaborative privacy management in online social networks. Journal of the American Society for Information Science and Technology 62(3) (March 2011)
13. Warren, S.D., Brandeis, L.D.: The right to privacy. Harward Law Review 4(5), 193–220 (1890)

A Additional Details to Participants' Answers

A.1 Online Photo Awareness

Figure 7 shows the answers concerning different decision-making criteria for sharing photos on the Web as discussed in Sect. 2.1. Figure 8 shows answers about the estimated chance that someone anytime in the future finds photos that may raise privacy concerns. The items differentiate the ways how a photo is connected to a person as described in Sect. 2.2.

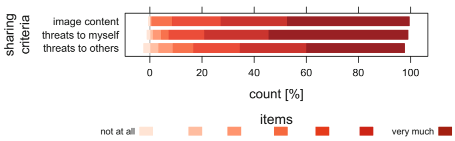


Fig. 7. (q11) Rate the influence of the items as criteria for sharing a photo

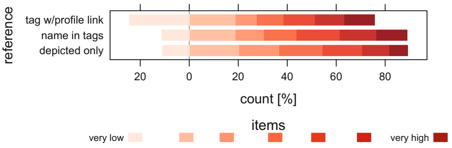


Fig. 8. (q25) Estimate the risk that someone finds an unwanted photo anytime in the future

A.2 Metadata Privacy

As presented in Sect. 3.2, we asked our participants to rate the possible privacy impact of adding metadata to media depicting others on a 7-point scale from *very low* to *very high*. Additionally, we asked them to rate the privacy impact of metadata added by others to media depicting themselves using the same scale. Figure 9 shows the answers to both questions and Table 1 a summary of them.

Table 1. Estimation of the impact of metadata

metadata added by with impact to	myself others		others myself		*Wilcoxon* *signed ranks*		*Spear-* *man's* ρ
	mean	*sd*	*mean*	*sd*	*Z*	*p*	($p < .001$)
headline, description, tags	2.94	1.66	3.23	1.75	−4.739	.000	0.73
date & time of creation	3.63	1.70	3.59	1.67	−0.478	.632	0.69
photographer's name	3.49	1.79	3.28	1.83	−3.274	.001	0.65
depicted peoples' names	5.08	1.70	4.76	1.87	−4.204	.000	0.64
broad location (city, region)	3.95	1.62	3.90	1.74	−0.902	.367	0.68
exact location (address, GPS)	5.31	1.68	5.17	1.75	−2.102	.036	0.68

Spearman's ρ indicates that participants' answers to both questions correlate positively (ρ between 0.64 and 0.73, $p < .001$): those who see a higher impact on their own privacy also see a higher impact on other's privacy. We also found a trend that respondents who stated to use location metadata more frequently also saw less privacy impact through that kind of metadata. This may indicate that people who add a particular kind of metadata are more open for the benefits of such information and thus have less concerns about their privacy impact.

The extent of the estimated impact on privacy appears to be independent from the direction of a threat, i. e. regardless of whether a participants's own metadata harms others or foreign metadata harms the participant. For most kinds of metadata, participants perceived that their own metadata has a higher privacy impact on others than others' metadata has on themselves. While some differences were statistically significant, the differences were only slight.

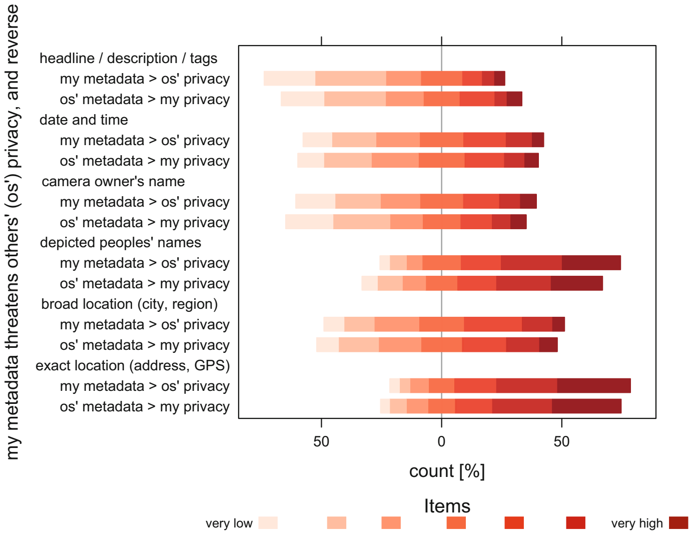


Fig. 9. (q23) Estimate the impact of metadata you add to shared photos on others. (q24) Estimate the impact of metadata others add to shared photos on you.

To compare the privacy levels of different metadata, we asked participants to rate the privacy of different metadata in the context of a privacy-privacy-tradeoff. We used a 7-point scale from *completely public* to *completely private*. The major results as shown in Fig. 10 are congruent with the question about the impact of metadata (cf. Fig. 9). Exact location information is considered to be the most private kind of data, with GPS-based coordinates being more sensitive (91.1% somehow private, 60.4 % completely private, $m = 6.26, sd = 1.2$) than addresses or location names (88.9 % somehow private, 45.7 % completely private, $m = 6.02, sd = 1.4$). Broad locations, like city names, have a mean value of $m = 4.21$ ($sd = 1.5$, median $= 4$, mode $= 5$). People depicted in the image are the second most private group of metadata, where tags with bounding boxes ($m = 5.46, sd = 1.4$, median $=$ mode $= 6$) in the image are regarded as slightly more private as those without (mean = 5.14, sd = 1.4, median = mode = 5). Again the name of the photographer is regarded as less private (mean $=$ median $=$ mode $= 4, sd = 1.7$). The unique id of a camera is also perceived to be more private, with a mean value of 4.72 ($sd = 2$, median $= 5$, mode $= 7$).

Figure 11 shows feelings about photos with embedded location information that someone might stumble upon as discussed in Sect. 3.2.

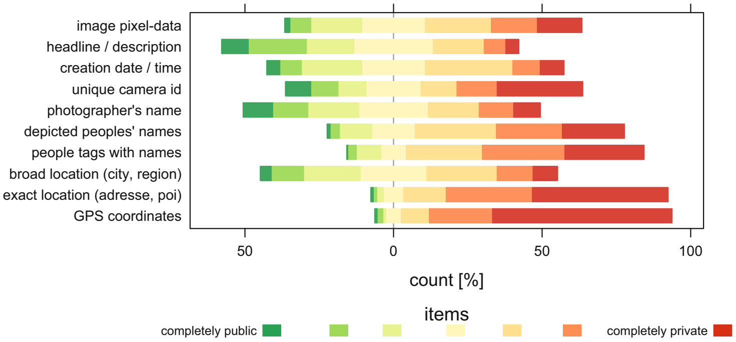


Fig. 10. (q29) How do you feel about the privacy of photo metadata?

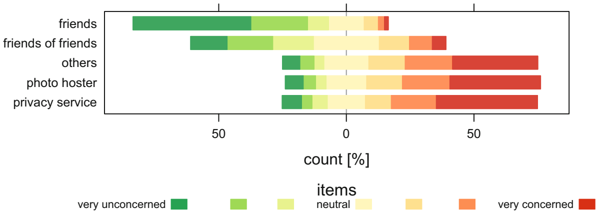


Fig. 11. (q26) How do you feel when these people get to see a photo of yourself that includes location information?

A.3 Privacy-Privacy-Tradeoff

Figure 12 shows answers to the three explicit privacy-privacy-tradeoffs as presented in Sect. 4.

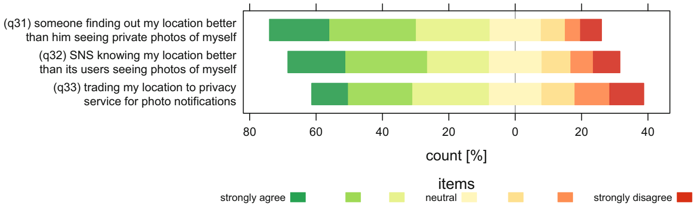


Fig. 12. (q31 - q33) Three explicit privacy-privacy-tradeoffs (cf. Sect. 4)

Bootstrapping Trust in Online Dating: Social Verification of Online Dating Profiles

Gregory Norcie[1,⋆], Emiliano De Cristofaro[2], and Victoria Bellotti[2]

[1] Indiana University
[2] PARC (A Xerox Company)

Abstract. Online dating is an increasingly thriving business which boasts billion-dollar revenues and attracts users in the tens of millions. Notwithstanding its popularity, online dating is not impervious to worrisome trust and privacy concerns raised by the disclosure of potentially sensitive data as well as the exposure to self-reported (and thus potentially misrepresented) information. Nonetheless, little research has, thus far, focused on how to enhance privacy and trustworthiness. In this paper, we report on a series of semi-structured interviews involving 20 participants, and show that users are significantly concerned with the veracity of online dating profiles. To address some of these concerns, we present the user-centered design of an interface, called Certifeye, which aims to bootstrap trust in online dating profiles using existing social network data. Certifeye verifies that the information users report on their online dating profile (e.g., age, relationship status, and/or photos) matches that displayed on their own Facebook profile. Finally, we present the results of a 161-user Mechanical Turk study assessing whether our veracity-enhancing interface successfully reduced concerns in online dating users and find a statistically significant trust increase.

1 Introduction

In the last few years, social networking has remarkably altered our social ecosystem. *Online Social Networks* (OSNs) offer highly efficient means for establishing or maintaining social connections. A 2011 survey of 6,000 people found that about 35% of respondents reported spending more time socializing online than face-to-face and that 33% of users were more likely to speak to someone new online than offline [3].

Alas, the resulting ubiquitous gathering and dissemination of personal information also prompts some important privacy and trust concerns. The research community has begun to investigate how publicly sharing some kinds of personal information can help malicious entities launch various schemes such as creating personalized phishing attacks [23] or guess Social Security numbers [2].

Motivated by the significance and increasing recognition of trust and privacy issues [6], researchers have applied principles from human-computer interaction to the design of security software, e.g., for social network chats [15], anonymizing social graphs [16], and file/email encryption [41]. One of the most difficult challenges in enhancing trust, security, privacy in most networked systems has always been, and still

[⋆] Work done while the author was an intern at PARC.

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 149–163, 2013.

remains, the human factor [22,36]. Awareness, perceptions, and reactions of non-tech-savvy users are subjective, and security experts and researchers have often struggled to gain understanding of the concerns of everyday usersfor whom security is not a primary task [13]. Making this situation more difficult to contend with, study volunteers have a tendency towards social desirability distortion, wishing to present themselves in a favorable light to the experimenter by omitting information or even misrepresenting their behaviors [33]. Thus, it is not surprising that participants report safer practices and higher privacy concerns than they actually demonstrate in practice [1].

As a result, we have been inspired to develop means for end users of social media to side-step awkward security and privacy mechanisms in achieving their social objectives. In particular, we focus on *Online Dating Services* (ODSs), attracted by their growing popularity and distinctive anthropological characteristics. ODSs differ from typical OSNs, as they are almost exclusively used to connect with strangers. boyd [4] defines OSNs as services allowing users to create and view web profiles, as well as to search other profiles to connect and communicate with. boyd also mentons that these connections almost exclusively involve existing or shared social contacts – *not* strangers. Conversely, the typical ODS *does* aim to connect strangers.

Online dating has become remarkably popular: 20% of heterosexual couples and 60% of same-sex couples now report having met online [34]. In 2012, more than 40 million users were estimated to be part of a $1.9 billion revenue industry [11]. Notwithstanding its popularity, online dating raises some worrisome privacy and trust issues. Because of the inherent need to engage with and reveal potentially sensitive information to unknown others, ODSs amplify many of traditional social-networking security and privacy issues. And yet, somewhat surprisingly, very little research has focused on the problem of how to make them more trustworthy and privacy-respecting.

Motivated by the above concerns, this paper explores a number of potential ways to enhance trust and privacy in ODSs. We seek ODS users' input by conducting a series of semi-structured interviews (involving 20 participants) and concentrate on concerns that are particularly relevant to non-security-savvy users. We found that participants were particularly worried about the veracity of ODS profiles. Following these semi-structured interviews, we present a user-centered design of a mock-up interface, called *Certifeye*, that allows users to certify some attributes, e.g., age, relationship status, photos, in their ODS profile. Certifeye does so by attesting that the information reported on the ODS profile matches that same information on the user's Facebook profile. Certifeye can be plugged on any existing ODS to add profile certification and bootstrap trust among users. In theory, profiles can be certified using a number of sources; however, driven by our user-centered design approach, and following a series of semi-structured interviews, we choose to use existing social network data for this purpose. For instance, if a user is listed as "in a relationship" on Facebook, it would be cumbersome to change her status to "single", as friends and, most embarrassingly, the user's partner would likely notice this change. Profile certification is performed seamlessly, by granting the ODS provider access to one's Facebook profile through Facebook API, which results into obtaining a "certification badge". Certifeye does not require users to mutually "open" their Facebook profiles and does not aim to replace the ODS credentials with Facebook credentials in the vein of OAuth [29] or OpenID [30] technologies – we only need a mechanism to verify the matching of information without accessing/storing any private information.

In order to assess whether or not a mock-up of our veracity-enhancing capability can successfully reduce trust concerns, we conducted a Mechanical Turk study involving 161 participants. The ODS users were asked to rank their concern on a scale from 1 to 7 (1 being not at all concerned, 7 extremely concerned), and we found a statistically significant increase in trust present when users were shown our proposed interface.

2 Related Work

This section reviews prior work on trust and privacy in Online Dating Services (ODSs). We also survey previous efforts toward the design of usable, private, and trustworthy Online Social Networks (OSNs) and analyze whether solutions applicable to OSNs can be adopted to ODSs.

2.1 Trust and Privacy Concerns in Online Dating

Misrepresentation in OSNs has been recently investigated by Sirivianos et al. [38], who introduced “FaceTrust”, a system using social tagging games to build assertion validity scores for profile information. Likewise, misrepresentation is an issue for ODSs as well and was first analyzed by Brym et al. [5], who reported that 89% of participants (ODS users) felt that “people online might not tell you the truth about themselves” and 85% agreed that “people you meet online might be hiding something.” Ellison et al. [14] also pointed out that online daters must balance two conflicting goals – they need to present oneself in the most positive light in order to attract a mate, while simultaneously knowing that one must be honest if one wants their relationship to progress past the first meeting. Thus, ODS users must balance positive self-presentation with transparency. However, while ODS users might not lie about crucial traits, they do alter attributes that they consider minor, such as age or height. Toma et al. [39] reported consistent, conscious misrepresentation. In their study, eighty participants were asked to recall the height, weight, and age listed on their ODS profiles. Participants were generally able to accurately recall the information on their ODS profile. The information on the participants profiles was found to be significantly different from what was reported on their driver’s licenses. The deception was found to be *not* subconscious, since participants were able to recall the false values.

Recent research also shows that online daters daters take action when they suspect misrepresentation. Gibbs, Ellison, and Lai [17] found that many participants often engaged in information seeking activities, such as “Googling” a potential date.

In addition to trust issues, there seems to be some privacy concerns associated with using ODSs, including disclosing one’s presence on a dating site. Couch et al. [10] described how the risk of “exposure” – i.e., a coworker or acquaintance stumbling across one’s profile. However, Couch’s participants who reported exposure concerns were users of specialty fetish sites and/or sites geared specifically towards extremely short-term relationships. Conversely, our interviews focused on how users looked for medium to long term relationships, and we feel there is no longer a social bias against users seeking long term, monogamous relationships using ODSs.

Finally, Motahari et al. [26] discussed the concept of *social inference*, finding that that 11% of study participants could correctly identify who they were communicating with, by utilizing out-of-band knowledge. For instance, someone may know there is only one female, hispanic local soccer team member, and use that outside knowledge to de-anonymize a seemingly anonymous profile.

2.2 Usable Trust and Privacy in Online Social Networks

As we discussed earlier, although ODSs and OSNs share several similar traits (e.g., the ability of viewing profiles, listing interests, and exchanging messages), they actually provide different types of services and incur different challenges. Nonetheless, by examining prior work on OSNs, we aim to derive best practices for ODSs.

Privacy, trust, and security issues are often associated with the collection, retention, and sharing of personal information. One reason privacy concerns are pervasive in OSNs is because security is not a primary task. As Dourish et al. [13] pointed out, users often view security as a barrier preventing them from accomplishing their goals. Furthermore, users may be unaware of the risks associated with sharing personal information. Data posted on social networks can be subject to subpoena or, even after years, can regrettably re-surface, e.g., during job hunting or an electoral campaign. Furthermore, social networking data can be used for social engineering scams. For instance, Jagatic et al. [23] showed that extremely effective phishing messages could be constructed by data mining social networking profiles to personalize phishing messages.

Motivated by the significance of associated threats, a considerable amount of work has been dedicated to user-centered design of privacy and trust enhanced OSNs. Privacy and trust are similar, but separate concepts. Nissenbaum [28] discussed the concept of "contextual integrity", pointing out that personal information is not simply private or public – privacy depends on context.

As pointed out by Camp [7], *trust* is separated from *privacy*, in that trust is the belief in the integrity or authority of the party being trusted. Thus, trust is extremely closely connected to *veracity* and *reputation*. Nonetheless, trust is similar to privacy in that it must be incorporated into software's design. Naturally, the task of effectively incorporating privacy and trust into a design may be challenging. Cavoukian [9] described 7 principles of "privacy by design." These principles aim to embed privacy and data protection a throughout the entire life cycle of technologies, from the early design stage to their deployment, use and ultimate disposal.

Similarly, Murayama et al. [27] discussed how the Japanese concept of "*anshin*" – the emotional component of trust – can be taken into account when building systems. While the concept of *anshin* may not be known by name to westerners, the concept itself is not new. For example, Bruce Schneier used the term "security theater" [37] to describe security measures taken by the TSA to increase the public's feeling of trust in flying post 9/11 which serve no useful purpose.

It could be argued that acts of security theater are attempts to create *anshin*. By operationalizing *anshin* in this manner, we can see that it is important to increase trust in a system, as well as see that failed attempts to increase trust can lead to user frustration.

Another issue typical of OSN is *over-sharing*. When social networks do not embed privacy into their designs, users tend to over-share and make dangerous errors.

For instance, Wang et al. [40] surveyed 569 Facebook users and found that 21% of users had regretted posting information on Facebook, with regrets usually centering around sensitive content, strong sentiments, or because the post exposed a lie or secret. Moreover, even content that users do not regret posting can have privacy implications. Gross and Acquisti [21] crawled the profiles of Carnegie Mellon University's Facebook population in 2005, and found that 90.8% of profiles publicly displayed images, 39.9% publicly displayed phone numbers, and 50.8% publicly displayed their current residence. They also found that most users had not changed their privacy settings from Facebook's defaults. Sharing this kind of information can be harmful, aiding an attacker in various re-identification attacks, such as guessing a user's Social Security numbers based on publicly available information [2].

In summary, while prior work has focused on privacy and trust in OSNs, or analyzed misrepresentation in ODSs, our work is the first, to the best of our knowledge, to present a user-driven and user-centered design of an ODS interface that enhances trust by leveraging information that is already available in OSNs.

3 Study Part 1: Ideation and Interviews

As mentioned in Section I, a remarkable amount of people who classify themselves as "single and looking" use Online Dating Services (ODSs) today. Naturally, online dating presents numerous privacy and trust issues, yet, little work has focused on them. In order to avoid unnecessary effort on unfocused interviews (e.g., covering *all* possible concerns about online dating), our first step was to brainstorm on a few concepts for privacy and trust enhanced online dating, informed by what prior research suggests as promising problem areas. Two security and privacy researchers held a series of informal brainstorming sessions and then reviewed and evolved the ideas with a third researcher experienced in user-centered design.

3.1 Initial Concepts

Based on our expert brainstorming session, we came up with four possible privacy and/or trust enhancements to existing ODSs.

1. ***Identifying potential romantic partners in social circle without trusted third parties:*** Consider the following scenario: Alice is attracted to Bob, but she does not want to reveal her sentiment, unless Bob also likes Alice. Alice belongs to a social networking service (e.g., Facebook), and installs an application that lets her list the people that she would like to date. By utilizing appropriate privacy-enhancing technologies (e.g., [12]), this information could be exchanged and stored in such a way that: (i) users only learn whether there is a match, and (ii) the provider does not obtain any information about users' interests and/or matches.
2. ***Identifying potential partners based on matching interests, without trusted third parties:*** Many ODSs ask extremely personal questions, the answers to which users would like to reveal only to prospective romantic partners. Similar to the above idea, users' answers could be exchanged and stored in a privacy-preserving manner, so that only users with a minimum number of matching interests would disclose personal information.

Table 1. Age breakdown for interviewees

Age	N	%
18–25	6	30%
26–35	4	20%
36–55	8	40%
55–70	2	10%

Table 2. Education breakdown for interviewees

Degree	N	%
Some college, no degree	4	20%
Associate's	1	5%
Bachelor's	6	30%
Master's	8	40%
PhD	1	5%

3. ***Automatic Exclusion of Coworkers/Friends:*** It may often be embarrassing to admit to friends that one is using ODSs to find someone for a certain kind of relationship. Therefore, it might be useful to grant the dating service access to their list of Facebook friends, and exclude friends, coworkers, and/or family members from seeing their profile.
4. ***Certification of Profile Data:*** A natural concern in ODSs is related to veracity of user profiles. It seems possible that some service could pull data from another social source (e.g., from Facebook) and certify on the dating site that the information is accurate.

Having developed these four ideas, our next step was to determine which of them (if any) matched the users' needs. Therefore, we conducted a series of semi-structured interviews where we asked users to describe their Facebook habits and their ODS habits. We also posed several questions specifically designed to explore our initial ideas. (For example, would users be willing to link their social networking profile to their ODS profile?) We discuss these interviews in details below.

3.2 Interview Methodology

We recruited 20 users from a local classified advertisements website, as well as from mailing lists at a local university.[1] Interviewees were required to be past or present users of ODSs. The male to female ratio was roughly equal (55% female). Participants ranged in ages from 24 to 70 and were mostly educated, with 75% of users possessing at least a bachelor's degree. Age and education breakdowns are reported in Table 1 and 2.

Prior to the interview, participants were asked to fill out a short online survey. Besides demographic information, they were presented with some multiple-choice question covering their ODS habits, the services they use, the types of information they post on Facebook, and whether or not they refrain from posting certain types of information on Facebook. Participants were also asked about where they meet partners offline. After the computer survey was filled out, a semi-structured face-to-face interview session was arranged where users were asked to log into their Facebook account. We then asked the interviewee to scroll down their "feed" until he or she came across an embarrassing, controversial, or "edgy" post, either by themselves or others. Such "provocative" items

[1] Our studies obtained the Exempt Registration status from PARC's Institutional Review Board (FWA Number: FWA00018829, Expiration 5/2/2017).

were used to drive the discussion, aiming to tap into our interviewees real responses to social actions (e.g., disclosures within the social networking environment), rather than have them imagine how they might feel about abstract privacy and trust issues. While examining the mini-feed, users were asked questions such as:

- Why did you choose to post this?
- Who can see this post? Did you consider that when posting this?
- Your friend made this update: would you post something like this? Why/why not?

We then asked the interviewees to discuss whether or not they would be willing to use a theoretical online dating application which existed within Facebook, and whether they would be willing to share their Facebook information with an ODS. The total time spent on the computer survey and interview was, on average, approximately 45 minutes.

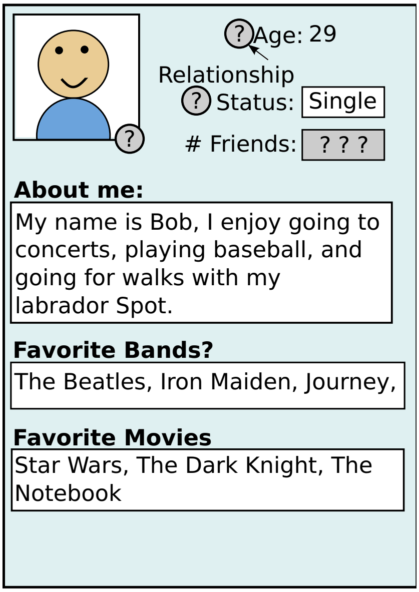
3.3 Interview Results

We anticipated that a relevant concern for our interviewees would be related to disclosing, e.g., to their social circle, the fact that they dated online. Previous work [10] had found that some ODS users feared exposure of their ODS habits, however, this concern was limited to users of "an online dating website focused on sexual interests", whereas, our interviews were focused on how interviewees used more typical romance-oriented ODSs – with fewer potentially embarrassing connotations. We found that a majority of interviewees (15/20) did not see online dating as very embarrassing. While many participants (14/20) did not want their entire social graph (including loose ties such as coworkers and/or acquaintances) knowing they used online dating, these interviewees were fine with close friends and family knowing that they used an ODS.

However, the participants still had some concerns. 75% of interviewees (15/20) did not want their mini-feeds to reflect their use of dating apps. Further, 25% of users expressed a belief that using an online dating app on Facebook would mix social circles in an undesirable fashion. While some participants did not wish to mix social circles and date within their social network (4/20), other users simultaneously complained that online dating "didn't work' (3/20) or that "chemistry is more than a profile" (7/20). Nonetheless, the users did not mind if their close friends or acquaintances saw them on an ODS. One interviewee compared being seen on an ODS to being seen at a gay bar: *"They can't really judge me, cause, hey, they're here too!"* Thus, we conclude that users were not particularly concerned with excluding coworkers and/or friends.

One problem that users did express was related to the veracity of ODS profiles. Users overwhelmingly felt that they could not trust that the data they found in ODSs was accurate. Older users were concerned that that profiles misrepresented age and/or relationship status, while younger participants were more concerned that the photos posted were either altered, out of date, or taken from an especially flattering angle (often referred to as the "MySpace Angle".)

Since we had initially considered creating a privacy-preserving dating application for Facebook, our pre-interview survey asked users if there was any information the interviewee refrained from posting to Facebook, then followed up with a question asking "What security measures would Facebook have to take for you to be willing to share this information?" All participants indicated that they had such information and about half



? Age: 29

Relationship ? Status: Single

Friends: ? ? ?

About me:

My name is Bob, I enjoy going to concerts, playing baseball, and going for walks with my labrador Spot.

Favorite Bands?

The Beatles, Iron Maiden, Journey,

Favorite Movies

Star Wars, The Dark Knight, The Notebook

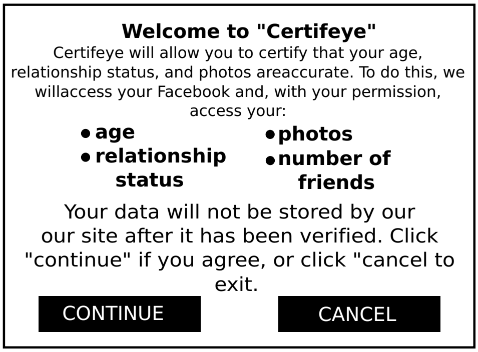
Fig. 1. An example profile shown to users in our study

(11/20) responded that they "simply refuse to post this information" and that no technical measure could change their minds. In the interviews, participants also revealed that certain information, such as address and/or phone number, was simply too private to be entrusted to Facebook. Overall, our analysis clearly showed that concerns about information disclosure were actually less relevant to users than the issue of veracity of information contained in profiles. Interestingly, all participants agreed that a Facebook profile contained enough information to make a decision about whether to date someone. This suggested that Facebook would be an ideal source of information to bootstrap trust in ODS.

Note: Our analysis of the interviews focused on discovering common themes and complaints to drive the design of our application (which was tested in a rigorous manner, as noted later). Thus, any observations above do not necessarily extend to the general population.

4 Study Part 2: The Certifeye Interface

Based on our interviews, we found that the direction of certifying profile data direction would be the most valuable to ODS users. We called our system concept "Certifeye" and designed it as a Facebook and ODS application allowing users to certify that their ODS profile information was accurate. Aiming to address the problem of doubtful profile veracity, Certifeye users could pull their relationship status, age, and photos from their Facebook account, and receive green badges to show that this information had been certified.



Welcome to "Certifeye"

Certifeye will allow you to certify that your age, relationship status, and photos areaccurate. To do this, we willaccess your Facebook and, with your permission, access your:

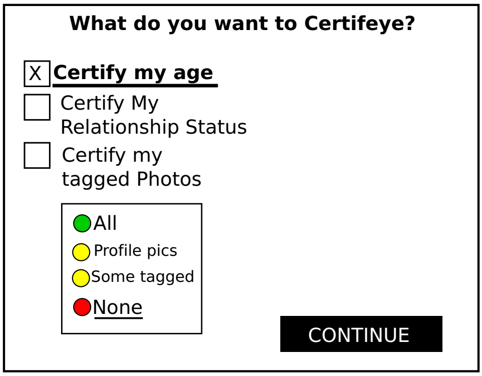
- **age**
- **relationship status**
- **photos**
- **number of friends**

Your data will not be stored by our our site after it has been verified. Click "continue" if you agree, or click "cancel to exit.

Fig. 2. The Certifeye consent screen

We developed a visual mock-up of Certifeye for use in end-user evaluations. The interface consists of three main screens. First, users are greeted with a generic ODS profile (Figure 1) containing various personal information items such as profile photo, age and relationship status, each with a certification badge displayed next to it. At the outset, all certification badges are greyed out and display question marks.Upon clicking on a certification badge, Certifeye displays a consent screen (Figure 2). It should be noted that the Certifeye UI also displays the numerical total of a user's Facebook friends. As discussed in Section 5, this is an important feature, since fake profiles will be harder to create if they must have large numbers of friends. Creating fake profiles requires technical expertise, time, and cost – barriers that may push dishonest users to other, less trustworthy ODS services.

Finally, after consenting to the syncing of her Facebook account with her ODS profile, the user is presented with the main Certifeye interface, where (s)he can choose to certify relationship status, age, and/or photos (Figure 3). Observe that we embed trust in our design by envisioning periodic user interaction to "renew" the certification. After a user has gone through the Certifeye interface once, badges are not updated when her Facebook information changes. Therefore, after a set period of time, the badges revert



What do you want to Certifeye?

☒ **Certify my age**

□ Certify My Relationship Status

□ Certify my tagged Photos

- 🟢 All
- 🟡 Profile pics
- 🟡 Some tagged
- 🔴 None

Fig. 3. The main Certifeye interface

Table 3. Age breakdown for Turkers

Age	N	%
18–22	19	11.8 %
23–30	65	40.4%
31–50	67	41.6%
51–70	20	6.2%

Table 4. Education breakdown for Turkers

Education Level	N	%
9th – 12th, no diploma	2	1.2%
HS (includes GED)	27	16.8%
Some college (no degree)	43	26.7%
Associates	25	15.5%
Bachelor's Degree	40	24.8%
Master's Degree	20	12.4%
PhD	4	2.5%

to grey question marks. The Facebook API would actually grants ongoing access to a user's profile information once a user initially consents, however, we prefer to respect users' privacy and let them explicitly consent to a new information access.

4.1 The Turk Experiment

After creating our prototype interface, we then showed this interface to 161 Mechanical Turk users. Users were asked if they were current ODS users, and users who were not ODS users were not allowed to proceed with the survey. Users who were not screened out were asked to rank their concern with the veracity of several types of ODS information with and without our trust enhancements. Participants were even more representative of the general population than those in our initial interviews. (Age and education breakdowns are reported in Table 3 and 4.) Also, remind that Mechanical Turk workers have been shown to be as reliable as traditional participant pools for human subject research [31].

Participants were asked for basic demographic information, then asked to rate their comfort with three scenarios on a scale from 1 to 7 (1 = Not at all concerned, 7 = Extremely concerned), in response to the following questions:

1. How concerned are you with people misrepresenting their relationship status on online dating sites?
2. How concerned are you about people misrepresenting their age on online dating sites?
3. How concerned are you about users on online dating sites misrepresenting themselves using old, altered, or engineered photos? (Examples: Photoshopped pictures, the "MySpace Angle")

Participants were also asked other questions (e.g., "How concerned are you about computer viruses?") to avoid biasing the participants towards trying to give "correct" answers. The questions were also presented in random order to eliminate any positioning effects. After participants had answered the questions, they were displayed a series of mocked-up screenshots from the Certifeye interface. Users were taken through as "Bob", a hypothetical user of Certifeye, exploring the functionality of the software via a series of annotated screenshots. Along the way, the functionality of Certifeye was revealed to Bob (and thus, to the participant). After being shown the interface, Turkers were asked to rank how concerned they were with misrepresentation *assuming an ODS used the interface they had just seen.* As Kittur et al. [24] pointed out, Mechanical Turkers often try to cheat at tasks. To guard against this, a series of short "sanity-check" questions ensured that users had paid attention to the interface. For example, a screen might state that Bob clicked on "None", and all the faces in the interface turned red. We would then ask the user "What color did the faces in the interface turn?" Users who did not answer all sanity checks correctly were not included in the analysis. Out of an initial pool of 200 users, 39 users failed their sanity check, leaving 161 valid responses for analysis.

4.2 Turk Study Results

Since levels of comfort are not assumed to be normally distributed, we used a Mann-Whitney U test to check that the changes in average Likert scores presented in Table 5 were statistically significant. We asked participants to rank on a scale from 1 to 7 how concerned they were that the ages, relationship statuses, and photos on an ODS which used Certifeye were accurate (1 = Not at all concerned, 7 = Extremely concerned). We found that users felt more comfortable with the Certifeye interface, and that the difference was statistically significant. Specifically, for, age, relationship status, and photos, respectively, Mann-Whitney U values are equal to 7978, 8326, and 7693.5. $n1 = n2 = 161$ and $P < 0.01$ (2-tailed). As 1 represented not concerned at all, and 7 extremely concerned, this means that new level of concern is positively below neutral (4/7), while the previous was not, thus, we conclude that the features provided by Certifeye—verification of age, photos, and relationship status—reduce users' concerns with respect to information's veracity.

Table 5. Results of Mechanical Turk Likert questions

Type of Info	Concern Pre	Concern Post	P-Value
Photos	4.7	3.5	$< .001$
Rel. Status	4.6	3.4	$< .001$
Age	4.1	3.1	$< .001$

In our Mechanical Turk study, before being shown our interface, users were asked to rank some criteria as for how important they are when deciding if a stranger's Facebook profile is genuine. (1 = Most important, 8 = Least important.) The order of options was randomized for each participant to avoid biasing respondents to any particular item. Below, we report suggested criteria, ordered from most important to least important (according to participants in our study):

1. Number of friends
2. Location (city)
3. Workplace
4. College / Grad School
5. Mutual interests
6. Attractiveness / appearance
7. Other

As per "other", common responses included having mutual friends, and whether the profile appeared "spammy". As a result, we conclude that "Number of friends" was the most preferred criteria to assess whether or not a social network profile was genuine.

5 Discussion

As mentioned earlier, our exploration of privacy and trust in ODSs has been user-driven. We interviewed users based on a few preliminary concepts, such as, (1) Identifying potential romantic partners within one's social circle without trusted third parties, (2) Listing sensitive information without a trusted third party, (3) Automatic exclusion of coworkers/friends, and (4) Certification of profile data.

While we only pursed the certification of profile data (motivated by a stronger interest of users), we now report on some lessons learned during our interviews. As mentioned in Section 3, we asked users whether they would like an application identifying potential romantic partners, both (within and outside one's social circle), without the need to disclose "private", possibly sensitive, information to any third party OSN or ODS sites. While such application would naturally enhance users' privacy, we found that users did not see the point of using a computer system to meet people they knew in real life. Thus, while interviewees were comfortable with linking Facebook profiles to their ODS accounts, they did not see the point in doing so. For instance, one user reported *"If I want to ask someone out, I'll ask them out – I don't need a computer to do that"*. Even when prompted as to whether being connected with "friends of friends" would be a useful function, interviewees were dubious as their friends could just introduce them to potential matches. In general, interviewees seemed reasonably comfortable with online dating but agreed that they did not want their online dating information sent to the general public. Some participants specifically mentioned that they would not use a dating application that posted information to their Facebook mini-feed.

As mentioned in the interview results, users did not mind if their friends or acquaintances knew they used an ODS. Also, even if users did find the idea of automatic exclusion of coworkers/friends useful, terms of service of most social network sites, e.g., Facebook, do not actually allow APIs to access social graph data of users who have not opted in (therefore, this approach would be infeasible). We also found that, while misrepresentation is a widespread concern, the specific kind of information actually concerning users varied with age. Older users were concerned about misrepresentation of age and/or relationship status. Relatively older women were especially wary that older males may misrepresent their marital status, with many citing bad experiences which led to such distrust. However, regardless of gender, older users wanted to verify that age and relationship status were accurate, whereas, younger users were more concerned about accurate, recent photos being present on the ODS.

As a result, trust being a major concern, we decided to explore the user-centered and usable design of mechanisms to bootstrap trust in ODS, by leveraging social-network based reputation. This approach presented several interesting challenges. First, a user could set up a fake social network profile, link it to her ODS profile, and "Certifeye" it despite the fact that her information is actually false. As mentioned in Section 4.2, study participants listed "number of friends" as the most preferred criteria to assess whether a stranger's Facebook profile was genuine or fake. So, we designed our software to always include the number of friends along with the certification data. Specifically, whenever a user certifies her profile, in addition to green or red badges, the profile is also populated with a small box listing how many friends she has. Also, note that Facebook actively takes steps to detect and remove accounts that it deems to be secondary or fake. This makes it hard for ordinary users to maintain fake profiles long enough to gain a substantial number of friends (Facebook looks for suspicious patterns and blocks suspect accounts but, unsurprisingly, will not reveal how it does this [19].)

Although users may friend unfamiliar people [35], Pew [20] has shown that most Facebook users have an average of 229 friends. Spammers (i.e., malicious entities with an economic incentive to abuse the system) tend to have more friends than average [32], however, this requires a non-trivial time and economic commitments that are unrealistic for ordinary people only trying to make their dating profile more appealing. Also, creating fake profiles also increases the risk of the profile being blocked by Facebook. Thus, these barriers will arguably push dishonest users to other less trustworthy ODSs. Previous work – for instance, in the context of spam [25] – has showed that raising costs for malicious users causes them to move to easier-to-abuse services.

Nonetheless, if we were to develop an actual service based on the Certifeye design, we would further enhance trust by also taking into account additional information, such as date of most recent status update or other automated fake account detection mechanisms, such as, the one proposed by Sirivianos et al. [8]. Also, we would need provide the users with information about how to interpret what data they see in another profile.

Certifeye also allowed the certification of ODS pictures. Users could earn a yellow badge by sharing all of their profile pictures, and a green badge by sharing all of their tagged photos. While, theoretically, users could untag themselves from unflattering photos before syncing their Facebook profile with our certification application, we feel this is not a limitation, as already pointed out in previous work. Facebook profiles serve as a form of self-presentation [4], thus, if Facebook users wish to share one facet of themselves with an ODS, and another facet of themselves with their Facebook friends, they would be forced to untag their unflattering photos, certify their profile, then retag themselves. Once again, raising the effort required to create fake profile data will likely prompt "malicious" users to move on to other, less well-protected ODSs.

Finally, while we describe our social verification based on Facebook profiles, note that our techniques are not limited to one specific OSN. We use Facebook as an example, motivated by its widespread penetration (900 million users in 2012 [18].) Nonetheless, our techniques could work with any ODS and with any OSN providing APIs to share age, relationship status, and pictures. We could easily modify our workflow so that Certifeye could interface with other services, such as Google Plus, Orkut, Diaspora, etc.

6 Conclusion

This paper analyzed concerns of Online Dating Services (ODSs) users about the veracity of information presented in the profiles of potential dates. Motivated by the results of semi-structured interviews (involving 20 users), we designed an interface, called Certifeye, that lets users certify some attributes, such as age, relationship status, and photos, in their ODS profile. Our prototype does so by attesting that the information reported on the ODS profile corresponds to that on the user's own Facebook profile. We ran a Mechanical Turk study with 161 users to assess whether or not a mock-up of this veracity-enhancing capability successfully reduced trust concerns and, indeed, we found a statistically significant reduction when users were presented with Certifeye.

Naturally, our work does not end here. We plan to develop our Certifeye interface and integrate it in an actual ODS. Also, we intend to further explore privacy and trust concerns in the context of both ODSs and OSNs, and deploy usable privacy-enhancing technologies, following similar user-driven approaches.

References

1. Acquisti, A., Gross, R.: Imagined communities: Awareness, information sharing, and privacy on the Facebook. In: Danezis, G., Golle, P. (eds.) PET 2006. LNCS, vol. 4258, pp. 36–58. Springer, Heidelberg (2006)
2. Acquisti, A., Gross, R.: Predicting Social Security numbers from public data. Proceedings of the National Academy of Sciences 106(27), 10975–10980 (2009)
3. Badoo. Social Lives Vs Social Networks (2012), http://corp.badoo.com/lv/entry/press/53/
4. Boyd, D., Ellison, N.B.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication 13(1), 210–230 (2007)
5. Brym, R., Lenton, R.: Love online: A report on digital dating in Canada. MSN. ca (February 6, 2001)
6. Byrne, C.: Users do care about location privacy (2011), http://www.venturebeat.com/2011/04/21/smartphone-users-location-privacy/
7. Camp, L.J.: Designing for trust. In: Falcone, R., Barber, S.K., Korba, L., Singh, M.P. (eds.) AAMAS 2002. LNCS (LNAI), vol. 2631, pp. 15–29. Springer, Heidelberg (2003)
8. Cao, Q., Sirivianos, M., Yang, X., Pregueiro, T.: Aiding the Detection of Fake Accounts in Large-scale Social Online Services. In: USENIX Symposium on Networked Systems Design and Implementation (2012)
9. Cavoukian, A.: S. U. S. of Law, C. S. C. for Computers, and Law. Privacy by design. Stanford Law School (2010)
10. Couch, D., Liamputtong, P.P.: Online dating and mating: Perceptions of risk and health among online users. Health, Risk & Society 9(3), 275–294 (2007)
11. Current Online Dating and Dating Services Facts & Statistics (2012), http://preview.tinyurl.com/254pum6
12. De Cristofaro, E., Tsudik, G.: Practical private set intersection protocols with linear complexity. In: Sion, R. (ed.) FC 2010. LNCS, vol. 6052, pp. 143–159. Springer, Heidelberg (2010)
13. Dourish, P., Grinter, R., Delgado de la Flor, J., Joseph, M.: Security in the wild: user strategies for managing security as an everyday, practical problem. Personal and Ubiquitous Computing 8(6), 391–401 (2004)
14. Ellison, N., Heino, R., Gibbs, J.: Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. Journal of Computer-Mediated Communication 11(2), 415–441 (2006)
15. Fahl, S., Harbach, M., Muders, T., Smith, M., Sander, U.: Helping Johnny 2.0 to encrypt his Facebook conversations. In: Symposium on Usable Privacy and Security, p. 11 (2012)
16. Felt, A., Evans, D.: Privacy protection for social networking APIs. In: Web 2.0 Security and Privacy (2008)
17. Gibbs, J., Ellison, N., Lai, C.: First comes love, then comes Google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. Communication Research 38(1), 70–100 (2011)
18. Goldman, D.: Facebook tops 900 million users (2012), http://money.cnn.com/2012/04/23/technology/facebook-q1/index.htm

19. Goo, S.K.: Explaining Facebook's Spam Prevention Systems (2010), http://preview.tinyurl.com/23gmv7r
20. Goo, S.K.: Facebook: A Profile of its 'Friends' (2012), http://preview.tinyurl.com/7hfpgjj
21. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: ACM Workshop on Privacy in the Electronic Society, pp. 71–80 (2005)
22. Huang, D.-L., Rau, P.-L.P., Salvendy, G.: A survey of factors influencing people's perception of information security. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4553, pp. 906–915. Springer, Heidelberg (2007)
23. Jagatic, T., Johnson, N., Jakobsson, M., Menczer, F.: Social phishing. Communications of the ACM 50(10), 94–100 (2007)
24. Kittur, A., Chi, E., Suh, B.: Crowdsourcing user studies with Mechanical Turk. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 453–456 (2008)
25. Liu, D., Camp, L.: Proof of work can work. In: Workshop on the Economics of Information Security (2006)
26. Motahari, S., Ziavras, S., Schuler, R., Jones, Q.: Identity inference as a privacy risk in computer-mediated communication. In: International Conference on System Sciences, pp. 1–10 (2009)
27. Murayama, Y., Hikage, N., Hauser, C., Chakraborty, B., Segawa, N.: An Anshin Model for the Evaluation of the Sense of Security. In: Hawaii International Conference on System Sciences (2006)
28. Nissenbaum, H.: Protecting privacy in an information age: The problem of privacy in public. Law and Philosophy 17(5), 559–596 (1998)
29. OAuth 2.0, http://oauth.net/
30. OpenID, http://openid.net/
31. Paolacci, G., Chandler, J., Ipeirotis, P.: Running Experiments On Amazon Mechanical Turk. Judgment and Decision Making 5(5), 411–419 (2010)
32. Protalinski, E.: How to spot a fake Facebook profile (2012), http://preview.tinyurl.com/87hmqyr
33. Richman, W., Kiesler, S., Weisband, S., Drasgow, F.: A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. Journal of Applied Psychology 84(5), 754–775 (1999)
34. Rosenfeld, M., Thomas, R.: Searching for a Mate The Rise of the Internet as a Social Intermediary. American Sociological Review 77(4), 523–547 (2012)
35. Ryan, T., Mauch, G.: Getting in bed with Robin Sage. In: Black Hat Conference (2010)
36. Sasse, M., Brostoff, S., Weirich, D.: Transforming the weakest link – Human/Computer Interaction Approach to Usable and Effective Security. BT Technology Journal 19(3), 122–131 (2001)
37. Schneier, B.: Beyond Fear: Thinking Sensibly about Security in an Uncertain World. Springer (2003)
38. Sirivianos, M., Kim, K., Gan, J.W., Yang, X.: Assessing the veracity of identity assertions via osns. In: 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS), pp. 1–10. IEEE (2012)
39. Toma, C., Hancock, J., Ellison, N.: Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. Personality and Social Psychology Bulletin 34(8), 1023–1036 (2008)
40. Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P., Cranor, L.: I regretted the minute I pressed share: A qualitative study of regrets on Facebook. In: Symposium on Usable Privacy and Security (2011)
41. Whitten, A., Tygar, J.: Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In: USENIX Security Symposium (1999)

SHADE: Secure HAmming DistancE Computation from Oblivious Transfer[*]

Julien Bringer[1], Hervé Chabanne[1,2], and Alain Patey[1,2]

[1] Morpho
[2] Télécom ParisTech
Identity and Security Alliance (The Morpho and Télécom ParisTech Research Center)

Abstract. We introduce two new schemes for securely computing Hamming distance in the two-party setting. Our first scheme is a very efficient protocol, based solely on 1-out-of-2 Oblivious Transfer, that achieves full security in the semi-honest setting and one-sided security in the malicious setting. Moreover we show that this protocol is significantly more efficient than the previous proposals, that are either based on garbled circuits or on homomorphic encryption. Our second scheme achieves full security against malicious adversaries and is based on Committed Oblivious Transfer. These protocols have direct applications to secure biometric identification.

Keywords: Secure Multi-Party Computation, Hamming Distance, Oblivious Transfer, Biometric Identification.

1 Introduction

Secure Multiparty Computation (SMC) [35,13] enables a set of parties to jointly compute a function of their inputs while keeping the inputs private. We here focus on the 2-party case [14], also known as Secure Function Evaluation. Several generic constructions exist in this setting, which apply SMC to any function computed by two parties. In the semi-honest setting, where security is ensured against adversaries following the protocol but trying to gain more information than they should, the Yao's protocol [35,24] can be used to achieve this purpose using Oblivious Transfers and Garbled Circuits. In the malicious model, where adversaries can follow any strategy, many generic constructions have been proposed [19,23,17,28,18,25]. The problem of generic constructions is that they are often far from being optimal when one wants to securely compute specific functions of interest. However, it may happen that generic constructions can be more efficient than specific ones [15].

We here consider the secure computation of the Hamming distance. Concretely, two parties P_1 and P_2 hold bit strings of the same length n, resp. $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ and want to jointly compute $d_H(X, Y) =$

[*] This work has been partially funded by the ANR SecuLar project and the European FP7 FIDELITY project (SEC-2011-284862).

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 164–176, 2013.

$\Sigma_{i=1}^{n}(x_i \oplus y_i)$, without P_1 (resp. P_2) revealing X (resp. Y) to P_2 (resp. P_1). For now, let us consider this problem in the semi-honest setting. It has first been solved using additive homomorphic encryption [20,29]. Using this technique, each bit of P_1's input has to be encrypted in one Paillier ciphertext [30] and sent to the other part who can then compute a ciphertext corresponding to the Hamming distance, using homomorphic encryptions. Since Paillier ciphertexts must be at least 2048 bit-long and homomorphic encryptions are multiplications and exponentiations in large groups, this technique is inefficient. However, they also propose in [20] an adaptation of their protocol to the malicious setting. Recently, Huang *et al.* [15] showed that the generic Yao algorithm applied to Hamming distance was more efficient in terms of computation time and bandwidth consumption. Using the Yao algorithm, one needs to describe the function as a binary circuit and then "garble" every gate of this circuit to a table of 4 symmetric ciphertexts. However, using the techniques of [22] and [32], XOR gates do not need to be garbled and garbled gates can be reduced to 3 items. The circuit used in [15] is the succession of n bit-wise (free) XOR's and a *Counter* circuit that adds the results of these XOR's. This Counter circuit is the bottleneck of their protocol.

The first proposal of our paper achieves full security in the semi-honest model. We almost only rely on 1-out-of-2 oblivious transfer (OT_1^2). This primitive enables a receiver to obtain 1 out of 2 elements held by a sender without the sender learning the choice of the receiver and without the receiver learning information on the other element held by the sender. In the Yao algorithm, using OT_1^2's, party P_2 gets his input keys for a garbled circuit of the function to compute. However, the keys sent by P_1 are independent of P_1's inputs. Here we design our scheme such that, in our OT's, the elements sent by P_1 also depend on the input bits of P_1 in such a way that the element obtained by P_2 during the i^{th} OT_1^2 depends on $x_i \oplus y_i$. Moreover using the technique of [26, Third Variant], we avoid the use of a costly Counter circuit. We prove, using the OT-hybrid model [6,23,14], that our protocol is fully secure in the semi-honest setting or one-sided secure in the malicious setting, depending on the security level of the underlying OT_1^2. This protocol is significantly more efficient than the previous proposals for secure Hamming distance in the semi-honest model [20,29,15,2].

We next extend our first proposal to a second protocol that is fully secure in the malicious setting. Therefore, we use Committed Oblivious Transfer (COT) [8] instead of basic OT_1^2. In particular, we use a COT on bit strings with homomorphic commitments, as in [21]. COT enforces that the parties are committed to their inputs to the oblivious transfers and moreover that the receiver is committed to his output. The homomorphic commitment scheme enables us to guarantee that the inputs of the sender are consistent and that the computation run by the receiver on these inputs after the OT's follows the protocol.

The proofs of security of our protocol secure in the malicious setting and extensions to secure computation of weighted Hamming distance, of biometric identification and of any linear combination of bit-wise independent functions, appear in the extended version of this paper [3].

2 SMC and Oblivious Transfer

2.1 Oblivious Transfer

Oblivious Transfer was first introduced by Rabin [33] as a two-party protocol where a sender has a secret message that he sends to a receiver, which receives it with probability 1/2, without the sender knowing if the message has been received or not. This is however not the version that is now used in secure protocols, but a slightly different primitive called 1-out-of-2 Oblivious Transfer (OT_1^2). We here describe this primitive, some extensions to improve its use and a derived version called Committed Oblivious Transfer (COT) [21], used in our second proposal.

1-out-of-2 Oblivious Transfer. A 1-out-of-2 Oblivious Transfer is a cryptographic primitive that enables a receiver R to obtain 1 out of 2 elements held by a sender, without learning information on the other element and without the sender knowing which element has been chosen. This kind of protocol is stronger than a Private Information Retrieval (PIR) protocol [7] where only the choice of the receiver remains hidden from the sender. The functionality enabled by a OT_1^2 is described in Figure 1. For more details on implementations, see for instance [14, Chapter 7]. For instance, the oblivious transfers of [27] and of [31] can be used, respectively, in the semi-honest and in the malicious setting (see Section 2.2 for the security definitions).

– **Inputs:** • Sender S inputs two n-bit strings X_0 and X_1
  • Receiver R inputs a choice bit b
– **Output:** • S learns nothing on b
  • R obtains X_b but learns nothing on X_{1-b}

Fig. 1. The OT_1^2 functionality

Extensions. Several kinds of optimizations can be applied to Oblivious Transfers, independently of the implementation. Two optimizations introduced in [16] are of interest for our proposals. The first one [16, Section 3] enables, in the random oracle model, to compute many OT's with a small elementary cost from k OT's at a normal cost, where k is a security parameter. The second one [16, Appendix B] enables to reduce oblivious transfers of long strings to oblivious transfers of short strings using a pseudo-random generator.

Committed Oblivious Transfer. Committed Oblivious Transfer (COT) is a combination of OT_1^2 and bit commitment, first introduced by Crépeau [8] under the name Verifiable Oblivious Transfer. In this variant, both sender and receiver are committed to their inputs before the oblivious transfer. Moreover, the sender receives a commitment to the receiver's output, and the receiver obtains the randomness for this commitment. To our knowledge, the only scheme

- **Inputs:** • S inputs two n-bit strings X_0 and X_1 and two random values r_0 and r_1 used for commitment.
  • R inputs a choice bit b and a random r used for commitment
  • The common inputs are $Com(b, r)$, $Com(X_0, r_0)$ and $Com(X_1, r_1)$
- **Output:** • S learns nothing on b and r
  • R obtains X_b and a random u but learns nothing on X_{1-b}, r_0 and r_1.
  • Both parties obtain $Com(X_b, u)$.

Fig. 2. The *COT* functionality

that considers COT of bit strings is the one of Kiraz *et al.* [21], which uses an homomorphic cryptosystem as commitment scheme. COT is described in Figure 2, where *Com* denotes a commitment scheme.

2.2 Secure Two-Party Computation

Overview. Secure Multi-Party Computation [35] enables a set of parties to jointly compute a function of their inputs while keeping their inputs private. Different kinds of adversaries are considered:
• *semi-honest* adversaries who follow the protocols and try to gain more information than they should on the other parties' inputs,
• *malicious* adversaries who use any kind of strategy to learn information.

There also exists a notion of *covert* adversaries [1] who are malicious but averse to being caught. Notice that we only consider static adversaries.

Security Definitions. Informally, security in SMC is ensured by simulating the secure protocol in an ideal model where the inputs of both parties are sent to a trusted party who takes care of the computation and sends the outputs back to the respective parties and showing that all adversarial behaviours in a real execution are simulatable in this ideal model. Full definitions and explanations can be found in [13,14].

We quickly recall how full security is proven in the malicious setting. Let π be a protocol for computing $f(x, y) = (f_1(x, y), f_2(x, y))$. In the real world, a probabilistic polynomial-time (PPT) adversary A sends messages on behalf of the corrupted party and follows an arbitrary strategy while the honest party follows the instructions of π. In the ideal world, the honest party sends his genuine input x to a trusted party. The adversary sends any input y', of the appropriate size to the trusted party. The trusted party first sends his output $f_1(x, y')$ to the adversary and, if the adversary does not abort, also sends his output $f_2(x, y')$ to the honest party. The adversary is also allowed to abort the protocol at any time. Full Security against a malicious party P_i is ensured if, for any PPT adversary in the real world, there is a PPT adversary in the ideal world such that the distribution of the outputs in the real world is indistinguishable from the distribution of the outputs in the ideal world.

A weaker notion is *Privacy* against a malicious party P_i, for $i = 1, 2$, that guarantees that P_i cannot learn any information on the other party's input.

However, the execution in the real model might not be simulatable in the ideal model. We say that a protocol achieves *One-Sided Security* in the malicious model if it is fully-secure against a malicious P_i and private against a malicious P_{3-i}. See [14, Section 2.6] for further details.

In this paper, we prove security of our schemes in the *OT-hybrid setting* [6,23,14]. In this setting, the execution in the real model is slightly modified. The parties have access to a trusted party that computes oblivious transfers for them. We only need to prove indistinguishability between executions in this hybrid model and the ideal model to ensure security.

3 Secure Hamming Distance Computation

In the following, the $+$ and $-$ operators respectively denote modular additions and subtractions, we assume that the context is explicit enough and do not recall the moduli in the description of the algorithms. $\bar{x}$, where x is a bit value, denotes $1 - x$. The Hamming distance is denoted by d_H.

3.1 The Basic Scheme

We here introduce our new scheme based on oblivious transfers. The Yao algorithm [35] also uses oblivious transfers but the inputs of the sender are random keys that are independent of the actual inputs of the sender for the secure computation. In the protocol we propose, the inputs of the sender P_1 to the OT's depend on P_1's input bits. Consequently, the output of each oblivious transfer depends on the input bits x_i of P_1 and y_i of P_2. We adjust our scheme so that this output depends on $x_i \oplus y_i$. Then, we use a technique inspired by [26, Third

- **Inputs:**
  - P_1 inputs a n-bit string $X = (x_1, \ldots, x_n)$
  - P_2 inputs a n-bit string $Y = (y_1, \ldots, y_n)$
- **Output:**
  - 1^{st} Option: P_1 obtains $d_H(X, Y)$ and P_2 obtains nothing
  - 2^{nd} Option: P_2 obtains $d_H(X, Y)$ and P_1 obtains nothing
- **Protocol:**
  1. P_1 generates n random values $r_1, \ldots, r_n \in_R \mathbb{Z}_{n+1}$ and computes $R = \Sigma_{i=1}^{n} r_i$
  2. For each $i = 1, \ldots, n$, P_1 and P_2 engage in a OT_1^2 where
     - P_1 acts as the sender and P_2 as the receiver.
     - P_2's selection bit is y_i.
     - P_1's input is $(r_i + x_i, r_i + \bar{x}_i)$.
     - The output obtained by P_2 is consequently $t_i = r_i + (x_i \oplus y_i)$.
  3. P_2 computes $T = \Sigma_{i=1}^{n} t_i$
  4. 1^{st} Option: (1) P_2 sends T to P_1 (2)P_1 computes and outputs $T - R$
  2^{nd} Option: (1) P_1 sends R to P_2 (2) P_2 computes and outputs $T - R$

Fig. 3. The Basic Scheme

Variant] to count the number of bits such that $x_i \oplus y_i = 1$, *i.e.* to compute the Hamming distance.

We assume that parties P_1 and P_2 respectively hold inputs $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$. Party P_1 prepares n random values $r_1, \ldots, r_n \in_R \mathbb{Z}_{n+1}$ and prepares n oblivious transfers, as a sender. The inputs of the i^{th} transfer are arranged in such a way that a receiver with bit input y gets $r_i + (y \oplus x_i) \bmod n+1$. To do so, input 0 of P_1 is set to $r_i + x_i$ and input 1 to $r_i + \bar{x_i}$. Indeed, if $y = 0$, $x_i \oplus y = x_i$ and if $y = 1$, $x_i \oplus y = \bar{x_i}$. P_2 acts as a receiver for all these n OT's, with bit inputs $y_1, \ldots, y_n$ and gets $(r_i + (x_i \oplus y_i))_{i=1,\ldots,n}$. Then, P_2 adds all these values and gets $T = \Sigma_{i=1}^{n} r_i + \Sigma_{i=1}^{n}(x_i \oplus y_i) = R + d_H(X,Y)$, where $R = \Sigma_{i=1}^{n} r_i$. Finally, depending on the party that is supposed to know the output, either P_1 sends R to P_2 or P_2 sends T to P_1, the final output being $D = T - R = d_H(X,Y)$. The protocol is described in Figure 3.

Theorem 1 (Security of the Basic Scheme)
Assuming that the underlying OT_1^2 is secure in the semi-honest setting, the Basic Scheme achieves full security in the semi-honest setting.

Assuming that the underlying OT_1^2 is secure in the malicious setting, the Basic Scheme achieves, in the malicious setting:

- *one-sided security, for the 2^{nd} option: privacy against a malicious P_1 and full security against a malicious P_2,*
- *privacy against a malicious P_2, for the 1^{st} option.*

The proofs are detailed in Section 4.1.

3.2 The Fully Secure Scheme

Requirements on the Commitment Scheme. We assume that the commitment scheme used in the Committed Oblivious Transfer we use in our scheme fulfills the following requirements.

First, it must be additively homomorphic, *i.e.* there exist efficient operations $\boxplus$ and $\odot$, such that $Com(x_1, r_1) \odot Com(x_2, r_2) = Com(x_1 + x_2, r_1 \boxplus r_2)$, for any x_1, x_2, r_1, r_2.

Second, there must exist a zero-knowledge proof of knowledge π_1^2, where both parties know a commitment $C = Com(x, r)$ and two values x_1 and x_2. In this proof, the prover knows x, r and proves that x is either x_1 or x_2. Using the notations of Camenisch and Stadler [4], $\pi_1^2 = PK\{(\alpha, \beta) : C = Com(\alpha, \beta) \wedge (\alpha = x_1 \vee \alpha = x_2)\}$.

Let us consider the commitment scheme used in [21]. This commitment consists of using a (2,2)-threshold homomorphic cryptosystem, *i.e.* $Com(x, r) = Enc(x, r)$ for a homomorphic cryptosystem where the public key is known by both parties and the secret key is shared between the parties. By definition, the first condition is fulfilled (usually $\odot$ is a product and $\boxplus$ an addition). The used cryptosystem can be an additive ElGamal [11] or a Paillier [30] encryption. In both cases, the second condition can be fulfilled (see resp. [5] and [10]). This confirms that our requirements are reasonable.

More details on the COT scheme of [21] and details on the π_1^2 proofs can be found in the extended version of this paper[3].

Our Proposal. Our second scheme adapts the Basic Scheme to the malicious setting. We use a COT with a commitment scheme fulfilling the requirements previously introduced. The commitment, together with the proofs of knowledge of the inputs helps to ensure that the inputs are consistent and that the same values are used along the protocol.

First, P_1 and P_2 commit to the oblivious transfer inputs and prove that these inputs are well-formed. P_2 proves that his inputs are bits and P_1 proves that his inputs differ by 1, *i.e.* for each input pair (a_i, b_i), there exists r_i such that $(a_i, b_i) = (r_i, r_i + 1)$ or $(a_i, b_i) = (r_i + 1, r_i)$. COT's are then run with the same inputs as in the basic scheme. Party P_2 receives committed outputs, performs the addition of these outputs and a commitment to this addition, thanks to the homomorphic properties of the commitment scheme. P_2 can prove, using the commitments, that the value T obtained by adding the results of the COT's is consistent. In the same way, party P_1 can prove that the value R is consistent with his inputs to the COT's. Indeed, $\Sigma_{i=1}^n a_i + b_i = \Sigma_{i=1}^n (r_i + r_i + 1) = 2\Sigma_{i=1}^n r_i + \Sigma_{i=1}^n 1 = 2R + n$. Using the commitments to the a_i's and to the b_i's, P_2 is then able

- **Inputs:** P_1 inputs $X = (x_1, \ldots, x_n)$; P_2 inputs $Y = (y_1, \ldots, y_n)$.
- **Output:** 1^{st} (resp. 2^{nd}) Option: P_1 (resp. P_2) obtains $d_H(X, Y)$ and P_2 (resp. P_1) obtains nothing
- **Protocol:** 1. P_2 commits to all his bits y_i: he computes and publishes $Com(y_i, \chi_i)$ for each $i = 1 \ldots n$. He also proves, using π_1^2 proofs on the commitments, that $y_i = 0$ or $y_i = 1$.
  2. P_1 generates n random values $r_1, \ldots, r_n$, uniformly from the plaintext space of Com, and computes $R = \Sigma_{i=1}^n r_i$
  3. For each $i = 1, \ldots, n$, P_1 computes $(a_i, b_i) = (r_i + x_i, r_i + \bar{x}_i)$ and commits to a_i and b_i. He computes and publishes $(A_i = Com(a_i, \alpha_i))_{i=1,\ldots,n}$ and $(B_i = Com(b_i, \beta_i))_{i=1,\ldots,n}$
  4. P_1 proves to P_2, using π_1^2 proofs on the commitments, that $|b_i - a_i| = 1$, for each $i = 1, \ldots, n$.
  5. For each $i = 1, \ldots, n$, P_1 and P_2 engage in a COT where P_1 acts as the sender and P_2 as the receiver, P_2's selection bit is y_i, P_1's input is (a_i, b_i). The output obtained by P_2 is $t_i = r_i + (x_i \oplus y_i)$ and τ_i. Both parties obtain $C_i = Com(t_i, \tau_i)$
  6. P_2 computes $T = \Sigma_{i=1}^n t_i$,
  7. 1^{st} Option: (1) P_2 computes $C = Com(T, \tau) = C_1 \odot \ldots \odot C_n$ (2) P_2 sends T and a zero-knowledge proof that C commits to T to P_1 (3) P_1 computes $C = C_1 \odot \ldots \odot C_n$ and checks the proof. (4) P_1 computes and outputs $T - R$
  2^{nd} Option: (1) P_1 computes $K = Com(2R+n, \rho) = A_1 \odot \ldots \odot A_n \odot B_1 \odot \ldots \odot B_n$ (2) P_1 sends R and a zero-knowledge proof that K commits to $2R + n$ to P_2 (3) P_2 computes $K = A_1 \odot \ldots \odot A_n \odot B_1 \odot \ldots \odot B_n$ and checks that $K = Com(2R + n, \rho)$. (4) P_2 computes and outputs $T - R$.

Fig. 4. The Fully Secure Scheme

to check if the value R is consistent with the inputs of the COT's. The protocol is described in Figure 4. At any step, if a check fails, the party computing the check should halt the protocol and output $\perp$.

Theorem 2 (Security of the Fully Secure Scheme). *Assuming that the underlying COT is secure in the malicious setting, the Fully Secure Scheme achieves full security in the malicious setting.*

4 Security Proofs

4.1 The Basic Scheme

We here give the proof of security against a malicious P_2 in the case of the 2^{nd} option. The guarantees of privacy against a malicious P_2 for the 1^{st} option, or against a malicious P_1 for the 2^{nd} option are easily deduced from the privacy of the OT's, since no other messages are sent to these parties during the protocol.

Theorem 3 (Full Security against a Malicious P_2-2^{nd} option). *Assuming that the underlying OT_1^2 is secure in the malicious setting, the Basic Scheme, following the 2^{nd} option, is fully-secure against a malicious P_2 in the OT-hybrid setting.*

The following proof is partially inspired from the proofs of [26]. Indeed, our scheme can be viewed as a reduction of the third variant of their Oblivious Automata Evaluation, with only one state per line of the matrix, but where the lines of the matrix are not identical.

Proof. Let B be a PPT adversary controlling P_2 in the real world, we describe a simulator S_B who simulates the view of B in the ideal world.

S_B runs B on input Y. Since we operate in the OT-hybrid model, B sends $Y' = (y'_1, \ldots, y'_n)$ to the OT oracle. S_B sends Y' to the trusted party and obtains $D = d_H(X, Y')$. S_B picks n random values $t_1, \ldots, t_{n-1}, T \in_R \mathbb{Z}_{n+1}$ and computes $t_n = T + D - \Sigma_{i=1}^{n-1} t_i$. S_B sends the t_i's to B as results of the oblivious transfer. He then sends T. S_B then outputs whatever B outputs.

Let us now prove the indistinguishability between the real and the simulated views. Let V be a random subset of size t of $\{1, \ldots, n\}$. (V represents the bit positions where $x_i \oplus y_i = 1$.) Consider the distributions:

- (D_V): Choose n uniformly random values $\{r_1, \ldots, r_n\} \in \mathbb{Z}_{n+1}$. For every $i \in \{1, \ldots, n\}$, let $r'_i = r_i + 1$ if $i \in V$ and $r'_i = r_i$ otherwise. Output $(r'_1, \ldots, r'_n)$.
- (D'_V): Choose n uniformly random values $R, r'_1, \ldots, r'_{n-1} \in \mathbb{Z}_{n+1}$. Let $R' = R + t$ and $r'_n = R' - \Sigma_{i=1}^{n} r'_i$. Output $(r'_1, \ldots, r'_n)$.

It is easy to show that D_V and D'_V are identically distributed and that sampling from D'_V only requires the knowledge of t. The distribution D_V represents the view of B in a real execution of the protocol while our simulator S_B samples from D'_V, with the only knowledge of the final output. Thus, the view of P_2 in the real world and the simulated view of P_2 in the ideal world are indistinguishable, which ensures full security against a malicious P_2. □

Remark 1. The proofs of security in the semi-honest setting are straightforward, given the security guarantees of the Oblivious Transfer and the arguments explained in the previous proof proving that the outputs of the OT's give no information on the inputs of P_2.

4.2 The Fully Secure Scheme

We use an adaptation of the OT-hybrid model to Committed Oblivious Transfer. When the parties engage a COT in the COT-hybrid model, parties interact with each other and have access to a trusted party that computes the COT for them. Concretely, the receiver sends $b, Com(b,r)$ to the trusted party, the sender sends $x_0, Com(x_0, r_0)$ and $x_1, Com(x_1, r_1)$ to the trusted party. The trusted party sends x_b and r' back to the receiver and $Com(x_b, r')$ to both parties. This model, for a slightly different COT, has already been used in the proof of security of the binHDOT protocol [20] for malicious adversaries.

Since we use zero-knowledge proofs of knowledge, our protocol cannot be proved secure in the UC model [6] but in the stand-alone setting only. The proofs of Theorem 4 and Theorem 5 appear in the extended version of this paper [3].

Theorem 4 (Full Security Against a Malicious P_1). *Assuming that the underlying COT is secure in the malicious setting, the Fully Secure Scheme is fully-secure against a malicious P_1 in the COT-hybrid setting.*

Theorem 5 (Full Security Against a Malicious P_2). *Assuming that the underlying COT is secure in the malicious setting, the Fully Secure Scheme is fully-secure against a malicious P_2 in the COT-hybrid setting.*

5 Efficiency

5.1 The Basic Scheme

The cost of the basic scheme described in Figure 3 is essentially the cost of n OT_1^2's of inputs of $\log(n)$ bits. Using the OT extension of [16], when many OT's are performed, the workload turns out to be two evaluations of a hash function for P_1 and one for P_2 per input bit. The bandwidth requirement is then roughly $2n \cdot \log(n)$ bits.

Comparison to Previous Schemes. Let us compare our Basic Scheme to two previous protocols [15], [20,29] for semi-honest secure Hamming Distance computation, previously known as the most efficient proposals.

Other techniques, like Private Set Intersection Cardinality [9] or Private Scalar Product Computation [12] can be easily adapted to perform secure Hamming distance computation. However, in these proposals, use of homomorphic encryption and/or a linear number of exponentiations leads to schemes that are less efficient than our proposal in the semi-honest model.

We first compare to the application of the Yao algorithm to Hamming distance computation described in [15]. In this setting, the Hamming distance function has to be represented as a binary circuit. To get an idea of the cost of the computation, we need to count the number of non-XOR gates in this circuit. Let us assume that the size n of the inputs is a power of 2: $n = 2^N$. The number G of non-free gates is obtained (see the description of the Counter circuit in [15]) by $G = \Sigma_{i=1}^{N}(2^{N-i}.i) \approx 2^{N+1} = 2n$. Let k be the security parameter of the scheme. For the generation of the circuit, party P_1 has to perform $4G$ hash function evaluations. Then, P_1 sends the circuit ($3k \cdot G$ bits) and his keys for the circuit ($n \cdot k$ bits). Then P_1 and P_2 perform n OT_1^2's on k-bit strings. P_2 has then to perform G hash functions evaluations. Using the OT extension of [16], the workload of P_1 is roughly $10n$ hash functions evaluations, the workload of P_2 is $3n$ hash function evaluations and the bandwidth is $6kn$ bits. When m Hamming distances on the same input of P_2 are evaluated, all these operations but the oblivious transfers of P_2's inputs have to be computed m times.

We now evaluate the workload and bandwidth requirements of the [20,29] algorithm. The binHDOT protocol presented in [20] enables evaluation of a class of functions depending on Hamming distance. We here consider its reduction to the evaluation of the Hamming distance only. We describe the corresponding protocol in the extended version of this paper. We moreover take into account, in our evaluations, the optimizations presented in [29]. Party P_2 prepares n homomorphic ciphertexts, encrypting each of his inputs bits. These ciphertexts are sent to P_1 who homomorphically adds and subtracts them to obtain the encryption of the Hamming distance. Taking into account the optimizations of [29] (although we do not separate off-line and on-line phases), P_1 has to perform n homomorphic encryptions and P_2 n homomorphic additions. They mainly exchange n ciphertexts. When m distances are computed, with the optimizations of [29], P_2's work is almost the same and P_1 has to perform $mn/2$ homomorphic additions, once n subtractions and $3.5n$ additions are preprocessed. The bandwidth depends on the option and on the receiver of the result.

The comparison of these 3 protocols is summed up in Table 1, where **hash** means hash function evaluations and k is the security parameter of the Yao algorithm of [15]. We extrapolate to the simultaneous computation of m Hamming distances (see [3, Section 5.2]) in Table 2. In the first line of Table 2, the $(+m)$ hom. ciphertexts corresponds to the case where P_2 gets the result instead of P_1.

For concrete estimations, k should be at least 80 and Paillier ciphertexts at least 2048-bit long. It is easy to see that, for reasonable sizes of n, our scheme is more efficient and requires significantly less bandwidth. In these tables, we do not mention the k base OT's that are needed in our basic scheme and in the scheme of [15] for OT extension. They can be preprocessed.

Implementation Results. To prove our allegations regarding efficiency improvements in terms of computational workload, we ran the implementation of secure Hamming distance used in [15] and an implementation of our basic scheme using the same framework [34] on the same computer. The framework is

Table 1. Secure Computation of One Hamming Distance in the Semi-Honest Model

	P_1	P_2	Bandwidth (bits)
[20,29]	n hom.add.	n hom.enc.	n hom.ciphertexts
[15]	$10n$ **hash**	$3n$ **hash**	$6kn$
The Basic Scheme	$2n$ **hash**	n **hash**	$2n\log(n)$

Table 2. Secure Computation of m Hamming Distances in the Semi-Honest Model

	P_1	P_2	Bandwidth (bits)
[20,29]	$mn/2$ hom.add.	n hom.enc.	$n(+m)$ hom.ciphertexts
[15]	$(2+8m)n$ **hash**	$(1+2m)n$ **hash**	$(2+4m)kn$
The Basic Scheme	$2n$ **hash**	n **hash**	$2mn\log(n)$

implemented in Java and we ran it on a single computer with a 2 GHz Intel Core i7 processor and a 4 GB RAM. We think that the ratio of computation times between the protocols is more relevant than an absolute value of the time of execution of our process. This comparison is illustrated in Figure 5. For inputs with a few thousands bits size, the computation time required for our Basic scheme is approximately 22% of the time required to compute the protocol of [15].

5.2 The Fully Secure Scheme

We assume that the *COT* of the Fully Secure Scheme is the one of [21], using a threshold El-Gamal cryptosystem. According to [21], 24 exponentiations are required per *COT*, once the inputs are committed.

P_1 performs $2n$ commitments and runs n π_1^2 proofs on the commitments. He participates in n COT's as a sender. He finally computes a product of n ciphertexts (or $2n$ for the 2^{nd} option). P_2 performs n commitments and runs n π_1^2 proofs on the commitments. He participates in n COT's as a receiver.

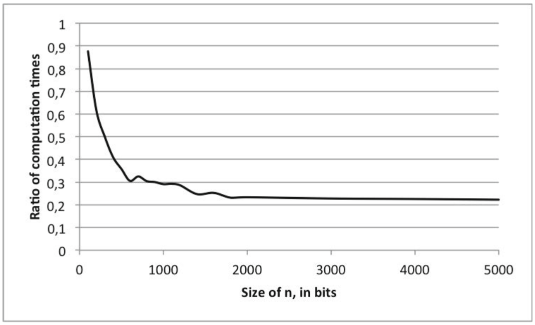


Fig. 5. Ratio computation times between our Basic Scheme and the protocol of [15]

He finally computes a product of n ciphertexts (or $2n$ for the 2^{nd} option). The bandwidth mainly comprises $3n$ commitments and n COT's.

In [20], Jarrous and Pinkas also propose an adaptation of their binHDOT protocol to the malicious setting. They also use a particular Committed Oblivious Transfer functionality, with proofs that the inputs differ by a constant number Δ, while we prove that our inputs always differ by 1. However, their protocol (for a more generic functionality) ends with an oblivious polynomial evaluation.

References

1. Aumann, Y., Lindell, Y.: Security against covert adversaries: Efficient protocols for realistic adversaries. In: Vadhan, S.P. (ed.) TCC 2007. LNCS, vol. 4392, pp. 137–156. Springer, Heidelberg (2007)
2. Blanton, M., Gasti, P.: Secure and efficient protocols for iris and fingerprint identification. In: Atluri, V., Diaz, C. (eds.) ESORICS 2011. LNCS, vol. 6879, pp. 190–209. Springer, Heidelberg (2011)
3. Bringer, J., Chabanne, H., Patey, A.: Shade: Secure hamming distance computation from oblivious transfer. IACR Cryptology ePrint Archive 2012, 586 (2012)
4. Camenisch, J., Stadler, M.: Efficient group signature schemes for large groups (extended abstract). In: Kaliski Jr., B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 410–424. Springer, Heidelberg (1997)
5. Camenisch, J., Stadler, M.: Proof systems for general statements about discrete logarithms. Tech. rep., Dept. of Computer Science, ETH Zurich (1997)
6. Canetti, R.: Security and composition of multiparty cryptographic protocols. J. Cryptology 13(1), 143–202 (2000)
7. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. In: FOCS, pp. 41–50 (1995)
8. Crépeau, C.: Verifiable disclosure of secrets and applications (abstract). In: Quisquater, J.-J., Vandewalle, J. (eds.) EUROCRYPT 1989. LNCS, vol. 434, pp. 150–154. Springer, Heidelberg (1990)
9. Cristofaro, E.D., Gasti, P., Tsudik, G.: Fast and private computation of set intersection cardinality. IACR Cryptology ePrint Archive 2011, 141 (2011)
10. Damgård, I., Jurik, M.: A generalisation, a simplification and some applications of paillier's probabilistic public-key system. In: Kim, K.-C. (ed.) PKC 2001. LNCS, vol. 1992, pp. 119–136. Springer, Heidelberg (2001)
11. El Gamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 10–18. Springer, Heidelberg (1985)
12. Goethals, B., Laur, S., Lipmaa, H., Mielikäinen, T.: On private scalar product computation for privacy-preserving data mining. In: Park, C.-S., Chee, S. (eds.) ICISC 2004. LNCS, vol. 3506, pp. 104–120. Springer, Heidelberg (2005)
13. Goldreich, O.: The Foundations of Cryptography. Basic Applications, vol. 2. Cambridge University Press (2004)
14. Hazay, C., Lindell, Y.: Efficient Secure Two-Party Protocols. Springer (2010)
15. Huang, Y., Evans, D., Katz, J., Malka, L.: Faster secure two-party computation using garbled circuits. In: USENIX Security Symposium (2011)
16. Ishai, Y., Kilian, J., Nissim, K., Petrank, E.: Extending oblivious transfers efficiently. In: Boneh, D. (ed.) CRYPTO 2003. LNCS, vol. 2729, pp. 145–161. Springer, Heidelberg (2003)

17. Ishai, Y., Prabhakaran, M., Sahai, A.: Founding cryptography on oblivious transfer – efficiently. In: Wagner, D. (ed.) CRYPTO 2008. LNCS, vol. 5157, pp. 572–591. Springer, Heidelberg (2008)
18. Ishai, Y., Prabhakaran, M., Sahai, A.: Secure arithmetic computation with no honest majority. In: Reingold, O. (ed.) TCC 2009. LNCS, vol. 5444, pp. 294–314. Springer, Heidelberg (2009)
19. Jarecki, S., Shmatikov, V.: Efficient two-party secure computation on committed inputs. In: Naor, M. (ed.) EUROCRYPT 2007. LNCS, vol. 4515, pp. 97–114. Springer, Heidelberg (2007)
20. Jarrous, A., Pinkas, B.: Secure hamming distance based computation and its applications. In: Abdalla, M., Pointcheval, D., Fouque, P.-A., Vergnaud, D. (eds.) ACNS 2009. LNCS, vol. 5536, pp. 107–124. Springer, Heidelberg (2009)
21. Kiraz, M.S., Schoenmakers, B., Villegas, J.: Efficient committed oblivious transfer of bit strings. In: Garay, J.A., Lenstra, A.K., Mambo, M., Peralta, R. (eds.) ISC 2007. LNCS, vol. 4779, pp. 130–144. Springer, Heidelberg (2007)
22. Kolesnikov, V., Schneider, T.: Improved garbled circuit: Free XOR gates and applications. In: Aceto, L., Damgård, I., Goldberg, L.A., Halldórsson, M.M., Ingólfsdóttir, A., Walukiewicz, I. (eds.) ICALP 2008, Part II. LNCS, vol. 5126, pp. 486–498. Springer, Heidelberg (2008)
23. Lindell, Y., Pinkas, B.: An efficient protocol for secure two-party computation in the presence of malicious adversaries. In: Naor, M. (ed.) EUROCRYPT 2007. LNCS, vol. 4515, pp. 52–78. Springer, Heidelberg (2007)
24. Lindell, Y., Pinkas, B.: A proof of security of Yao's protocol for two-party computation. J. Cryptology 22(2), 161–188 (2009)
25. Lindell, Y., Pinkas, B.: Secure two-party computation via cut-and-choose oblivious transfer. In: Ishai, Y. (ed.) TCC 2011. LNCS, vol. 6597, pp. 329–346. Springer, Heidelberg (2011)
26. Mohassel, P., Niksefat, S., Sadeghian, S., Sadeghiyan, B.: An efficient protocol for oblivious DFA evaluation and applications. In: Dunkelman, O. (ed.) CT-RSA 2012. LNCS, vol. 7178, pp. 398–415. Springer, Heidelberg (2012)
27. Naor, M., Pinkas, B.: Efficient oblivious transfer protocols. In: SODA, pp. 448–457 (2001)
28. Nielsen, J.B., Orlandi, C.: LEGO for two-party secure computation. In: Reingold, O. (ed.) TCC 2009. LNCS, vol. 5444, pp. 368–386. Springer, Heidelberg (2009)
29. Osadchy, M., Pinkas, B., Jarrous, A., Moskovich, B.: Scifi - A system for secure face identification. In: IEEE Symposium on Security and Privacy, pp. 239–254 (2010)
30. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
31. Peikert, C., Vaikuntanathan, V., Waters, B.: A framework for efficient and composable oblivious transfer. In: Wagner, D. (ed.) CRYPTO 2008. LNCS, vol. 5157, pp. 554–571. Springer, Heidelberg (2008)
32. Pinkas, B., Schneider, T., Smart, N.P., Williams, S.C.: Secure two-party computation is practical. In: Matsui, M. (ed.) ASIACRYPT 2009. LNCS, vol. 5912, pp. 250–267. Springer, Heidelberg (2009)
33. Rabin, M.O.: How to exchange secrets with oblivious transfer. Tech. Rep. TR-81, Aiken Computation Lab, Harvard University (1981)
34. University of Maryland, University of Virginia: Might be evil: Privacy-preserving computing, http://mightbeevil.com
35. Yao, A.C.C.: How to generate and exchange secrets (extended abstract). In: FOCS, pp. 162–167 (1986)

Garbled Circuits via Structured Encryption*

Seny Kamara[1] and Lei Wei[2]

[1] Microsoft Research
[2] UNC-Chapel Hill

Abstract. The garbled circuit technique transforms a circuit in such a way that it can be evaluated on encrypted inputs. Garbled circuits were originally introduced by Yao (FOCS '86) for the purpose of secure two-party computation but have since found many applications.

In this work, we consider the problem of designing *special-purpose* garbled circuits, which are garbled circuits that handle only a specific class of functionalities. Special-purpose constructions are usually smaller than general-purpose ones and lead to more efficient two-party protocols.

We propose a design framework for constructing special-purpose garbled circuits based on structured encryption schemes, which are encryption schemes that encrypt data structures in such a way that they can be queried through the use of a token. Using our framework, we show how to design more efficient garbled circuits for several graph-based functionalities (with applications to online social network analysis), Boolean circuits, deterministic finite automata, and branching programs.

1 Introduction

Yao's garbled circuit technique transforms circuits in such a way that they can be evaluated on encrypted inputs. While garbled circuits were originally introduced for the purpose of two-party secure function evaluation (SFE) [19], they have since found many applications, some of which include the design of homomorphic encryption schemes, one-time programs, circular-secure encryption, non-interactive verifiable computation, functional encryption, and single-server-aided SFE.

At a high level, the garbled circuit technique consists of: (1) a garbling procedure that transforms a circuit C that computes a function f, and a set of inputs $\mathbf{x} = (x_1, \ldots, x_n)$ into a garbled circuit $\widetilde{\mathrm{C}}$ and an encoded input $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$; (2) an evaluation procedure that computes a garbled output $\widetilde{\mathbf{y}}$ given $\widetilde{\mathrm{C}}$ and $\widetilde{\mathbf{x}}$; and (3) a decoding procedure that, given $\widetilde{\mathbf{y}}$ and a set of decoding keys $\mathbf{dk}$ returns $f(x)$. The main security property provided by garbled circuits is *input privacy*, which guarantees that, given $(\widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}, \mathbf{dk})$, no information about $\mathbf{x}$ is revealed by the garbled circuit evaluation beyond what can be inferred from $f(\mathbf{x})$. As shown by Yao, combining garbled circuits with oblivious transfer results in constant-round two-party SFE secure against semi-honest adversaries.

The importance of the garbled circuit technique in cryptography can be attributed to several factors, including its security properties, its relative efficiency

* Work done while at Microsoft Research.

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 177–188, 2013.

and, most importantly, its generality. In fact, like fully-homomorphic encryption, garbled circuits are one of the few general-purpose primitives in cryptography. While generality is crucial for establishing completeness theorems and for understanding the power of cryptographic techniques, it is well-known that it often comes at the price of efficiency. In fact, it is common for special-purpose constructions (i.e., constructions that handle only a a sub-class of functionalities) to be more efficient than general-purpose constructions.

Our Contributions. In this work, we consider the problem of designing special-purpose garbling schemes. Given the importance of garbled circuits and the efficiency improvements enjoyed by special-purpose constructions, this is a natural and well-motivated problem. We make the following contributions.

We introduce a general framework for designing special-purpose garbling schemes. Our framework is based on a connection between garbled circuits and the notion of *structured encryption* [8] which is a generalization of index-based searchable symmetric encryption (SSE) [18,10,7,9]. Roughly speaking, a structured encryption scheme encrypts a data structure in such a way that it can be queried through the use of a query-specific token that does not reveal information about the query. Our approach essentially reduces the problem of designing special-purpose garbled circuits to the problem of designing structured encryption schemes. Consequently, improvements in either the efficiency or functionality of structured encryption can lead to similar improvements in the design of special-purpose two-party protocols in the semi-honest model and other cryptographic primitives that rely on input-private garbled circuits.

While our main contributions are conceptual, we demonstrate the utility of our approach by constructing special-purpose garbling schemes for several useful functionalities. For example, using our framework with the structured encryption schemes of [8], we get special-purpose garbling schemes (and therefore two-party protocols) for several graph-based functionalities that have applications to online social networks. In addition, in the full version of this work we use our framework to construct garbling schemes for other functionalities like branching programs (BP), deterministic finite automata (DFA) and even Boolean circuits. In all cases, the garbled circuits resulting from our approach are more efficient (i.e., either smaller or with faster evaluation) than the garbled circuits that would result from applying Yao's general-purpose construction.

The main building block we need to handle DFAs, BPs and Boolean circuits is a matrix encryption scheme that supports lookups, i.e., a structured encryption scheme that encrypts matrices in such a way that a location (i, j) can be queried using a token. While such a scheme is described in [8], that particular construction is not appropriate for our purposes. The problem is that the scheme from [8] is only 1-dimensional in the sense that it generates a single token for a location (i, j) in the matrix. For our purposes, however, we need a 2-dimensional scheme that generates two independent tokens, i.e., one for i and one for j that can be combined to lookup location (i, j). We show how to construct such a scheme based on the 1-dimensional construction of [8] and pseudo-random synthesizers [17].

1.1 Background on Structured Encryption

Several variants of structured encryption were described in [8] but for our purposes we need the *structure-only* variant which only encrypts data structures as opposed to the standard variant which also encrypts messages. A structured encryption scheme is a tuple of four polynomial-time algorithms $\mathsf{SE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Token}, \mathsf{Query_e})$ such that Gen is a probabilistic algorithm that takes as input a security parameter k and outputs a private key K. Let $\mathscr{T}$ be an abstract data type that maps queries q from a query space $\mathcal{Q}$ to an answer a from a response space $\mathcal{R}$. Enc is a probabilistic algorithm that takes as input a key K, a data structure $\delta \in \mathscr{T}$ and outputs an encrypted data structure γ. Token is a (possibly probabilistic) algorithm that takes as input a private key K and a query $q \in \mathcal{Q}$ and outputs a token τ. $\mathsf{Query_e}$ is a deterministic algorithm that takes as input an encrypted data structure γ and a token τ and outputs an answer $a \in \mathcal{R}$. Informally, a structured encryption scheme is secure against chosen-query attacks (CQA1) if no useful information about q and δ can be recovered from γ and τ beyond what can be deduced from a. We say that a structured encryption scheme is secure against *adaptive* chosen-query attacks (CQA2) if this holds even when queries are made adaptively (i.e., as a function of the encrypted data structure γ and the results of previous queries and tokens).

As a concrete example, consider a graph encryption scheme $\mathsf{Graph} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Token}, \mathsf{Neigh}_e)$ that supports neighbor queries (we refer the reader to [8] for a concrete construction). With such a scheme one can encrypt the edges E of a graph $G = (V, E)$ by computing $\gamma \leftarrow \mathsf{Enc}(K, E)$. A token for a vertex $v \in V$ can be created as $\tau \leftarrow \mathsf{Token}(K, v)$ and the neighbors of v, denoted $\Gamma(v)$, can be recovered by computing $\mathsf{Neigh}_e(\gamma, \tau)$.

Associative Structured Encryption. For our purposes, we need *associative* structured encryption schemes which allow one to associate arbitrary strings to each output. So, with respect to our previous example, an associative graph encryption scheme supporting neighbor queries would: (1) allow the encryptor to associate arbitrary strings to each vertex of the graph during the encryption step; and (2) reveal these strings whenever the associated vertex is in $\Gamma(v)$. More precisely, in addition to the secret key sk and the edges E, the Enc algorithm would also take as input a set of strings $(s_{v_1}, \ldots, s_{v_{|V|}})$, where s_{v_i} is associated with vertex v_i. Then, the algorithm Neigh_e would return, in addition to $\Gamma(v)$, the set $\{s_w\}_{w \in \Gamma(v)}$.

Due to space restrictions, we refer the reader to [8] for formal definitions of (associative) structured encryption and of the relevant security definitions.

1.2 Overview of Our Framework

At a high level, our framework consists of two steps. In the first step, the function f is represented as a *structured circuit* which is a circuit-like computational model where each gate g can query a data structure δ and where the input and output wires of g carry queries for the structures of g and g's descendent, respectively. Our notion of structured circuits is reminiscent of Naor and Nissim's circuits with lookup tables [16] though, in our setting, the contents of the data structure cannot be set during computation. In the second step, at a very high level, the structured circuit is garbled by encrypting each data structure δ with an appropriate structured encryption scheme. These encrypted structures are viewed as the garbled gates and the tokens used to query them are viewed as the encoded wire values.

Note that the functionality and security properties needed to construct a garbled gate are precisely what is provided by associative structured encryption schemes. Indeed, a garbled gate must: (1) privately store the encodings for its outgoing wires; (2) reveal those encodings when presented with encodings for its input wires (according to the operation implemented by the gate); and (3) not reveal anything about a wire value given only its encoding. Similarly, an associative structured encryption scheme encrypts a data structure in such a way that: (1) arbitrary strings can be stored in the encrypted structure and associated with any answer; (2) these strings are only revealed when presented with an appropriate query token; and (3) no information is revealed about the query from the token.

2 Related Work

Garbled Circuits. Garbled circuits were introduced by Yao in his seminal work on SFE [19]. Since, they have found many additional applications as discussed in Section 1. Due to their wide applicability, several garbling techniques have been introduced over the years. Recently, Applebaum, Ishai and Kushilevitz proposed the first garbling scheme for arithmetic circuits [1].

Formalizations of garbled circuits have been proposed in the past. The most notable is the notion of randomized encodings (RE), which was introduced by Ishai and Kushilevitz [12,13]. While REs have found many applications in cryptography and even in complexity theory, the RE abstraction is not appropriate for certain applications. Recently, Bellare, Hoang and Rogaway [4,3] provided a formal treatment of garbled circuits. The formalization we use here is similar to that of [4,3] but does not capture function privacy (which, intuitively, guarantees that a garbled circuit does not reveal information about its functionality) since we are mostly concerned here with applications to two-party computation (as opposed to private function evaluation).

Special-Purpose Garbled Circuits. In addition to the general-purpose constructions described above, several works in the past have proposed two-party protocols for various classes of functions that (sometimes implicitly) relied on special-purpose garbled circuits. Some examples are [14,6,2], which construct efficient two-party protocols for evaluating ordered binary decision diagrams (OBDDS); and [15] which gives an efficient protocol for evaluating DFAs. All these protocols can be viewed as a combination of a special-purpose garbling scheme with OT, just as Yao's general-purpose two-party protocol is a combination of a general-purpose garbling scheme with OT.

Structured Encryption. Structured encryption was introduced in [8] as a generalization of the notion of a secure index considered by Song, Wagner and Perrig in [18] and Goh in [10] for the purpose of building searchable symmetric encryption (SSE) schemes. SSE was first considered explicitly in [18].

3 Preliminaries

We use oracles in some of our definitions for conciseness. In each case, these oracles only allow the adversary to make a *single* query. To stress this, we will often say that an oracle is a single-query oracle.

An *abstract data type* is a collection of objects together with a set of operations defined on those objects. We recall the formalization of an abstract data type given in [8]. Formally, a data type $\mathscr{T}$ is defined by a universe $\mathcal{U} = \{\mathcal{U}_k\}_{k\in\mathbb{N}}$ and an operation $\mathsf{Query} : \mathcal{U} \times \mathcal{Q} \rightarrow \mathcal{R}$, where $\mathcal{Q} = \{\mathcal{Q}_k\}_{k\in\mathbb{N}}$ is the operation's query space and $\mathcal{R} = \{\mathcal{R}_k\}_{k\in\mathbb{N}}$ is its response space. The universe, query and response spaces are ensembles of finite sets indexed by the security parameter k. We assume that the universe is a totally ordered set and that the response space includes special elements $\perp$ and ϵ denoting failure and the empty string, respectively. Given a data structure δ we sometimes write $\mathscr{T}(\delta)$ to refer to its type.

4 Syntactic and Security Definitions

A garbling scheme Garb consists of four algorithms $(\mathsf{Grb}, \mathsf{Enc}, \mathsf{Eval}, \mathsf{Dec})$. The algorithm Grb is used to garble a circuit and to generate a secret key sk and a set of decoding keys $\mathbf{dk}$. The algorithm Enc is used with the secret key to encode inputs, and the Eval algorithm is used to evaluate a garbled circuit on a set of encoded inputs. Evaluation results in an encoded output which can be decoded into the real output using the decoding algorithm Dec and an appropriate subset of the decoding keys.

Definition 1 (Garbling scheme). *A garbling scheme* $\mathsf{Garb} = (\mathsf{Grb}, \mathsf{Enc}, \mathsf{Eval}, \mathsf{Dec})$ *consists of four polynomial-time algorithms that work as follows:*

- $(\widetilde{\mathrm{C}}, \mathbf{dk}, \mathrm{sk}) \leftarrow \mathsf{Grb}(1^k, \mathrm{C})$: *is a probabilistic algorithm that takes as input a circuit* C *with n inputs and ℓ outputs and returns a garbled circuit* $\widetilde{\mathrm{C}}$, *a set of decoding keys* $\mathbf{dk} = (\mathrm{dk}_1, \ldots, \mathrm{dk}_\ell)$ *and a secret key* sk.
- $\widetilde{x} := \mathsf{Enc}(\mathrm{sk}, x)$: *is a deterministic algorithm that takes as input a secret key* sk, *an input x and returns an encoded input $\widetilde{x}$. We sometimes write* $\widetilde{\mathbf{x}} := \mathsf{Enc}(\mathrm{sk}, \mathbf{x})$ *to denote the algorithm that takes multiple inputs* $\mathbf{x} = (x_1, \ldots, x_n)$, *runs* $\mathsf{Enc}(\mathrm{sk}, \cdot)$ *on each x_i and returns the garbled inputs* $\widetilde{x}_1$ *through* $\widetilde{x}_n$.

- $\widetilde{\mathbf{y}} := \mathsf{Eval}(\widetilde{\mathrm{C}}, \widetilde{\mathbf{x}})$ *: is a deterministic algorithm that takes as input a garbled circuit* $\widetilde{\mathrm{C}}$ *and encoded inputs* $\widetilde{\mathbf{x}}$ *and returns encoded outputs* $\widetilde{\mathbf{y}}$*.*
- $\{\bot, y_i\} := \mathsf{Dec}(\mathrm{dk}_i, \widetilde{y}_i)$ *: is a deterministic algorithm that takes as input a decoding key* dk_i *and an encoded output* $\widetilde{y}_i$ *and returns either the failure symbol* $\bot$ *or an output* y_i*. We sometimes write* $\{\bot, \mathbf{y}\} := \mathsf{Dec}(\mathbf{dk}, \widetilde{\mathbf{y}})$ *to denote the algorithm that takes multiple garbled outputs* $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_\ell)$*, runs* $\mathsf{Dec}(\mathrm{dk}_i, \cdot)$ *on each* $\widetilde{y}_i$ *and returns the outputs* y_1 *through* y_ℓ*.*

We say that Garb *is correct if for all* $k \in \mathbb{N}$*, for all polynomial-size circuits* C*, for all inputs* $\mathbf{x}$ *for in the domain of* C*, for all* $(\widetilde{\mathrm{C}}, \mathbf{dk}, \mathrm{sk})$ *output by* $\mathsf{Grb}(1^k, \mathrm{C})$*, for* $\widetilde{\mathbf{x}} := \mathsf{Enc}(\mathrm{sk}, \mathbf{x})$ *and* $\widetilde{\mathbf{y}} := \mathsf{Eval}\big(\widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}\big)$ *and for all* $i \in [\ell]$*,* $\mathsf{Dec}\big(\mathrm{dk}_i, \widetilde{y}_i\big) = y_i$*.*

Non-Adaptive Input Privacy. Most applications of garbled circuits rely on a simple notion of security that guarantees that a garbled circuit $\widetilde{\mathrm{C}}$ together with encoded inputs $\widetilde{\mathbf{x}}$ and the decoding keys $\mathbf{dk}$ reveal at most $f(\mathbf{x})$. The following simulation-based definition guarantees that the garbled circuit, the encoded inputs and the decoding keys are all simulatable given the result of the computation. Intuitively, this implies that for some set of inputs $\mathbf{x}$, an efficient adversary that holds $(\widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}, \mathbf{dk})$ will not learn anything beyond $f(\mathbf{x})$.

Definition 2 (Sim1-security). *A garbling scheme* Garb = (Grb, Enc, Eval, Dec) *is* SIM1*-secure with respect to a circuit* C *if, for all polynomial-size adversaries* $\mathcal{A}$*, there exists a polynomial-size simulator* $\mathcal{S}$ *such that the following distributions are computationally indistinguishable:*

$$\Big\{\big\langle \widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}, \mathbf{dk} \big\rangle : (\widetilde{\mathrm{C}}, \mathbf{dk}, \mathrm{sk}) \leftarrow \mathsf{Grb}\big(1^k, \mathrm{C}\big); \mathbf{x} \leftarrow \mathcal{A}(1^k); \widetilde{\mathbf{x}} \leftarrow \mathsf{Enc}(\mathrm{sk}, \mathbf{x})\Big\},$$

$$\Big\{\big\langle \widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}, \mathbf{dk} \big\rangle : \mathbf{x} \leftarrow \mathcal{A}(1^k); (\widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}, \mathbf{dk}) \leftarrow \mathcal{S}\big(\mathrm{C}, f(\mathbf{x})\big)\Big\}.$$

Adaptive Input Privacy. While non-adaptive privacy is sufficient for some applications (e.g., secure two-party computation in the semi-honest model) there are other useful applications for which it falls short. This typically occurs in situations where the adversary can choose its inputs *as a function of the garbled circuit* (for example in one-time programs [11]). The following simulation-based definition of adaptive input privacy guarantees that the garbled circuit, the encoded input and the decoding keys are all simulatable given only the circuit and the result of the computation. Like the non-adaptive definition, this holds for adversarially-chosen inputs; but, unlike the non-adaptive definition, the inputs can be chosen as a function of the garbled circuit.

Definition 3 (Sim2-security). *A garbling scheme* Garb = (Grb, Enc, Eval, Dec) *is* SIM2*-secure with respect to a circuit* C *if, for all polynomial-size adversaries* $\mathcal{A}$*, there exists a polynomial-size stateful simulator* $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2)$ *such that the following distributions are computationally indistinguishable:*

$$\Big\{\big\langle \widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}, \mathbf{dk}, st_{\mathcal{A}} \big\rangle : (\widetilde{\mathrm{C}}, \mathbf{dk}, \mathrm{sk}) \leftarrow \mathsf{Grb}\big(1^k, \mathrm{C}\big); st_{\mathcal{A}} \leftarrow \mathcal{A}^{\mathsf{Enc}(\mathrm{sk}, \cdot)}(\widetilde{\mathrm{C}}, \mathbf{dk})\Big\},$$

$$\left\{ \left\langle \widetilde{\mathrm{C}}, \widetilde{\mathbf{x}}, \mathbf{dk}, st_{\mathcal{A}} \right\rangle : (\widetilde{\mathrm{C}}, \mathbf{dk}) \leftarrow \mathcal{S}_1(\mathrm{C}); st_{\mathcal{A}} \leftarrow \mathcal{A}^{\mathrm{OSIM}^{\mathcal{S}_2}(\cdot)}(\widetilde{\mathrm{C}}, \mathbf{dk}) \right\},$$

where $\mathrm{OSIM}^{\mathcal{S}_2}$ *is a single-query oracle that takes as input* $\mathbf{x}$ *and returns* $\widetilde{\mathbf{x}} \leftarrow \mathcal{S}_2(\mathrm{C}, f(\mathbf{x}))$.

5 Garbling Schemes via Structured Encryption

The first step in our framework is to describe the functionality f as a *structured circuit* which, roughly speaking, is a circuit with gates that can query data structures that support a given set of operations. Given a structured circuit representation of f we then garble it using an appropriate set of structured encryption schemes.

Structured Circuits. An n input and m output structured circuit C over a basis $\mathcal{B} = \{\mathscr{T}_1, \ldots, \mathscr{T}_\beta\}$ is a directed acyclic graph with n input wires and m output wires such that each gate g has access to a data structure of type $\mathscr{T} \in \mathcal{B}$ which supports an operation $\mathsf{Query} : \mathcal{U} \times \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_\nu \to \mathcal{R}$. We say that g is a $(\mathscr{T}, \nu)$-gate if: (1) it has access to a structure δ of type $\mathscr{T}$; and (2) it has ν input wires that carry queries $(q_1, \ldots, q_\nu) \in \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_\nu$ and an output wire that carries answers in $\mathcal{R}$. We require that if g_1's output wire is g_2's ith input wire, then $\mathcal{R}_1 = \mathcal{Q}_{2,i}$ where $\mathcal{R}_1$ refers to the response space of g_1 and $\mathcal{Q}_{2,i}$ denotes the ith query space of g_2.

Throughout, we assume a topological ordering on C and denote its ith gate by g_i. For notational convenience, we sometimes write $\mathscr{T}(g)$, $\mathcal{Q}(g)$, $\mathcal{R}(g)$ to refer to a gate g's type, query space and answer space, respectively.

A structured circuit C is evaluated on input $(q_1, \ldots, q_n)$ from the input wires to the output wires. When the inputs to the incoming wires of a gate g have been obtained, the output wire of g is set to $a := \mathsf{Query}(\delta, q_1, \ldots, q_\nu)$. The output of the circuit are the values obtained on the output wires of the circuit.

5.1 Our Framework

We now describe our approach to designing special-purpose garbled circuits. Let C be a structured circuit over the basis $\mathcal{B}$ and let $\{\mathsf{SE}_1, \ldots, \mathsf{SE}_\beta\}$ be a set of structured encryption schemes for each abstract data type in $\mathcal{B}$. Our approach is described in detail in Fig. 1 and, at a high level, works as follows.

If g has access to a data structure δ of type $\mathscr{T}$ (e.g., a graph or a matrix) then δ is encrypted using a structured encryption scheme for type $\mathscr{T}$. The resulting encrypted structure γ is the garbled gate and the tokens for queries q are used as encodings for the wires. To allow for the connection of gates to one another, the underlying structured encryption schemes must be associative.

Intuitively, the input privacy of the resulting garbled (structured) circuit is guaranteed by the security of the structured encryption scheme. This approach results in garbled circuits that have the same size as the *structured* circuit for f. For certain functions, the structured circuit representation can be much smaller than the boolean circuit representation (we discuss this further in Section 6).

Let $\mathcal{B} = \{\mathscr{T}_1, \ldots, \mathscr{T}_\beta\}$ be a basis and $(\mathsf{SE}_1, \ldots, \mathsf{SE}_\beta)$ be associative structured encryption schemes for the types in $\mathcal{B}$. Construct a garbling scheme $\mathsf{Garb} = (\mathsf{Grb}, \mathsf{Enc}, \mathsf{Eval}, \mathsf{Dec})$ for the class of n input and m output structured circuits over $\mathcal{B}$ as follows:

- $\mathsf{Grb}(1^k, \mathrm{C})$:
  1. **(output gates)** let $\textsc{out} = (o_1, \ldots, o_m)$ be the set of output gates and for each o_i,
     (a) generate a key $K_i \leftarrow \mathsf{SE}_{\mathscr{T}(o_i)}.\mathsf{Gen}(1^k)$
     (b) for all $a \in \mathcal{R}(o_i)$, sample $\lambda_{i,a} \xleftarrow{\$} \{0,1\}^k$
     (c) compute $\gamma_i \leftarrow \mathsf{SE}_{\mathscr{T}(o_i)}.\mathsf{Enc}_{K_i}(\delta, \boldsymbol{\lambda})$, where δ is the structure of o_i and $\boldsymbol{\lambda} = \{\lambda_{i,a}\}_{a \in \mathcal{R}(o_i)}$
     (d) set dk_i to be a lookup table that maps $\lambda_{i,a}$ to a.
  2. **(non-output gates)** let $\overline{\textsc{out}} = (g_1, \ldots, g_\ell)$ be the set of non-output gates and for each g_i,
     (a) generate a key $K_i \leftarrow \mathsf{SE}_{\mathscr{T}(g_i)}.\mathsf{Gen}(1^k)$
     (b) let d be the descendant of g_i and let K_d be the key generated for it
     (c) for all $q \in \mathcal{Q}(d)$, compute $\tau_q := \mathsf{SE}_{\mathscr{T}(d)}.\mathsf{Token}_{K_d}(q)$
     (d) compute $\gamma_i \leftarrow \mathsf{SE}_{\mathscr{T}(g_i)}.\mathsf{Enc}_{K_i}(\delta, \boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{|\mathcal{Q}(d)|})$.
  3. let $\widetilde{\mathrm{C}} = (\gamma_1, \ldots, \gamma_{|\mathrm{C}|})$, where γ_j is the garbling of the jth gate in C
  4. if $\textsc{in} = (g_1^\star, \ldots, g_n^\star)$ are the inputs gates, let $\mathrm{sk} = (K_1^\star, \ldots, K_n^\star)$ be the keys generated for these gates
  5. let $\mathbf{dk} = (\mathrm{dk}_1, \ldots, \mathrm{dk}_m)$,
  6. output $(\widetilde{\mathrm{C}}, \mathbf{dk}, \mathrm{sk})$
- $\mathsf{Enc}(\mathrm{sk}, x)$: compute $\tau \leftarrow \mathsf{SE}_{\mathscr{T}(x)}.\mathsf{Token}_{K(x)}(x)$ and output $\widetilde{x} = \tau$.
- $\mathsf{Eval}(\widetilde{\mathrm{C}}, \widetilde{\mathbf{x}})$: evaluate $\widetilde{\mathrm{C}}$ from the input wires to the output wires as follows: when the tokens τ_1 and τ_2 of the incoming wires to a garbled gate γ have been obtained, set the output wire of γ to $\tau_3 \leftarrow \mathsf{SE}_{\mathscr{T}(\gamma)}.\mathsf{Query}_{\mathsf{e}}(\gamma, \tau_1, \tau_2)$. After processing all gates, output $\widetilde{\mathbf{y}} = (\lambda_{o_1}, \ldots, \lambda_{o_m})$, where λ_{o_i} is the value obtained on the output wire of the ith output gate o_i.
- $\mathsf{Dec}(\widetilde{\mathbf{x}}, \mathrm{dk}_i, \widetilde{y}_i)$: parse $\widetilde{y}_i$ as λ_i and output $y_i := \mathrm{dk}_i[\lambda_i]$.

Fig. 1. A framework for designing special-purpose garbled circuits

Non-adaptive Input Privacy. We show that if the underlying structured encryption schemes are CQA1-secure, then our construction results in a garbling scheme that provides non-adaptive input privacy. Due to space restrictions, the proof is deferred to the full version of this work.

Theorem 1. *If* $(\mathsf{SE}_1, \ldots, \mathsf{SE}_\beta)$ *are* CQA1*-secure, then the scheme described in Fig.1 is* SIM1*-secure.*

Adaptive Input Privacy. In the full version of this work, we also show that if the underlying schemes are CQA2-secure, then the resulting garbling scheme provides the stronger notion of adaptive input privacy.

Theorem 2. *If* $(\mathsf{SE}_1, \ldots, \mathsf{SE}_\beta)$ *are* CQA2*-secure, then the construction described in Fig.1 is* SIM2*-secure.*

A Remark on Yao's Construction. We observe that Yao's garbled circuit construction can be viewed as an instantiation of our framework using 2×2 matrix encryption schemes that support lookup queries. Recall that in Yao's construction, garbled circuits are constructed as follows. Each gate g in the circuit C is replaced with a garbled gate $\widetilde{g}$. Here, we assume without loss of generality that g has two input wires w_a and w_b and one output wire w_c. The bit values conducted by each wire are replaced with a randomly chosen encoding. So the 0 and 1 bits on wire w_a are encoded as ω_0^a and ω_1^a which are sampled uniformly at random. The encodings for all the bits of w_b and w_c are generated similarly. The garbled gate $\widetilde{g}$ is constructed such that, given (ω_0^a, ω_0^b) it returns $\omega_{g(0,0)}^c$, given (ω_0^a, ω_1^b) it returns $\omega_{g(0,1)}^c$, and so on. Notice that because the encodings are chosen uniformly at random, they do not reveal any information about the real wire values.

These garbled gates can be viewed as structured encryption schemes for 2×2 matrices that support lookups. To illustrate this, we briefly sketch how each implies the other (we defer a more formal treatment to the full version of this work). Given a Boolean gate g we can construct a garbled gate $\widetilde{g}$ using any associative 2-dimensional matrix encryption scheme for 2×2 matrices. The 0 and 1 labels for w_a are tokens for the first and second row, respectively; and the 0 and 1 labels for w_b are tokens for the first and second column, respectively. The garbled gate $\widetilde{g}$ is then the encryption of the matrix M defined as $M[i,j] = \tau_{g(i-1,j-1)}^c$, where τ_0^c and τ_1^c are the tokens used as encodings for g's output wire (alternatively, for one of its descendent's input wires). In the other direction, we can construct an associative 2×2 matrix encryption scheme from any garbled gate construction. It suffices to view the garbled gate as the encrypted matrix (replacing the output wire encodings with the associated data) and the input wire encodings as the tokens for lookup queries.

6 Concrete Constructions

In the previous Section, we showed how to construct special-purpose garbled circuits for any function f that can be written as a structured circuit over a basis $\mathcal{B}$. This requires, however, that we have structured encryption schemes for the data types in $\mathcal{B}$. In [8], several structured encryption schemes were proposed including a matrix encryption scheme that supports lookup queries, a graph encryption scheme that supports adjacency queries (i.e., given two nodes, test whether they are adjacent), a graph encryption scheme that supports neighbor queries (i.e., given a node, return all of its neighbors) and a web-graph encryption scheme that supports focused subgraph queries. All the schemes in [8] were shown CQA2-secure so, using our framework, we get adaptively-secure special-purpose garbling schemes for any structured circuit over the basis $\mathcal{B}$ consisting of the data types mentioned above.

This leads to garbling schemes and (when combined with OT in the natural way) special-purpose two-party protocols in the semi-honest model for several graph-based functionalities. Note that in all these functionalities there is a set of *public* vertices V, and one player holds a private set of edges E over V that the second player wants to query in some way. This captures several real-life scenarios, e.g., in online social network analysis where the identities of users is public (e.g., Facebook, LinkedIn, Google+) but the relationships between users (i.e., friendships, connections, relationships) is private. In particular, this leads to two-party protocols for the following functionalities:

- (neighbor queries) $f_V(E, v) = (\perp, \varGamma(v))$, where $\varGamma(v)$ are the neighbor of v.
- (adjacency queries) $f_V(E, (v_1, v_2)) = (\perp, M[v_1, v_2])$, where M_G is the adjacency matrix of $G = (V, E)$.
- (focused subgraph queries) $f_V((E, D_1, \ldots, D_{|V|}), w) = (\perp, \varSigma(w))$, where D_1 through $D_{|V|}$ are documents (e.g., user profiles) associated with the vertices in V and $\varSigma(w) = \{v_i \in V : w \in D_i\} \cup \{\varGamma(v_i) \subseteq V : w \in D_i\}$, i.e., the vertices whose documents contain the keyword w and their neighbors.

We briefly note that in the context of online social networks, focused subgraph queries (FSQ) allow P_2 to make queries of the type "search for all users who are friends with someone that likes product X", which is particularly compelling for marketing applications. In the context of healthcare (i.e., the vertices are patients and the documents are their medical records), FSQs allow P_2 to query for all patients who are related to someone who has a particular disease or symptom.

In addition to the schemes mentioned above, we can use our framework to design special-purpose garbling schemes (and therefore two-party protocols) for functionalities not handled by the structured encryption schemes of [8]. This includes Boolean circuits, DFAs and BPs. In fact, in the full version, we show that all these functionalities can be handled using a 2-dimensional matrix encryption scheme. Due to space restrictions, we only describe this new matrix encryption construction and leave its application to Boolean circuits, DFAs and BPs—which is straightforward—to the full version of this work.

While [8] show how to construct an associative matrix encryption scheme that is CQA2-secure, their particular construction is not appropriate for our purpose. More precisely, their scheme is only one-dimensional, in the sense that it only generates a single token for a lookup query (i, j). On the other, for our purposes, we need a scheme that generates independent tokens for i and j that can later be combined to do a lookup at location (i,j) on the encrypted matrix.

1-D Matrix Encryption. At a high level, the scheme from [8] works as follows: given an $n \times m$ matrix M a new matrix C is constructed such that each element $M[i,j]$ is stored in C at location $(\alpha,\beta) := P_{K_1}(i,j)$ encrypted under key $K_{\alpha,\beta} := F_{K_2}(\alpha,\beta)$, where $P : \{0,1\}^k \times [n] \times [m] \to [n] \times [m]$ is a pseudo-random permutation [1] $F : \{0,1\}^k \times [n] \times [m] \to \{0,1\}^{\ell(k)}$ is a pseudo-random function. The encrypted matrix is C and a lookup token for location (i,j) consists of the tuple $(\alpha,\beta,F_{K_2}(\alpha,\beta))$.

2-D Matrix Encryption. We sketch here how to make the scheme from [8] two-dimensional. Note that this approach only yields a CQA1-secure scheme. This, however, implies a SIM1-secure garbling scheme which is sufficient for important applications like two-party computation.

Let $\ell(k)$ be an upper bound on the length of the information stored in the matrix (e.g., the associated data). We use a primitive introduced by Naor and Reingold in [17] called a pseudo-random synthesizer, which can be built from weak pseudo-random functions. A synthesizer Synth is an efficiently computable function such that

$$\left\{\left\langle \mathsf{Synth}(x_i,y_j)\right\rangle_{1\le i,j\le m} : \mathbf{x} \overset{\$}{\leftarrow} X^n; \mathbf{y} \overset{\$}{\leftarrow} X^n\right\} \overset{c}{\approx} \left\{\langle \mathbf{r}\rangle : \mathbf{r} \overset{\$}{\leftarrow} X^{n^2}\right\}.$$

Let P and Q be two pseudo-random permuations and let F be a pseudo-random function. In the new scheme the element $M[i,j]$ is stored at location $(\alpha,\beta) := (P_{K_1}(i), Q_{K_2}(j))$ in C and XORed with the pad $K_{i,j} := \mathsf{Synth}(F_{K_3}(0\|\alpha), F_{K_3}(1\|\beta))$. Lookup tokens for location (i,j) are simply $(\alpha, F_{K_3}(0\|\alpha))$ and $(\beta, F_{K_3}(1\|\beta))$. It is easy to show that this scheme is CQA1-secure (we defer a proof to the full version) so by, Theorem 1, it can be used to construct a SIM1-secure garbling schemes. If SIM2-security is needed one can use the transformation of [3].

References

1. Applebaum, B., Ishai, Y., Kushilevitz, E.: How to garble arithmetic circuits. In: Symposium on Foundations of Computer Science (FOCS 2011), pp. 120–129. IEEE Computer Society (2011)
2. Barni, M., Failla, P., Kolesnikov, V., Lazzeretti, R., Sadeghi, A.-R., Schneider, T.: Secure evaluation of private linear branching programs with medical applications. In: Backes, M., Ning, P. (eds.) ESORICS 2009. LNCS, vol. 5789, pp. 424–439. Springer, Heidelberg (2009)
3. Bellare, M., Hoang, V.T., Rogaway, P.: Adaptively secure garbling with applications to one-time programs and secure outsourcing. In: Wang, X., Sako, K. (eds.) ASIACRYPT 2012. LNCS, vol. 7658, pp. 134–153. Springer, Heidelberg (2012)

[1] Note that pseudo-random permutations over small domains can be constructed using techniques from [5].

4. Bellare, M., Hoang, V.T., Rogaway, P.: Foundations of garbled circuits. In: ACM Conference on Computer and Communications Security (CCS 2012), pp. 784–796 (2012)
5. Black, J., Rogaway, P.: Ciphers with arbitrary finite domains. In: Preneel, B. (ed.) CT-RSA 2002. LNCS, vol. 2271, pp. 114–130. Springer, Heidelberg (2002)
6. Brickell, J., Porter, D., Shmatikov, V., Witchel, E.: Privacy-preserving remote diagnostics. In: ACM Conference on Computer and Communications Security (CCS 2007), pp. 498–507. ACM (2007)
7. Chang, Y.-C., Mitzenmacher, M.: Privacy preserving keyword searches on remote encrypted data. In: Ioannidis, J., Keromytis, A.D., Yung, M. (eds.) ACNS 2005. LNCS, vol. 3531, pp. 442–455. Springer, Heidelberg (2005)
8. Chase, M., Kamara, S.: Structured encryption and controlled disclosure. In: Abe, M. (ed.) ASIACRYPT 2010. LNCS, vol. 6477, pp. 577–594. Springer, Heidelberg (2010)
9. Curtmola, R., Garay, J., Kamara, S., Ostrovsky, R.: Searchable symmetric encryption: Improved definitions and efficient constructions. In: ACM Conference on Computer and Communications Security (CCS 2006), pp. 79–88. ACM (2006)
10. Goh, E.-J.: Secure indexes. Technical Report 2003/216, IACR ePrint Cryptography Archive (2003), http://eprint.iacr.org/2003/216
11. Goldwasser, S., Kalai, Y.T., Rothblum, G.N.: One-time programs. In: Wagner, D. (ed.) CRYPTO 2008. LNCS, vol. 5157, pp. 39–56. Springer, Heidelberg (2008)
12. Ishai, Y., Kushilevitz, E.: Randomizing polynomials: A new representation with applications to round-efficient secure computation. In: IEEE Symposium on Foundations of Computer Science (FOCS 2000), pp. 294–304. IEEE Press (2000)
13. Ishai, Y., Kushilevitz, E.: Perfect constant-round secure computation via perfect randomizing polynomials. In: Widmayer, P., Triguero, F., Morales, R., Hennessy, M., Eidenbenz, S., Conejo, R. (eds.) ICALP 2002. LNCS, vol. 2380, pp. 244–256. Springer, Heidelberg (2002)
14. Kruger, L., Jha, S., Goh, E.-J., Boneh, D.: Secure function evaluation with ordered binary decision diagrams. In: ACM Conference on Computer and Communications Security (CCS 2006), pp. 410–420. ACM (2006)
15. Mohassel, P., Niksefat, S., Sadeghian, S., Sadeghiyan, B.: An efficient protocol for oblivious DFA evaluation and applications. In: Dunkelman, O. (ed.) CT-RSA 2012. LNCS, vol. 7178, pp. 398–415. Springer, Heidelberg (2012)
16. Naor, M., Nissim, K.: Communication preserving protocols for secure function evaluation. In: Symposium on Theory of Computing (STOC 2001), pp. 590–599. ACM (2001)
17. Naor, M., Reingold, O.: Synthesizers and their application to the parallel construction of pseudo-random functions. In: Symposium on Foundations of Computer Science (FOCS 1995), pp. 170–181. IEEE Computer Society (1995)
18. Song, D., Wagner, D., Perrig, A.: Practical techniques for searching on encrypted data. In: IEEE Symposium on Research in Security and Privacy, pp. 44–55. IEEE Computer Society (2000)
19. Yao, A.: How to generate and exchange secrets. In: IEEE Symposium on Foundations of Computer Science (FOCS 1986), pp. 162–167. IEEE Computer Society (1986)

On the Minimal Number of Bootstrappings in Homomorphic Circuits

Tancrède Lepoint[1,2] and Pascal Paillier[1]

[1] CryptoExperts, France
[2] École Normale Supérieure, France
{tancrede.lepoint,pascal.paillier}@cryptoexperts.com

Abstract. We propose a method to compute the *exact* minimal number of bootstrappings required to homomorphically evaluate any circuit. Given a circuit (typically over $\mathbb{F}_2$ although our method readily extends to circuits over any ring), the maximal noise level supported by the considered fully homomorphic encryption (FHE) scheme and the desired noise level of circuit inputs and outputs, our algorithms return a minimal subset of circuit variables such that boostrapping these variables is enough to perform an evaluation of the whole circuit. We introduce a specific algorithm for 2-level encryption (first generation of FHE schemes) and an extended algorithm for $\ell_{\max}$-level encryption with arbitrary $\ell_{\max} \geqslant 2$ to cope with more recent FHE schemes. We successfully applied our method to a range of real-world circuits that perform various operations over plaintext bits. Practical results show that some of these circuits benefit from significant improvements over the naive evaluation method where all multiplication outputs are bootstrapped. In particular, we report that a circuit for the AES S-box put forward by Boyar and Peralta admits a solution in 17 bootstrappings instead of 32, thereby leading to a 88% faster homomorphic evaluation of AES for any 2-level FHE scheme.

Keywords: Fully Homomorphic Encryption, Bootstrapping, Boolean Circuits, AES S-box.

1 Introduction

Fully homomorphic encryption (FHE) allows a worker to evaluate any circuit on plaintext values while manipulating their encryption in a public fashion *i.e.* with no knowledge of the decryption key. Gentry's original proposal [13] introduced a design principle that was later followed by a lot of FHE schemes [13,12,20,9,5,10,8]. Inherent to this design principle is the property that ciphertexts contain some noise which grows with successive homomorphic multiplications; thus ciphertexts need to be *refreshed* to maintain a low level of noise and allow subsequent homomorphic operations. In order to refresh ciphertexts, Gentry's key idea, referred to as *bootstrapping*, consists in homomorphically evaluating the decryption circuit of the FHE scheme using the decryption key bits in encrypted form, thus resulting in a different encryption of the same plaintext but with reduced noise. One ensures

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 189–200, 2013.

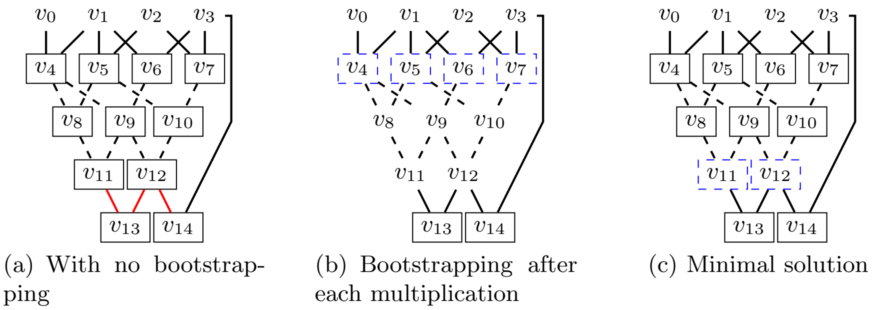
that the scheme parameters are such that the refreshed ciphertexts can handle at least one additional homomorphic multiplication. By repeating this procedure, the number of homomorphic operations becomes unlimited, thereby yielding a *fully* homomorphic encryption scheme.

Noise Levels. In all known FHE schemes, a ciphertext c_i contains a noise r_i which grows along with homomorphic multiplications and decryption is ensured as long as r_i does not exceed a given bound, *i.e.* $r_i < r_{\max}$. Without loss of generality, we can assume that the noise is lower-bounded by the noise after a bootstrapping operation[1]. We adopt a simplified approach by associating with each ciphertext c_i a *discretized noise level* $\ell_i = 1, 2, \ldots,$ where 1 is the noise level of ciphertexts resulting from a bootstrapping operation. Let c_1 (resp. c_2) be a ciphertext with noise level ℓ_1 (resp. ℓ_2). Gentry-like FHE schemes are such that $c_3 = c_1 + c_2$ has noise level $\ell_3 = \max(\ell_2, \ell_1)$ and $c_3 = c_1 \times c_2$ has noise level $\ell_3 = \ell_1 + \ell_2$, where + and × respectively denote homomorphic addition and multiplication. Therefore in these schemes, the noise level grows exponentially with the number of homomorphic multiplications: to evaluate a circuit with L sequential layers of multiplications, one must impose the maximum noise level $\ell_{\max}$ to be larger than 2^L. This is practically unacceptable even for small values of L and one must resort to bootstrapping periodically as the circuit is being evaluated.

Note that our definition of noise levels neglects the logarithmic increase of the noise size after a homomorphic addition. This approximation is often considered in the literature and remains valid as long as the proportion of additions does not become overwhelming in the circuit. Clearly, our simplified model would become invalid outside of this context.

Exponential vs. Linear FHE Schemes. For the purpose of this work, the above schemes will be referred to as being *exponential*. Recently, Brakerski, Gentry and Vaikuntanathan described a different framework where the ciphertext noise grows only linearly with the number of performed multiplications instead of exponentially [4]. This framework was used in several subsequent works [15,18,16] and even improved [3]. These FHE schemes are said to be *linear* throughout the paper. In linear schemes, homomorphic addition still outputs ciphertexts of level $\ell_3 = \max(\ell_1, \ell_2)$. However, a homomorphic multiplication $c_3 = c_1 \times c_2$ now results in a noise level $\ell_3 = \max(\ell_1, \ell_2) + 1$. Thus, to evaluate a circuit with L layers of multiplications, one only requires $\ell_{\max} \geqslant L$. However, when the depth of the circuit is not known at key generation time, this improvement is not strong enough to completely eliminate the need for intermediate bootstrappings.

[1] Notice that in most FHE schemes, freshly generated ciphertexts have a smaller noise than the noise obtained after a bootstrapping operation, allowing the circuit evaluator to save several bootstrappings at the beginning of the circuit. However, it is very likely that in *real-world* applications, data to be evaluated homomorphically will have been pre-processed and will not contain the smallest possible noise anymore.



(a) With no bootstrapping

(b) Bootstrapping after each multiplication

(c) Minimal solution

Fig. 1. Different bootstrapping solutions in a FHE scheme with $\ell_{\max} = 2$. Plain lines represent homomorphic multiplications while dashed lines represent homomorphic additions. The red lines in (a) reveal that the ciphertext noise will exceed the noise limit. Variables in a plain rectangle have a "large" noise ($\ell_i = \ell_{\max} = 2$) and the ones in a dashed blue rectangle are bootstrapped *i.e.* are re-encrypted to convert a "large" noise ($\ell_i = 2$) into a "small" noise ($\ell_i = 1$).

Minimizing Bootstrappings. Overall, both exponential and linear FHE schemes must resort to boostrapping in homomorphic circuit evaluation, either periodically or once in a while. However, the bootstrapping operation is reported as being the most drastic computational bottleneck in all known FHE implementations [14,9,10,16,8]. Worse, most of them merely perform a bootstrapping operation right after each multiplication, as suggested in [13,12]. It is easily seen though, as shown by the toy example depicted on Fig. 1, that this simple approach is often not optimal and that *fewer* bootstrappings may be sufficient to evaluate the whole circuit if positioned more judiciously.

Note that, even though finding a minimal solution is trivial and easily done by hand in Fig. 1, this optimization problem seems to become far more difficult with (even slightly) more complex circuits. Automated tools are therefore necessary to identify (one of) the smallest possible set of circuit variables whose bootstrapping will ensure a complete circuit evaluation in minimal time.

Contributions and Outline. We propose two efficient algorithms that automatically find an *exact* minimal solution for any given circuit *i.e.* output a minimal list of circuit variables to which bootstrapping can be applied to evaluate the circuit. Section 2 introduces a first algorithm specific to the case of FHE schemes with a maximum noise level set to $\ell_{\max} = 2$. This covers both exponential and linear schemes since the two categories collide in this particular case. In Section 3, we extend our algorithm to support exponential FHE schemes handling up to $\ell_{\max} \geqslant 3$ noise levels. We show that the same extended algorithm can also be used with linear schemes via a problem reformulation. Finally, Section 4 reports a number of experimental results on a range of real-world circuits, namely the benchmarking circuits for MPC and FHE proposed by Smart and Tillich [19], as well as circuits implementing the AES S-box suggested by Boyar and Peralta [2,1].

2 Homomorphic Schemes with 2 Noise Levels

In this section, we consider a FHE scheme that can only handle two levels of randomness in ciphertexts, *i.e.* level-1 ciphertexts can either be added (yielding a level-1 ciphertext) or multiplied (yielding a level-2 ciphertext); however only addition can be performed on ciphertexts with levels $(1,2),(2,1)$ or $(2,2)$ since the result of a multiplication would not be decryptable. As a result, the scheme can only handle a single multiplication after each bootstrapping operation. This framework was heavily considered [13,12,9,5,10,8] and implementations are available [11,7].

2.1 Stating the Problem

Let $\mathsf{C} = \mathsf{C}(n_1, n_2)$ be a Boolean circuit made of AND, XOR and NOT gates which takes as input n_1 bits and outputs n_2 bits. We denote by $\mathsf{C}^\dagger$ the same circuit as C where gates are replaced with homomorphic additions and multiplications[2]. Feeding $\mathsf{C}^\dagger$ with n_1 encrypted bits (under the FHE scheme), it will then output n_2 encrypted bits corresponding to the outputs of C applied on the same input bits in the clear. We denote by $\mathsf{V} = \{v_i : 1 \leqslant i \leqslant n\}$ the set of all single-assignment variables (ciphertexts) used in $\mathsf{C}^\dagger$ where $v_1, \ldots, v_{n_1}$ are the input variables and $v_{n-n_2+1}, \ldots, v_n$ the output variables. Now we assign a noise level $\ell_i \in \{1, 2, \ldots\}$ to each v_i as follows: the noise levels $\ell_1, \ell_2, \ldots, \ell_{n_1} \in \{1, 2\}$ are already fixed by the input variables $v_1, \ldots, v_{n_1}$. Using the two rules $\ell_{i_3} = \max(\ell_{i_1}, \ell_{i_2})$ when $v_{i_3} = v_{i_1} + v_{i_2}$ and $\ell_{i_3} = \ell_{i_1} + \ell_{i_2}$ when $v_{i_3} = v_{i_1} \times v_{i_2}$, we let noise levels automatically propagate throughout the circuit down to some output levels $\ell_{n-n_2+1}, \ldots, \ell_n$. Note that the noise levels of intermediate and output variables are left totally unbounded during that initial propagation and may therefore exceed by far the maximum level $\ell_{\max} = 2$ supported by the FHE scheme, meaning that the corresponding variables are in fact not decryptable. However, bootstrapping some variable v_i resets ℓ_i to 1 and it is easily seen that bootstrapping all variables $v_1, \ldots, v_n$ makes them all decryptable again: we then say that $\mathsf{C}^\dagger$ is evaluatable. What we are after is a minimal subset $I \subseteq \{1, n\}$ such that bootstrapping v_i for all $i \in I$ has the same effect.

A Boolean Reformulation. To each $v_i \in \mathsf{V}$ is assigned a Boolean value $b_i \in \{\mathsf{True}, \mathsf{False}\}$ that tells whether v_i is to be bootstrapped or not when evaluating $\mathsf{C}^\dagger$. We also define a Boolean mapping $\mathsf{B}(v_i)$ such that

$$\mathsf{B}(v_i) = \mathsf{True} \quad \text{if and only if} \quad \ell_i = 1 \ .$$

We see that if $v_{i_3} = v_{i_1} + v_{i_2}$ then

$$\mathsf{B}(v_{i_3}) = b_{i_3} \vee \big(\mathsf{B}(v_{i_1}) \wedge \mathsf{B}(v_{i_2})\big) \ . \tag{1}$$

[2] XOR and NOT gates correspond to homomorphic additions and AND gates to homomorphic multiplications.

This is because $\ell_{i_3} = 1$ only if $\ell_{i_1} = \ell_{i_2} = 1$ or, as an alternate case, ℓ_{i_3} equals 2 when v_{i_3} is computed but bootstrapping v_{i_3} afterwards resets ℓ_{i_3} to 1. Moreover, if $v_{i_3} = v_{i_1} \times v_{i_2}$ then

$$\mathsf{B}(v_{i_3}) = b_{i_3} \ . \tag{2}$$

Indeed as the result of a multiplication v_{i_3} has level $\ell_{i_3} = 2$. The only way to get $\ell_{i_3} = 1$ is therefore to bootstrap v_{i_3} after computing it. We also see that $\mathsf{B}(v_i)$ is already determined for input variables since for $i = 1, \ldots, n_1$,

$$\mathsf{B}(v_i) = \begin{cases} \mathsf{True} & \text{if } \ell_i = 1, \\ b_i & \text{if } \ell_i \neq 1. \end{cases} \tag{3}$$

Overall, we see that the Boolean predicate B can also be propagated (as a multivariate Boolean expression) across the circuit using the above rules (1)–(3). This operation can be done statically given the description of the circuit and will result in a list of formal Boolean expressions for $\mathsf{B}(v_1), \ldots, \mathsf{B}(v_n)$ that only involve the "bootstrapping" variables $b_1, \ldots, b_n$.

We now capture the fact that $\mathsf{C}^\dagger$ is evaluatable or not as a Boolean predicate ϕ^2_{C}. In order to ascertain the correctness of all variables of $\mathsf{C}^\dagger$, one must just ensure that all variables entering a multiplication have noise level 1. Hence

$$\phi^2_{\mathsf{C}} = \bigwedge_{v_k = v_i \times v_j \in \mathsf{C}^\dagger} \left(\mathsf{B}(v_i) \wedge \mathsf{B}(v_j)\right) \ . \tag{4}$$

Obviously, ϕ^2_{C} is a predicate involving $b_1, \ldots, b_n$ (or a subset thereof) and can be computed once B has been propagated throughout the circuit. All in all, evaluating $\mathsf{C}^\dagger$ with a minimal number of bootstrappings is reformulated as a Boolean satisfiability problem: ϕ^2_{C} must be satisfied with a minimal number of variables $b_1, \ldots, b_n$ set to True.

DNF and Monotone Predicates. We observe that the Boolean predicate $\phi^2_{\mathsf{C}} = \phi^2_{\mathsf{C}}(b_1, \ldots, b_n)$ is monotone since no negated literal $\neg b_i$ appears in ϕ^2_{C}. A monotone predicate is trivially satisfiable by setting all its variables to True. What we want, however, is to satisfy ϕ^2_{C} with as few b_i's set to True as possible. An exact solution to our problem would be to represent ϕ^2_{C} in Disjunctive Normal Form (DNF) *i.e.* as an OR of ANDs. Given a DNF representation of ϕ^2_{C}, it is easy to identify an AND involving a minimal number of variables, thus providing a minimal bootstrapping configuration for $\mathsf{C}^\dagger$. However, noting $\mu(\phi^2_{\mathsf{C}}) \in [1, n]$ this minimal number, even just deciding whether $\mu(\phi^2_{\mathsf{C}}) \leqslant t$ for some $t \in [1, n]$ is a priori intractable:

Theorem 1 ([17], Th. 3.4). *Let ϕ be an n-variate Boolean monotone predicate and $t \in [1, n]$. Let $\mu(\phi)$ be the size of its smallest prime implicant. Deciding whether $\mu(\phi) \leqslant t$ is NP-complete.*

We therefore circumvent this obstacle by adopting a heuristic approach and further validate its effectiveness experimentally as reported later in the paper.

2.2 A Heuristic Solver

We observe that ϕ^2_{C} is computed in Eq. 4 as an accumulated conjunction: thus when propagating B across $\mathsf{C}^{\dagger}$, we systematically put each $\mathsf{B}(v_i)$ in minimal Conjunctive Normal Form (min-CNF) *i.e.* as an AND of ORs with as few terms as possible. Obviously $\mathsf{B}(v_i)$ becomes more complex (involves more b_i's) as the variable v_i is taken deeper in the circuit. However, the complexity increase remains incremental from $\mathsf{B}(v_{i_1}), \mathsf{B}(v_{i_2})$ to $\mathsf{B}(v_{i_3})$ for $v_{i_3} = v_{i_1} \mathsf{~op~} v_{i_2}$ and computing the min-CNF of $\mathsf{B}(v_{i_3})$ given the min-CNF of $\mathsf{B}(v_{i_1}), \mathsf{B}(v_{i_2})$ therefore requires a moderate computational effort. ϕ^2_{C} is then aggregated along the way as a min-CNF of other min-CNFs, which is easy to program. Once we are done collecting parts and putting together the multivariate predicate ϕ^2_{C}, we apply heuristic transformations on its min-CNF until it becomes small enough to allow a conversion to DNF using a standard algorithm. A minimal bootstrapping configuration is then selected from one of the smallest conjunctive clauses in the resulting DNF.

We apply 3 independent transformations on the min-CNF of ϕ^2_{C}:

1. **Bootstrap required variables:** if $\phi^2_{\mathsf{C}} = (\cdots) \wedge b_i \wedge (\cdots)$ for some b_i then set $b_i = \mathsf{True}$ and repeat the operation until no longer applicable;
2. **Remove redundant variables:** a variable b_i is *redundant* w.r.t. a variable b_j if every occurrence of b_i in a clause of ϕ^2_{C} appears together with an occurrence of b_j (but the converse might not be true). In other words, any clause c containing b_i is of the form $c = (\cdots) \vee b_i \vee (\cdots) \vee b_j \vee (\cdots)$. Setting $b_i = \mathsf{True}$ would of course lead to $c = \mathsf{True}$ but this will only remove all such clauses c from ϕ^2_{C}, whereas setting $b_j = \mathsf{True}$ instead might induce additional simplifications in other clauses of ϕ^2_{C}. Therefore, we set $b_i = \mathsf{False}$, propagate simplifications in the CNF of ϕ^2_{C}, repeat the operation until no longer applicable and restart with Step 1;
3. **Maintain minimal CNF:** Eliminate any clause that is tautologically implied by another clause of ϕ^2_{C}; repeat the operation until no longer applicable and restart with Step 1.

In practice, these transformations are reasonably efficient and allow us to reduce the min-CNF of ϕ^2_{C} in such proportions that converting it to DNF afterwards is either immediate or unnecessary (depending on the circuit C, ϕ^2_{C} sometimes reduces to True by itself along the way, which terminates our algorithm). Therefore, even though our method is unproven, we validated its practical effectiveness. We refer to Sections 4 and 4.2 for experimental results.

Remark 1. Note that one might may also want to ensure that some output variables v_{n-n_2+j} for $j \in J \subseteq [1, n_2]$ have noise level 1 instead of 2. Now, resolving ϕ^2_{C} and bootstrapping these output variables might not yield a minimal solution. To address this case, we simply accumulate the predicates $\mathsf{B}(v_{n-n_2+j})$ for $j \in J$ into ϕ^2_{C} and apply the exact same strategy as above.

3 Extension to FHE Schemes with Many Noise Levels

Assume we are now given a FHE scheme that can handle $\ell_{\max} \geqslant 2$ levels of noise. Let c_1, c_2 and c_3 be ciphertexts with noise levels ℓ_1, ℓ_2 and ℓ_3 respectively. As discussed earlier, there exists essentially two different formulas for ℓ_3 when $c_3 = c_1 \times c_2$:

- $\ell_3 = \ell_1 + \ell_2$: this corresponds to the settings of exponential schemes [13,12,9,10,5,3]. In these schemes, the modulus remains the same after a multiplication but the noise increase depends on the amount of initial noises in the input ciphertexts[3]. At most $\log_2(\ell_{\max})$ layers of homomorphic multiplications can be evaluated before resorting to bootstrapping;
- $\ell_3 = \max(\ell_1, \ell_2)+1$: this corresponds to linear FHE schemes found in [6,4,10]. The noise grows negligibly after a homomorphic multiplication, but the modulus is modified after each multiplication (therefore the relative amount of noise increases). This technique is known as modulus switching, wherein $\ell_{\max}$ different moduli are used to evaluate $\ell_{\max}$ layers of homomorphic multiplications without bootstrapping. Moreover two ciphertexts can only be added or multiplied when they have exactly the same noise level so that their underlying rings become identical. In the following, we assume that the cost of modulus switching for a variable v_i, *i.e.* incrementing its noise level, is negligible compared to the cost of a bootstrapping operation.

We generalize the method of Section 2 to FHE schemes with $\ell_{\max} \geqslant 2$ noise levels: Section 3.1 focuses on a extended algorithm that works with exponential schemes, and we show in Section 3.2 how to slightly modify $\mathsf{C}^\dagger$ in order to reuse the very same algorithm as a black-box to address linear schemes.

We recall that our goal is to minimize the number of bootstrappings needed to homomorphically evaluate the circuit $\mathsf{C}^\dagger$ on input $(v_i, \ell_i)_{1 \leqslant i \leqslant n_1}$. As above, we associate to every circuit variable $v_i \in \mathsf{V}$ a Boolean variable $b_i \in \{\mathsf{True}, \mathsf{False}\}$ that tells whether v_i is to be bootstrapped or not. Again, we construct a Boolean predicate $\phi_{\mathsf{C}}^{\ell_{\max}}$ as a function of $b_1, \ldots, b_n, \ell_1, \ldots, \ell_{n_1}$ that tells whether $\mathsf{C}^\dagger$ is evaluatable. We then rely on our heuristic solver of Section 2 to issue a minimal set $I \subseteq [1, n]$ such that $b_i = \mathsf{True}$ for all $i \in I$ implies $\phi_{\mathsf{C}}^{\ell_{\max}} = \mathsf{True}$.

3.1 Extension to Exponential FHE Schemes

To any variable $v_i \in \mathsf{V}$, we now associate a vector $\mathbf{B}(v_i) = (\mathsf{B}_{i,1}, \ldots, \mathsf{B}_{i,\ell_{\max}-1})$ with $(\ell_{\max} - 1)$ Boolean coefficients such that $\ell_i = j$ if and only if $B_{i,j}$ is the first coefficient set to True as j ranges from 1 to $\ell_{\max} - 1$, and $\ell_i = \ell_{\max}$ if none of the coefficients is True. We make use of the Boolean vector $\mathbf{B}(v_i)$ to encode the noise level ℓ_i of v_i and propagate it throughout the circuit as we did with $\mathsf{B}(v_i)$ in the binary case $\ell_{\max} = 2$. Let us describe in more detail how $\mathbf{B}(v_i)$ evolves when being propagated across the circuit:

[3] Notice that the order of noise increase is quite different between [13,12,9,10,5] and [3], but this does not change our high-level description.

- for $1 \leqslant i \leqslant n_1$ *i.e.* for input variables, set

$$\mathsf{B}_{i,j} = \mathsf{False} \quad \text{for } j \neq \ell_i \qquad \text{and} \qquad \mathsf{B}_{i,\ell_i} = \mathsf{True} \quad \text{if } \ell_i < \ell_{\max} \,.$$

- when $v_k = v_i + v_j$, set

$$\mathbf{B}(v_k) = \begin{pmatrix} b_k \vee \big(\mathsf{B}_{i,1} \wedge \mathsf{B}_{j,1}\big) \\ \big(\mathsf{B}_{i,1} \wedge \mathsf{B}_{j,2}\big) \vee \big(\mathsf{B}_{j,1} \wedge \mathsf{B}_{i,2}\big) \vee \big(\mathsf{B}_{i,2} \wedge \mathsf{B}_{j,2}\big) \\ \vdots \end{pmatrix} , \tag{5}$$

  Indeed, $\ell_k = 1$ if and only if v_k is bootstrapped or $(\ell_i, \ell_j) = (1,1)$, otherwise $\ell_k = 2$ if $(\ell_i, \ell_j) \in \big\{(1,2),(2,1),(2,2)\big\}$, etc. All vector coefficients $\mathsf{B}_{k,3}, \ldots, \mathsf{B}_{k,\ell_{\max}-1}$ are formed in the same fashion.
- when $v_k = v_i \times v_j$, set

$$\mathbf{B}(v_k) = \begin{pmatrix} b_k \\ \mathsf{B}_{i,1} \wedge \mathsf{B}_{j,1} \\ \big(\mathsf{B}_{i,1} \wedge \mathsf{B}_{j,2}\big) \vee \big(\mathsf{B}_{j,1} \wedge \mathsf{B}_{i,2}\big) \\ \vdots \end{pmatrix} . \tag{6}$$

  This multiplication expresses the fact that $\ell_k = \ell_i + \ell_j$. Indeed, $\ell_k = 1$ if and only if v_k is bootstrapped, $\ell_k = 2$ if and only if $\ell_i = \ell_j = 1$, and so forth.

Remark 2. Before explaining how to construct the Boolean formula $\phi_{\mathsf{C}}^{\ell_{\max}}$, let us give a couple of remarks on our representation. First of all, this representation does not imply that $\big(\mathsf{B}_{i,j} = \mathsf{True} \text{ and } \mathsf{B}_{i,m} = \mathsf{False} \text{ for } m \neq j\big) \iff \ell_i = j$, but that

$$\big(\mathsf{B}_{i,j} = \mathsf{True} \text{ and } \mathsf{B}_{i,m} = \mathsf{False} \text{ for } 1 \leqslant m < j\big) \iff \ell_i = j \,.$$

This allows us to simplify the formulas for homomorphic addition and multiplication as we do not need to check whether $\mathsf{B}_{i,m} = \mathsf{False}$ for $m > \ell_i$ (see $\mathsf{B}_{k,2}$ in Eq. (5) and $\mathsf{B}_{k,3}$ in Eq. (6)). Secondly, when all the elements of $\mathbf{B}(v_i)$ are False, this means that v_i is at the maximum level of noise $\ell_i = \ell_{\max}$. Therefore this representation nicely generalizes the one of Section 2.

We now construct the Boolean formula $\phi_{\mathsf{C}}^{\ell_{\max}}$ which tells whether the circuit is evaluatable by setting

$$\phi_{\mathsf{C}}^{\ell_{\max}} = \bigwedge_{v_i \in \mathsf{V}} (\ell_i \leqslant \ell_{\max}) = \bigwedge_{v_k = v_i \times v_j \in \mathsf{C}^\dagger} \Bigg(\bigvee_{1 \leqslant m \leqslant \ell_{\max}} \mathsf{B}_{k,m} \Bigg) \,.$$

Note that the clauses of $\phi_{\mathsf{C}}^{\ell_{\max}}$ encode the fact that to properly evaluate a homomorphic operation $v_k = v_i \ \mathsf{op}\ v_j$, one must just have $\ell_k \leqslant \ell_{\max}$. This is automatically guaranteed by induction for all additions; expressed on all multiplications, this constraint precisely gives the above expression. As before, we use minimal CNF representation to propagate $\mathbf{B}(v_i)$ throughout the circuit and aggregate all the clauses of $\phi_{\mathsf{C}}^{\ell_{\max}}$ on the way. This results in a min-CNF for $\phi_{\mathsf{C}}^{\ell_{\max}}$ to which we apply the same 3 simplifying transformations. We finally convert the resulting predicate to DNF (if necessary) to identify a minimal configuration.

Remark 3. Notice that one might want to ensure that (a subset of) the output variables have noise levels bounded by some $\ell \leqslant \ell_{\max}$. One then aggregates in $\phi_{\mathsf{C}}^{\ell_{\max}}$ the clauses $\bigvee_{i \leqslant \ell} \mathsf{B}_{n-n_2+j,i}$ for $j \in [1, n_2]$ before solving the system.

3.2 Extension to Linear FHE Schemes

In this section, we explain how to deal with the case where $\ell_3 = \max(\ell_1, \ell_2) + 1$ when $c_3 = c_1 \times c_2$. Instead of adapting the previous method, we apply it as a black box to a modified version of the homomorphic circuit $\mathsf{C}^\dagger$. The modified circuit will no longer be consistent with its specification but can be treated by our algorithm regardless. The key idea is to see that one can simulate the linear framework in the exponential framework by replacing every homomorphic multiplication $c_3 = c_1 \times c_2$ with a subcircuit $c_3 = (c_1 + c_2) \times c_{1,2}$ where $c_{1,2}$ is a fixed ciphertext with noise level $\ell_{1,2} = 1$. Indeed, we get

$$\ell_3 = \max(\ell_1, \ell_2) + \ell_{1,2} = \max(\ell_1, \ell_2) + 1,$$

which is the wanted value in linear schemes. As mentioned, the correctness of the modified circuit as a homomorphic version of C is destroyed, but our extended algorithm remains applicable to it and will compute a minimal bootstrapping configuration in an oblivious fashion.

Note however that we need to slightly twitch the extended solver, otherwise solutions might suggest to bootstrap the newly introduced variables $v_{i,j}$. This would not make any sense as these variables have no real existence and only serve as helper variables in our simulation. We can easily circumvent this by not assigning a Boolean $b_{i,j}$ (or equivalently by forcing it to be False in $\mathsf{B}_{i,j}$) to the variables $v_{i,j}$. This eliminates the undesired collateral effect of seeing these variables being bootstrapped when solving $\phi_{\mathsf{C}}^{\ell_{\max}}$. We then successfully compute a minimal bootstrapping configuration from $\phi_{\mathsf{C}}^{\ell_{\max}}$ as previously described.

4 Practical Experiments

In this section, we discuss practical results obtained by applying our algorithms on several circuits (see Table 1). We implemented our basic and extended solvers using Mathematica 9 running on a 2.6 GHz Intel Core i7 with 16 GB of RAM. Although we did not specifically measure execution times, these range from a few seconds to a few hours depending on the circuit size and $\ell_{\max}$ (timings tend to grow exponentially with $\ell_{\max}$). We focused on the benchmarking circuits for MPC/FHE proposed by Smart and Tillich [19], and on circuits put forward by Boyar and Peralta for the AES S-box [2,1]. For each circuit, we computed the minimal number of bootstrappings needed to evaluate homomorphically that circuit with an exponential FHE scheme supporting $\ell_{\max} = 2$ or $\ell_{\max} = 4$ noise levels and with level-1 inputs and outputs *i.e.*

$$\ell_1 = \cdots = \ell_{n_1} = 1 \qquad \text{and} \qquad \ell_{n-n_2+1} = \cdots = \ell_n = 1.$$

Table 1 reports the results we obtained by applying our algorithms to the selected circuits.

Table 1. Minimal number of bootstrappings with level-1 inputs and outputs

Circuit $\mathsf{C}^\dagger$	$\boldsymbol{\ell_{\max}}$	**Number of hom. multiplications in** $\mathsf{C}^\dagger$	**Exact minimal number of bootstappings**
Adder 32 bits [19]	2	127	127
Adder 32 bits [19]	4	127	64
Comparator 32 bits [19]	2	150	146
Comparator 32 bits [19]	4	150	74
DES (expanded key) [19]	2	18175	18041
DES (expanded key) [19]	4	18175	8997
AES S-box [2]	2	32	19
AES S-box [2]	4	32	12
AES S-box [1]	2	32	17
AES S-box [1]	4	32	12

4.1 MPC/FHE Benchmark Circuits

Our results show that circuits given as reference by [19] tend to be disappointing when $\ell_{\max} = 2$ as we find that the minimal number of bootstrapping required to evaluate them is nearly equal to the number of homomorphic multiplications, thus being very close to the (trivial) upper bound. This can be explained by the fact that these circuits are automatically generated from hardware components, and clearly not optimized: they were not constructed to be small in terms of gate count, or have a significantly smaller depth, etc. Their linear parts were not optimized either [2]. Also note that setting $\ell_{\max} = 4$ instead of 2 divides the number of required bootstrappings by a factor nearly two.

4.2 The Boyar-Peralta AES S-Box

To the best of our knowledge, the first real-life circuit evaluated by a fully homomorphic encryption scheme is a circuit for AES encryption proposed by Gentry, Halevi and Smart [16]. However the authors decided to completely get rid of the bootstrappings by choosing a FHE scheme with $\ell_{\max} = 100$ so that the entire circuit can be evaluated at once. The drawback of this choice is that the public key becomes *prohibitively large* and required a server with 256GB of RAM to run the implementation and issue performance benchmarks. The authors suggested that bootstrapping might certainly be used as an optimization, *i.e.* as a way to balance the running time and the memory requirements.

The non-linear part of AES, computing the S-box, cannot be performed by table lookups in an homomorphic implementation. We considered circuits for the AES S-box already optimized by Boyar and Peralta with respect to gate count or depth [2,1]. Our practical results are detailed on Table 1. Contrarily to the circuits of [19], the Boyar-Peralta circuits were optimized and we found that their minimal number of bootstrappings is nearly half the number of homomorphic multiplications when $\ell_{\max} = 2$. As a result, homomorphically evaluating an AES encryption with a 2-level FHE scheme can be boosted by a factor 1.88 by just choosing the circuit from [1] and use our 17-bootstrapping optimal configuration $\{t_{21}, t_{22}, t_{23}, t_{24}, t_{26}, t_{29}, t_{33}, t_{36}, t_{40}, s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$ as described in [8].

However, when $\ell_{\max}$ grows, the gain of minimal bootstrapping operations with respect to the case $\ell_{\max} = 2$ is smaller than for the circuits of [19] (and even lower bounded by 8) due to the structure of these circuits[4]. Since the output variables are required to have a minimal noise level, the last reduction phase implies that the minimal solution consists in bootstrapping these output variables.

5 Conclusion

We introduced a method that computes the exact minimal number of bootstrappings required to homomorphically evaluate any circuit using any known FHE scheme. When $\ell_{\max} = 2$, the number of homomorphic multiplications is a strict upper bound on the minimal number of bootstrappings but significantly better figures can be found using our approach as exemplified by the circuit from [1]. We see, however, that most commonly used circuits are disappointingly unoptimized with respect to their "bootstrapping complexity". As an avenue for future research, we suggest to explore algorithmic strategies to *build* bootstrapping-efficient circuits *i.e.* to decrease their boostrapping complexity by a specific design effort. Finally, it would be interesting to refine our definition of noise levels to take into account the additional logarithmic effects induced by homomorphic operations, especially in the case of linear FHE schemes.

References

1. Boyar, J., Matthews, P., Peralta, R.: Logic minimization techniques with applications to cryptology. Journal of Cryptology 26(2), 280–312 (2013)
2. Boyar, J., Peralta, R.: A depth-16 circuit for the AES s-box. Cryptology ePrint Archive, Report 2011/332 (2011), `http://eprint.iacr.org/`
3. Brakerski, Z.: Fully homomorphic encryption without modulus switching from classical GapSVP. In: Safavi-Naini, R., Canetti, R. (eds.) CRYPTO 2012. LNCS, vol. 7417, pp. 868–886. Springer, Heidelberg (2012)
4. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping. In: Goldwasser, S. (ed.) Innovations in Theoretical Computer Science 2012, pp. 309–325. ACM (2012)
5. Brakerski, Z., Vaikuntanathan, V.: Efficient fully homomorphic encryption from (standard) LWE. In: Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, pp. 97–106. IEEE Computer Society (2011)
6. Brakerski, Z., Vaikuntanathan, V.: Fully homomorphic encryption from Ring-LWE and security for key dependent messages. In: Rogaway, P. (ed.) CRYPTO 2011. LNCS, vol. 6841, pp. 505–524. Springer, Heidelberg (2011)
7. Brenner, M., Perl, H., Smith, M.: Implementation of the fully homomorphic encryption schemes of Gentry and Smart and Vercauteren, `https://hcrypt.com/`

[4] Notice that these circuits are composed of three phases: top linear transformations, shared non-linear component, and bottom linear transformations.

8. Cheon, J.H., Coron, J.-S., Kim, J., Lee, M.S., Lepoint, T., Tibouchi, M., Yun, A.: Batch fully homomorphic encryption over the integers. In: Johansson, T., Nguyen, P.Q. (eds.) EUROCRYPT 2013. LNCS, vol. 7881, pp. 315–335. Springer, Heidelberg (2013)
9. Coron, J.-S., Mandal, A., Naccache, D., Tibouchi, M.: Fully homomorphic encryption over the integers with shorter public keys. In: Rogaway, P. (ed.) CRYPTO 2011. LNCS, vol. 6841, pp. 487–504. Springer, Heidelberg (2011)
10. Coron, J.-S., Naccache, D., Tibouchi, M.: Public key compression and modulus switching for fully homomorphic encryption over the integers. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 446–464. Springer, Heidelberg (2012)
11. Coron, J.-S., Tibouchi, M.: Implementation of the fully homomorphic encryption scheme over the integers with compressed public keys in sage, `https://github.com/coron/fhe`
12. van Dijk, M., Gentry, C., Halevi, S., Vaikuntanathan, V.: Fully homomorphic encryption over the integers. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 24–43. Springer, Heidelberg (2010)
13. Gentry, C.: A fully homomorphic encryption scheme. PhD thesis, Stanford University (2009), `http://crypto.stanford.edu/craig`
14. Gentry, C., Halevi, S.: Implementing Gentry's fully-homomorphic encryption scheme. In: Paterson, K.G. (ed.) EUROCRYPT 2011. LNCS, vol. 6632, pp. 129–148. Springer, Heidelberg (2011)
15. Gentry, C., Halevi, S., Smart, N.P.: Fully homomorphic encryption with polylog overhead. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 465–482. Springer, Heidelberg (2012)
16. Gentry, C., Halevi, S., Smart, N.P.: Homomorphic evaluation of the AES circuit. In: Safavi-Naini, R., Canetti, R. (eds.) CRYPTO 2012. LNCS, vol. 7417, pp. 850–867. Springer, Heidelberg (2012)
17. Goldsmith, J., Hagen, M., Mundhenk, M.: Complexity of DNF and isomorphism of monotone formulas. In: Jedrzejowicz, J., Szepietowski, A. (eds.) MFCS 2005. LNCS, vol. 3618, pp. 410–421. Springer, Heidelberg (2005)
18. López-Alt, A., Tromer, E., Vaikuntanathan, V.: On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In: Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, pp. 1219–1234. ACM (2012)
19. Smart, N.P., Tillich, S.: Circuits of basic functions suitable for MPC and FHE, `http://www.cs.bris.ac.uk/Research/CryptographySecurity/MPC/`
20. Smart, N.P., Vercauteren, F.: Fully homomorphic encryption with relatively small key and ciphertext sizes. In: Nguyen, P.Q., Pointcheval, D. (eds.) PKC 2010. LNCS, vol. 6056, pp. 420–443. Springer, Heidelberg (2010)

Privacy Preserving Data Processing with Collaboration of Homomorphic Cryptosystems

Shigeo Tsujii[1], Hiroshi Doi[2], Ryo Fujita[1], Masahito Gotaishi[1],
Yukiyasu Tsunoo[3], and Takahiko Syouji[4]

[1] Research and Development Initiative, Chuo University
1–13–27 Kasuga, Bunkyo-ku, Tokyo, 112–8551 Japan
[2] Institute of Information Security
2–14–1 Tsuruya-cho, Kanagawa-ku, Yokohama, 221–0835 Japan
[3] Knowledge Discovery Research Laboratories, NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, 211–8666 Japan
[4] YDK Technologies Company
1288 Oshitate, Inagi-shi, Tokyo, 206–0811 Japan

Abstract. We propose a privacy-preserving data processing system using homomorphic cryptosystem. Proposed system consists of several functionalities corresponding to addition and multiplication of plaintexts encrypted in ciphertexts. Using these functionalities repeatedly, any multivariate polynomial evaluation of secret inputs can be achieved. We clarify the role and the function of each organization participating in the process — custodians of personal data, processing center of cryptographic function, and computing center. The cooperation of several entities makes arbitrary times of the calculations, which is a requirement of fully homomorphic encryption, more efficient. We give security proofs of the scheme and show the result of implementation of the scheme.

Keywords: e-administration, healthcare network, personal data protection, homomorphic cryptosystem.

1 Introduction

1.1 Background

Systems such as e-administration or medical information system have to read personal data. Although utilization of personal data generates many benefits for society there is an undeniable risk that the personal data stored in the systems might be abused by operators or system administrators. Against this background, technological demands for preventing these abuses have been growing. "Homomorphic system," which enables it, has been actively developed. That trend would have been triggered by Gentry [5], who discussed the possibility of fully homomorphic cryptosystem. With the "cloud computing," enabling to share wide range of data and to utilize them among multiple entities, the demand for performing arithmetic operation on encrypted personal data (hereinafter referred to as 'encrypted data processing'), which enables "to allow utilization of

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 201–212, 2013.

the data without disclosing them" is rapidly growing. Besides the one which was proposed by Gentry, several other encrypted data processing systems including multi-party protocol have been actively developed by various researchers. Nonetheless, existing systems would still be in the theoretical stage, because of some challenges to overcome, including efficiency and resource requirement [10,6,7,2].

In our study, a practical encrypted data processing system is proposed, using conventional homomorphic cryptosystems such as Paillier [8,9,3] and a strict but still practically feasible role-based access control. The roles and the functions of these entities are as follows:

1. Custodians of personal data (Custodian) do not disclose the data which they are responsible for. Therefore they are responsible for encrypting/masking the specified data.
2. Processing Center of Cryptographic Function (PCCF) has two divisions. Random Number Generation (RNG) division tells which data is necessary and gives random numbers to mask each of them. It knows which random number is used to mask which data, but cannot access any of the masked/encrypted data. Decryption (DEC) division is responsible for decrypting the cipher data which is sent from CC. They know the encrypted result and every secret key to decrypt the result (result of operating encrypted data), but cannot access any of the encrypted raw data.
3. Computing Center (CC) performs the calculations which are requested by the client. It has full access to the encrypted/masked data but cannot access to any other data including secret keys.

1.2 Our Contributions

Our goal is computing $f(\boldsymbol{M})$ where f is a function composed of addition and multiplication, and $\boldsymbol{M} = (M_1, \ldots, M_j)$ is messages. Similar to the multi-party protocol, this computation is executed while $\boldsymbol{M}$ is kept secret. In order to compute $f(\boldsymbol{M})$, the entity is divided into Custodian, PCCF, and CC, computes cooperatively, and gets the final result. Also, computing $f(\boldsymbol{M})$ (e.g. $f(\boldsymbol{M}) = f_1(\boldsymbol{M})\cdot f_2(\boldsymbol{M})+f_3(\boldsymbol{M})$) is divided into several additions and multiplications, and the computation is executed by the step-by-step approach. Then, encryption of the value of partial polynomial function f_i of f is computed, and their combination gives encryption of the final result of f. $f_i(\boldsymbol{M})$ and its encryption $E(f_i(\boldsymbol{M}))$ do not appear throughout the process, and only encryption $E(\alpha \cdot f_i(\boldsymbol{M}))$ of the value $\alpha \cdot f_i(\boldsymbol{M})$ masked by a random number α only appears. Therefore, the entity with the decryption key are prohibited to access the encrypted data and the ones with the access to the encrypted data do not have the decryption key, thereby enabling the processing of data without giving information of the plaintext $\boldsymbol{M}$ or the values of partial polynomial functions at all. While our proposed scheme requires different random numbers with respect to each computation of $f(\boldsymbol{M})$ every time, re-encryption step or refreshing ciphertexts, which are applied to the fully homomorphic encryption schemes, are not needed.

2 Preliminary

2.1 Homomorphic Public Key Encryption

Homomorphic public key encryption plays an important role in our proposed scheme. First, we introduce a notion of public key cryptosystem and its security.

Definition 1 (Public Key Encryption). *A public key encryption (PKE) scheme $\mathcal{E} = (G, E, D)$ consists of three algorithms.*

- *The key generation algorithm G accepts the security parameter λ in unary form and outputs the plaintext space $\mathcal{M}$, the ciphertext space $\mathcal{C}$, a pair (PK, SK), the public key and secret key for E. That is $(PK, SK) \leftarrow G(1^\lambda)$.*
- *The encryption algorithm E takes as input a public key PK and a message M and outputs a ciphertext C. That is, $C \leftarrow E(PK, M)$.*
- *The decryption algorithm D takes as input a secret key SK and a ciphertext C and outputs a message M. That is $M \leftarrow D(SK, C)$.*

The correctness property of the PKE scheme $\mathcal{E}$ is defined as:

$$\forall (PK, SK) \leftarrow G(1^\lambda), \quad \forall M \in \mathcal{M} : D(SK, E(PK, M)) = M.$$

The most basic notion of security for a PKE scheme $\mathcal{E}$ is indistinguishability. The advantage of the adversary $\mathcal{A}_{\mathsf{PKE}}$ against an encryption scheme $\mathcal{E}$ is defined as follows:

$$Adv_{\mathsf{PKE}}[\mathcal{E}] \overset{\text{def}}{=} \left| \Pr[\mathcal{A}_{\mathsf{PKE}}(M_0, M_1, c) = \bar{b} \mid \bar{b} \overset{R}{\leftarrow} \{0,1\}, c = E(M_{\bar{b}})] - \frac{1}{2} \right|$$

The encryption scheme $\mathcal{E}$ has indistinguishable property if for all probabilistic polynomial time adversaries $\mathcal{A}_{\mathsf{PKE}}$, the advantage $Adv_{\mathsf{PKE}}[\mathcal{E}]$ is negligible.

We give the definition of homomorphic PKE (e.g. [4], [1], [8], [9]) as follows.

Definition 2 (Homomorphic PKE). *A homomorphic public key encryption scheme $\mathcal{E} = (G, E, D)$ is defined by three algorithms as in Definition 1, together with at least one pair of operations $\langle +, \cdot \rangle$ on the plaintext space $\mathcal{M}$ and the ciphertext space $\mathcal{C}$, respectively. The encryption algorithm E is a "homomorphism" between the plaintext space $\mathcal{M}$ and the ciphertext space $\mathcal{C}$ if $M_1 + M_2 = D(SK, C_1 \cdot C_2)$ where $C_1 = E(PK, M_1)$ and $C_2 = E(PK, M_2)$ for arbitrarily $M_1, M_2 \in \mathcal{M}$.*

2.2 Multivariate Polynomial

The inputs of our proposed scheme are values of each variable in a multivariate polynomial, and the output can be directly computed by polynomial evaluation at the given values without any cryptographic schemes. Multivariate polynomial f is generally expressed as sum of several monomials. However, it is not always efficient while the number of terms in f increases exponentially as total degree of polynomial increases. On the other hand, f may have very simple expression.

For example, if f can be decomposable into some factors of polynomials, we should adopt this type of expression of f. In this case, in order to evaluate the polynomial, by using this expression, fast and efficient computation is achieved, while somewhat complicated computation (addition and multiplication of polynomials) is required than in the case of general expression.

3 Model

Fig. 1 illustrates the model of our proposed scheme. Hereafter, we explain the components of the model. We assume that secure channels, for example, which use SSL, are used in our model.

3.1 Entities

In our proposed scheme, there are four entities: client, Custodians of the sensitive data, processing center of cryptographic function (PCCF), and computing center (CC). PCCF has two divisions: random number generation and decryption. We assume that these entities are passive adversary and do not collude with each other. Every entities is obliged to follow the protocol given as follows, but cannot intentionally forget knowledge that it learns during the execution of the protocol. The roles and functions of client, Custodian, PCCF, and CC are as follows:

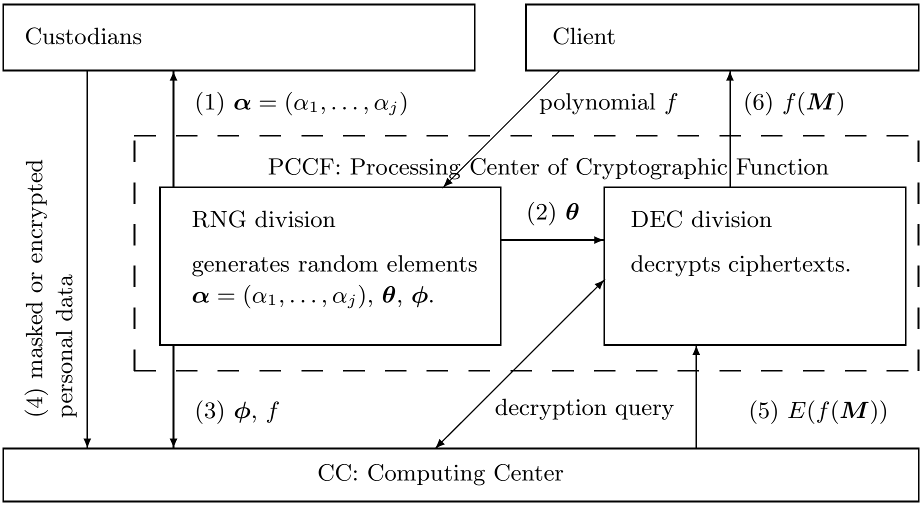

Custodians
Client
(1) $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_j)$
polynomial f
(6) $f(\boldsymbol{M})$
PCCF: Processing Center of Cryptographic Function
RNG division
generates random elements $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_j)$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$.
(2) $\boldsymbol{\theta}$
DEC division
decrypts ciphertexts.
(4) masked or encrypted personal data
(3) $\boldsymbol{\phi}$, f
decryption query
(5) $E(f(\boldsymbol{M}))$
CC: Computing Center

Fig. 1. System Construction Overcoming Contradiction between the Protection and Utilization of Personal Data

Client. Client has an intention of computing some statistical information from personal data. For obtaining a specific value, such as by utilizing statistics and personal data, it defines calculation formula (multivariate polynomial whose variables are personal data M_i [1]). Client gets to know the final result only by the equation above, and cannot know the value of the middle course such sums and products of personal data.

Custodians of Sensitive Data. Each custodian keeps personal data secret (including the ability to deposit in the cloud), and sends only masked/encrypted data to CC. Each custodian does not transmit and receive personal data mutually even it is encrypted. An example in the real society is a hospital containing charts of patients.

PCCF: Processing Center of Cryptographic Function. PCCF has two divisions: random number generation (RNG) and decryption (DEC). Their divisions both cannot know plaintext (e.g. personal data).

RNG division generates random numbers $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and sends to other entities. RNG division requests CC to compute specified function f on encrypted data.

DEC division has the secret key of the cryptosystem used in the scheme. It receives the encrypted data from CC and once computes the decrypted value. Then, it randomizes the value by the random element given from the RNG division. After that, it sends the randomized value back to CC. It also decrypts the encryption of the final result sent from CC, and sends the decrypted data to the client.

CC: Computing Center. CC computes on encrypted data without plaintext (e.g. personal data). CC receives ciphertexts, calculation formula and auxiliary ciphertexts and computes on encrypted data as promotion center operates. CC computes addition and multiplication of the encrypted data, and calls the DEC division to decrypt some masked data as required. After all these computations, CC sends the encryption of the final result to the DEC division.

4 Functionalities of Proposed System and Their Security

Proposed system consists of several functionalities corresponding addition and multiplication of plaintexts encrypted in ciphertexts. Using these functionalities repeatedly, any multivariate polynomial evaluation of secret inputs can be achieved.

In this paper, we denote by $\mathcal{M}$, the plaintext space and by $\mathcal{C}$, the ciphertext space. We assume that $\mathcal{M}$ is a ring where an addition and a multiplication are

[1] For practical purposes, personal data is expressed as M_{ij}, where i corresponds to personal ID, j corresponds to the type such as pension, income etc. For simplicity, we denote the personal data by M_i throughout this paper.

defined. On the other hand, only a multiplication is defined in $\mathcal{C}$. In our construction, $\mathcal{M} = \mathbb{Z}_n$ where $n = pq$, both p, q are large secret primes. Furthermore, we use a homomorphic public key encryption (G, E, D) defined in section 2.1.

4.1 Masked Plaintext

Given a plaintext $M \in \mathcal{M}$ and a random element $\alpha \in \mathcal{M}$, we denote by $\alpha \cdot M$, the masked plaintext of M (masked by α).

If masked plaintext $C = \alpha \cdot M$ is invertible i.e. $C \in \mathbb{Z}_n^*$, C has no information of the plaintext $M \in \mathbb{Z}_n^*$. That is, given $C \in \mathbb{Z}_n^*$, there exists an element $\alpha \in \mathbb{Z}_n^*$ which satisfies $C = \alpha \cdot M$ for arbitrarily $M \in \mathbb{Z}_n^*$.

4.2 Functionalities

Our scheme consists of four functionalities AMC, ACC, MMC, and MCC as follows.

Addition from Masked Plaintexts to Ciphertext (AMC) on input two masked plaintexts of M_1, M_2 and two auxiliary ciphertexts, outputs a ciphertext of the masked plaintext of $M_1 + M_2$. In our construction,

$$\left(\alpha_1 M_1, \alpha_2 M_2, E\left(\frac{\gamma_1}{\alpha_1}\right), E\left(\frac{\gamma_1}{\alpha_2}\right)\right) \longmapsto E(\gamma_1(M_1 + M_2))$$

is realized by multiplications of ciphertext space, i.e. $E\left(\frac{\gamma_1}{\alpha_1}\right)^{\alpha_1 M_1} \cdot E\left(\frac{\gamma_1}{\alpha_2}\right)^{\alpha_2 M_2}$.

Addition from Ciphertext to Ciphertext (ACC) on input two ciphertexts of the masked plaintexts of M_1, M_2 and two auxiliary elements in $\mathcal{M}$, outputs a ciphertext of the masked plaintext of $M_1 + M_2$. In our construction,

$$\left(E(\beta_1 M_1), E(\beta_2 M_2), \frac{\gamma_2}{\beta_1}, \frac{\gamma_2}{\beta_2}\right) \longmapsto E(\gamma_2(M_1 + M_2))$$

is realized by multiplications of ciphertext space, i.e. $E(\beta_1 M_1)^{\frac{\gamma_2}{\beta_1}} \cdot E(\beta_2 M_2)^{\frac{\gamma_2}{\beta_2}}$.

Multiplication from Masked Plaintext to Ciphertext (MMC) on input two masked plaintexts of M_1, M_2 and an auxiliary ciphertext, outputs a ciphertext of the masked plaintext of $M_1 \cdot M_2$. In our construction,

$$\left(\alpha_1 M_1, \alpha_2 M_2, E\left(\frac{\delta_1}{\alpha_1 \alpha_2}\right)\right) \longmapsto E(\delta_1 M_1 M_2)$$

is realized by multiplications of ciphertext space, i.e. $E\left(\frac{\delta_1}{\alpha_1 \alpha_2}\right)^{\alpha_1 M_1 \cdot \alpha_2 M_2}$.

Multiplication from Ciphertext to Ciphertext (MCC) is two party protocol between CC and DEC division as follows. At the beginning of the protocol, CC has two ciphertexts of masked plaintexts of M_1, M_2 and an auxiliary ciphertext. On the other hand, DEC division has two auxiliary elements. After the execution of the protocol, CC outputs the ciphertext of masked plaintext of $M_1 \cdot M_2$. In our construction, two party protocol is realized as follows.

1. The inputs of CC are $E(\beta_1 M_1), E(\beta_2 M_2)$, and auxiliary ciphertext $E\left(\dfrac{\delta_2}{\beta_1\beta_2\theta_1\theta_2}\right)$.
2. The inputs of DEC division are $\theta_1, \theta_2 \in \mathcal{M}$.
3. CC sends $E(\beta_1 M_1), E(\beta_2 M_2)$ to DEC division.
4. DEC division obtains $\beta_1 M_1$ and $\beta_2 M_2$ by decrypting $E(\beta_1 M_1), E(\beta_2 M_2)$.
5. DEC division computes $\theta_1\beta_1 M_1, \theta_2\beta_2 M_2$ using $\beta_1 M_1, \beta_2 M_2$, and auxiliary elements θ_1, θ_2.
6. DEC divisions sends $\theta_1\beta_1 M_1, \theta_2\beta_2 M_2$ to CC.
7. CC computes $E(\delta_2 M_1 M_2) = E\left(\dfrac{\delta_2}{\beta_1\beta_2\theta_1\theta_2}\right)^{\theta_1\beta_1 M_1 \cdot \theta_2\beta_2 M_2}$.

4.3 Definitions of Security

We give the notion of security of $\mathsf{AMC}, \mathsf{ACC}, \mathsf{MMC}$, and MCC. In our model, we assume all the communication channels are secure, and any two entities do not collude. Because of the strong assumptions of our model, the security of functionalities can be defined simply (almost same of the security definition of public key encryption).

Definition 3. *The advantage of any probabilistic polynomial time algorithm $\mathcal{A}_{\mathsf{AMC}}$ against AMC is defined by*

$$Adv_{\mathsf{AMC}} = \Big|\Pr\Big[\mathcal{A}_{\mathsf{AMC}}\left(M_{0,1}, M_{0,2}, M_{1,1}, M_{1,2}, c_1, c_2, c_3, c_4\right) = b \mid b \xleftarrow{R} \{0,1\}, \\ \alpha_1, \alpha_2, \gamma_1 \xleftarrow{R} \mathcal{M},\ c_1 \leftarrow \alpha_1 M_{b,1}, c_2 \leftarrow \alpha_2 M_{b,2}, \\ c_3 \leftarrow E\left(\frac{\gamma_1}{\alpha_1}\right), c_4 \leftarrow E\left(\frac{\gamma_1}{\alpha_2}\right)\Big] - \frac{1}{2}\Big|. \tag{1}$$

AMC *is secure if Adv_{AMC} is negligible.*

Definition 4. *The advantage of any probabilistic polynomial time algorithm $\mathcal{A}_{\mathsf{ACC}}$ against ACC is defined by*

$$Adv_{\mathsf{ACC}} = \Big|\Pr\Big[\mathcal{A}_{\mathsf{ACC}}\left(M_{0,1}, M_{0,2}, M_{1,1}, M_{1,2}, c_1, c_2, c_3, c_4\right) = b \mid b \xleftarrow{R} \{0,1\}, \\ \beta_1, \beta_2, \gamma_2 \xleftarrow{R} \mathcal{M},\ c_1 \leftarrow E(\beta_1 M_{b,1}), c_2 \leftarrow E(\beta_2 M_{b,2}), \\ c_3 \leftarrow \frac{\gamma_2}{\beta_1}, c_4 \leftarrow \frac{\gamma_2}{\beta_2}\Big] - \frac{1}{2}\Big|. \tag{2}$$

ACC *is secure if Adv_{ACC} is negligible.*

Definition 5. *The advantage of any probabilistic polynomial time algorithm* $\mathcal{A}_{\mathsf{MMC}}$ *against* MMC *is defined by*

$$\begin{aligned}
Adv_{\mathsf{MMC}} = \Bigl| \Pr \bigl[& \mathcal{A}_{\mathsf{MMC}} \left(M_{0,1}, M_{0,2}, M_{1,1}, M_{1,2}, c_1, c_2, c_3 \right) = b \mid b \overset{R}{\leftarrow} \{0,1\}, \\
& \alpha_1, \alpha_2, \delta_1 \overset{R}{\leftarrow} \mathcal{M},\ c_1 \leftarrow \alpha_1 M_{b,1}, c_2 \leftarrow \alpha_2 M_{b,2}, \\
& c_3 \leftarrow E\left(\frac{\delta_1}{\alpha_1 \alpha_2} \right) \bigr] - \frac{1}{2} \Bigr| . \qquad (3)
\end{aligned}$$

MMC *is secure if* Adv_{MMC} *is negligible.*

Definition 6. *The advantage of any probabilistic polynomial time algorithm* $\mathcal{A}_{\mathsf{MCC,CR}}$ *in CC against* MCC *is defined by*

$$\begin{aligned}
Adv_{\mathsf{MCC,CR}} = \Bigl| \Pr \bigl[& \mathcal{A}_{\mathsf{MCC,CR}} \left(M_{0,1}, M_{0,2}, M_{1,1}, M_{1,2}, c_1, c_2, c_3, c_4, c_5 \right) = b \mid b \overset{R}{\leftarrow} \{0,1\}, \\
& \beta_1, \beta_2, \delta_2, \theta_1, \theta_2 \overset{R}{\leftarrow} \mathcal{M},\ c_1 \leftarrow E(\beta_1 M_{b,1}), c_2 \leftarrow E(\beta_2 M_{b,2}), \\
& c_3 \leftarrow \theta_1 \beta_1 M_{b,1},\ c_4 \leftarrow \theta_2 \beta_2 M_{b,2},\ c_5 \leftarrow E\left(\frac{\delta_2}{\beta_1 \beta_2 \theta_1 \theta_2} \right) \bigr] - \frac{1}{2} \Bigr| . \quad (4)
\end{aligned}$$

The advantage of any probabilistic polynomial time algorithm $\mathcal{A}_{\mathsf{MCC,Dd}}$ *in DEC division against* MCC *is defined by*

$$\begin{aligned}
Adv_{\mathsf{MCC,Dd}} = \Bigl| \Pr \bigl[& \mathcal{A}_{\mathsf{MCC,Dd}} \left(SK, M_{0,1}, M_{0,2}, M_{1,1}, M_{1,2}, c_1, c_2, \theta_1, \theta_2 \right) = b \mid b \overset{R}{\leftarrow} \{0,1\}, \\
& \beta_1, \beta_2, \theta_1, \theta_2 \overset{R}{\leftarrow} \mathcal{M},\ c_1 \leftarrow E(\beta_1 M_{b,1}), c_2 \leftarrow E(\beta_2 M_{b,2}) \\
& \bigr] - \frac{1}{2} \Bigr| . \qquad (5)
\end{aligned}$$

MCC *is secure if both* $Adv_{\mathsf{MCC,CR}}$ *and* $Adv_{\mathsf{MCC,Dd}}$ *are negligible.*

Definition 7. *Proposed scheme is secure if* AMC, ACC, MAC, *and* MCC *are secure.*

5 Proposed Scheme

Based on the model, functionalities given in previous sections, we explain our proposed scheme using Fig. 2.

Decision of Polynomial. The client decides the multivariate function f and sends it to the proposed system. Note that f is not secret.

(1)-(3) Random Number Generation. RNG division on inputs f, outputs random elements, auxiliary elements, and auxiliary ciphertexts to Custodian, DEC division, CC respectively. Note that the random elements and the auxiliary elements are nonce with respect to each computation of f.

(4) Masking/Encryption. Custodians computes masked plaintexts and ciphertexts. They send them to the CC. After sending them, Custodians do not need to participate in the computation.

Computation in CC CC calculates AMC, ACC, and MMC.

Protocols between DEC and CC. DEC division and CC calculate MCC cooperatively.

(5)-(6) Final Decryption. CC sends $E(f(\mathbf{M}))$ to DEC, and DEC sends the decryption results to Client.

Note that the RNG division coordinate the random elements to get $f(\mathbf{M})$ (i.e. the final masked element is 1). On the other hand, all the inputs/outputs of each entities are masked plaintexts using random elements, or their ciphertexts, any entities cannot know the intermediate computation results without masked element.

6 Security Proofs

Theorem 1. *If the homomorphic cryptosystem used in proposed scheme is indistinguishable against Chosen Plaintext Attack, the proposed scheme is secure.*

From the definition 7, it is sufficient to show that AMC, ACC, MMC, and MCC are secure. From the limitation of pages, we give the proof that Adv_{AMC} is negligible. Other proofs will be given in full version of this paper.

Proof. First, we construct an algorithm $\mathcal{B}$ that breaks indistinguishability of PKE (see section 2.1) using an adversary $\mathcal{A}_{\mathsf{AMC}}$ who breaks security of AMC of our scheme with the non-negligible probability. First, $\mathcal{C}$, the challenger of security game of PKE, sends PK to $\mathcal{B}$. $\mathcal{B}$ forwards PK to $\mathcal{A}_{\mathsf{AMC}}$. $\mathcal{A}_{\mathsf{AMC}}$ sends $(M_{0,1}, M_{0,2}, M_{1,1}, M_{1,2})$ to $\mathcal{B}$ as the challenge message. $\mathcal{B}$ chooses $\alpha_1, \alpha_2, \gamma_1 \overset{R}{\leftarrow} \mathcal{M}$ and sends $\left(\frac{\gamma_1}{\alpha_1}, \frac{\gamma_1}{\alpha_2}\right)$ to $\mathcal{C}$ as the challenge message of PKE game. $\mathcal{B}$ obtains the challenge ciphertext $C = E\left(\frac{\gamma_1}{\alpha_{1+\bar{b}}}\right)$ from $\mathcal{C}$. $\mathcal{B}$ chooses $\beta \overset{R}{\leftarrow} \{0,1\}$, and return

$$(c_1, c_2, c_3, c_4) = \left(\alpha_1 M_{\beta,1}, \alpha_2 M_{\beta,2}, C, E\left(\frac{\gamma_1}{\alpha_2}\right)\right) \quad \text{if} \quad \beta = 0,$$
$$(c_1, c_2, c_3, c_4) = \left(\alpha_1 M_{\beta,1}, \alpha_2 M_{\beta,2}, E\left(\frac{\gamma_1}{\alpha_1}\right), C\right) \quad \text{if} \quad \beta = 1,$$

to $\mathcal{A}_{\mathsf{AMC}}$. After obtaining b as the guess of $\mathcal{A}_{\mathsf{AMC}}$, $\mathcal{B}$ outputs b. If $\beta = \bar{b}$, the random coin of the challenger of PKE game, $\mathcal{B}$'s simulation is perfect. Otherwise, the challenge ciphertext (c_1, c_2, c_3, c_4) has no information of message because of the randomness of $\alpha_1, \alpha_2, \gamma_1$. Therefore, $Adv_{\mathsf{PKE}}[\mathcal{E}] = \frac{1}{2} Adv_{\mathsf{AMC}}$.

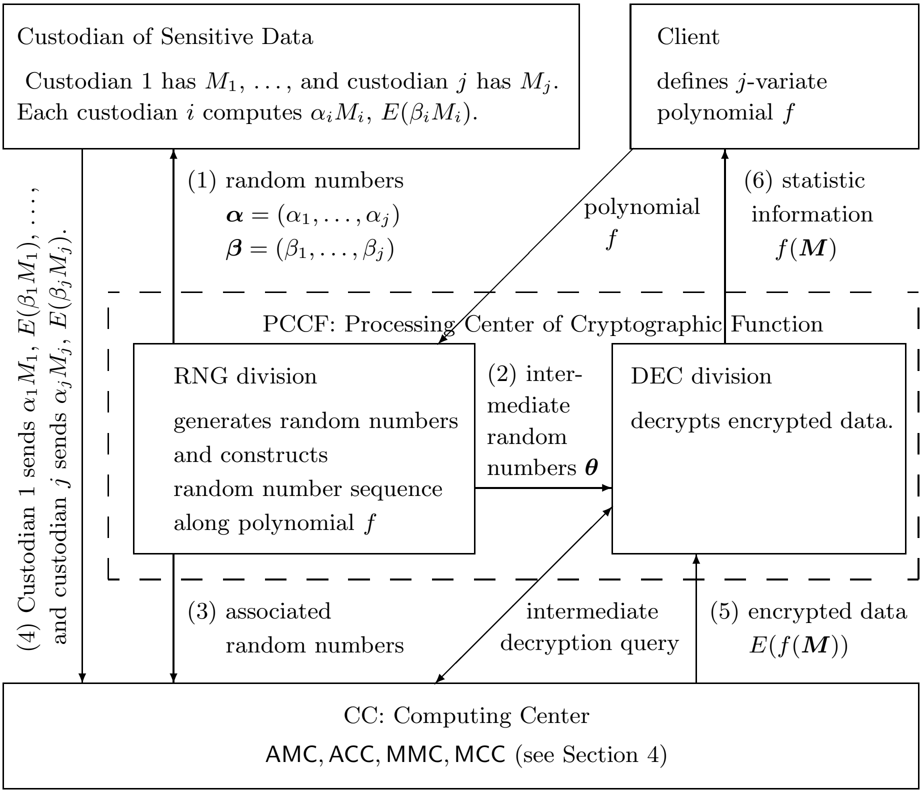

Custodian of Sensitive Data
Custodian 1 has M_1, ..., and custodian j has M_j. Each custodian i computes $\alpha_i M_i$, $E(\beta_i M_i)$.
Client
defines j-variate polynomial f
(1) random numbers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_j)$ $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j)$
polynomial f
(6) statistic information $f(\boldsymbol{M})$
(4) Custodian 1 sends $\alpha_1 M_1$, $E(\beta_1 M_1)$, ..., and custodian j sends $\alpha_j M_j$, $E(\beta_j M_j)$.
PCCF: Processing Center of Cryptographic Function
RNG division
generates random numbers and constructs random number sequence along polynomial f
(2) intermediate random numbers $\boldsymbol{\theta}$
DEC division
decrypts encrypted data.
(3) associated random numbers
intermediate decryption query
(5) encrypted data $E(f(\boldsymbol{M}))$
CC: Computing Center
AMC, ACC, MMC, MCC (see Section 4)

Fig. 2. General Form of System Construction to Enhance Compatibility between the Protection and Utilization of Personal Data

7 Performance of Implementation

We confirmed that whether our proposed scheme works in realistic time with computational experiments. Table 1 shows a result of our implementation of the proposed scheme with Paillier cryptosystem [9]. Any homomorphic public key encryption scheme is applicable to our scheme, and Paillier cryptosystem has large plaintext space while that space is one bit in many practical fully homomorphic encryption schemes. The computation times are evaluated on PC with Intel Core i3 at 3GHz and 4GB of RAM. In our setting, each party has two data $M_{1,i}$ and $M_{2,i}$ for $i = 1, \ldots, j$. Utilizing the proposed scheme, we computed average, variance and covariance of these data. In order to compute these values, the value of multivariate polynomials $f_1 = \sum_{i=1}^{j} M_{1,i}$, $f_2 = \sum_{i=1}^{j} M_{2,i}$, $f_3 = \sum_{i=1}^{j} M_{1,i}^2$, $f_4 = \sum_{i=1}^{j} M_{2,i}^2$, and $f_5 = \sum_{i=1}^{j} M_{1,i} M_{2,i}$ are required. Table 1 shows that the

Table 1. Implementation Results of Proposed Scheme

	Encryption Scheme:			
	1024 bits Paillier Cryptosystem		2048 bits Paillier Cryptosystem	
	Parameters:		Parameters:	
	$j = 20$	$j = 100$	$j = 20$	$j = 100$
random number generation: (1)(2)	0.15 sec.	0.67 sec.	0.16 sec.	0.74 sec.
encryption: (3)(4)	18.8 sec.	93.7 sec.	141 sec.	722 sec.
computing $\mathsf{AMC}, \mathsf{ACC}, \mathsf{MMC}, \mathsf{MCC}$	9.95 sec.	51.3 sec.	69 sec.	365 sec.
computing $f_1, \ldots, f_5$: decryption of (5)	0.33 sec.	1.40 sec.	2.22 sec.	9.63 sec.
computing statistical values: (6)	$<$ 0.01 sec.	$<$ 0.01 sec.	$<$ 0.01 sec.	$<$ 0.01 sec.

statistical processing on encrypted data using our proposed is feasible. The total computation time is less than 20 minutes even in the case of the number j of personal data is 100. We remark that the times of the encryption (3)(4) and the computing $\mathsf{AMC}, \mathsf{ACC}, \mathsf{MMC}, \mathsf{MCC}$ are higher estimates and reducible when they are performed in parallel. These computations are performed by each custodian or the computing center, which have enormous computational resource. Since the computation time of our scheme does not depend on the bound on the maximum number of multiplications, our scheme enables the other kind of computation than the statistical processing as we computed. Comparison with other schemes is a subject of future investigation.

8 Concluding Remarks

We proposed a system with collaboration of homomorphic cryptosystems. Using several functionalities defined on this paper enables us to compute multivariate polynomial evaluation of secret inputs. Although our proposed scheme requires slightly strong assumptions on the entities, the security definitions are very simple and security proofs of the scheme is given in an intuitive manner. Composing a scheme based on weaker assumptions on communication channels and the entities is an issue in the future. Formal proofs of the security of the whole protocol is also our future study. Towards the future practical implementation, we will be urged the consideration on the system management scheme.

Acknowledgement. This work is supported by the Ministry of Economy, Trade, and Industry (METI) of Japanese Government. The authors would like to thank the anonymous reviewers for helpful comments.

References

1. Benaloh, J.: Dense Probabilistic Encryption. In: Proceedings of the Workshop on Selected Areas of Cryptography (SAC 1994), pp. 120–128 (1994)
2. Cramer, R., Damgård, I., Nielsen, J.B.: Secure Multiparty Computation — Book Draft, `http://daimi.au.dk/~ivan/MPCbook.pdf`
3. Damgård, I., Jurik, M.: A generalisation, a simplification and some applications of Paillier's probabilistic public-key system. In: Kim, K.-C. (ed.) PKC 2001. LNCS, vol. 1992, pp. 119–136. Springer, Heidelberg (2001)
4. Elgamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory 31(4), 469–472 (1985)
5. Gentry, C.: A fully homomorphic encryption scheme. Ph. D. Thesis, Stanford University (2009)
6. Goldreich, O., Micali, S., Wigderson, A.: How to play any mental game or a completeness theorem for protocols with honest majority. In: Proc. STOC 1987, pp. 218–229 (1987)
7. Lauter, K., Naehrig, M., Vaikuntanathan, V.: Can homomorphic encryption be practical? In: Proc. CCSW 2011, pp. 113–124. ACM Press (2011)
8. Okamoto, T., Uchiyama, S.: A new public-key cryptosystem as secure as factoring. In: Nyberg, K. (ed.) EUROCRYPT 1998. LNCS, vol. 1403, pp. 308–318. Springer, Heidelberg (1998)
9. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
10. Yao, A.C.-C.: How to generate and exchange secrets (extended abstract). In: Proc. FOCS 1986, pp. 162–167 (1986)

Parallel Homomorphic Encryption

Seny Kamara[1] and Mariana Raykova[2,⋆]

[1] Microsoft Research
[2] IBM Research

Abstract. In the problem of private outsourced computation, a client wishes to delegate the evaluation of a function f on a private input x to an untrusted worker without the latter learning anything about x and $f(x)$. This problem occurs in many applications and, most notably, in the setting of cloud computing.

In this work, we consider the problem of privately outsourcing computation to a *cluster* of machines, which typically happens when the computation needs to be performed over massive datasets, e.g., to analyze large social networks or train machine learning algorithms on large corpora. At such scales, computation is beyond the capabilities of any single machine so it is performed by large-scale clusters of workers.

To address this problem, we consider *parallel* homomorphic encryption (PHE) schemes, which are encryption schemes that support computation over encrypted data through the use of an evaluation algorithm that can be efficiently executed in parallel. More concretely, we focus on the MapReduce model of parallel computation and show how to construct PHE schemes that can support various MapReduce operations on encrypted datasets including element testing and keyword search. More generally, we construct schemes that can support the evaluation of functions in NC^0 with locality 1 and $\mathsf{polylog}(k)$ (where k is the security parameter).

Underlying our PHE schemes are two new constructions of (local) randomized reductions (Beaver and Feigenbaum, *STACS* $'90$) for univariate and multivariate polynomials. Unlike previous constructions, our reductions are not based on secret sharing and are *fully-hiding* in the sense that the privacy of the input is guaranteed even if the adversary sees *all* the client's queries.

Our randomized reduction for univariate polynomials is information-theoretically secure and is based on permutation polynomials, whereas our reduction for multivariate polynomials is computationally-secure under the multi-dimensional noisy curve reconstruction assumption (Ishai, Kushilevitz, Ostrovsky, Sahai, *FOCS '06*).

1 Introduction

In the problem of private outsourced computation, a client wishes to delegate the evaluation of a function f on a private input x to an untrusted worker without the latter learning anything about x and $f(x)$. This problem occurs in many applications and, most notably, in the setting of cloud computing, where a provider makes its computational resources available to clients "as a service".

[⋆] Supported by NSF Grant No.1017660. Work done while at Microsoft Research.

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 213–225, 2013.

One approach to this problem is via the use of homomorphic encryption (HE). An encryption scheme is homomorphic if it supports computation on encrypted data, i.e., in addition to the standard encryption and decryption algorithms it also has an evaluation algorithm that takes as input an encryption of some message x and a function f and returns an encryption of $f(x)$. If a HE scheme supports both addition and multiplication, then it can evaluate any arithmetic circuit over encrypted data and we say that it is a fully homomorphic encryption (FHE) scheme [10].

The problem of outsourced computation occurs in various forms. For instance, in addition to the simple client/worker setting described above, clients often wish to outsource their computation to *clusters* of workers. This typically occurs when the computation is to be performed over massive datasets, e.g., to analyze large social networks or train machine learning algorithms on large corpora. At such scales, computation is beyond the capabilities of any single machine so it is performed on clusters of machines, i.e., large-scale distributed systems often composed of low-cost unreliable commodity hardware. For our purposes, we will view such a cluster as a system composed of w workers and one controller. Given some input, the controller generates n jobs (where typically $n \gg w$) which it distributes to the workers. Each worker executes its job in parallel and returns some value to the controller who then decides whether to continue the computation or halt.

In this work, we consider the problem of *privately* outsourcing computation to a cluster of machines. To address this, we introduce *parallel* homomorphic encryption (PHE) schemes, which are encryption schemes that support computation over encrypted data through the use of an evaluation algorithm that can be efficiently executed in parallel. Using a PHE scheme, a client can outsource the evaluation of a function f on some private input x to a cluster of w machines as follows. The client encrypts x and sends the ciphertext and f to the controller. Using the ciphertext, the controller generates n jobs that it distributes to the workers and, as above, the workers execute their jobs in parallel. When the entire computation is finished, the client receives a ciphertext which it decrypts to recover $f(x)$.

Applications of PHE. As discussed above, the most immediate application of PHE is to the setting of outsourced computation where a weak computational device wishes to make use of the resources of a more powerful cluster. Clearly, to be useful in this setting it is crucial that either: (1) running the encryption and decryption operations of the PHE scheme take less time than evaluating f on the input x directly; or (2) the PHE scheme is *multi-use* in the sense that the evaluations of several (different) functions can be done on a single ciphertext (this is also referred to as the online/offline setting). In this work we focus on the latter and present several multi-use PHE schemes. Using our schemes a client can encrypt a large database during an offline phase and then, have the workers evaluate many different functions on its data during the online phase. In particular, at the time of encryption, the client does not need to know the functions it will want to evaluate during the online phase.

Parallel Computation. Most computations are not completely parallelizable and require some amount of communication between machines. The specifics of how the computation and communication between processors are organized leads to particular architectures, each having unique characteristics in terms of computational and communication complexity. This has motivated the design of several architecture-independent models of parallel computation, including NC circuits [5], the parallel RAM (PRAM) [8,14], Valiant's bulk synchronous parallel (BSP) model [18], LogP [6] and, more recently, the MapReduce [7] and Dryad models [12]. It follows that an important consideration in the design of PHE schemes is the parallel model in which the function will be evaluated. In this work, we focus on the MapReduce model (which we describe below) but note that our choice is due mainly to practical considerations (e.g., the emergence of cloud-based MapReduce services such as Amazon's Elastic MapReduce) and that PHE can also be considered with respect to other models of parallel computation. As an example, note that any FHE scheme yields an NC-parallel HE scheme for any function f in NC.

1.1 Overview of Techniques

Designing PHE Schemes. We propose a general approach to designing PHE schemes. Roughly speaking, our approach yields PHE schemes for any function f that can be randomly reduced to another function g. A randomized reduction (RR) [2,3] from a function f to a function g transforms an input x in the domain of f to a set of n inputs $S = (s_1, \ldots, s_n)$ in the domain of g such that $f(x)$ can be efficiently reconstructed from $(g(s_1), \ldots, g(s_n))$. In addition, a RR guarantees that no information about x or $f(x)$ can be recovered from any subset of $t \leq n$ elements of S.

A natural approach to constructing a PHE scheme (ignoring the particular model of parallel computation) is therefore to encrypt x by using a RR to transform it into a set $(s_1, \ldots, s_n)$ and have each worker i evaluate g on s_i independently. The results can then be sent back to the client who can recover $f(x)$ using the reduction's reconstruction algorithm. As long as at most t workers collude, the RR will guarantee the confidentiality of x and $f(x)$. Unfortunately, there are two problems with this approach. First, as far as we know, the best hiding threshold achieved by any RR is $t \leq (n-1)/q$, which is for univariate polynomials of degree q [2,3]. In the context of cloud computing, however, this is not a reasonable assumption as the cloud provider owns *all* the machines in the cluster. [1] Another limitation is that the client has to run the RR's reconstruction algorithm which can represent a non-trivial amount of work depending on the particular scheme and the parameters used.

We address these limitations in the following way. First, we show how to construct *fully-hiding* RRs, i.e., reductions with a hiding threshold of $t = n$. Our first construction is for the class of univariate polynomials while the second is for multivariate polynomials with a "small" (i.e., poly-logarithmic in the security parameter) number of variables. As far as we know, these are the first RRs to achieve a threshold of $t = n$. Towards handling the second limitation, we observe that if the recovery algorithm of the RR can be evaluated homomorphically, then the reconstruction step can also be outsourced to the workers. Clearly, using FHE any recovery algorithm can be outsourced, but our goal here is to avoid the use of FHE so as to have practical schemes. Our approach therefore

[1] Of course one could use the above approach with more than one cloud providers if they do not collude.

will be to design RRs with recovery algorithms that are either (1) simple enough to be evaluated without FHE; or (2) efficient enough to be run by the client. We note that in cases where the reconstruction algorithm can be outsourced to the workers, we can make use of RRs with reconstruction algorithms that are more expensive than evaluating $f(x)$ directly.

Designing Fully-Hiding RRs. The best known RRs for polynomials [2,3] work roughly as follows. Let $\mathbf{Q}$ be the polynomial of degree q that we wish to evaluate and $\mathbf{x} \in \mathbb{F}^m$ be the input. First, each element of $\mathbf{x}$ is shared into $q \cdot t + 1$ shares using Shamir secret sharing with a sharing polynomial of degree t (i.e., the hiding threshold). This yields m sets of shares $(\mathbf{s}_1, \ldots, \mathbf{s}_m)$, where $\mathbf{s}_i = (\mathbf{s}_i[1], \ldots, \mathbf{s}_i[q{\cdot}t{+}1])$. Each worker $j \in [q{\cdot}t{+}1]$ is then given $(\mathbf{s}_1[j], \ldots, \mathbf{s}_m[j])$ and evaluates $\mathbf{Q}$ on his shares. Given the results of all these evaluations, the client interpolates at 0 to recover $\mathbf{Q}(\mathbf{x})$. This approach yields a hiding threshold of up to $t = (n-1)/q$. Note that this construction works equally as well for $m = 1$. As shown in [2,3], this can be improved to $t = n \cdot c \log(m)/m$ for any constant $c > 0$ and $m > 1$.

Due to their reliance on secret sharing, it is not clear how to extend the techniques from [2,3] to achieve $t = n$ and (informally) it seems hard to imagine using any technique based on secret sharing to achieve full hiding. Instead, we introduce two new techniques for designing RRs. The first works for univariate polynomials and makes use of permutation polynomials over finite fields (i.e., bijective families of polynomials). The resulting RR is information-theoretically secure and very efficient. Our second approach is only computationally-secure but works for multivariate polynomials. The security of the RR is based on the multi-dimensional noisy curve reconstruction assumption [13,17].

Resulting PHE Schemes. Using our fully-hiding RRs we get PHE schemes for univariate and multi-variate polynomials (with a small number of variables). We stress, however, that PHE schemes for univariate polynomials can be constructed without going through our RR-based approach. In fact, in the full version of this work we give an example of such a construction based only on HE schemes that support addition and a single multiplication [4,11]. This particular construction is very simple and slightly more efficient (i.e., by a constant factor) with respect to client-side work than our PHE scheme for univariate polynomials. We stress, however, that our RR-based approach is more general and yields schemes for more than just univariate polynomials. Since the focus of our work is on our RR-based approach to PHE, we only describe here the construction that results from our RR for univariate polynomials and omit the "simple" construction.

1.2 Our Contributions

While (sequential) homomorphic encryption constitutes an important step towards private outsourced computation, an increasing fraction of the computations performed "in the cloud" is on massive datasets and therefore requires the computation to be performed on clusters of machines. To address this, we make the following contributions:

1. We initiate the study of PHE . In particular, we consider the MapReduce model of parallel computation and formalize MapReduce-parallel HE schemes. Given the

practical importance of the MapReduce model and the emergence of cloud-based MapReduce clusters, we believe the study of MapReduce-parallel HE to be important and well motivated.

2. We construct new RRs for univariate and multivariate polynomials with a small number of variables (i.e., polylogarithmic in the security parameter). Our reduction for univariate polynomials is information theoretically secure while our reduction for multivariate polynomials is secure based on the multi-dimensional noisy curve reconstruction assumption [13]. Both our constructions achieve a hiding threshold of $t = n$ and are, as far as we know, the first constructions to do so.
3. We give a general transformation from any RR to a MR-parallel HE scheme given any public-key HE scheme that can evaluate the reductions' recovery algorithm. If the RR works for any function within a class $\mathcal{C}$, then the resulting MR-parallel scheme is $\mathcal{C}$-homomorphic.

Due to space limitations, we are not able to include all our results. In the full version of this work, we also consider and formalize the notion of *delegated* PHE (which also hides the function being evaluated) and give a delegated construction for any function with output values that can be computed by evaluating a (fixed) univariate polynomial over the input values. We also give optimized variants of our (non-delegated) MR-parallel HE schemes for both univariate and multi-variate polynomials. Finally, we show how, using techniques from [15] and [9], our MR-PHE schemes can be used to perform various queries over encrypted databases like set membership testing, disjunctions queries and keyword search.

2 Preliminaries and Notation

Polynomials. If p is a univariate polynomial of degree d over a field $\mathbb{F}$, then it can be written as $p(x) = \sum_{\alpha \in S} p(\alpha) \cdot \mathcal{L}_\alpha(x)$, where S is an arbitrary subset of $\mathbb{F}$ of size $d+1$ and $\mathcal{L}_\alpha$ is the Lagrangian coefficient defined as $\mathcal{L}_\alpha(x) = \prod_{i \in S, i \neq \alpha}(x-i)/(\alpha - i)$. A permutation polynomial $p \in \mathbb{F}[x]$ is a bijection over $\mathbb{F}$. One class of permutation polynomials which will make use of in this work are the Dickson polynomials (of the first kind) which are a family of polynomials $\mathbf{D} = \{\mathbf{D}_{d,\beta}\}$ over a finite field $\mathbb{F}$ indexed by a degree $d > 0$ and a non-zero element $\beta \in \mathbb{F}$. If $|\mathbb{F}|^2 - 1$ is relatively prime to d and if $\beta \neq 0$, then the Dicskon polynomial $\mathbf{D}_{d,\beta}$ defined as

$$\mathbf{D}_{d,\beta}(x) \overset{def}{=} \mathbf{D}_d(x,\beta) = \sum_{\lambda=0}^{\lfloor d/2 \rfloor} \frac{d}{d-\lambda} \cdot \binom{d-\lambda}{\lambda} \cdot (-\beta)^\lambda x^{d-2\lambda},$$

is a permutation over $\mathbb{F}$. For $d = 2$ and any $\beta \neq 0$, we have $\mathbf{D}_{2,\beta}(x) = x^2 - 2\beta$ which is a permutation over any $\mathbb{F}$ such that $|\mathbb{F}|^2 - 1$ is odd.

Homomorphic Encryption. Let $\mathcal{F}$ be a family of n-ary functions. A $\mathcal{F}$-homomorphic encryption scheme is a set of four polynomial-time algorithms $\mathsf{HE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Eval}, \mathsf{Dec})$ such that Gen is a probabilistic algorithm that takes as input a security parameter k and outputs a secret key K; Enc is a probabilistic algorithm that takes as input a key

K and an n-bit message m and outputs a ciphertext c; Eval is a (possibly probabilistic) algorithm that takes as input a function $f \in \mathcal{F}$ and n encryptions $(c_1, \ldots, c_n)$ of messages $(m_1, \ldots, m_n)$ and outputs an encryption c of $f(m_1, \ldots, m_n)$; and Dec is a deterministic algorithm takes as input a key K and a ciphertext c and outputs a message m. In this work, we make use of 2DNF-HE schemes which support an arbitrary number of additions and a single multiplication. Concrete instantiations of such schemes include [4] and [11].

3 MapReduce-Parallel Homomorphic Encryption

In this section, we first give an overview of the MapReduce model of computation together with an example of a simple MapReduce algorithm. We refer the reader to [7,16] for a more detailed exposition. After formalizing the MapReduce model, we define MapReduce-parallel HE schemes and present our security definitions for standard and delegated MR-parallel HE schemes.

3.1 The MapReduce Model of Computation

At a high level, MapReduce works by applying a map operation to the data which results in a set of label/value pairs. The map operation is applied in *parallel* and the resulting pairs are routed to a set of reducers. All pairs with the same label are routed to the same reducer which is then tasked with applying a reduce operation that combines the values into a single value for that label.

A MapReduce algorithm $\Pi = (\mathsf{Parse}, \mathsf{Map}, \mathsf{Red}, \mathsf{Merge})$ is executed on a cluster of w workers and one controller as follows. The client provides a function f and an input x to the controller who runs Parse on (f, x), resulting in a sequence of input pairs $(\ell_i, v_i)_i$. Each pair is then assigned by the controller to a worker that evaluates the Map algorithm on it. This results in a sequence of intermediate pairs $\{(\lambda_j, \gamma_j)\}_j$. Note that since the Map algorithm is stateless, it can be executed in parallel. Typically the number of input pairs is much larger than the number of workers so this stage may require several rounds. When all the input pairs have been processed, the controller partitions all the intermediate pairs and each set of the partition is then assigned to a worker that applies the Red algorithm on it. Again, since Red is stateless it can be executed in parallel (though it can be sequential on its own partition). The outputs of all these Red executions are then processed using Merge and the final result is returned to the client. At any time, a worker is either executing the Map algorithm (in which case it is a *mapper*) or the Red algorithm (in which case it is a *reducer*).

An Example. A simple example of a MapReduce algorithm is to determine frequency counts, i.e., the number times a keyword occurs in a document collection. The parse algorithm takes the document collection $(D_1, \ldots, D_n)$ as input and outputs a set of input pairs $(i, D_i)_i$. Each mapper receives an input pair (i, D_i) and outputs a set of intermediate pairs $(w_j, 1)_j$ for each word w_j found in D_i. All the intermediate pairs are then partitioned by the partition operation into sets $\{P_l\}$, where P_l consists of all the intermediate pairs with label w_l. The reducers receive a set P_l of intermediate pairs and sum the values of each pair. The result is a count of the number of times the word

w_l occurs in the document collection. The merge algorithm then concatenates all these counts and returns the result.

3.2 Syntax and Security Definitions

An MR-parallel HE scheme is a HE whose evaluation operation can be computed using a MapReduce algorithm.

Definition 1 (MR-parallel HE). *A private-key MR-parallel $\mathcal{F}$-homomorphic encryption scheme is a tuple of polynomial-time algorithms* $\mathsf{PHE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Eval}, \mathsf{Dec})$, *where* $(\mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$ *are as in a private-key encryption scheme and* $\mathsf{Eval} = (\mathsf{Parse}, \mathsf{Map}, \mathsf{Red}, \mathsf{Merge})$ *is a MapReduce algorithm. More precisely we have:*

$K \leftarrow \mathsf{Gen}(1^k)$*: is a probabilistic algorithm that takes as input a security parameter k and that returns a key K.*
$c \leftarrow \mathsf{Enc}(K, x)$*: is a probabilistic algorithm that takes as input a key K and an input x from some message space* X*, and that returns a ciphertext c. We sometimes write this as* $c \leftarrow \mathsf{Enc}_K(x)$.
$(\ell_i, v_i)_i \leftarrow \mathsf{Parse}(f, c)$*: is a deterministic algorithm that takes as input a function $f \in \mathcal{F}$ and a ciphertext c, and that returns a sequence of input pairs.*
$(\lambda_j, \gamma_j)_j \leftarrow \mathsf{Map}(\ell, v)$*: is a (possibly probabilistic) algorithm that takes an input pair (ℓ, v) and that returns a sequence of intermediate pairs.*
$(\lambda, z) \leftarrow \mathsf{Red}(\lambda, P)$*: is a (possibly probabilistic) algorithm that takes a label λ and a partition P of intermediate values and returns an output pair (λ, z).*
$c' \leftarrow \mathsf{Merge}\big((\lambda_t, z_t)_t\big)$*: is a deterministic algorithm that takes as input a set of output pairs and returns a ciphertext c'.*
$y \leftarrow \mathsf{Dec}(K, c')$*: is a deterministic algorithm that takes a key K and a ciphertext c' and that returns an output y. We sometimes write this as* $y \leftarrow \mathsf{Dec}_K(c')$.

We say that PHE *is correct if for all $k \in \mathbb{N}$, for all $f \in \mathcal{F}_k$, for all K output by* $\mathsf{Gen}(1^k)$, *for all $x \in$* X*, for all c output by* $\mathsf{Enc}_K(x)$, $\mathsf{Dec}_K\big(\mathsf{Eval}(f, c)\big) = f(x)$.

To be usable in the setting of private outsourced computation, a PHE scheme should guarantee that its ciphertexts reveal no useful information about the input x or the output $f(x)$. We note that in this setting it is sufficient for this to hold with respect to a *single* input. In the context of outsourced computation, as opposed that of secure communication, the cost of generating a new key per input is negligible. As such, our security definitions only guarantee security for a single input (which could be, e.g., a massive dataset).

Definition 2 (CPA^1**-security).** *Let* $\mathsf{PHE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Parse}, \mathsf{Map}, \mathsf{Red}, \mathsf{Merge}, \mathsf{Dec})$ *be a MR-parallel $\mathcal{F}$-homomorphic encryption scheme and consider the following probabilistic experiments where $\mathcal{A}$ is an adversary and $\mathcal{S}$ is a simulator:*

$\mathbf{Real}_{\mathsf{PHE},\mathcal{A}}(k)$*: the challenger begins by running* $\mathsf{Gen}(1^k)$ *to generate a key K. $\mathcal{A}$ outputs an input x and receives a ciphertext* $c \leftarrow \mathsf{Enc}_K(x)$ *from the challenger. $\mathcal{A}$ returns a bit b that is output by the experiment.*

Ideal$_{\mathsf{PHE},\mathcal{A},\mathcal{S}}(k)$*:* $\mathcal{A}$ *outputs an input* x*. Given* $|x|$*,* $\mathcal{S}$ *generates and returns a ciphertext* c *to* $\mathcal{A}$*.* $\mathcal{A}$ *returns a bit* b *that is output by the experiment.*

We say that PHE *is secure against a single-message chosen-plaintext attack if for all* PPT *adversaries* $\mathcal{A}$*, there exists a* PPT *simulator* $\mathcal{S}$ *such that*

$$|\Pr\left[\,\mathbf{Real}_{\mathsf{PHE},\mathcal{A}}(k) = 1\,\right] - \Pr\left[\,\mathbf{Ideal}_{\mathsf{PHE},\mathcal{A},\mathcal{S}}(k) = 1\,\right]| \leq \mathsf{negl}(k),$$

where the probabilities are over the coins of Enc*,* $\mathcal{A}$ *and* $\mathcal{S}$*.*

4 Randomized Reductions for Polynomials

In this section, we formally define randomized reductions [1,2,3] and then present our fully-hiding constructions for univariate and multivariate polynomials. Our definitions follow closely the ones given by Beaver, Feigenbaum, Killian and Rogaway [3].

Let $t, n \in \mathbb{N}$ such that $t \leq n$. A function $f : X \to Y$ is (t,n)-locally random reducible to a function $g : \widetilde{X} \to \widetilde{Y}$ if there exists two polynomial-time algorithms $\mathsf{RR} = (\mathsf{Scatter}, \mathsf{Recon})$ that work as follows. Scatter is a probabilistic algorithm that takes as input an element $x \in X$ and a parameter $n \in \mathbb{N}$, and returns a sequence $\mathbf{s} \in \widetilde{X}^n$ and some state information st. Recon is a deterministic algorithm that takes as input some state st and a sequence $\mathbf{y} \in \widetilde{Y}^n$ and returns an element $y \in Y$. In addition, we require that RR satisfy the following properties:

- (Correctness) for all $x \in X$,

  $$\Pr\Big[\,\mathsf{Recon}\big(st, g(s_1), \ldots, g(s_n)\big) = f(x) : (\mathbf{s}, st) \leftarrow \mathsf{Scatter}(x, n)\,\Big] \geq 3/4,$$

  where the probability is over the coins of Scatter. We depart slightly from the original definition [3] in that here Recon does not need to take x as input.
- (t-hiding) for all $I \subseteq [n]$ such that $|I| = t$, and all x_1 and x_2 in X such that $|x_1| = |x_2|$,

  $$\left\{\langle s_i\rangle_{i\in I} : (\mathbf{s}, st) \leftarrow \mathsf{Scatter}(x_1, n)\right\} \approx \left\{\langle s_i\rangle_{i\in I} : (\mathbf{s}, st) \leftarrow \mathsf{Scatter}(x_2, n)\right\}$$

  where the distributions are over the coins of Scatter. If $t = n$, we sometimes say that f is *fully* hiding. If the distributions are identically distributed we say that f is *perfectly* hiding, and if the distributions are computationally indistinguishable we say f is *computationally* hiding.
- (Efficiency) for all $x \in X$ and all $\mathbf{s}$ and st output by $\mathsf{Scatter}(x, n)$, the time to evaluate $\mathsf{Recon}(st, g(s_1), \ldots, g(s_n))$ is less than the time to evaluate $f(x)$.

If $g \neq f$ then RR is a *local random reduction* (LRR). If $g = f$, then RR is a *randomized self reduction* (RSR). Furthermore, if there exists a pair of algorithms $\mathsf{RSR} = (\mathsf{Scatter}, \mathsf{Recon})$, such that for every function f in some class $\mathcal{C}$, RSR is a random self reduction for f, then we say that RSR is a *universal* random self reduction over $\mathcal{C}$. All of our constructions are universal.

A Note on Efficiency. For our purposes, the efficiency requirement is not necessary. This is because in our MR-PHE constructions, the Recon algorithm is not executed by the client but, instead, is executed homomorphically by the cluster. As such, a more important requirement for us is that Recon to be "simple" enough so that it can be evaluated homomorphically without making use of FHE.

4.1 A Perfect Randomized Self Reduction for Univariate Polynomials

In this section, we present a *fully*-hiding randomized reduction for univariate polynomials. As far as we know, the best hiding threshold previously achieved by any RR for univariate polynomials is $t \leq (n-1)/q$ which is achieved by the construction of Beaver, Feigenbaum, Killian and Rogaway [2,3]. Like the construction presented in [2,3], our randomized reduction is *universal* and *self-reducing*.

Let $\mathbf{Q}$ be a degree q univariate polynomial over a finite field $\mathbb{F}$ such that $|\mathbb{F}| \geq 2q+1$ and $|\mathbb{F}|^2 - 1 \equiv 1 \pmod 2$, and let $\delta[\mathbb{F}^n] \overset{def}{=} \{\mathbf{v} \in \mathbb{F}^n : v_i \neq v_j \text{ for all } i,j \in [n]\}$. Consider the random self reduction $\mathsf{Poly}^1_\mathsf{q} = (\mathsf{Scatter}_q, \mathsf{Recon}_q)$ for $\mathbf{Q}$ defined as follows:

- $\mathsf{Scatter}_q(x)$: let $n = 2q+1$ and sample a vector $\boldsymbol{\alpha}$ uniformly at random from $\delta\left[\mathbb{F}^n\right]$. For all $i \in [n]$, compute $s_i := \mathbf{D}_2(\alpha_i, -x/2) = \alpha_i^2 + x$. Output $(s_1, \ldots, s_n)$ and $st = \boldsymbol{\alpha}$.
- $\mathsf{Recon}_q(st, y_1, \ldots, y_n)$: output $y = \sum_{i=1}^n y_i \cdot \mathcal{L}_{\alpha_i}(0)$.

Theorem 1. $\mathsf{Poly}^1_\mathsf{q}$ *is a perfect and fully-hiding randomized self reduction.*

Proof. Towards showing correctness, let $\widehat{\mathbf{Q}}(\alpha) \overset{def}{=} \mathbf{Q}\big(\mathbf{D}_2(\alpha, -x/2)\big)$ (for some $x \in \mathbb{F}$) and note that $\widehat{\mathbf{Q}}(0) = \mathbf{Q}(x)$. We therefore have:

$$y = \sum_{i=1}^{2q+1} y_i \cdot \mathcal{L}_{\alpha_i}(0) = \sum_{i=1}^{2q+1} \mathbf{Q}\big(\mathbf{D}_2(\alpha_i, -x/2)\big) \cdot \mathcal{L}_{\alpha_i}(0) = \sum_{i=1}^{2q+1} \widehat{\mathbf{Q}}(\alpha_i) \cdot \mathcal{L}_{\alpha_i}(0) = \widehat{\mathbf{Q}}(0) = \mathbf{Q}(x),$$

since $\deg(\widehat{\mathbf{Q}}) = 2q$. We now consider perfect hiding. Let $n = 2q+1$ and note that for fixed $q \in \mathbb{N}$ and $x \in \mathbb{F}$, $\mathsf{Scatter}$ evaluates the vector-valued function $f_{x,q} : \delta[\mathbb{F}^n] \to \delta[\mathbb{F}^n]$ defined as

$$f_{x,q}(\boldsymbol{\alpha}) = \bigg(\mathbf{D}_2\big(\alpha_1, -x/2\big), ..., \mathbf{D}_2\big(\alpha_n, -x/2\big)\bigg),$$

for a random $\boldsymbol{\alpha}$. Note that $f_{x,q}$ is a permutation over $\delta[\mathbb{F}^n]$ since $\mathbf{D}_2(\alpha, \beta)$ is a permutation over $\mathbb{F}$ for any β (this follows from the fact that $|\mathbb{F}|^2 - 1 \equiv 1 \pmod 2$). Let $\mathcal{U}$ be the uniform distribution over $\delta[\mathbb{F}^n]$. In the following, for visual clarity we drop the subscript q and denote $f_{x,q}$ by f_x. For all x_1 and x_2 in $\mathbb{F}$,

$$\begin{aligned}
\mathsf{SD}\big(f_{x_1}(\mathcal{U}), f_{x_2}(\mathcal{U})\big) &= \max_{S \subset \delta[\mathbb{F}^n]} \left|\Pr\left[f_{x_1}(\mathcal{U}) \in S\right] - \Pr\left[f_{x_2}(\mathcal{U}) \in S\right]\right| \\
&= \max_{S \subset \delta[\mathbb{F}^n]} \left|\Pr\left[\mathcal{U} \in f_{x_1}^{-1}(S)\right] - \Pr\left[\mathcal{U} \in f_{x_2}^{-1}(S)\right]\right| \\
&\leq \max_{V, V' \subset \delta[\mathbb{F}^n]} \left|\Pr\left[\mathcal{U} \in V\right] - \Pr\left[\mathcal{U} \in V'\right]\right| \\
&= 0
\end{aligned}$$

where the last equality follows from the fact that $|V| = |V'|$ since f_{x_1} and f_{x_2} are permutations over $\delta[\mathbb{F}^n]$.

4.2 A Computational Randomized Self Reduction for Multivariate Polynomials

We now present a fully-hiding RSR for multi-variate polynomials. The best known hiding threshold previously achieved is from a construction of [2,3] which achieves $t \leq n \cdot c\log(m)/m$ for c and m greater than 1. Our construction is universal and self-reducing.

Let $\mathbf{Q}$ be a m-variate degree q polynomial over a finite field $\mathbb{F}$ such that $|\mathbb{F}| \geq n+1$, for $n \in \mathbb{N}$. Consider the randomized self reduction $\mathsf{Poly}_\mathsf{q}^\mathsf{m} = (\mathsf{Scatter}_q, \mathsf{Recon}_q)$ defined as follows:

- $\mathsf{Scatter}_q(\mathbf{x})$: let $n = 2q+1$ and sample m univariate polynomials $(p_1, \ldots, p_m)$ of degree 2 such that $p_i(0) = x_i$ for all $i \in [m]$. Let $N = \omega(n \cdot (n/q)^m)$ and $\boldsymbol{\alpha} \overset{\$}{\leftarrow} \delta[\mathbb{F}^n]$. For all $j \in [n]$, set $\mathbf{z}_j := \big(p_1(\alpha_j), \ldots, p_m(\alpha_j)\big)$ and for all $j \in [n+1, n+N]$ set $\mathbf{z}_j \overset{\$}{\leftarrow} \mathbb{F}^m$. Let $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_{n+N})$ be the sequence that results from permuting the elements of $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{n+N})$ at random and let Γ be the locations in $\mathbf{S}$ of the elements in $\mathbf{Z}$ that were chosen at random in $\mathbb{F}^m$. Output $\mathbf{S}$ and $st = (\pi(\boldsymbol{\alpha}), \Gamma)$, where π denotes the (random) permutation used to permute $\mathbf{Z}$.
- $\mathsf{Recon}_{m,q}(st, y_1, \ldots, y_{n+N})$: parse st as $(\boldsymbol{\alpha}, \Gamma)$ and output $y = \sum_{i \notin \Gamma} y_i \cdot \mathcal{L}_{\alpha_i}(0)$.

The security of our randomized reduction is based on the multi-dimensional noisy curve reconstruction assumption from Ishai, Kushilevitz, Ostrovsky and Sahai [13], which extends the polynomial reconstruction (PR) assumption from Naor and Pinkas [17].

Assumption 2 (Multi-dimensional noisy curve reconstruction [13,17]). *The multi dimensional noisy curve reconstruction (CR) assumption is defined in terms of the following experiment where* $\mathbf{x}$ *is a* m*-dimensional vector over a finite field* $\mathbb{F}$*,* $d > 1$*, and* $t = t(k)$ *and* $z = z(k)$ *are functions of* k*:*

$\mathbf{CurveRec}(k, \mathbf{x}, d, n, N, m)$*: sample a vector* $\boldsymbol{\alpha} \overset{\$}{\leftarrow} \mathbb{F}^n$ *and a random subset of* N *indices* Γ *chosen from* $[n+N]$*. Choose* m *random univariate polynomials* $(p_1, \ldots, p_m)$ *such that each* p_i *is of degree at most* d *and that* $p_i(0) = x_i$*. For all* $j \in [n]$*, set* $\mathbf{z}_j = (p_1(\alpha_j), \ldots, p_m(\alpha_j))$ *and for all* $j \in [n+1, n+N]$ *set* $\mathbf{z}_j \overset{\$}{\leftarrow} \mathbb{F}^m$*. Let* $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_{n+N})$ *be the sequence that results from permuting the elements of* $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{n+N})$ *uniformly at random. The output of the experiment is* $(\mathbf{s}_1, \ldots, \mathbf{s}_{n+N})$*.*

We say that the CR assumption holds over $\mathbb{F}$ *with parameters* (d, n, N, m) *if for all* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *in* $\mathbb{F}^m$*,*

$$\Big\{\mathbf{CurveRec}(k, \mathbf{x}_1, d, n, N, m)\Big\} \overset{c}{\approx} \Big\{\mathbf{CurveRec}(k, \mathbf{x}_2, d, n, N, m)\Big\}$$

We note that the CR assumption is believed to hold when N *is* $\omega(n \cdot (n/d)^m)$ *and* $|\mathbb{F}| = N$ *[13].*

Remark. Setting n and N to be polynomial in k, the CR assumption is believed to hold as long as $m = \mathsf{polylog}(k)$. We note, however, that the parameters provided in [13] and used in this work are for the *stronger* "augmented CR" assumption which outputs, in addition to the vectors $(\mathbf{s}_1, \ldots, \mathbf{s}_{n+N})$, the evaluation points $(\alpha_1, \ldots, \alpha_n)$ together with N random values. It is therefore plausible that the CR assumption could hold for a wider range of parameters and, in particular, for $m = \mathsf{poly}(k)$.

In the following theorem, we show that $\mathsf{Poly}_\mathsf{q}^\mathsf{m}$ is a fully-hiding and universal RSR for the class of multivariate polynomials with a poly-logarithmic number of variables.

Theorem 3. $\mathsf{Poly}_\mathsf{q}^\mathsf{m}$ *is a computational and fully-hiding random self reduction.*

The proof follows almost directly from Assumption 2, so due to space limitations, it is deferred to the full version of this work.

5 MR-Parallel HE from Randomized Reductions

We now show how to construct a MR-parallel HE scheme from any $\mathcal{F}$-homomorphic encryption scheme and any fully-hiding RR between functions f and g whose reconstruction algorithm is in $\mathcal{F}$. At a high-level, the construction works as follows.

The RR's scatter algorithm is applied to each element x_i of the input $\mathbf{x}$. This results in a sequence $\mathbf{s}_i$ and a state st_i. The latter is encrypted using the $\mathcal{F}$-homomorphic encryption scheme and each mapper receives a pair composed of a label $\ell = i$ and a value v of the form $(s_i[j], e_i)$ for $i \in [\#\mathbf{x}]$ and $j \in [n]$ and where e_i is an $\mathcal{F}$-homomorphic encryption of st_i. The mapper evaluates g on $s_i[j]$ and returns an intermediate pair with label $\lambda = i$ and value $\gamma = \big(g(s_i[j]), e_i\big)$. After the shuffle operation, each reducer receives a pair composed of a label i and a partition

$$P = \bigg(\big(y_{i,j}, e_i\big), \ldots, \big(y_{i,n}, e_i\big) \bigg),$$

where $y_{i,j} = g(s_i[j])$ for $j \in [n]$. Since Recon is in $\mathcal{F}$, the reducer can evaluate $\mathsf{Recon}(e_i, y_{i,1}, \ldots, y_{i,n})$ homomorphically which results in an encryption of $f(x_i)$.

Theorem 4. *If* HE *is CPA-secure and if* RR *is fully-hiding, then* PHE *as described in Figure 1 is secure against single-message chosen-plaintext attacks.*

We sketch a proof of Theorem 4 and leave a full proof to the full version of this work. Consider the simulator $\mathcal{S}$ that simulates ciphertexts in an $\mathbf{Ideal}(k)$ experiment as follows. Given $\#\mathbf{x}$ it generates $(pk', sk') \leftarrow \mathsf{Gen}(1^k)$ and, for all $i \in [\#\mathbf{x}]$, it computes $(\mathbf{s}'_i, st'_i) \leftarrow \mathsf{Scatter}(0)$ and $e'_i \leftarrow \mathsf{HE.Enc}_{pk'}(st'_i)$. It outputs $\mathbf{c}' = (pk', \mathbf{s}'_1, \ldots, \mathbf{s}'_{\#\mathbf{x}}, e'_1, \ldots, e'_{\#\mathbf{x}})$. The fully-hiding property of RR guarantees that the $\mathbf{s}'_i$'s are indistinguishable from the $\mathbf{s}_i$'s generated in a $\mathbf{Real}(k)$ experiment. Similarly, the CPA-security of HE guarantees that the e'_i's are indistinguishable from the e_i's generated in a $\mathbf{Real}(k)$ experiment.

Direct Constructions. By instantiating the RR and the HE scheme in our general construction with our fully-hiding RSR for univariate polynomials (from section 4.1) and

Let $\mathsf{HE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Eval}, \mathsf{Dec})$ be a public-key $\mathcal{F}$-homomorphic encryption scheme and let $\mathsf{RR} = (\mathsf{Scatter}, \mathsf{Recon})$ be a $\mathcal{C}$-universal (t, n)-local randomized reduction from f to g such that $\mathsf{Recon} \in \mathcal{F}$. Consider the multi-use MR-parallel $\mathcal{C}$-homomorphic encryption scheme $\mathsf{PHE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Eval}, \mathsf{Dec})$, where $\mathsf{PHE.Eval} = (\mathsf{Parse}, \mathsf{Map}, \mathsf{Red}, \mathsf{Merge})$, defined as follows:

- $\mathsf{Gen}(1^k)$: compute $(pk, sk) \leftarrow \mathsf{HE.Gen}(1^k)$. Output $K = (sk, pk)$.
- $\mathsf{Enc}(K, \mathbf{x})$: for all $i \in [\#\mathbf{x}]$, compute $(\mathbf{s}_i, st_i) \leftarrow \mathsf{Scatter}(x_i)$ and $e_i \leftarrow \mathsf{HE.Enc}_{pk}(st_i)$. Output $\mathbf{c} = (pk, \mathbf{s}_1, \ldots, \mathbf{s}_{\#\mathbf{x}}, e_1, \ldots, e_{\#\mathbf{x}})$.
- $\mathsf{Parse}(f, \mathbf{c})$: for all $i \in [\#\mathbf{x}]$ and $j \in [n]$, set $\ell_{i,j} := i$ and $v_{i,j} := (f, pk, \mathbf{s}_i[j], e_i)$. Output $(\ell_{i,j}, v_{i,j})_{i,j}$.
- $\mathsf{Map}(\ell, v)$: parse v as (f, s, e) and compute $a \leftarrow \mathsf{HE.Enc}_{pk}(g(s))$. Output $\lambda := \ell$ and $\gamma := (a, e)$.
- $\mathsf{Red}(\lambda, P)$: parse P as $(a_r, e_r)_r$ and compute $z \leftarrow \mathsf{HE.Eval}(\mathsf{Recon}, e_r, (a_r)_r)$. Output (λ, z).
- $\mathsf{Merge}\big((\lambda_t, z_t)_t\big)$: output $\mathbf{c}' := (z_t)_t$.
- $\mathsf{Dec}(K, \mathbf{c}')$: for all $i \in [\#\mathbf{c}']$, compute $y_i := \mathsf{HE.Dec}_{sk}(z_i)$. Output $\mathbf{y} = (y_1, \ldots, y_{\#\mathbf{c}'})$.

Fig. 1. MR-parallel HE from RR and HE

an FHE scheme, we get a multi-use MR-parallel HE scheme for the class of functions whose output values can be computed by evaluating a (fixed) univariate polynomial of the inputs. In addition, the resulting construction can be made delegated by encrypting the coefficients of the polynomial using the FHE scheme and having the mappers perform their computations homomorphically. Current FHE constructions, however, are not yet practical enough for our purposes so, in the full version, we present a direct construction based only on additively homomorphic encryption. The construction can be made delegated if we use 2DNF-HE. The direct construction also has the advantage that the input pairs sent to the mappers are smaller than what would result from our general construction.

Similarly, if we instantiate our general construction with our RR for multi-variate polynomials (from Section 4.2) and an FHE scheme, we get an MR-parallel HE scheme for the class of functions whose output values can be computed by evaluating a (fixed) multi-variate polynomial on the inputs (with small number of variables). To avoid the use of FHE, however, we present in the full version of this work a direct construction that only makes use of additively HE.

References

1. Beaver, D., Feigenbaum, J.: Hiding instances in multioracle queries. In: Choffrut, C., Lengauer, T. (eds.) STACS 1990. LNCS, vol. 415, pp. 37–48. Springer, Heidelberg (1990)
2. Beaver, D., Feigenbaum, J., Kilian, J., Rogaway, P.: Security with low communication overhead. In: Menezes, A., Vanstone, S.A. (eds.) CRYPTO 1990. LNCS, vol. 537, pp. 62–76. Springer, Heidelberg (1991)
3. Beaver, D., Feigenbaum, J., Kilian, J., Rogaway, P.: Locally random reductions: Improvements and applications. Journal of Cryptology 10(1) (1997)

4. Boneh, D., Goh, E.-J., Nissim, K.: Evaluating 2-DNF formulas on ciphertexts. In: Kilian, J. (ed.) TCC 2005. LNCS, vol. 3378, pp. 325–341. Springer, Heidelberg (2005)
5. Borodin, A.: On relating time and space to size and depth. SIAM J. of Comp. 6(4) (1977)
6. Culler, D., Karp, R., Patterson, D., Sahay, A., Santos, E., Schauser, K., Subramonian, R., von Eicken, T.: Logp: A practical model of parallel computation. Comm. of the ACM 39(11) (1996)
7. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. In: Symposium on Opearting Systems Design & Implementation (2004)
8. Fortune, S., Wyllie, J.: Parallelism in random access machines. In: ACM Symposium on Theory of Computing, STOC 1978 (1978)
9. Freedman, M.J., Ishai, Y., Pinkas, B., Reingold, O.: Keyword search and oblivious pseudorandom functions. In: Kilian, J. (ed.) TCC 2005. LNCS, vol. 3378, pp. 303–324. Springer, Heidelberg (2005)
10. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: ACM Symposium on Theory of Computing, STOC 2009 (2009)
11. Gentry, C., Halevi, S., Vaikuntanathan, V.: A simple BGN-type cryptosystem from LWE. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 506–522. Springer, Heidelberg (2010)
12. Isard, M., Budiu, M., Yu, Y., Birrell, A., Fetterly, D.: Dryad: distributed data-parallel programs from sequential building blocks. In: ACM SIGOPS/EuroSys European Conference on Computer Systems, EuroSys 2007 (2007)
13. Ishai, Y., Kushilevitz, E., Ostrovsky, R., Sahai, A.: Cryptography from anonymity. In: IEEE Symposium on Foundations of Computer Science, FOCS 2006 (2006)
14. Karp, R.M., Ramachandran, V.: Parallel algorithms for shared-memory machines (1990)
15. Kissner, L., Song, D.: Privacy-preserving set operations. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 241–257. Springer, Heidelberg (2005)
16. Lin, J., Dyer, C.: Data-Intensive Text Processing with MapReduce. M. & C. (2010)
17. Naor, M., Pinkas, B.: Oblivious polynomial evaluation. SIAM J. of Comp. 35(5) (2006)
18. Valiant, L.: A bridging model for parallel computation. Comm. of the ACM 33(8) (1990)

Targeting FPGA DSP Slices for a Large Integer Multiplier for Integer Based FHE

Ciara Moore, Neil Hanley, John McAllister, Máire O'Neill,
Elizabeth O'Sullivan, and Xiaolin Cao

Centre for Secure Information Technologies (CSIT),
Queen's University Belfast, Northern Ireland
{cmoore50,n.hanley,e.osullivan,xcao03}@qub.ac.uk,
{j.Mcallister,m.oneill}@ecit.qub.ac.uk

Abstract. Homomorphic encryption offers potential for secure cloud computing. However due to the complexity of homomorphic encryption schemes, performance of implemented schemes to date have been unpractical. This work investigates the use of hardware, specifically Field Programmable Gate Array (FPGA) technology, for implementing the building blocks involved in somewhat and fully homomorphic encryption schemes in order to assess the practicality of such schemes. We concentrate on the selection of a suitable multiplication algorithm and hardware architecture for large integer multiplication, one of the main bottlenecks in many homomorphic encryption schemes. We focus on the encryption step of an integer-based fully homomorphic encryption (FHE) scheme. We target the DSP48E1 slices available on Xilinx Virtex 7 FPGAs to ascertain whether the large integer multiplier within the encryption step of a FHE scheme could fit on a single FPGA device. We find that, for toy size parameters for the FHE encryption step, the large integer multiplier fits comfortably within the DSP48E1 slices, greatly improving the practicality of the encryption step compared to a software implementation. As multiplication is an important operation in other FHE schemes, a hardware implementation using this multiplier could also be used to improve performance of these schemes.

1 Introduction

Cloud computing offers numerous advantages to users, such as computing as a service, storage and management of large amounts of data. Yet this requires the trust of the public cloud service provider to maintain an adequate level of security and prevent leakage of private data. Data security has been shown to be the greatest concern of clients who use the cloud [1]. If users could encrypt their data before storing it in an (untrusted) cloud server and still be able to compute on these ciphertexts, they could take advantage of the benefits of cloud computation without the risk of leaking their private data.

Secure cloud computing could be achieved by the use of an efficient fully homomorphic encryption scheme. Homomorphic encryption is a method of encryption

A.A. Adam, M. Brenner, and M. Smith (Eds.): FC 2013, LNCS 7862, pp. 226–237, 2013.

featuring four steps: {*key-gen, encrypt, evaluate, decrypt*}, where the step *evaluate* enables the correct computation, such as addition and multiplication, on ciphertexts without the use of decryption. Traditionally, homomorphic encryption schemes were either additively or multiplicatively homomorphic; such schemes are also known as partially homomorphic encryption schemes. Examples include the multiplicatively homomorphic ElGamal [2] and the additively homomorphic Paillier [3] cryptosystems. In 2005 Boneh-Goh-Nissam introduced a scheme which allowed a combination of additions and one multiplication on encrypted data [4].

The area of homomorphic encryption leapt forward in 2009 however, with Gentry's ground-breaking work on a *fully* homomorphic encryption (FHE) scheme based on ideal lattices, which introduced the first technique to allow an arbitrary number of operations (both additions and multiplications) to be employed on ciphertexts [5]. A FHE scheme is created by extending a *somewhat* homomorphic encryption (SHE) scheme, which allows a limited number of multiplications and additions. In the last few years there has been much research to improve the efficiency of homomorphic encryption schemes [6], [7], [8]. The theory behind homomorphic encryption is developing at a quick pace; however there are few published results of timings from implementations of these schemes. Moreover, from the results that have been published, it is clear that improvements in the efficiency of these schemes are still needed. For example, in the SHE implementation of the largest lattice-based scheme in [9], bitwise encryption is reported to take 3.2 minutes. In addition, the FHE implementation of the integer-based scheme for the large implementation in [10], bitwise encryption takes 7 minutes 15 seconds. The recent FHE implementation of AES [8] requires approximately 36 hours and 256 GB RAM to evaluate AES; this shows there is still much to be done before such schemes are practical and comparable to existing cryptographic encryption schemes. It also highlights the complexity of homomorphic encryption and underlines the demand for more efficient implementations. In this paper we investigate implementing a hardware building block, which in some form features in all of the SHE and FHE schemes, in order to improve their performance and hence their practicality.

Three main structures have been proposed for FHE/SHE schemes: lattice-based, integer-based and schemes based on learning with errors (LWE) or ring learning with errors (RLWE). The current focus of the research community is on RLWE schemes, as these promise greater efficiency due to recent optimisations to support batching, for example in [7]. However the integer-based schemes, introduced by van Dijk, Gentry, Halevi and Vaikuntanathan (DGHV) in [11], have a relatively simple structure in comparison to the RLWE schemes and lattice-based methods introduced by Gentry. The efficiency of the latest integer-based schemes [10], [12] is comparable to the lattice-based schemes.

As a first step in our investigation into a hardware implementation of SHE or FHE schemes, we consider the proposed parameter sizes and the main underlying computations involved in the encryption step of the integer-based FHE scheme proposed by Coron et al [10], a scheme similar to the original DGHV integer-based FHE scheme [11]. The main computations are modular reduction and large integer multiplication, and are used in all of the FHE schemes. Therefore an efficient hardware implementation of these crypto-primitives can be used in future real time hardware implementations of any FHE scheme to improve performance. We focus on considering a hardware implementation of large integer multiplication and highlight some of the major issues involved. We begin to address these implementation issues by selecting a suitable large integer multiplication algorithm for hardware implementation. Due to the computational complexity of large integer multiplication, it is likely that a custom circuit architecture exploiting an Application Specific Integrated Circuit (ASIC) or a high-end FPGA technology in the form of a Xilinx Virtex 7 device will be required to enable real-time implementation. Considering the reconfigurable nature and quick development time of FPGAs we base our implementations on these. These devices also have exceptional levels of on-chip multiplication capability in the form of DSP48E1 slices.

To our knowledge, there are no current hardware implementations of complete FHE schemes; however there has been work on FPGA implementation of primitives for a SHE scheme using Mathworks® Simulink [13]. There has also been research in similar areas, for example [14] discusses the practicality of existing applications of homomorphic encryption by an empirical evaluation based on the lattice-based scheme by Smart and Vercauteren [15], and highlights implementation issues such as memory access. Another related publication [16] considers the hardware building blocks for the LWE cryptosystem and uses Fast Fourier Transform (FFT) multiplication in polynomial rings. Although it is stated that there may be more suitable multiplication algorithms for this purpose, it is shown that this hardware implementation of LWE still outperforms the software implementation. The Comba multiplication algorithm, introduced in 1962 [17], has been implemented in an FPGA using DSP slices to carry out multiplications required in the area of elliptic curve cryptography [18]. We look at using this multiplication method for large integer multiplication required in FHE schemes, as this type of multiplication has been shown to be very suitable for use on DSP slices. We estimate the performance of using Comba multiplication in DSP slices for the parameter sizes in the integer-based scheme by Coron et al [10] in order to establish the feasibility of a FPGA implementation of FHE schemes, and whether a hardware implementation of a multiplier would enable practical performance of the encryption step in [10], therefore offering a significant improvement to the existing implementations of large integer multiplication in FHE schemes.

We find in this initial evaluation for the toy-sized version of the encryption step of the FHE scheme in [10], the large integer multiplier fits comfortably within the DSP48E1 slices in a FPGA and would improve the practicality of the encryption step in [10], compared to a software implementation. Moreover, the large integer multiplier for the specified small, medium and large versions of the encryption step also fits comfortably within the DSP48E1 slice, though in these versions off-chip memory must be used to cope with the large parameter sizes. Indeed, as multiplication is an important operation in this type of encryption scheme, a hardware implementation using this multiplier, could be used to improve the performance of all FHE schemes. To our knowledge, there has been little previous analysis into the practicality of an FPGA based implementation of crypto primitives for FHE schemes.

In Section 2 of this paper, the selected integer-based scheme is introduced and we justify our approach. Section 3 presents a very brief survey of some multiplication methods and introduces the Comba multiplication method. A suitable hardware architecture and rough estimates for timings and resource requirements is given in Section 4. Some of the major implementation issues are also highlighted in this section.

2 Overview of Integer-Based FHE Scheme by Coron et al.

We focus on the proposed FHE scheme by Coron et al [10], based on the original integer-based FHE scheme [11], for its simple approach, detailed parameter sizes and reasonable performance in comparison to other implemented schemes, such as [9], [15]. We focus in particular on the encryption step, as this is one of the key steps in a FHE scheme which may need to be performed multiple times, unlike key generation which is only required initially. Moreover the encryption step in [10] involves two important cryptographic building blocks: multiplication of large integers and modular reduction, which are also used in all other FHE schemes. We explain the encryption step in the integer based FHE scheme in detail because of its relevance to this work. However, we refer the reader to [10] for details of the other steps in the scheme.

The encryption step for a given message $m \in \{0,1\}$ is given as:

$$c \leftarrow m + 2 \cdot r + 2 \cdot \sum_{i=1}^{\tau} b_i \cdot x_i \; mod \; x_0 \qquad (1)$$

where r is an integer from a specified range $(-2^{\rho'}, 2^{\rho'})$ and is used as random noise; $x_0 = q_o \cdot p$, where q_0 is a random odd integer in the range $[0, 2^{\gamma}/p)$ and p is a random prime integer of η bits; x_i for $1 \leq i \leq \tau$ is an array of large random integers; and $b_i, \forall 1 \leq i \leq \tau,$ is an array of random integers selected from a smaller range $[0, 2^{\alpha})$. The parameters γ, ρ', η and α in Equation (1) vary according to the size of scheme implemented. Hence we refer the reader to [10] for full details on these parameters and further information on the generation of x_i.

We target in particular the toy-sized FHE scheme; the parameter sizes for the four versions of the FHE scheme are listed in Table 1. In the toy-sized scheme 158 multiplications of $b_i \cdot x_i$ are required where the bit sizes for b_i and x_i are 936 bits and 150,000 bits respectively. In this paper we focus on the multiplier and establish a suitable approach to deal with these large parameter sizes. As can be seen in Table 1, the parameter sizes are very large, which is common in FHE schemes. For a discussion of security of this scheme, we again refer the reader to [10].

Table 1. Parameter Sizes (bits) for Encryption step in FHE Scheme in [10]

Parameter	Toy	Small	Medium	Large
b_i	936	1,476	2,016	2,556
x_i	150,000	830,000	4,200,000	19,350,000
x_0	150,000	830,000	4,200,000	19,350,000
τ	158	572	2110	7659

The two main bottlenecks in the selected scheme are large integer multiplication and modular reduction. These operations are also required in many other FHE schemes, such as the lattice based schemes [9], [15]. We have chosen to focus initially on multiplication as most efficient hardware implementations of modular reduction also require the use of a multiplier, for example Barrett reduction and Montgomery reduction both require multiplications [19]. Moreover, one of the main motivations for FHE and SHE schemes is to compute, using additions and multiplications, on encrypted data. Therefore an efficient multiplier for large parameter sizes is essential for such schemes.

Multiplication is only one of the issues to be addressed to implement this type of encryption scheme in hardware. Other major issues in the hardware implementation of homomorphic encryption schemes exist, such as the transfer of large blocks of data to and from the board, memory access and efficient scheduling of operations. In this initial study, we focus our attention on the multiplication bottleneck to establish the viability of an FPGA implementation of a FHE scheme and thus to justify continuing research to address the other important issues for a hardware implementation.

3 Overview of the Comba Multiplication Algorithm

Many multiplications with large multiplicands are required for implementation of the selected encryption scheme. There are various different algorithms available to deal with larger multiplicands and multipliers. Karatsuba multipliers [20] can be used to reduce the number and size of multiplications for large numbers by representing the large numbers, X and Y, as additions of two smaller numbers, for example $X = X_1 2^k + X_0$, $Y = Y_1 2^k + Y_0$ where X and Y are numbers of bit length $2k$. Then the multiplications are reduced from 4 multiplications (and 3 additions) to 3 multiplications (and 1 addition and 3 subtractions) as shown in Equation (2):

$$XY = (X_1 2^k + X_0) \cdot (Y_1 2^k + Y_0)$$

$$= 2(X_0 \cdot Y_0 + X_1 \cdot \ Y_1 2^{2k}) - (X_1 2^k - X_0) \cdot (Y_1 2^k - Y_0) \qquad (2)$$

However, Karatsuba requires intermediate storage of multiplication and subtraction results and is therefore not ideal for mapping to DSP slices, especially when considering such large parameter sizes. Fast Fourier transforms (FFTs) can also be used for multiplications, particularly when many multiplications are required. The use of FFTs has also been suggested in previous homomorphic encryption implementations [13]. Another alternative is Montgomery multiplication, commonly used in asymmetric cryptosystems. However, this technique requires multiplications for both post- and pre-computation. This method is more suitable when repeating multiplications such as in exponentiation algorithms, for example in RSA [21]. As we propose to target the DSP slices on a FPGA for large integer multiplication, we select a multiplication algorithm particularly suitable for the underlying FPGA platform for our initial investigation. The Comba multiplication method introduced in [17] is used for hardware-based large integer multiplication in [18] and it is very suitable for use on DSP slices as it can be easily broken down into partial products, therefore making efficient use of resources. Moreover, when these partial products are accumulated, they are retained within the DSP block. This method of multiplication involves a reversal of the order of words in the multiplicand, several shifts and multiplications with each shift. For example, to multiply two 3-word numbers, $A \times B$ for $A = A_2A_1A_0$ and $B = B_2B_1B_0$, reverse B => $B' = B_0B_1B_2$ and calculate the partial products PP_i by multiplying and adding:

$$\begin{aligned}
PP_0 &<= A_0 \times B_0 \\
PP_1 &<= A_1 \times B_0 + A_0 \times B_1 \\
PP_2 &<= A_2 \times B_0 + A_1 \times B_1 + A_0 \times B_2 \\
PP_3 &<= A_2 \times B_1 + A_1 \times B_2 \\
PP_4 &<= A_2 \times B_2
\end{aligned}$$

Each of the partial products PP_i are shifted left by i words ($\ll i$) and summed together to give the final product, giving:

$$A \times B = (PP_4 \ll 4) + (PP_3 \ll 3) + (PP_2 \ll 2) + (PP_1 \ll 1) + PP_0.$$

For a generalised multiplication of $A \times B$, let the word-length of A equal m and the word-length of B equal n and without loss of generality let $m \geq n$. There will be $m + n - 1$ required partial products in the Comba multiplication. When $i < m$, the i^{th} partial product PP_i requires $i + 1$ multiplications. The partial products can therefore have a maximum of m multiplications. When $i \geq m$, the i^{th} partial product requires $m + n - 1 - i$ multiplications. As suggested in [18] we can combine the partial products into m steps which have m multiplications in each step. Continuing the above example, we then have three steps which combine all of the partial products:

$$\begin{aligned}
PP'_0 &<= A_0 \times B_0 \;\; \boldsymbol{A_2} \times \boldsymbol{B_1} + \boldsymbol{A_1} \times \boldsymbol{B_2} \\
PP_1{}' &<= A_1 \times B_0 + A_0 \times B_1 \;\; \boldsymbol{A_2} \times \boldsymbol{B_2} \\
PP'_2 &<= A_2 \times B_0 + A_1 \times B_1 + A_0 \times B_2
\end{aligned}$$

We refer the reader to [18] for further details on this optimisation and their hardware implementation.

The choice of multiplier greatly depends on the size of the multiplication. In the particular case of the implementation of the toy scheme mentioned previously, we have a multiplier of 936 bits and a multiplicand of 150000 bits. We therefore propose to use a 936 bit multiplier and this can then be used several times and the partial products can be added to achieve the overall large multiplier. When we consider an FPGA implementation of Comba multiplication, we can run each of the steps in a separate parallel DSP slice, and then the number of clock cycles required per multiplication is m, the number of words in the largest multiplicand, and a few extra clocks for the summation of the partial products. The number of DSP slices required for the multiplication is equal to the number of steps after combining the partial products which is also m.

4 DSP Slice Usage and Estimated Timings for Large Integer Multiplier

FPGAs are a suitable target technology for hardware for implementations of SHE and FHE. They are cheaper and offer greater flexibility than ASIC devices. This makes them suitable for cryptographic purposes, as they can be re-programmed in-situ when protocols are changed and updated. The latest FPGA devices offer a large amount of embedded hardware blocks, which can be used to carry out optimised operations, such as addition and multiply-accumulate steps. The inclusion of dedicated DSP slices on an FPGA allows for very efficient multiplication and multiply-accumulate (MAC) operations. For example, on current Xilinx Virtex 7 FPGAs there are up to 3600 DSP48E1 slices, each with the capacity of a 18×25 bit signed multiplication and 48 bit accumulation; see Table 2 for further examples. Furthermore, the 18×25 bit signed multiplier and the 48bit accumulator are capable to run at frequency of up to 741 MHz [22].

Table 2. Examples of Available DSP Slices in Selected Xilinx Virtex 7 FPGA Devices

Xilinx Virtex 7 FPGA	No. of DSP Slices	No. Columns
XC7VX415T	2,160	18
XC7VX485T	2,800	20
XC7VX980T	3,600	20

The compact size imposes significant constraints on input word sizes and storage, which is problematic for FHE and SHE implementations with large key and ciphertext sizes. To circumvent this in so far as possible, we target one of the largest FPGAs, the Xilinx Virtex 7XC7VX980T.

Our goal is to implement a 936 bit multiplier. An un-optimised initial inference of a multiplier using ISE Design Suite reveals that a 936 bit multiplier requires 76% of the targeted device's DSP slices. Therefore this highly un-optimised large multiplier fits on to the FPGA device. However, inferring and cascading multipliers in ISE Design Suite requires exponentially many DSP slices for increasingly large multiplications. If an un-optimised implementation of a 936 bit multiplier is designed using partial products and shifts for example, this will most likely occupy over 2700 DSP slices just to implement the toy-sized multiplier. The Xilinx Core Generator can only generate multipliers for up to 64 bits long, which is much smaller than our required multiplier. Therefore an indirect approach must be taken.

In [18] the dedicated hard core functions on FPGAs are targeted to produce efficient implementations of both AES and ECC. An efficient multiplier is presented, which firstly calculates the partial products using the Comba method and then these are added to generate the final result. This technique is very suitable for large integers, as the DSP slices can be used in parallel, which allows for less device usage for the same multiplier size. Although our target multiplier is 936 bits, we use a 16 bit unsigned multiplier, as the DSP slice has an 18-bit signed multiplier and thus a 16 bit multiplier is the most suitable size to work with that fits within a DSP slice. Using this approach, a 944 bit multiplier can be designed using 59 DSP slices, where a few of the multiplications are redundant. Each of these 59 DSP slices will calculate 16×16 multiplications up to 59 times and thus a full 944 bit multiplication can be calculated in around 60 clock cycles. This multiplier can then be run multiple times to reach the appropriate multiplication size. Therefore the 150000 bit x_i is represented in 9575 16 bit words and the 944 bit multiplier is used approximately 159 times. After each multiplication in the DSP slice the x_i are shifted right, and further multiplications with the shifted x_i are carried out and accumulated in the DSP slices. The partial product adder combines these partial products. The least significant word of each of the partial products accumulated in the DSP slices is saved in a register and the remainder is added to the next partial product. This process is continued with all of the partial products consecutively to give the final output. Additionally, several multipliers of this size can be implemented in parallel to increase the performance of the multiplier in the encryption step.

Figure 1 shows a basic hardware architecture design of the Comba multiplier as proposed in [18]. The chosen 944 bit multiplicand and multiplier are both represented by 59 16 bit words, as shown by registers A and B in Figure 1; A and B represent the x_i and b_i from the selected FHE scheme. The value s is equal to the number of words, in this case 59, and w is equal to the word size, 16. This can be extended to larger sizes: $s = Multiplier\ Size\ /\ Word\ Size.$ Each of these 59 words from both A and B is input into a separate DSP slice, again as shown in Figure 1. The product of these two terms is accumulated within the DSP slices, using the internal 48 bit accumulator logic. The accumulation output is a maximum of $2w + \log_2(s)$ bits. After each multiplication, the x_i in Figure 1 are shifted left by 16 bits and a new word is input to each of the 59 DSP slices to be multiplied by b_i which is also shifted one word to the right. This process accumulates all of the partial products. These partial products are then added together as previously described. After the final output is stored in memory; the multiplier is used again 158 more times to calculate all of the parts of x_i and the output is combined to achieve the final result.

Table 3 gives conservative estimates of timings for the multiplications required in the encryption step in all four versions of the FHE scheme in [10] without considering parallel implementation of multipliers, which would considerably speed up timings. We also assume a conservative estimate of a 500MHz clock frequency for the multiplier, as the critical path goes through the DSP block. The published software timings for the encryption step in [10] requires, for example, 1 second for the toy sized encryption step in the FHE scheme. The multiplication step is one of the two bottlenecks in the encryption scheme and this suggests that the use of hardware could greatly improve the practicality of such encryption steps or indeed any step in FHE schemes which requires large integer multiplication.

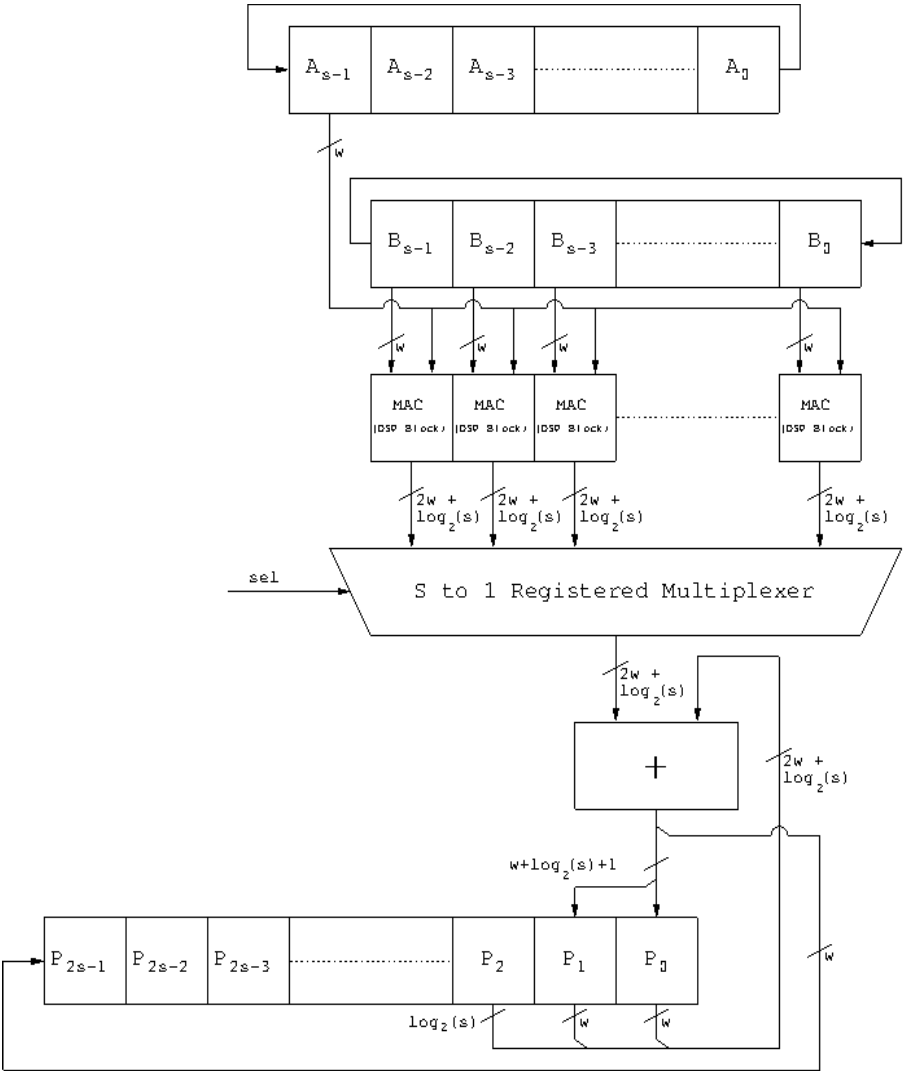


Fig. 1. Hardware Architecture of Comba Multiplier

We make the assumption that we can access the off-chip memory storage; storage of the products is an issue, especially with the larger versions of the FHE schemes but for the toy size this is not a major issue, as only 60 DSP slices are required per 944 bit multiplier. Moreover 158 of these 0.15Mbit-sized products require a total of 23.7 Mbits memory. This is manageable on the targeted Xilinx Virtex 7 XC7VX980T, as there is 68 Mbits block RAM (BRAM) available. For the large FHE scheme, each multiplication is around 19.35Mbits long and 7659 of these are required to be added, which highlights the storage issues associated with the large scheme sizes.

Table 3. DSP Slices required and estimated timings for large integer multiplier in encryption step using Comba multiplication at 500 MHz

Size of Scheme:	Toy	Small	Medium	Large
No. of multiplications required in encryption τ	158	572	2110	7659
Size of required multiplier (bits)	936× 150000	1476× 830000	2016× 4200000	2556× 19350000
Target multiplier (bits)	944×944	1488×1488	2016×2016	2560× 2560
No. of DSP slices required for target multiplier	60	94	127	161
Estimation of Clock Cycles required (multiplications not run in parallel)	1507320	30002544	558449480	9320995341
Estimated timing of all multiplications required in encryption step (secs)	0.00301	0.06001	1.11690	18.64200
Published Timing (secs) of Encryption Step in [10]	0.05	1	21	435

Additionally the transfer of data must also be considered. Not only do the parameter sizes increase with the larger versions of the FHE schemes but the number of required multiplications for encryption also increase; the issue of memory storage and access becomes a major issue and it is impossible to store the partial products or intermediate values within the memory storage on the FPGA. Obviously there is a need to make use of off-chip memory, which will require careful management so as not to become the architecture bottleneck.

We give an estimation of the timing for the multiplications required in each of these four versions using the number of multiplications required, an estimate of the number of required clock cycles to achieve the target multiplier size and the number of cycles required to achieve the full size multiplier. We do not consider parallelising the multiplications in this estimation, although this is possible, as the number of required DSP slices for the selected target multiplier for all four versions occupies less than 5% of the target FPGA DSP slices. Furthermore, we do not fully utilise the DSP slice multiplier of 18×25 bits; we could extend the 16×16 bit multiplier to a 17×24 bit unsigned multiplier for example, which would improve performance of the multiplier. Therefore Table 3 lists conservative estimates. From these results however, we can still see that the toy size version will fit on an FPGA and this could be parallelised to give an even better performance. Moreover the estimated timings for the small, medium and large schemes suggest that a hardware implementation of FHE could offer significant improvements to the practicality of FHE schemes.

Preliminary synthesis results[1] of a 944 bit multiplier show that it requires 59 DSP48E1s, and has a latency of 121 clocks: 1 for loading, 60 for multiply accumulate and 60 for partial product addition and shifting. The overall latency of the multiplier is $2s + 3$ clock cycles, where s is the number of words. To our knowledge, this is one of the first analyses into the practicality of an FPGA based implementation of crypto primitives for use in FHE schemes.

5 Conclusions

We have considered one of the most important building blocks involved in FHE schemes, large integer multiplication. We have looked at the Comba multiplication method and the possibility of targeting DSP48E1 slices on a Xilinx Virtex 7 FPGA to perform the large integer multiplication to ultimately improve the performance of FHE schemes. From the preliminary results we establish that the large integer multiplication in the encryption step for the toy scheme will fit comfortably on a single FPGA device. Furthermore the conservatively estimated timings suggest using a hardware implementation of this multiplication algorithm should improve performance of FHE schemes compared to the software implementations, especially for the larger versions of the FHE schemes [10]. This establishes the potential and justification for continuing research into hardware implementations of crypto-primitives, such as large integer multiplication, to improve the performance and hence the practicality of FHE schemes. There will however be issues with memory storage with these large versions of the FHE schemes. As this is a relatively recent area of research, there is a lot of future work still to be carried out and we are currently pursuing a hardware implementation of a complete encryption step of a FHE scheme.

References

1. Cloud Industry Forum: UK cloud adoption and trends for 2013 (2013), `http://www.cloudindustryforum.org/white-papers/uk-cloud-adoption-and-trends-for-2013`
2. El Gamal, T.: A Public Key Cryptosystem and a Signature Scheme based on Discrete Logarithms. IEEE Transactions on Information Theory 31(4), 473–481 (1985)
3. Paillier, P.: Public-Key Cryptosystems based on Composite Degree Residuosity Classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
4. Boneh, D., Goh, E.-J., Nissim, K.: Evaluating 2-DNF Formulas on Ciphertexts. In: Kilian, J. (ed.) TCC 2005. LNCS, vol. 3378, pp. 325–341. Springer, Heidelberg (2005)
5. Gentry, C.: A Fully Homomorphic Encryption Scheme. Stanford: PhD Dissertation (2009)
6. Brakerski, Z., Vaikuntanathan, V.: Efficient Fully Homomorphic Encryption from (standard) LWE. In: FOCS, pp. 97–106 (2011)

[1] Post place and route results are not presented, due to over-mapping of the i/o pins.

7. Gentry, C., Halevi, S., Smart, N.P.: Fully Homomorphic Encryption with Polylog Overhead. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 465–482. Springer, Heidelberg (2012)
8. Gentry, C., Halevi, S., Smart, N.P.: Homomorphic Evaluation of the AES Circuit. In: Safavi-Naini, R., Canetti, R. (eds.) CRYPTO 2012. LNCS, vol. 7417, pp. 850–867. Springer, Heidelberg (2012)
9. Gentry, C., Halevi, S.: Implementing Gentry's Fully-Homomorphic Encryption Scheme. In: Paterson, K.G. (ed.) EUROCRYPT 2011. LNCS, vol. 6632, pp. 129–148. Springer, Heidelberg (2011)
10. Coron, J.-S., Naccache, D., Tibouchi, M.: Public Key Compression and Modulus Switching for Fully Homomorphic Encryption over the Integers. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 446–464. Springer, Heidelberg (2012)
11. van Dijk, M., Gentry, C., Halevi, S., Vaikuntanathan, V.: Fully Homomorphic Encryption over the Integers. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 24–43. Springer, Heidelberg (2010)
12. Coron, J.-S., Mandal, A., Naccache, D., Tibouchi, M.: Fully Homomorphic Encryption over the Integers with Shorter Public Keys. In: Rogaway, P. (ed.) CRYPTO 2011. LNCS, vol. 6841, pp. 487–504. Springer, Heidelberg (2011)
13. Cousins, D., Rohloff, K., Peikert, C., Schantz, R.: An update on SIPHER (Scalable Implementation of Primitives for Homomoprhic Encryption) - FPGA implementation using Simulink. In: HPEC, pp. 1–5 (2012)
14. Brenner, M., Perl, H., Smith, M.: Practical Applications of Homomorphic Encryption. In: SECRYPT, pp. 5–14 (2012)
15. Smart, N.P., Vercauteren, F.: Fully Homomorphic Encryption with Relatively Small Key and Ciphertext Sizes. In: Nguyen, P.Q., Pointcheval, D. (eds.) PKC 2010. LNCS, vol. 6056, pp. 420–443. Springer, Heidelberg (2010)
16. Göttert, N., Feller, T., Schneider, M., Buchmann, J., Huss, S.: On the Design of Hardware Building Blocks for Modern Lattice-Based Encryption Schemes. In: Prouff, E., Schaumont, P. (eds.) CHES 2012. LNCS, vol. 7428, pp. 512–529. Springer, Heidelberg (2012)
17. Comba, P.G.: Exponentiation Cryptosystems on the IBM PC. IBM Systems Journal 29, 526–538 (1990)
18. Güneysu, T.: Utilizing Hard Cores of Modern FPGA Devices for High-Performance Cryptography. J. Cryptographic Engineering 1, 37–55 (2011)
19. Bosselaers, A., Govaerts, R., Vandewalle, J.: Comparison of Three Modular Reduction Functions. In: Stinson, D.R. (ed.) CRYPTO 1993. LNCS, vol. 773, pp. 175–186. Springer, Heidelberg (1994)
20. Karatsuba, A., Ofman, Y.: Multiplication of Many-Digit Numbers by Automatic Computers. Doklady Akad. Nauk SSSR 145, 293–294 (1962)
21. Rivest, R., Shamir, A., Adleman, L.: A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM 21, 120–126 (1978)
22. 7 Series FPGAs Overview, `http://www.xilinx.com` (accessed December 28, 2012)
23. Brenner, M., Perl, H., Smith, M.: How Practical is Homomorphically Encrypted Program Execution? An Implementation and Performance Evaluation. In: TrustCom, pp. 375–382 (2012)

Author Index