Proceedings

Learning from Authoritative
Security Experiment Results

LASER 2017

Arlington, VA, USA ◆ October 18-19, 2017

usenix
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

Proceedings of LASER 2017
Learning from Authoritative Security Experiment Results

Arlington, VA, USA
October 18–19, 2017

ISBN 978-1-931971-41-6

Table of Contents

Program

Organizing Committee

Terry Benzel (USC ISI), General Chair
Brendan Dolan-Gavitt (NYU), Program Chair
David Balenson (SRI International), Funding/Local Arrangements/Scholarships
Laura S. Tinnel (SRI International), Publicity/Web/IT Services
Fanny Lalonde Levesque (Ecole Polytechnique de Montreal), Publications
Carrie Gates (Consultant), Advisor
Greg Shannon (CMU/CERT), Advisor

Program Committee

Brendan Dolan-Gavitt (NYU), Chair
Christian Collberg (University of Arizona)
Kovila Coopamootoo (Newcastle University)
Eric Eide (Utah)
Lori Flynn (SEI/CMU)
Thomas Gross (Newcastle University)
Bart Knijnenburg (Clemson)
Jelena Mirkovic (USC Information Sciences Institute (ISI))
Pradeep Murukannaiah (RIT)
Daniela Oliveira (University of Florida)
Konrad Rieck (TU Braunschweig)
John Seymour (UMBC)
Xinyu Xing (Penn State)
Michael Zhivich (Akamai)

Workshop Sponsors

NSF

ADVANCING CYBERSECURITY SINCE 1985
ACSA
APPLIED COMPUTER SECURITY ASSOCIATES

SRI International

usenix®
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

Message from the General Chair

Welcome to the 2017 Workshop on Learning from Authoritative Security Experiment Results (LASER).

Each year, LASER focuses on an aspect of experimentation in cyber security. The 2017 workshop focus was on improving the rigor and quality of security experimentation through experimental methods and research that exemplifies sound scientific practice.

The event was structured as a workshop with invited talks and a variety of guided group discussions in order to best meet the overall workshop goals.

LASER 2017 sought research papers exemplifying the practice of science in cyber security, whether the results were positive or negative. Papers documenting experiments with a well-reasoned hypothesis, a rigorous experimental methodology for testing that hypothesis, and results that proved, disproved, or failed to prove the hypothesis were sought.

This year, many of the papers and talks for the 2017 LASER Workshop included aspects of measurement and analysis of experimental approaches. This theme was highlighted in the invited talk “The Advancement of Science in Cyber Security” by Dr. Laurie Williams from North Carolina State University, who gave a report on the NSA Lablet program efforts designed to more aggressively advance the science of cyber security. A key motivation of this work is to catalyze a shift in relevant areas towards a more organized and cohesive scientific community for a science of cyber security.

Invited speaker Dr. Josiah Dykstra’s talk “She Blinded Me with Science: Understanding Misleading, Manipulative, and Deceptive Cybersecurity” described how people are often misled, manipulated, or deceived by real and bogus science, wild claims, and marketing trickery. Dykstra’s work explores the dangers of vendor-sponsored studies, surveys, and spurious (false) correlations. Drawing on his book *Essential Cybersecurity Science,* Dykstra discussed how researchers can improve communication with security practitioners and the dangers of manipulative graphics and visualizations that work through mental shortcomings and perception or because of the data they omit.

The workshop received 15 submissions, which were each reviewed by at least 3 members of the Program Committee. The Program Committee accepted 8 full papers, which they believed embodied the workshop spirit and focus of LASER 2017.

This year, the LASER Workshop returned to its roots and was held in October and was hosted by SRI International at their Arlington, VA facility.

LASER recognizes that the future of cyber security lies with the next generation of researchers. As such, LASER sponsors students who are working to become researchers to attend and participate in the workshop. In 2017, four students received full sponsorship.

On behalf of LASER 2017, I wish to thank the many people who made this workshop possible:

- Our program chairs, who worked diligently to put together a strong technical program that would benefit the community
- The authors, who submitted papers to this workshop
- The members of the Program Committee, who carefully reviewed the submissions and participated in paper discussions
- Our organizing committee, who provided guidance and donated their time to handle everything from publicity to logistics
- The National Science Foundation, ACSA, SRI and USENIX , who provided the funding and facilities necessary to make the workshop a reality
- The attendees, without whom there would be no workshop at all. We look forward to meeting everyone at LASER 2018!

Terry Benzel, *USC Information Sciences Institute*
LASER 2017 General Chair

Understanding Malware's Network Behaviors using Fantasm

Xiyue Deng
xiyueden@isi.edu
Information Sciences Institute

Hao Shi
shihao@isi.edu
Information Sciences Institute

Jelena Mirkovic
mirkovic@isi.edu
Information Sciences Institute

Abstract

Background: There is very little data about how often contemporary malware communicates with the Internet and how essential this communication is for malware's functionality.

Aim: We aim to quantify what fraction of contemporary malware samples are environment-sensitive and will exhibit very few behaviors when analyzed under full containment. We then seek to understand the purpose of the malware's use of communication channel and if malware communication patterns could be used to understand its purpose.

Method. We analyze malware communication behavior by running contemporary malware samples on bare-metal machines in the DeterLab testbed, either in full containment or with some limited connectivity, and recording and analyzing all their network traffic. We carefully choose which communication to allow, and we monitor all connections that are let into the Internet. This way we can guarantee safety to Internet hosts, while exposing interesting malware behaviors that do not show under full containment.

Results. We find that 58% of samples exhibit some network activity within the first five minutes of running. We further find that 78% of these samples exhibit more network behaviors when ran under our limited containment, than when ran under full containment, which means that 78% of samples are environment-sensitive. Most common communication patterns involve DNS, ICMP ECHO and HTTP traffic toward mostly non-public destinations. Likely purpose of this traffic is botnet command and control. We further show that malware's network behaviors can be used to determine its purpose with 85–89% accuracy.

Conclusions. Ability to communicate with outside hosts seems to be essential to contemporary malware. This calls for better design of malware analysis environments, which enable safe and controlled communication to expose more interesting malware behaviors.

1 Introduction

Malware today evolves at an amazing pace. Kaspersky lab [1] reports that more than 300,000 new malware samples are found each day. While many have analyzed malware binaries to understand its purpose [7, 9], little has been done on analyzing and understanding malware communication patterns [17, 22]. Specifically, we do not know how much malware needs outside connectivity and what impact limited connectivity has on malware's functionality. We further do not understand which application and transport protocols are used by contemporary malware, and what is the purpose of this communication. Understanding these issues is necessary for two reasons. First, much malware analysis occurs in full containment due to legal and ethical reasons. If communication is essential to malware, then analyzing it in full containment makes what defenders observe very different from how malware behaves in the wild. Second, understanding malware communication patterns may be useful to understand its functionality, even when malware code is obfuscated or encrypted.

We hypothesize that communication may be essential to malware for multiple reasons. First, contemporary malware is becoming *environment-sensitive* and may test its environment before it reveals its functionality [7, 14]. If constrained environment is detected, malware may modify or abort its behavior. Second, much of malware functionality today relies on a functional network [13, 24]. Malware often downloads binaries needed for its functionality from the Internet, or connects into command and control channel to receive instructions on its next activity [25]. Without network access such malware is an empty shell, containing no useful code. Third, malware functionality itself may require network access. Advanced persistent threats [15] and keyloggers collect sensitive information on users' computers, but need network access to transfer it to the attacker. DDoS attack tools, scanners, spam and phishing malware require network access to send malicious traffic to their targets. Without connectivity, such malware will become dormant.

We test our hypothesis by analyzing 2,994 contemporary malware samples, chosen to represent a wide variety of functional behaviors (e.g., key loggers, ransomware, bots, etc.). We analyze each sample under full and under partial containment, for five minutes, and record all network traffic. Our partial containment is designed to carefully allow select malware communication attempts into the Internet, when we believe this is necessary to reveal more interesting behaviors. All traffic is monitored for signs of malicious intent (e.g., DDoS or scanning) and quickly aborted if these are detected. This way we can guarantee safety to the Internet from our experimentation.

We find that 58% of samples exhibit some network behavior, and that 78% of these samples exhibit more network behaviors when ran under our partial containment, than when ran under full containment, which means they are environment-sensitive. Most malware samples send DNS, ICMP ECHO and HTTP traffic, and contact obscure destinations rather than popular servers. Likely purpose of these malware communication attempts is command and control communication, and new binary download. We further show that malware's network behaviors can be used to determine its purpose with 85–89% accuracy. We also show that our partial containment is safe for the Internet. In twelve weeks of running, we have received no abuse complaints and our IP addresses have not been blacklisted.

All the code developed in our work and the materials used in our evaluation are available at our project website: `https://steel.isi.edu/Projects/fantasm/`

2 Related Work

In this section, we summarize related work on understanding malware behaviors.

Most malware analysis works focus on analyzing system traces and malware binaries [20, 21]. There are fewer efforts on analyzing the semantics of malware's network behavior. The Sandnet article [22] provides a detailed, statistical analysis of malware's network traffic. The authors give an overview of the popularity of each protocol that malware employs. However, they do not attempt to understand the high-level semantics of malware's network conversations, and this is the contribution we make. Our work also updates results from [22] with communication patterns of contemporary malware. For example, we observe that ICMP ECHO has become the second most popular protocol used by malware. Morales et al. [17] define seven network activities based on heuristics and analyze malware for prevalence of these behaviors. Yet this work does not provide insight into a malware sample's purpose (e.g., worm, scanner, etc.) and it may miss behaviors other than those seven select ones. Our work complements this work and covers a richer set of behaviors, composed out of some basic communication patterns discussed in Section 5.3.

3 Fantasm

In this section, we describe the goals for our Fantasm system, our partial containment rules and how we ensure safety to the Internet from our experimentation.

3.1 Goals

Our goal in designing the Fantasm system was to support safe and productive malware experimentation. *Safe* means that we wanted to ensure that we do no harm to other Internet hosts with our experiments. *Productive* means that we wanted to ensure that as many as possible outgoing communication requests, launched by malware, receive a reply to that malware may move on to its next activity.

3.2 Partial Containment

One could achieve safety in full containment, without letting any traffic out of the environment. But because malware is environment-sensitive this would not lead to productive experimentation. One could also experiment in an open environment, where all the traffic is let out. But this would not be safe since the analysis environment could become a source of harmful scans, DDoS attacks and worm infections, which harm other Internet hosts. Due to ethical consideration, no organization would support such analysis for long.

To meet our goals we decided to experiment with malware in *partial* containment, where we selectively decide which malware flows to allow to reach into the Internet based on our assessment of their potential risk to the Internet, which is conformant to the ethical principles for information and communication technology research [11]. We also attempt to handle each outgoing flow in full containment first, by impersonating remote servers and crafting generic replies. This further reduces the amount of traffic we must let out and improves experimentation safety. We now explain how we assessed this risk and how we enforced the containment rules.

Based on a malware flow's purpose we distinguish between the following flow categories: benign (e.g., well-formed requests to public servers at a low rate), e-mail (spam or phishing), scan, denial of service, exploit and C&C (command and control). Potential harm to Internet hosts depends on the flow's category. Spam, scans

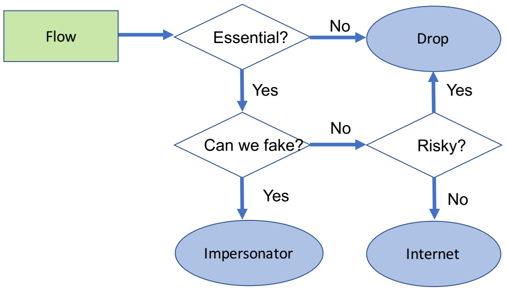


Figure 1: Flow handling: how we decide if an outgoing flow will be let out, redirected to our impersonators or dropped.

and denial of service are harmful only in large quantities – letting a few such packets out will usually not cause severe damages to their targets, but it may generate complaints from their administrators. On the other hand, binary and text-based exploits are destructive, even in a single flow. The C&C and benign communications are not harmful and usually must be let out to achieve productive malware experimentation.

The challenge of handling the outside communication with a fixed set of rules lies in the fact that the flow's purpose is usually not known a priori. For example a SYN packet to port 80 could be the start of a benign flow (e.g., a Web page download to check connectivity), a C&C flow (to report infection and receive commands for future activities), an exploit against a vulnerable Web server, a scan or a part of denial-of-service attack. We thus have to make a decision how to handle a flow based on incomplete information, and revise this decision when more information is available. Our initial decision depends on how essential we believe the flow is to the malware's continued operation, how easy it is for us to fabricate responses without letting the flow out of our analysis environment, and how risky it may be to let the flow out into the Internet. For essential flows whose replies are predictable, we develop generic services that provide these predictable responses and do not allow these flows into the Internet. We call these services "impersonators". Essential flows whose replies are not predictable, and which are not risky, are let out into the Internet, and closely observed lest they exhibit risky behavior in the future. Non-essential flows and essential but risky flows are dropped. Figure 1 illustrates our flow handling.

Traffic that we let out could be misused for scanning or DDoS if we let it out in any quantity. We actively monitor for these activities and enforce limits on the number of suspicious flows that a sample can initiate. We define a *suspicious* flow as a flow, which receives no replies from the Internet. For example, a TCP SYN to port 80 that does not receive a TCP SYN-ACK would be a part of a suspicious flow. Similarly a DNS query that receives no reply is a suspicious flow. Suspicious flows will be present if a sample participates in DDoS attacks or if it scans Internet hosts. If the sample exceeds its allowance of suspicious flows, we abort this sample's analysis.

We summarize our initial decisions and revision rules in Table 1. We consider DNS, HTTP and HTTPS flows as essential and non-risky, whose replies we cannot fake. We make this determination because many benign and C&C flows use these services to obtain additional malware executables, report data to the bot master and receive commands. Among our samples, DNS is used by 62%, HTTP by 35%, and HTTPS by 10% of samples (Section 4).

We consider FTP, SMTP and ICMP flows as essential flows with predictable replies. We forward these to our corresponding impersonators (Figure 1). These are machines in our analysis environment that run the given service, and are configured to provide generic replies to service requests. We redirect ICMP ECHO requests to our service impersonators and fake positive replies. We drop other ICMP traffic.

Our FTP service impersonator is a customized, permissive FTP service that positively authenticates when any user name and password are supplied. This setting can handle all potential connection requests from malware. If malware tries to download a file, we will create one with the same extension name, such as `.exe`, `.doc`, `.jpg`, and others. We save uploaded files for further analysis. For SMTP service, we set up an Email server that can reply with a "250 OK" message to any request. Our ICMP impersonator sends positive replies to any ICMP ECHO request.

4 Experimentation Goals Environment and Design

In this section, we discuss our experimentation goals, environment and experiment design.

4.1 Experimentation Goals

We wanted to observe and analyze communication patterns of malware. This necessitated identification of a relatively recent, representative set of malware binaries and running them in partial containment, while recording their communication. We further needed a way to quickly and automatically restore "clean state" of machines between malware samples

Goal	Action	Targeted Services
Elicit malware behavior	Forward	DNS, HTTP, HTTPS
	Redirect	FTP, SMTP, ICMP ECHO
Restrict forwarded flows	Drop	Other services
	Limit	Number of suspicious flows

Table 1: Flow policies for partial containment

4.2 Experimentation Environment

We experiment with malware samples in the DeterLab testbed [8]. DeterLab [8] enables remote remote experimentation and automated setup. An experimenter gains exclusive access and sudoer privileges to a set of physical machines and may connect them into custom topologies. The machines run an operating system and applications of a user's choice. Experimental traffic is usually fully contained, and does not affect other experiments on the testbed, nor can it get out into the Internet. In our experiments, we leverage a special functionality in the DeterLab testbed, called "risky experiment management", which allows exchange of some user-specified traffic between an experiment and the Internet. We specify that all DNS, HTTP and HTTPS traffic should be let out.

We run malware samples on several machines in a DeterLab experiment, which we will call Inmates. We hijack default route on Inmates and make all their traffic to the Internet pass through a special machine in our experiment, called Gateway. This Gateway implements our partial containment rules. We implement all of the service impersonators on a single physical machine. Each machine has a 3GHz Intel processor, 2GB of RAM, one 36Gb disk, and 5 Gigabit network interface cards.

To hide the fact that our machines reside within DeterLab from environment-sensitive malware we modify the system strings shown in Table 2. For example, we replace the default value ("Netbed User") of "Registered User" with a random name, e.g., – "Jack Linch". Therefore, malware will not detect the existence of DeterLab by searching for such strings.

4.3 Experiment Design

We run each malware sample under a given containment strategy (full or partial) for five minutes and record all network traffic at the Gateway. After analyzing each malware sample, we must restore Inmates to a clean state. We take advantage of the OS setup functionality provided by DeterLab to implement this function. We first perform certain OS optimization to reduce the size of OS image and thus shorten the time needed to load the image when restoring clean state. This modified OS is saved into a snapshot using the disk imaging function of DeterLab. This step takes a few minutes but is carried out only once for our experimentation. Later, whenever we need to restore the system after analyzing a malware sample, we reload the OS image using DeterLab's `os_load` command.

Our environment could also be used to study behavior of benign code, but this is outside of the scope of this research.

4.4 Malware Dataset

We obtained a recent set of malware samples by downloading 29,319 malware samples between March 4th and March 17th, 2017 from OpenMalware [2]. In order to obtain a balanced dataset we establish ground truth about the purposes of these samples by submitting their md5 hashes to VirusTotal [5]. We retrieve 28,495 valid reports. Each report contains the analysis results of about 50∼60 anti-virus (AV) products for a given sample. We keep the samples that were labeled as malicious by more than 50% AV products. This leaves us with 19,007 samples.

Concise Tagging. Each AV product tags a binary with vendor-specific label, for example, "worm.win32.allaple.e.", "trojan.waski.a", "malicious_confidence_100% (d)", or just "benign". As demonstrated in [6], AV vendors disagree not only on which tag to assign to a binary, but also how many unique tags exist. To overcome this limitation, we devise a translation service that translates vendor-specific tags into a nine concise, generic tags, such as: worm, trojan, virus, etc. We learn the translation rules by first taking a union of all the tags assigned by the AV products (74,443 in total), and then manually extracting common keywords out of them that signify a given concise category. Finally, we tag the sample with the concise category that is assigned by the majority of the AV products. Table 3 shows the breakdown of our samples over our concise tags.

We then randomly select 2,994 out of the 19,007 samples, trying to select equal number of samples from each category, to achieve diversity and form a representative malware set. We continue working with this malware set.

Key Name	Default in DeterLab	Our Modification
Registered User	"Netbed User"	Random name, e.g., "Jack Linch"
Computer Name	"pc.isi.deterlab.net"	Random name, e.g., "Jack's PC"
Workgroup	"EMULAB"	"WORKGROUP"

Table 2: Minimizing artifacts of DeterLab.

Table 3: Concise Tagging of Malware Samples

Categories	Samples	Categories	Samples
Virus	6,126/32%	Riskware	409/2%
Trojan	6,040/32%	Backdoor	197/1%
Worm	4,227/22%	Bot	45/<1%
Downloader	984/5%	Ransomware	17/<1%
Adware	962/5%	Total	19,007

Protocols	Samples	Protocols	Samples
DNS	1081/62%	1042	65/4%
ICMP echo	818/47%	799	33/2%
HTTP	600/35%	6892	25/1%
65520	237/14%	11110	17/1%
HTTPS	173/10%	11180	17/1%
SMTP	75/4%	FTP	12/1%

Table 4: Top 12 application protocols used by malware, and the number and percentage of samples that use them.

5 Results

In our evaluation, out of 2,994 malware samples in our malware set 1,737 samples exhibited some network activity during a run. The remaining samples may be dormant, waiting for some trigger or may simply exhibit too small communication frequency, which we cannot observe given our experiment duration (5 minutes).

5.1 Partial Containment Exposes More Malware Behavior

We measure the quantity of observable malware behavior by counting the number of network flows recorded during experimentation. Out of 1,737 samples that exhibit any network behavior, 1,354 (78%) generate more flows under partial containment than under full containment. This supports our hypothesis that network connectivity is essential for malware functionality, and that most malware samples are environment-sensitive. Out of 1,737 that exhibit network behavior there were 9,304,083 outgoing flows generated during our 5 minute experimentation interval. Out of these 9,304,083 flows, our impersonators could fake replies to 9,270,831 (99.64%) of them. We had to let 2,295 flows (0.02%) out into the internet because we could not fake their replies and they were deemed essential. Finally 30,957 flows (0.33%) were dropped because we did not have an impersonator for their protocol, but they were deemed too risky to be let out. We hope to develop more impersonators in the future, and thus further reduce risk to the Internet.

As a proof of how safe our experimentation was, during twelve weeks that we ran, we received no abuse complaints. We also analyzed 203 IP blacklists from 56 well-known maintainers (e.g., [3]), which contain 178 million IPs and 34,618 /16 prefixes for our experimentation period. Our external IP was not in any of the blacklists, which further supports our claim that no harmful traffic was let out.

5.2 Malware Communication Patterns

Table 4 shows the top 12 application protocols used by our malware dataset. DNS is used by 62% of samples and its primary use seemed to resolve the IPs of the domains that malware wishes to contact. ICMP was used by 47% of samples, likely to test reachability, either to detect if malware is running in a contained environment or to identify live hosts that may later be infected, if vulnerable. HTTP (35% of samples) and HTTPS (10% of samples) are likely used to retrieve new binaries, as we find many of these connections going out to file-hosting services. Port 65520 is mostly used by a virus that infects executable files and opens a back door on the compromised computers. The SMTP protocol is used to spread spam.

Samples in our malware dataset queried a total of 5,548 different domains, among which `zief.pl` (14%) and `google.com` (11%) are the most popular domains. We query these domains from `alexa.com`[1] , which has the records for 341 (6%) domains, as shown in Figure 2. We find that only 1% of the domains have ranks lower than 10,000, 5% have higher ranks and 94% of domains are not recorded by `alexa`. For the domains whose rank is lower than 10,000, most are web portals, such as YouTube and many are file storage services, like Dropbox. We manually check 20 domains that have no record in `alexa`, and none had a valid DNS record. This suggests that malware may use portal websites either test

[1] In our future work we will look to use a more robust representation of popular domains, like proposed by Metcalf et al in [16]

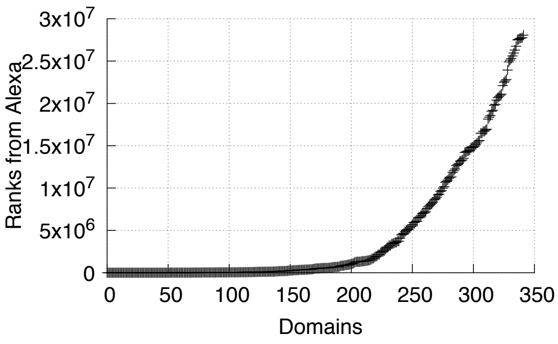


Figure 2: Ranks of domains from `alexa.com`.

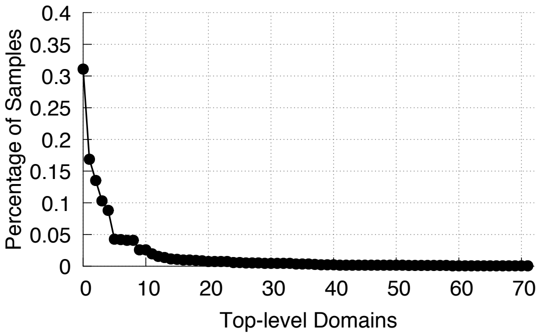


Figure 3: Popularity of top-level domains in our observed malware communications.

network reachability or for file transfer, and it may use private servers for file transfer or for C&C communication.

We classify the queried names based on their top-level domain, e.g., `.com` or `.net`. We find a total of 72 distinct top-level domains, as shown in Figure 3. The Top 3 of these domains are shown in Table 5. The `.com` is the most popular top-level domain, which is queried by 540 (31%) samples. The third column in Table 5 shows the top 3 queried domains in each top-level category. These domains contain 53 country codes, with Poland, Germany, and Netherlands being the top three countries. This means that malware in our dataset predominantly targeted European victims.

5.3 Summarizing Malware Communication

We now explore how to summarize malware communication so we can further investigate common patterns in how malware uses the Internet. Our goal was to create a concise and human-readable digest of malware's communication starting from recorded tcpdump logs. We call this representation *NetDigest*.

Top-level	Samples	Second-level	Samples
.com	540/31%	google.com	187/35%
		msrl.com	73/14%
		ide.com	73/14%
.pl	293/17%	zief.pl	244/83%
		brenz.pl	26/9%
		ircgalaxy.pl	22/8%
.net	235/14%	secureserver.net	73/31%
		surf.net	68/29%
		aol.net	65/28%

Table 5: Popularity of domains in malware DNS queries.

Protocol	[Attribute: Value]
All	[LocalPort: `integer`],‡ [NumPktSent: `integer`],‡ [NumPktRecv: `integer`],‡ [PktSentTS: `float_list`],‡ [PktRecvTS: `float_list`],‡ [PayloadSize: `integer_list`]†
DNS	[Server: `IP_address`],‡ [QueryType: `string`],‡ [CNAME: `string`],‡ [ResponseType: `IP_address list`]‡
HTTP/FTP	[Server: `IP_address`],‡ [Proactive: `boolean`],‡ [GotResponse: `boolean`],‡ [Code: `integer`]‡, [Download: `file_type`],† [Upload: `file_type`]†
SMTP	[Server: `IP_address`],‡ [EmailTitle: `string`],* [Recipients: `string`],* [BodyLength: `integer`],* [ContainAttachment: `boolean`],* [AttachmentType: `string`]*
ICMP	[RequestIP: `IP_address`],* [NumRequests, `integer`]*

‡ Occur exactly once
† May have zero or more occurrences
* Have at least one occurrence

Table 6: NetDigest of a session.

We start by splitting a malware's traffic into flows based on the communicating IP address and port number pairs, and the transport protocol. We call each such flow a "session". Then, for each session, we extract the application protocol employed and devise a list of {attribute: value} pairs for this protocol, as shown in Table 6.

The first row of Table 6 shows the information that we will extract for all types of application protocols. For example, "LocalPort" denotes the local IP port used by malware, which is an integer. This attribute appears only once for a single session, and is derived from the definition of a session. The "NumPktSent" means the total number of packets sent by malware in an individual session. The "PktSentTS" is a list of Unix epoch time of all the packets sent by malware. Finally, we also maintain a list of each packet's payload size.

The DNS protocol has one attribute "Server", which has the value of `IP_address` that the query is sent to. For the domain queried by malware, the `QueryType` can be address record (`A`), mail exchange record (`MX`), pointer record (`PTR`), or others. For the response sent back by DNS server, we first save its canonical name, if any, in a CNAME field. Then, we extract the response type and corresponding values and assign them to the ResponseType field.

For an HTTP or FTP session, we first take note of the server's IP address in the ServerIP field. Then, we use boolean values to denote if this session is initiated by malware ("Proactive") and if malware receives any response from Internet host ("GotResponse"). If the outside server replies to malware, we classify the following packets as "Download" or "Upload" based on the direction of the bulk volume of data. We also extract the file type being transferred.

For an SMTP message, we extract the server IP address, Email title, recipients, and body length. We also use a boolean value to note whether the message has an attachment and save the attachment's file type in a string.

For the ICMP protocol, we extract the destination IP address into the RequestIP field. We also save the number of requests in NumRequests field.

After we build the lists of attribute-value pairs for all the sessions produced by a malware sample, we sort the lists based on their first timestamps. The final, sorted list of session abstractions is called the *NetDigest*.

One sample NetDigest is shown in Figure 4 for the sample tagged as Trojan by AV products. At the beginning, this sample queries a domain (`ic-dc.deliverydlcenter.com`) using the default DNS server that is part of our impersonator set. Our DNS server acts as a recursive resolver and obtains and returns the actual mapping. Then, this sample downloads a picture and blob files from the first IP address returned. However, for the remaining Internet hosts, this sample just establishes connections with them but does not download or upload any information. For example, the second domain (`www.1-ads.com`) suggests that it is an advertising website, but no payload is downloaded from this website (session starting at timestamp `1488068896.977464`). In addition, some IPs are unreachable at the time of our execution, such as `52.85.83.112`.

5.4 Classifying Malware by Its Network Behavior

We now explore if unknown malware could be classified based on its communication patterns. Current malware classification relies on binary analysis. Yet, this approach has a few challenges. First, malware may use packing or encryption to obfuscate its code, thus defeating binary analysis. Second, malware may be environment-sensitive and may not exhibit interesting behavior and code if ran in a virtual machine or debugger, which are usually used for binary analysis. We thus explore malware classification based on its communication behavior, reasoning that malware may obfuscate its code but it must exhibit certain key behaviors to achieve its basic functionality. For example, a scanner must scan its targets and cannot significantly change this behavior without jeopardizing its functionality.

In our classification we divide our malware set into a training and a testing set. We then apply machine learning to learn associations on the training set between some features of malware communication, which we describe next, and our concise labels denoting malware purpose. Finally, we attempt to classify the malware in the testing set and report our success rate.

Extracting Features. We start with 83 select features, extracted out of the malware's NetDigest, as shown in Table 7.

We abstract malware's network traffic into four broad categories: Packet, Session, Protocol, and Content. For the Packet category, we divide it into three subgroups: Header, Payload, and Statistics. In the Header subgroup, we count the number of distinct IPs that a sample's packets have been sent to. In addition, we also look up the geographical locations of the IPs from the GeoLite [4] database, including the countries and continent they reside in. We chose these features because it is known that certain classes of malware target Internet hosts in different countries. In the Payload subgroup, we calculate the total size of payload in bytes. Furthermore, we compute the following statistics for both sent and received volume, the packet counts and the packet timing: minimum, maximum, mean, and standard deviation.

For the Session category, we consider all packets that are exchanged between malware and a single IP address. For these packets, we divide them into different *sessions* according to the local ports used by malware. For each session, we determine if its direction is proactive or passive, depending on whether the malware initiates the session or not. We say the Result of a session is successful if malware initiates the session and receives any responses from the host. We further calculate the number of TCP SYN packets, which can be used to detect SYN flood attacks. We also record the number of sessions per IP, which can be useful to further establish communication purpose. For example, in our evaluation, we find that one sample launches one short session with the first IP and then initiates multiple sessions with the second one for download. This network behavior indicates that the first IP serves as a master, directing the malware sample to the second, which acts as a file server.

For the Protocol category, we extract features for different types of application protocols. For example, for the DNS we summarize the number of distinct domains queried by malware in their DNS query and response packets. For HTTP, we count the number of packets carrying specific HTTP status codes, such as 200

```
1488068895.052901: DNS - [Server: 10.1.1.3], [A: ic-dc.deliverydlcenter.com],
                         [CNAME: N/A], [A: 52.85.83.81, 52.85.83.112,
                          52.85.83.132, 52.85.83.4, 52.85.83.96, 52.85.83.56,
                          52.85.83.32, 52.85.83.37]
1488068895.154335: HTTP - [Server: 52.85.83.81], [Proactive: True], [GotResponse: True],
                          [Download: blob], [Download: .png], [Download: blob]
1488068895.948346: HTTP - [Server: 52.85.83.81], [Proactive: True], [GotResponse: True]
1488068896.767094: DNS - [Server: 10.1.1.3], [A: www.1-1ads.com], [CNAME: n135adserv.com],
                         [A: 212.124.124.178]
1488068896.977464: HTTP - [Server: 212.124.124.178], [Proactive: True], [GotResponse: True]
1488069110.044756: DNS - [Server: 10.1.1.3], [A: ic-dc.deliverydlcenter.com], [CNAME: N/A],
                         [A: 52.85.83.56, 52.85.83.112, 52.85.83.96, 52.85.83.37,
                          52.85.83.81, 52.85.83.4, 52.85.83.132, 52.85.83.32]
1488069110.049507: DNS - [Server: 10.1.1.3], [A: ic-dc.deliverydlcenter.com], [CNAME: N/A],
                         [A: 52.85.83.32, 52.85.83.37, 52.85.83.56, 52.85.83.112,
                          52.85.83.96, 52.85.83.132, 52.85.83.4, 52.85.83.81]
1488069110.338822: HTTP - [Server: 52.85.83.81], [Proactive: True], [GotResponse: False]
1488069110.342816: HTTP - [Server: 52.85.83.81], [Proactive: True], [GotResponse: False]
1488069131.273458: HTTP - [Server: 52.85.83.112], [Proactive: True], [GotResponse: False]
1488069131.277206: HTTP - [Server: 52.85.83.112], [Proactive: True], [GotResponse: False]
1488069152.304031: HTTP - [Server: 52.85.83.132], [Proactive: True], [GotResponse: False]
1488069152.308025: HTTP - [Server: 52.85.83.132], [Proactive: True], [GotResponse: False]
1488069173.334854: DNS - [Server: 10.1.1.3], [A: ic-dc.deliverydlcenter.com], [CNAME: N/A],
                         [A: 52.85.83.32, 52.85.83.132, 52.85.83.96, 52.85.83.81,
                          52.85.83.4, 52.85.83.56, 52.85.83.112, 52.85.83.37]
1488069173.338605: DNS - [Server: 10.1.1.3], [A: ic-dc.deliverydlcenter.com], [CNAME: N/A],
                         [A:52.85.83.32, 52.85.83.132, 52.85.83.112, 52.85.83.56,
                          52.85.83.81, 52.85.83.4, 52.85.83.37, 52.85.83.96]
1488069173.381571: HTTP - [Server: 52.85.83.4], [Proactive: True], [GotResponse: False]
1488069173.383566: HTTP - [Server: 52.85.83.4], [Proactive: True], [GotResponse: False]
```

Figure 4: Example NetDigest (md5: `0155ddfa6feb24c018581084f4a499a8`).

Categories	**Subgroups**	**Features (83 in total)**
Packet	Header	Distinct number of: IPs, countries, continent, and local ports
	Payload	Total size in bytes; Sent/received: total number, minimum, maximum, mean, and standard variance
	Statistics	Sent/received packets: total number, rate; Sent/received time interval: min, max, mean, and standard variance
Session	Direction	Proactive (initiated by malware) or passive (initiated by Internet servers)
	Result	Succeeded or failed
	Statistics	Total number of SYNs sent; Number of sessions per IP: minimum, maximum, mean, and standard variance
Protocol	DNS	Number of distinct domains queried by malware
	HTTP	Number of replies received per reply code: 200, 201, 204, 301, 302, 304, 307, 400, 401, 403, 404, 405, 409, 500, 501, 503; Method: GET, POST, HEAD
	ICMP	Total number of packets; Number per IP: min, max, mean, and standard variance
	Other	Ports: total number of distinct ports, top three used
Content	Files	php, htm, exe, zip, gzip, ini, gif, jpg, png, js, swf, xls, xlsx, doc, docx, ppt, pptx, blob
	Host info	OS id, build number, system language, NICs
	Registry	Startup entries, hardware/software configuration, group policy
	Keyword	Number of: “mailto”, “ads”, “install”, “download”, “email”

Table 7: Features extracted from a malware’s NetDigest for classification purpose.

Algorithms	Rank 1	Rank 2	Rank 3
Decision Tree	242/89%	257/94%	259/95%
Support Vector	231/85%	259/95%	265/97%
Multi-layer Perception	231/85%	257/94%	262/96%

Table 8: Classification results: Rank 1 – our label was the top label assigned by AV products, Rank 2 – our top label was in the top 2 labels assigned by AV products, Rank 3 – our top label was among top 3 assigned by AV products.

(OK). Some malware samples behave differently based on the returned status code. For ICMP, we calculate minimum, maximum, average and standard deviation of packet counts. For non-standard IP ports, we maintain a set of distinct port numbers and calculate the top three ports targeted by each malware sample.

For the Content category, we investigate the payload content carried in HTTP packets, because this is the top application protocol used by malware in our experiments. We then use regular expressions to extract files from hyperlinks in HTTP content, and interpret their extensions. Sometimes the content is binary, and we tag it as blob. We also attempt to identify, using regular expressions, if payload contains host information and Windows registries that are typically reported to bot masters. Finally, we collect the frequencies of select keywords that are may indicate a malware purpose, such as “ads”.

Classification Results. We investigate three popular classification methods in machine learning area – decision trees [10], support vector machines [12], and multilayer perception [23]. We implement these algorithms and standard data pre-processing (data scaling and feature selection) through a Python package Scikit [18].

We use 80% of this data set for training and the remaining 20% of samples for testing. The results are shown in Table 8. Since malware today has very versatile functionality, it may be possible that a sample exhibits behavior that matches multiple labels. We denote as “Rank 1” the case when our chosen label matches the top one concise label chosen by the majority of AV products. When it matches one of top two labels, we denote this as “Rank 2” and if it matches one of top three labels, we denote it as “Rank 3”. Our Rank 1 success rate ranged from 85 % (support vectors and multi-layer perception) to 89% (decision trees), which is very good performance. When we allow for a match between top two labels (Rank 2), our success rate climbs to 94–95%. And if we count match with any of the top three labels as a success (Rank 3), our rate climbs to 95–97%. Based on the typical performance of applying machine learning techniques in malware analysis [19], we conclude that our NetDigest representation can lead to very accurate malware classification, based only on observed communication patterns.

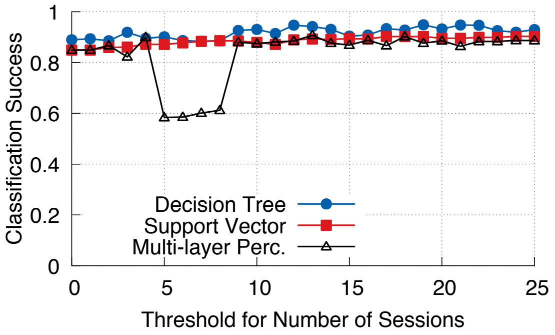


Figure 5: Classification precision as number of sessions grows.

We further investigated the root causes of our misclassifications in Rank 1 that later became a success under Rank 2 or Rank 3 criteria. Toward this goal, we manually examined `pcap` traces of related samples. We find that all these samples exhibit limited network behavior that was not sufficient for classification. For example, one sample queries a domain and then establishes a connection with the HTTP server. However, no payload is downloaded or uploaded, and thus this behavior may match any malware category.

To investigate the relationship between classification accuracy and the number of sessions observed in malware communication we perform several iterations of the classification experiment. In each iteration filter out samples that launched fewer than *N* sessions. We then divide the remaining samples into training and testing set in 80%/20% ratio, train on the training set, perform the classification on the testing set, and report the success rate. We vary *N* from 1 to 25. The evaluation results are shown in Figure 5. The x-axis of Figure 5 denotes our limit on the number of sessions in a given run – *N* and the y-axis shows the classification success rate for each algorithm, corresponding to our Rank 1 criterion, on the testing set. Figure 6 shows the number of samples that generated *N* or fewer sessions in the training and the testing set together. Overall, all three of the classification methods performed well and were stable, except for multi-layer perception when session quantity is between 5 to 8. After investigating these sessions, we found that they do not have enough distinguishing feature values for multi-layer perception algorithm. The small variance of the input are further reduced by the intermediate calculation (hidden layers) of the algorithm [18]. The classification success rate increased slightly as the limit on number of sessions increased, from 88% at 1 session to 93% at 25 sessions. Thus longer observations increase

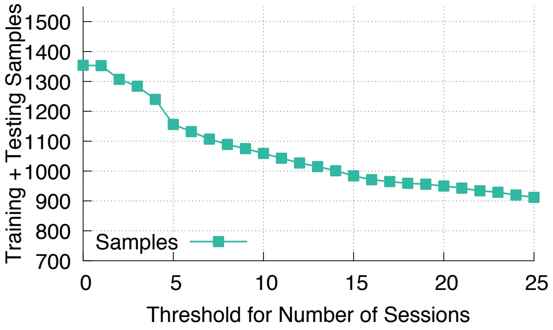


Figure 6: Number of samples as the limit on number of sessions grows.

classification accuracy but not by a lot.

6 Conclusions

In this work, we investigate how essential Internet connectivity is for malware functionality. We find that 58% of diverse malware samples initiate network connections within the first five minutes and that 78% of these samples will become dormant in full containment. We further provide breakdown of popular communication patterns and some evidence as to the purpose of these communications. Finally we show that malware communication behaviors ca be used for relatively accurate (85–89%) inference of a sample's purpose.

As future work, we will extend our framework to include analysis system-level activities for better understanding of a malware's purpose, and will seek to improve our generic impersonators to further reduce the cases when traffic must be let outside of the analysis environment.

7 Acknowledgments

This material is based upon work supported by the Department of Homeland Security under Contract No. HSHQDC-16-C-00024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Homeland Security.

References

[1] Kaspersky Lab, 323,000 New Malware Samples Found Each Day. http://www.darkreading.com/vulnerabilities---threats/kaspersky-lab-323000-new-malware-samples-found\\-each-day/d/d-id/1327655, 2016.

[2] ISC Tech Georgia, Open Malware. http://oc.gtisc.gatech.edu/, 2017.

[3] Master Feeds, Bambenek Consulting Feeds. http://osint.bambenekconsulting.com/feeds/, 2017.

[4] MaxMind, GeoLite Legacy Downloadable Databases. http://dev.maxmind.com/geoip/legacy/geolite/, 2017.

[5] VirusTotal. https://www.virustotal.com/en/, 2017.

[6] M. Bailey, J. Oberheide, J. Andersen, Z. M. Mao, F. Jahanian, and J. Nazario. Automated classification and analysis of internet malware. In *RAID*, volume 4637, pages 178–197. Springer, 2007.

[7] D. Balzarotti, M. Cova, C. Karlberger, E. Kirda, C. Kruegel, and G. Vigna. Efficient detection of split personalities in malware. In *NDSS*, 2010.

[8] T. Benzel. The Science of Cyber-Security Experimentation: The DETER Project. In *Annual Computer Security Applications Conference (ACSAC)*, 2011.

[9] P. M. Comparetti, G. Salvaneschi, E. Kirda, C. Kolbitsch, C. Kruegel, and S. Zanero. Identifying dormant functionality in malware programs. In *IEEE Symposium on Security and Privacy*, pages 61–76, 2010.

[10] G. De'ath and K. E. Fabricius. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192, 2000.

[11] D. Dittrich and E. Kenneally. The menlo report: Ethical principles guiding information and communication technology research. *US Department of Homeland Security*, 2012.

[12] I. Guyon, B. Boser, and V. Vapnik. Automatic capacity tuning of very large vc-dimension classifiers. In *Advances in neural information processing systems*, pages 147–155, 1993.

[13] T. Holz, M. Engelberth, and F. Freiling. Learning more about the underground economy: A case-study of keyloggers and dropzones. *Computer Security–ESORICS*, pages 1–18, 2009.

[14] M. Lindorfer, C. Kolbitsch, and P. Milani Comparetti. Detecting environment-sensitive malware. In *Recent Advances in Intrusion Detection*, pages 338–357. Springer, 2011.

[15] S.-T. Liu, Y.-M. Chen, and S.-J. Lin. A novel search engine to uncover potential victims for apt investigations. In *IFIP International Conference on Network and Parallel Computing*, pages 405–416, 2013.

[16] L. B. Metcalf, D. Ruef, and J. M. Spring. Open-source measurement of fast-flux networks while considering domain-name parking. In *Proceedings of the Learning from Authoritative Security Experiment Results Workshop*, 2017.

[17] J. A. Morales, A. Al-Bataineh, S. Xu, and R. Sandhu. Analyzing and exploiting network behaviors of malware. In *International Conference on Security and Privacy in Communication Systems*, pages 20–34, 2010.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[19] L. Portnoy, E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. In *Proceedings of ACM CSS Workshop on Data Mining Applied to Security*, 2001.

[20] K. Rieck, T. Holz, C. Willems, P. Düssel, and P. Laskov. Learning and classification of malware behavior. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 108–125, 2008.

[21] K. Rieck, P. Trinius, C. Willems, and T. Holz. Automatic analysis of malware behavior using machine learning. *Journal of Computer Security*, 19(4):639–668, 2011.

[22] C. Rossow, C. J. Dietrich, H. Bos, L. Cavallaro, M. Van Steen, F. C. Freiling, and N. Pohlmann. Sandnet: Network traffic analysis of malicious software. In *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, pages 78–88. ACM, 2011.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[24] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The underground economy of spam: A botmaster's perspective of coordinating large-scale spam campaigns. *LEET*, 11:4–4, 2011.

[25] K. Thomas, D. Yuxing, H. David, W. Elie, B. C. Grier, T. J. Holt, C. Kruegel, D. McCoy, S. Savage, and G. Vigna. Framing dependencies introduced by underground commoditization. In *Proceedings (online) of the Workshop on Economics of Information Security*, 2015.

Open-source Measurement of Fast-flux Networks While Considering Domain-name Parking

Leigh B. Metcalf
Software Engineering Institute
Carnegie Mellon University

Dan Ruef
Software Engineering Institute
Carnegie Mellon University

Jonathan M. Spring
Software Engineering Institute
Carnegie Mellon University

Abstract

Background: Fast-flux is a technique malicious actors use for resilient malware communications. In this paper, domain parking is the practice of assigning a nonsense location to an unused fully-qualified domain name (FQDN) to keep it ready for "live" use. Many papers use "parking" to mean typosquatting for ad revenue. However, we use the original meaning, which was relevant because it is a potentially confounding behavior for detection of fast-flux. Internet-wide fast-flux networks and the extent to which domain parking confounds fast-flux detection have not been publicly measured at scale.

Aim: Demonstrate a repeatable method for open-source measurement of fast-flux and domain parking, and measure representative trends over 5 years.

Method: Our data source is a large passive-DNS collection. We use an open-source implementation that identifies suspicious associations between FQDNs, IP addresses, and ASNs as graphs. We detect parking via a simple time-series of whether a FQDN advertises itself on IETF-reserved private IP space and public IP space alternately. Whitelisting domains that use private IP space for encoding non-DNS responses (e.g. blacklist distributors) is necessary.

Results: Fast-flux is common; usual daily values are 10M IP addresses and 20M FQDNs. Domain parking, in our sense, is uncommon (94,000 unique FQDNs total) and does not interfere with fast-flux detection. Our open-source tool works well at internet-scale.

Discussion: Real-time detection of fast-flux networks could help defenders better interrupt them. With our implementation, a resolver could potentially block name resolutions that would add to a known flux network if completed, preventing even the first connection. Parking is a poor indicator of malicious activity.

1 Introduction

Fast-flux service networks were first reported in 2007, identified as "a network of compromised computer systems with public DNS records that are constantly changing, in some cases every few minutes" [25, §1]. Criminals use the technique to "evade identification and to frustrate law enforcement and anti-crime efforts aimed at locating and shutting down web sites" that are used for abuse or illegal purposes [13, p. 2].

Despite this long history, and a variety of publications on detecting fast flux, there is no maintained, open-source tool that can detect it at scale. The Honeynet Project's own tool for the purpose, Tracker (`http://honeynet.org/project/Tracker`), has a defunct homepage. Tools from the time, such as ATLAS [20] and FluXOR [21], handle on the order of 400 domains. Our tool, Analysis Pipeline, handles networks on the order of 1 million fully-qualified domain names (FQDN, hereafter simply "domain" if the usage is unambiguous). Pipeline simultaneously tracks other network behavior, such as network flow records. Therefore Pipeline can detect when a host connects to an IP address in the fast-flux network in near-real time, for example.

We also measure a phenomenon mentioned but not measured in some older fast-flux detection papers: domain name parking. When a domain is *parked* on an IP address, the IP address to which the domain resolves is inactive or otherwise not controlled by the domain owner. Parking is common practice when a user first registers an effective second-level domain (eSLD) – the registrar supplies a nonsense IP address to prevents DNS errors. However, this parking pattern is distinctive and simple. We look for other, suspicious patterns.

There are multiple distinct senses of the term "domain parking," and our topic is not synonymous with any other study of which we are aware. Domain parking on private IP address space is, however, a relatively old phenomenon; it is mentioned in some fast-flux identification algorithm studies as an obstacle [35, 14]. This older usage of "domain parking" is our topic of study We use *domain parking on private IP address space* to differentiate it from the newer usage [2, 34] that more accurately is *domain parking on routeable IP addresses for advertisement revenue generation*.

Two of the first studies on parking domains for illicit ad revenue find large-scale use of 4 million to 8 million domains [2, 34]. However, from the authors' description this appears to be more like typosquatting (as described in Szurdi et. al. [31]) than resolution error suppression. We are not studying typosquatting or skimming ad revenue off user typos. Domain parking of the sort we study is a strategy for suppressing domain resolution errors, likely used to keep command and control infrastructure stealthy.

The domain name system permits a variety of different resiliency mechanisms for distributed architectures. Often these have legitimate uses, but malicious actors are equally able to adopt successful techniques. Fast-flux and domain parking of private IP space are both candidates for such abuse. ICANN responded to the security concerns from fast-flux networks in 2009 by stating there would be no policy response [15, p. 10]. Thus, of the six mitigation options outlined in the original Honeynet analysis, five are unlikely or inconsistently applied because they can only be enacted by ISPs or Registrars. The last, "passive DNS harvesting/monitoring to identify A or NS records advertised" as part of fast-flux networks [25, §10], is the approach we implement with Analysis Pipeline. Our detection method is inspired by the Mannheim score [12].

We find a mixture of positive and negative results. On one hand, our measurement of fast-flux networks confirms our hypothesis that this behavior remained prevalent. As one negative example, domain parking on private IP address space is not worth much concern, for fast-flux network detection or otherwise. Negative results are important and useful in shaping future work. Publication bias has been a documented concern in medical literature for 30 years [8]. Despite this attention, publication of negative results has generally dwindled across disciplines. The relative publication frequency of positive results over negative results grew by 22% from 1990 to 2007 [9]. We expect such publication bias away from negative results is a contributing factor to why there seems to be no public measurement of this phenomenon.

Section 2 discusses the common elements of the method between our parking and fast flux measurements, which primarily is the passive DNS data source and our open-source tool Analysis Pipeline. We present our measurement method for FQDNs that exhibit parking on private IP address space in Section 2.1. Section 2.2 describes our fast-flux measurement methods. Section 3 presents the full results. Section 4 interprets and discusses these results.

2 Method

Our measurements of domain parking of private IP address space and fast-flux networks use different algorithms over the same data set and implemented using the same open-source tool. We describe the common measurement period, data, and tool here. The methods specific to each parking and fast-flux measurement are described in the following subsections.

We measure activity during over five years of passive DNS data, from January 1, 2012 to June 30, 2017. The data source, the Security Information Exchange (SIE), has been demonstrated to be reasonably representative of the global Internet with a small North American collection bias [29]. This is high-volume passive DNS data, as well as being representative. Each month, the unique FQDNs observed range between 550 million and 1 billion. With the relatively small and stable zone `.edu`, the data source is sufficiently represetative to reconstruct 93% of the zone in five weeks [27].

We use our own data filtering and packing tools, independent from the SIE database. About 35-40 GB of data is ingested daily in compressed `nmsgtool` format [10], including source DNS server and precise time range the response was valid. Unique resource record sets (RRsets) are extracted for each 24-hour day, with a cutoff of 0000 UTC. We store just the fields for `rname`, `TTL`, `type`, and `rdata`. The nmsg data canonicalizes `rdata`, so this field is sorted set of all `rdata` in a single DNS message for a `rname,type,class` triple. The `rname` field is label-wise reversed, so www.example.com becomes com.example.www; this makes sorting and lookup easier, as the TLD is usually a more important key. The RRsets are then simply sorted and unique RRsets stored per day. When compressed with standard tools such as bzip or gzip, this ASCII storage format takes about 5 GB per day.

The rationale for this storage method is similar to that for why SiLK, a netflow analysis tool suite, stores flow in time-sorted, partitioned flat files via the file system rather than in a database [32]. We are interested in long trends, or retrospective analysis of poorly understood past events. This is different from the use case of many passive DNS users, who are looking for keywords indicating abuse of particular brand names. Our storage format provides details on what domains resolved to on particular days. This allows us to see interleaving changes in domain-IP mappings and large gaps of inactivity that are not possible in a database that only stores first- and last-seen times. Our parking measurement, in particular, requires such granularity.

A final benefit of this time-partitioned storage format is that it is highly parallelizable. Using an HDFS cluster, queries parallelize naturally with each node processing a day. We find that using a database can speed our analysis, especially of fast-flux. However, we parallelize even the database, loading each daily RRset file and then operating over it independently. The overhead of managing a database for all five years of data is superfluous.

Both parking and fast-flux measurements make use of context data to enrich IP addresses. Most importantly, we associate IP addresses with the autonomous system which is advertising it on the relevant day. Autonomous System Number (ASN) attribution is derived from the RouteViews [23] and RIPE NCC RIS [22] data. All ASN data are freely available online [3] under folders for the respective dates. The baseline mapping of ASNs across all IP space uses the open-source SiLK [5] tools for prefix maps and IP sets [32]. Strictly, we count unique routing profiles, not unique ASNs. If an IP address is dual-homed or the global BGP otherwise has consensus that two last-hop ASNs are viable, we mark that IP address with both ASNs. Since our goal is to identify when IP addresses are routed differently to identify stewardship changes, this interpretation is sensible. The geolocation data we use is the public MaxMind GeoLite2 [18].

We relate our analysis algorithms as Analysis Pipeline [24] configuration files. Pipeline is one of the open-source tools associated with SiLK [5]. It is a real-time traffic analyzer that works on IPFIX and network flow records. Pipeline is a Network Behavior Analyzer subtype of an Intrusion Detection System in NIST terminology [26]. As a sort of IDS, Pipeline can run in real time on a network to dynamically detect fast-flux or parking domains and then dynamically add them to a list to watch or block. Thus, with the configurations here and the published SiLK tools, our measurements are readily reproducible in the sense of Feitelson [11], where to reproduce means in a different setting with similar artifacts.

Algorithm 1 Analysis Pipeline command-line

```
/usr/local/sbin/pipeline  \
        --site-config-file=/usr/local/share/
            silk/silk.conf \
        --alert-log-file=~/AlertLog.txt \
        --aux-alert-file=~/AuxLog.txt \
        --ipfix \
        --time-is-clock \
        --configuration=~/parking.conf \
        --name-files input_files_list
```

Algorithm 1 is an example of how to execute Pipeline. Specifically it calls our parking measurement configuration, listed later in Algorithm 2. It reads DNS records encoded in the standard IPFIX format [6]. A sample python script to convert DNS records from CSV format to IPFIX is available in the pyfixbuf documentation [4]. Passive DNS data can also be converted to CSV or IPFIX directly using the nmsg python bindings [10].

In addition to rote results, we perform some simple summary and context operations. The main summary is based on the effective second level domain (eSLD) of the parked domains. The eSLD of `www.example.com` is simply the SLD `example.com`; however, the eSLD of `www.example.co.uk` includes a third label: `example.co.uk`. To identify eSLDs we use the Mozilla public suffix list (effective_tld_names.dat).

The main context-enrichment operation is to intersect parked domains and their publicly-routable IP addresses with fast-flux domains and IP addresses. The intersection is simple set intersection on the domain names. We do not report time slices of the intersection, simply the intersection between the union of all domains exhibiting parking behavior and the union of all fast-flux domains. We also summarize to eSLD and repeat the intersection.

For some context about malicious intent of parking and fast flux, we associate each set of domains with lists of malicious domains. While we have expressed our doubts about the soundness of evaluating an approach by comparing it to blacklists [19], we have mitigated this error by including as many lists as possible (over 100) and limiting our assumptions of the information provided by this comparison.

We perform blacklist comparisons within 6-month blocks. All blacklist entries over the whole timespan are unioned, and that set is intersected with all domains exhibiting the behavior of interset at any time during the timespan. This mitigates the possibility that it takes some time to identify and blacklist a malicious domain. It has proven logistically impractical to provide a sliding window for blacklist detection. But 6-month windows are pretty broad, as many malicious behaviors are consistently detected and vendors list names within a few hours. At this wide window, we already risk false-positives due to IP-address churn or other exogenous factors overwhelming true blacklist associations. We happen to have more IP-address based blacklists than domain-based ones; it is unclear whether this overestimates IP-address participation in blacklists or underestimates domain names, if either.

2.1 Parking Detection

We measure *domain parking on private IP address space* straightforwardly. The algorithm is summarized as follows; we expand the description through the rest of the section. First, we find all DNS IPv4-answer RRsets that refer to private address space to acquire a set of possible domains. We remove whitelisted domains that we have found to use private address space as an information carrier for other services. For all domains that remain after this subtraction, we find all their DNS RRsets for that day and a window three weeks into the future. These three-week time series are examined for transitions from public to private IP-space. We repeat this algorithm for each day in the five-year measurement period, evaluating 1807 three-week windows.

Our first step is to extract or mark all RRsets that contain a private IP address in the `rdata`. Private IP address space is exactly those addresses listed in Table 1. These blocks are selected because they are special-purpose assignments that RFC 6890 lists as either not forwardable by routers or not global, meaning only forwardable in specified administrative zones [7].

This provides a set of `rname` data that have been associated with a private IP address. Most are not parking. Private IP space used to encode various kinds of non-location data, such as responses to lookups on DNSBLs [17]. SURBL provides a good example of how and why their service does this [30]. Seeded by lists of threat intelligence and blacklist providers such as `intel.criticalstack.com`, and refined through human expert analysis, we whitelist 165 DNS zones that consistently encode non-location data.

The process so far yields a list of RRsets with `rdata` in private IP space whose `rname` zones do not have a whitelisted, known use. We next find all RRsets with the same `rname` values and publicly routeable IP addresses within 21 days. These domains transitioned between private and routeable IP address space some time in the 3-week window. We define FQDNs that exhibit such a transition as demonstrating *parking behavior* on private IP address space during our observation period.

CIDR block	Justification
0.0.0.0/8	RFC 1122
10.0.0.0/8	RFC 1918
100.64.0.0/10	RFC 6598
127.0.0.0/8	RFC 1700
169.254.0.0/16	RFC 3927
172.16.0.0/12	RFC 1918
192.0.0.0/24	RFC 6890
192.0.2.0/24	RFC 5737
192.168.0.0/16	RFC 1918
198.18.0.0/15	RFC 2544
198.51.100.0/24	RFC 5737
203.0.113.0/24	RFC 5737
224.0.0.0/3	RFC 1112

Table 1: Private IP address space

In order to cover the 5-year time period, we compute this rather simple algorithm over 1800 times. For each day's set of unique RRsets, we extract the domains mapping to private IP address space, remove whitelisted zones, and expand to those also mapping to a public address at some time within three weeks.

For each FQDN that has exhibited parking behavior, we can generate a course-grained time series of the behavior to categorize what occurred. Table 2 demonstrates some sample behavioral groupings. P indicates a day where the only `rdata` was in private IP address space, G indicates a day where the only `rdata` was in globally routeable IP address space, and X indicates a day where both address types were observed, indicating a day a change between parking and active occurred.

The configuration listed in Algorithm 2 runs our method in Pipeline. The SiLK IPset "priv.set" contains exactly the address blocks listed in Table 1.

2.2 Method: Fast-flux

Our fast-flux detection algorithm implements prior work, such as the Mannheim score [12]. The main novelty of our work is the scale and duration of our measurement and the use of open-source tools. Detection algorithms were sufficiently well-studied in 2010, and the same concept holds for detection of fast-flux today.

The basis of our fast-flux detection algorithm is that a legitimate administrator owns or rents their infrastructure in a relatively small number of places.

January:	1-8	9-16	17-24	25-31
Activation on Jan 19	PPPPPPPP	PPPPPPPP	PPXGGGGG	GGGGGGG
Deactivation on Jan 19	GGGGGGGG	GGGGGGGG	GGXPPPPP	PPPPPPP
alextringham.com	GGGGGGGG	GGGGGGGG	GGGGXPPX	PXPXPPP
proxyie.cn	GGXXXXXX	GGGXGXGG	GGPGGXGX	XGGGXGX
bnlv.homeip.net	GGGGGPGG	GGGGGGGG	PGPPGGGG	GGGGPGG

Table 2: Example parking behavior patterns per domain, January 2014. G := only globally routeable IPs observed that day. P := only privately reserved IPs observed. X := both observed on same day.

Algorithm 2 Parking detection in Pipeline

```
FILTER emptyDomainNames
  dnsRRIPv4Address IN LIST "priv.set"
END FILTER

INTERNAL FILTER emptyDomains
  FILTER emptyDomainNames
  dnsQName domainsWithNoIP 1 DAY
END INTERNAL FILTER

FILTER unparked
  dnsQName IN LIST domainsWithNoIP
  dnsRRIPv4Address NOT IN LIST "priv.set"
END FILTER

EVALUATION unparkedRecords
  FILTER unparked
  CHECK EVERYTHING PASSES
  END CHECK
  ALERT ALWAYS
  ALERT EVERYTHING
END EVALUATION
```

Algorithm 3 Fast-flux detection in Pipeline

```
PMAP asn "$today.ip2asn.pmap"
FILTER fluxwhitelist
  sourceIPv4Address NOT IN_LIST "priv.set"
  DNS_SLD+TLD(DNS_INVERT(dnsQName)) NOT
      IN_LIST "$today.whitelist"
END FILTER

EVALUATION ipfixFastFluxExample
  FILTER fluxwhitelist
  CHECK FAST FLUX
    IP_FIELD sourceIPv4Address 500
    ASN asn 23
    DNS dnsQName 667 TO myFastFluxDNSs
    NODE MAXIMUM 100000000
    VERBOSE ALERTS
  END CHECK
  CLEAR ALWAYS
END EVALUATION
```

We perform some pilot studies on January 2016 to get a sense of what counts as small. We represent fast-flux networks as graphs of each of three kinds of resource: IP addresses, ASN, and FQDN. The intuition is that shared hosting may have 10,000 domains on a single IP, and if changes to another IP address it is likely one the hosting provider owns. The network identifiers cluster when one or another resource is advertised on a new resource. If two domains both map to 192.168.0.1, and then one is changed to 10.0.1.1, then our algorithm considers both IP addresses as well as both domains to be part of a cluster. The information cluster would also include any domains that had previously mapped to 10.0.1.1 within the time frame, and any IP addresses they had been mapped to, and so on. For our example, imagine this linking brings in 20 more domains, all on the 16 IP addresses in 172.16.0.32/28. However, the AS for all 18 IP addresses is AS112. We do not consider this a fast-flux network, because all the resources are only related to one AS. They are probably related because of a the AS owner doing regular maintenance or load balancing, not fast flux. This example helps show why ASN is needed, rather than relying on CIDR block.

Our results are run with the conservative thresholds that a graph must contain 500 unique IPs, 23 ASNs, and 667 FQDNs to be marked fast flux. We also whitelist any IP addresses in our private IP address space (Table 1) and a FQDN whitelist that created from the Alexa most popular domains list. Resources on these whitelists will not be added to a graph, and so can never be marked fast flux.

The FQDN whitelist captures more stability than naïvely using the Alexa top X. For a given day, we find all names in the Alexa top 24,000 on 330 of the past 365 days. The whitelist functions as a wildcard, not as an FQDN perfect match. So if example.com is on the whitelist, *.example.com will not be watched for fast flux. For this reason, we remove any effective TLDs on the Mozilla public suffix list (version on June 15, 2017), of which there are routinely between 50-100 on the Alexa list. We also remove any well-known dynamic DNS providers from the whitelist. Finally, we make sure that no domains have subdomains on the list; this interferes with the wildcard function of pipeline whitelists. The aver-

[Distribution Statement A] Approved for public release and unlimited distribution.

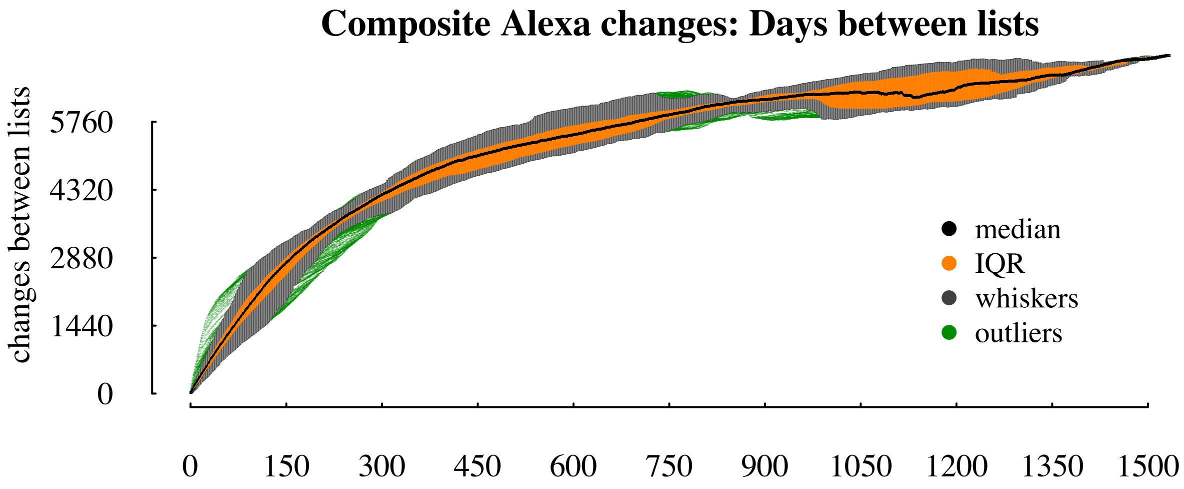


Figure 1: Statistics on changes between Alexa-based whitelists, calculated on the pairwise differences between lists. Changes includes additions and removals. Each value on the X-axis is a color-based box-and-whisker plot representing the distribution of changes for all pairwise sets of lists that many days apart.

age size of the whitelist is 14,847. The minimum is 13,780 (March 7, 2017); the maximum size is 15,946 (April 4, 2013). Even with this algorithm designed for some stability, there significant change, which warrants using a new list each day. Figure 1 displays the distribution of whitelist-entry changes calculated pairwise between all 1492 lists. The first day we have Alexa data available is April 1, 2012, so the first day a composite whitelist is possible is April 1, 2013. We use this April 1, 2013 list for any flux measurements before that date.

We selected a relatively high threshold (500-23-667) after exploratory analysis demonstrated some clearly legitimate uses that exceeded our initial 5-5-5 threshold. Before implementing the Alexa whitelist, we also found Tumblr, as 616 IPs, 10 ASNs, and 10,658,458 FQDNs. But not every large network like this is handled by a whitelist. For example, the network signature 216 IPs, 19 ASNs, 2,341,876 FQDNs is not in fact a fast-flux network, but Tekblue, an ad tracking company. But, because Tekblue is an ad tracker, it does not appear on our Alexa whitelist, on any day. Ampproject.net also consistently produced huge fast-flux-like clusters—such as 755 IPs, 15 ASNs, 533,082 FQDNs and 671 IPs, 12 ASNs, 534,956 FQDNS—but is never on our whitelist. These are evidence for increasing our ASNs threshold. We also find evidence to increase our FQDN threshold. The graph 9,051 IPs, 102 ASNs, 63 FQDNs is Akamai. NTP device pools also produce quite strange signatures, such as 2100 IPs, 845 ASNs, 122 FQDNs and 2,507 IPs, 910 ASNs, 267 FQDNs.

We did a sample run at 500-23-667 to check results. The graph with the fewest ASNs passing these thresholds was 27; its domains often have a suspicious pattern that appears machine-generated. There are some names that look human-generated; however, they may be compromised. At the least, it is not something to obviously exclude by increasing the thresholds. The test run reduced the results from 513 (with 5-5-5 thresholds) to 196 distinct, non-overlapping flux networks. But these 196 still capture almost all of the IP addresses from the 513; 99.7% of the unique IP addresses across flux networks remain in the results with the increased threshold. Therefore, while our thresholds are conservatively high, we still find significant malicious activity while reducing obvious false positives.

One may wonder if publishing such a detection threshold would benefit adversaries more than defenders. However, forcing adversaries to keep smaller, disjoint networks would reduce their reliability and increase their management effort. Adversaries would no longer be able to use any resource if its FQDN or IP is associated with a known, live flux network. Pipeline can issue such alerts in real time, as new resources are seen and added to known networks. Potentially, this means many communications can be blocked at the first instance, preventing even one use of the FQDN or IP if it is added to a known flux network. Such before-first-use blocking is recommended by Spring [28] as necessary to keep adversaries from profiting.

Analysis Pipeline (version 5.0 and later) includes a primitive data element for fast-flux networks [24].

FQDNs	intersection	of flux	of park
2012-1	14609	0.0008	0.2069
2012-2	12103	0.0009	0.0286
2013-1	10930	0.0006	0.1990
2013-2	11662	0.0006	0.1126
2014-1	11106	0.0006	0.1362
2014-2	30346	0.0007	0.0714
2015-1	61259	0.0010	0.1756
2015-2	40824	0.0008	0.0800
2016-1	32687	0.0004	0.0934
2016-2	46291	0.0004	0.1125
2017-1	56758	0.0006	0.0458

Table 4: Half-yearly intersections of domains exhibiting parking and fast-flux networks. Values range $[0, 1]$.

IPs	intersection	of flux	of park
2012-1	523056	0.0147	0.9127
2012-2	2565808	0.0717	0.8847
2013-1	220651	0.0052	0.8934
2013-2	257741	0.0083	0.8644
2014-1	364870	0.0127	0.9415
2014-2	730484	0.0149	0.9421
2015-1	478327	0.0086	0.9100
2015-2	557457	0.0109	0.8911
2016-1	1567197	0.0268	0.9039
2016-2	1562149	0.0169	0.7736
2017-1	837599	0.0141	0.7640

Table 5: Half-yearly intersections of IP addresses exhibiting parking and fast-flux networks. Values range $[0, 1]$.

Pipeline builds a connected graph of ASN, FQDN, IP address tuples. If the connected graph passes a threshold for all three resources, that graph is considered to be a fast flux network. Algorithm 3 is the Pipeline configuration that implements our detection algorithm. The alerts can be configured to report the whole connected graph, or just lists of domains or IP addresses. Algorithm 4 shows how to use such output lists.

3 Results

Fast-flux networks do not overlap much with resources exhibiting parking. Neither fast-flux nor parking overlap much with blacklisted resources. Parking behavior is present, but small in the scheme of the global internet. Fast-flux networks, on the other hand, appear to make use of a large number of internet resources.

Figure 2 plots the number of unique FQDNs and effective SLDs used each day for parking. Our parking algorithm uses a 3-week window to detect parking, which has an effect of smoothing out the day-to-day changes. The median FQDNs parking on a given day is 13,300, in a median of 4,750 eSLDs.

Figure 3 captures the fact that fast-flux networks have a much bigger footprint. Many days have over 10,000,000 IPs and 20,000,000 FQDNs involved in fast-flux.

Table 4 reports the overlap between fast-flux and parking FQDNs detected in each half-year. Table 5 reports the analogous overlap for IP addresses. Since there are fewer resources that evidence parking, these intersections make up a larger share of parking than of fast flux.

Given the conservative (i.e., large) definition of flux network size we set, it is unlikely these collections of internet resources have a benign purpose. Our initial pilot study (using 5-5-5 as a threshold) contained common internet services such as ad networks, content distribution networks, and the NTP servers. The threshold of 500-23-667 excludes such benign services, based on our expert analysis of the results.

Despite the fact we do not have a benign explanation for this behaviour, as Figure 4 and Figure 5 demonstrate, few IP addresses that have participated in a fast-flux network are on any blacklists. The median monthly blacklist intersections range between 100,000 and 400,000; roughly 4-8% of the flux networks. The exception seems to be overlap during 2017 between flux networks and FQDN blacklists.

Parking, likewise, is uncommonly blacklisted. Table 3 displays the results for both FQDNs and live IP addresses associated with parking behavior. Because so few domains park, the contribution of parking to blacklists is negligible. However, parking is also not a reliable indicator of blacklist membership. Fewer than 2% of domains and between 5-10% of IPs that exhibit parking end up on blacklists.

Our results for January 2014 are available for download (see `http://www.cert.org/downloads/name-parking-patterns-certcc-2014-57.txt`) in the format of a domain name followed by the behavior pattern encoded as in Table 2.

FQDNs from the Alexa top 100 occasionally were found on our parking list [1]. We manually removed about 30 Alexa top 100 domains from the results each period. The root cause for these anomalous DNS responses is not known.

	FQDNs	% lists	% parking	IPs	% lists	% parking
2012-1	3033	0.0469	4.2959	87069	0.2079	14.1214
2012-2	2513	0.0272	0.5929	85853	0.2525	2.5345
2013-1	1746	0.0139	3.1782	26479	0.0941	10.1526
2013-2	4439	0.0883	4.2842	13300	0.1473	4.1119
2014-1	2005	0.0723	2.4587	33654	0.1342	8.4516
2014-2	6596	0.0631	1.5530	51061	0.1702	5.7200
2015-1	5616	0.0459	1.6101	47535	0.1565	8.6189
2015-2	8339	0.0639	1.6332	49728	0.2010	7.3737
2016-1	6802	0.0279	1.9426	91716	0.2647	4.9624
2016-2	4295	0.0164	1.0442	100620	0.3417	4.5519
2017-1	5315	0.0136	0.4292	64859	0.1462	10.332

Table 3: Half-yearly intersections of resources exhibiting parking and blacklists. Percentages range $[0, 100]$

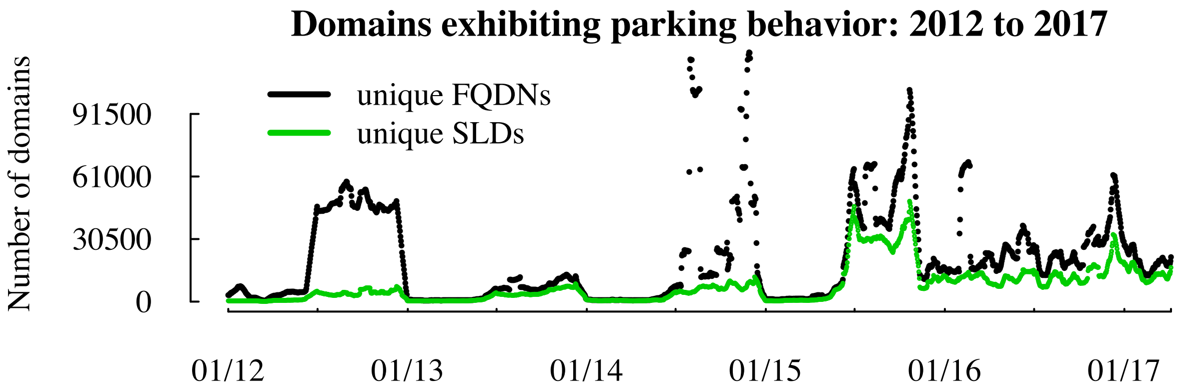


Figure 2: Three-week rolling window of unique domains and eSLDs exhibiting parking behavior from beginning of 2012 to April 2017.

Some parking domains used dynamic DNS services; however, usage is minimal. We compared the results to a list of 71 known dynamic DNS providers. The bulk were hosted on two providers: dyndns.org or on some name affiliated with no-ip. These are the two biggest providers, so this distribution is expected based on market share.

4 Conclusions

We shall discuss our fast-flux results first, and then our parking results.

Fast-flux networks remain remarkably common 10 years after first reported use by malicious actors. The lack of intersection with blacklists despite the obviously suspicious nature of this behavior is especially noteworthy. We suspect that fast-flux networks are used for intermediary malicious behavior, such as providing clandestine communication to already infected hosts. It is also possible that blacklist vendors do not bother to list something that they know will change within a very short interval. We also have not ruled out all possible alternative interpretations; for example, peer-to-peer networks. However, if this were the case, we would still expect our flux results to be a superset of malicious flux networks. Excessive poor detection precision would decrease the number of flux resources on blacklists, but it does not explain why so few of our blacklist entries are in flux networks.

We have designed our method to capture nonbenign fast-flux networks. Our first attempt at setting thresholds captured many recognizable, legitimate internet services. However, as described in Section 2.2, we eliminated all obvious and large false positives. Our observation is limited by the data source. However, again, this has a known bias, and it is known to be comprehensive. We do not make any claims that we can project our observations onto the parts of the internet we do not observed. However, the absolute numbers of resources participating in fast-flux we detect are quite large. We do not need

[Distribution Statement A] Approved for public release and unlimited distribution.

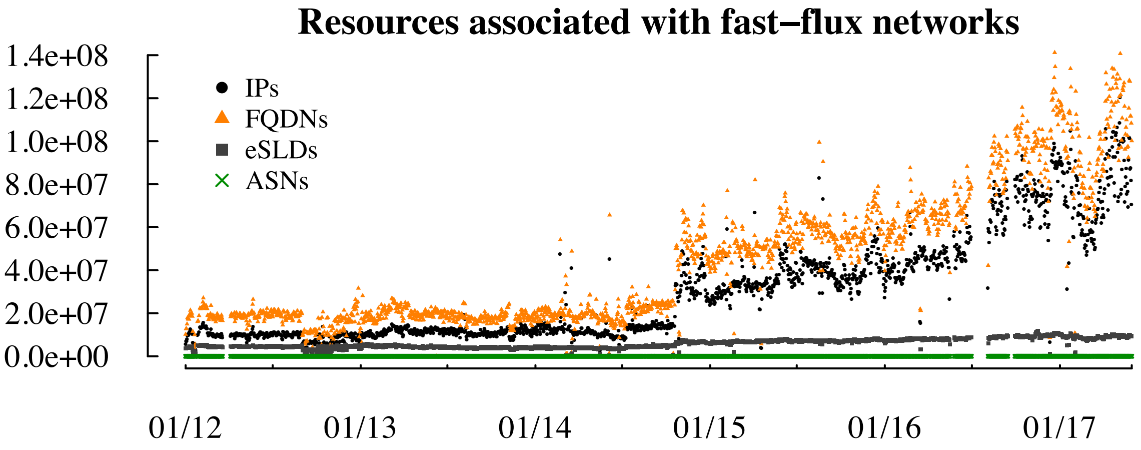


Figure 3: Total unique network resources of different types associated with fast-flux networks every day. Gap in July 2016 is a collection error.

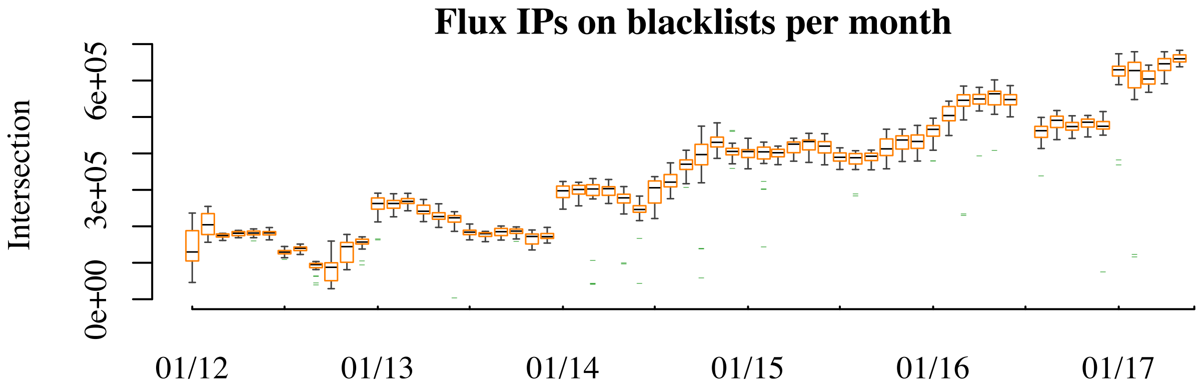


Figure 4: Summary statistics fast-flux–blacklist intersection. Each day's flux from January through June 2013 is intersected with all 1H2013 blacklist data, for example, and the daily intersections are summarized in monthly box plots. The whisker length is 1.5 times the inter-quartile range (IQR).

to project onto a target population. The measurement as-is finds over 100 million IPs and FQDNs some days in 2017. We have high confidence that few of these resources are participating for benign reasons. Even in the unlikely event that the rest of the internet sees no other unique resources participating in fast-flux, these values are worrisome.

As demonstrated repeatedly, the contents of individual blacklists rarely overlap [19, 16, 33]. One plausible explanation for this disjointedness is that blacklists track a lot of ephemeral IP addresses. However, if fast-flux would have a statistical impact on this disjointedness, it would need to represent more than 1% of blacklist identifiers: as we found in this study.

Our leading interpretation of the fact that our fast-flux results are mutually disjoint with the blacklists we have access to is the following. None of the blacklists are tracking fast-flux. This interpretation is consistent with prior interpretation of the blacklist disjointedness generally, which is that each list is good, but very precise about what it is following and from what sensor vantage [19]. This interpretation is strengthened by the long observation time, over many years, and the consistency of both the size of the fast-flux networks and the lack of blacklist overlap during that time.

Algorithm 4 is an example of how the results from fast-flux measurement can be applied to continuous network situational awareness. We do not report an evaluation on a live network due to data access and publication issues; however, we provide the configuration so that network operators can apply it consistently as a test and compare results privately. The result would be that once enough domain-IP pairs are looked up to form a fast-flux network, one alerts

[Distribution Statement A] Approved for public release and unlimited distribution.

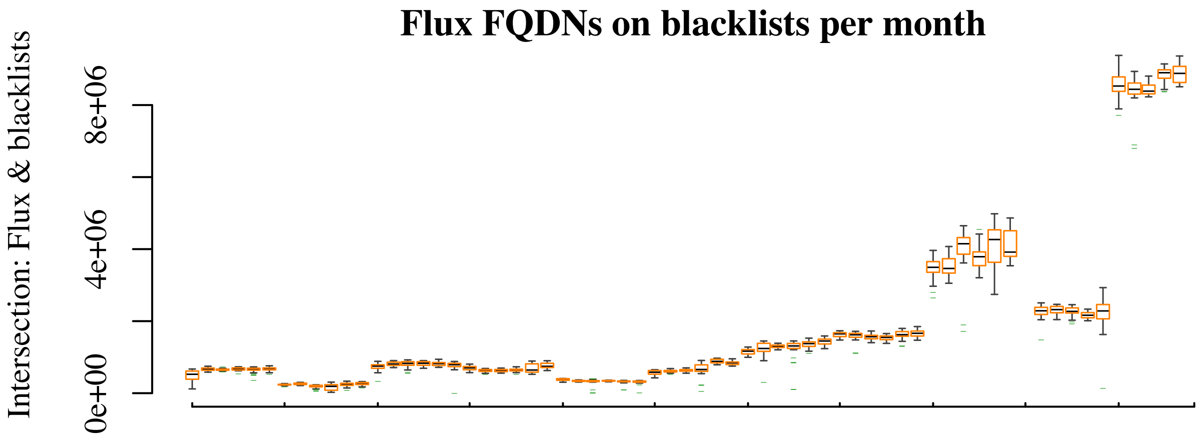


Figure 5: Summary statistics fast-flux–blacklist intersection. Plot follows the same conventions as Figure 4.

Algorithm 4 Sample fast-flux watch-list usage in Pipeline

```
FILTER watchlistFromFastFlux
 sourceIPv4Address IN LIST ipList
END FILTER
EVALUATION alertWatchlist
 FILTER watchlistFromFastFlux
 ALERT EVERYTHING
 CHECK EVERYTHING PASSES
 END CHECK
END EVALUATION
```

on new connections if, at the time of the first connection attempt, they would add to a known fast-flux network. Defenders can effectively cap the viable size of fast-flux networks, reducing their usefulness to adversaries.

We can confirm that domains exhibiting parking on private IP addresses does not likely confound fast-flux network detection. We can also help explain why the recent literature uses ‘parking’ to mean typo-squatting for revenue generation: the older usage we study is much less common. Although parking on private IP addresses is rare, it is still an odd behavior. Further analysis may evidence what these resources are used for. However, given how long it takes to detect such parking, it is likely not the best use of defender resources. This assessment may change domains parked on private IP address space are used for high-impact attacks; however, our observations do not evaluate this concern. Future work should occasionally re-validate this assessment.

There are possible alternative interpretations of our parking results. Perhaps adversarial capability in utilizing parked domains in this way is still in an early phase of development. Alternatively, the domains exhibiting this kind of parking may be malicious, but simply are not found by any detection method used by the blacklists we compare against.

We have presented a combination of surprising results, on fast-flux, and unsurprising results, on parking. More notable is the importance of the method of long-term internet measurement in detecting trends and making conclusions. Passive DNS remains a useful tool, because it supports such long time scales while still keeping wide coverage feasible. However, our results also rely on archiving many years of contextual data, for routing, blacklists, and Alexa’s top domains. The providers of these data do not have much incentive to store and archive long spans of data. Routeviews and RIPE RIS do a good job collecting this routing information. Similar initiatives for other internet metadata would improve the community’s ability to pursue research.

Acknowledgment

Thanks to Emily Sarneso for her help with IPFIX.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution. Internal use:* Permission to reproduce this material and to prepare derivative works from this material for internal use is granted, provided the copyright and "No Warranty" statements are included with all reproductions and derivative works. External use:* This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other external and/or commercial use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

* These restrictions do not apply to U.S. government entities.

References

[1] Alexa. Alexa Internet, inc. – top sites. `http://www.alexa.com/topsites`, January 13, 2013.

[2] Alrwais, S., Yuan, K., Alowaisheq, E., Li, Z., and Wang, X. Understanding the dark side of domain parking. In *23rd USENIX Security Symposium (USENIX Security 14)* (San Diego, CA, Aug 2014), USENIX Association.

[3] CERT/NetSA at Carnegie Mellon University. CERT/CC Route Views Project Page. `http://routeviews-mirror.cert.org`. [Accessed: Feb 13, 2017].

[4] CERT/NetSA at Carnegie Mellon University. pyfixbuf. `http://tools.netsa.cert.org/pyfixbuf/index.html`. [Accessed: Jan 24, 2017].

[5] CERT/NetSA at Carnegie Mellon University. SiLK (System for Internet-Level Knowledge). `http://tools.netsa.cert.org/silk`. [Accessed: Feb 4, 2017].

[6] Claise, B., Trammell, B., and Aitken, P. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information. RFC 7011 (INTERNET STANDARD), Sept. 2013.

[7] Cotton, M., Vegoda, L., Bonica, R., and Haberman, B. Special-Purpose IP Address Registries. RFC 6890 (Best Current Practice), Apr. 2013.

[8] Dickersin, K., Chan, S., Chalmersx, T., Sacks, H., and Smith, H. Publication bias and clinical trials. *Controlled clinical trials 8*, 4 (1987), 343–353.

[9] Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics 90*, 3 (2012), 891–904.

[10] FarSight Security, Inc. nmsgtool. `https://archive.farsightsecurity.com/nmsgtool/`, Sep 25, 2013. [Accessed: Aug 12, 2014].

[11] Feitelson, D. G. From repeatability to reproducibility and corroboration. *ACM SIGOPS Operating Systems Review 49*, 1 (2015), 3–11.

[12] Holz, T., Gorecki, C., Rieck, K., and Freiling, F. C. Measuring and detecting fast-flux service networks. In *Proceedings of the 15th Annual Network and Distributed System Security Symposium* (February 2008).

[13] ICANN. SSAC advisory on fast flux hosting and dns. Tech. Rep. SAC-025, Internet Corporation for Assigned Names and Numbers – Security and Stability Advisory Committee, March 2008.

[14] Knysz, M., Hu, X., and Shin, K. G. Charlatans' web: Analysis and application of global IP-usage patterns of fast-flux botnets. *University of Michigan Ann Arbor* (2011), 1–22.

[15] Konings, M. Final report of the gnso fast flux hosting working group. Tech. rep., Internet Corporation for Assigned Names and Numbers – Generic Names Supporting Organization, August 2009.

[16] Kührer, M., Rossow, C., and Holz, T. Paint it black: Evaluating the effectiveness of malware blacklists. Tech. Rep. TR-HGI-2014-002, Ruhr-Universität Bochum, Horst Görtz Institute for IT Security, June 2014.

[17] Levine, J. DNS Blacklists and Whitelists. RFC 5782 (Informational), Feb. 2010.

[18] MaxMind. Geolite2 free downloadable databases. `http://dev.maxmind.com/geoip/geoip2/geolite2/`, Jan 28, 2014.

[19] Metcalf, L. B., and Spring, J. M. Blacklist ecosystem analysis: Spanning Jan 2012 to Jun 2014. In *The 2nd ACM Workshop on Information Sharing and Collaborative Security* (Denver, Oct 2015), pp. 13–22.

[20] Nazario, J., and Holz, T. As the net churns: Fast-flux botnet observations. In *Malicious and Unwanted Software (MALWARE)* (Sep 2008), IEEE, pp. 24–31.

[21] Passerini, E., Paleari, R., Martignoni, L., and Bruschi, D. Fluxor: detecting and monitoring fast-flux service networks. *Detection of Intrusions and Malware, and Vulnerability Assessment* (2008), 186–206.

[Distribution Statement A] Approved for public release and unlimited distribution.

[22] RIPE Network Coordination Center. Routing information service (RIS). http://www.ripe.net/data-tools/stats/ris/routing-information-service, January 3, 2012.

[23] Route-Views. University of oregon route views project. http://www.routeviews.org, January 3, 2012.

[24] Ruef, D. Analysis pipeline v.5.6. http://tools.netsa.cert.org/analysis-pipeline5/index.html, Jan 7, 2017. [Accessed Jan 8, 2017].

[25] Salusky, W., and Danford, R. Know your enemy: Fast-flux service networks. Tech. rep., The Honeynet Project, July 13, 2007.

[26] Scarfone, K., and Mell, P. Guide to intrusion detection and prevention systems (IDPS). Tech. Rep. SP 800-94, U.S. National Institute of Standards and Technology, Gaithersburg, MD, Feb 2007.

[27] Spring, J. M. Large scale DNS traffic analysis of malicious internet activity with a focus on evaluating the response time of blocking phishing sites. Master's thesis, University of Pittsburgh, 2010.

[28] Spring, J. M. Modeling malicious domain name take-down dynamics: Why eCrime pays. In *eCrime Researchers Summit (eCRS)* (San Francisco, Sep 2013), IEEE, pp. 1–9.

[29] Spring, J. M., Metcalf, L. B., and Stoner, E. Correlating domain registrations and DNS first activity in general and for malware. In *Securing and Trusting Internet Names: SATIN* (Teddington, UK, Mar 2011).

[30] SURBL. Implementation guidelines. http://www.surbl.org/guidelines, Dec 9, 2011. [Accessed: Aug 1, 2014].

[31] Szurdi, J., Kocso, B., Cseh, G., Spring, J. M., Felegyhazi, M., and Kanich, C. The long "taile" of typosquatting domain names. In *23rd USENIX Security Symposium* (San Diego, Aug 2014), USENIX Association, pp. 191–206.

[32] Thomas, M., Metcalf, L., Spring, J. M., Krystosek, P., and Prevost, K. SiLK: A tool suite for unsampled network flow analysis at scale. In *IEEE BigData Congress* (Anchorage, Jul 2014), pp. 184–191.

[33] Verizon. 2015 data breach investigations report (DBIR). Tech. rep., 2015.

[34] Vissers, T., Joosen, W., and Nikiforakis, N. Parking sensors: Analyzing and detecting parked domains. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS)* (San Diego, CA, February 2015).

[35] Yadav, S., Reddy, A. K. K., Reddy, A., and Ranjan, S. Detecting algorithmically generated malicious domain names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement* (2010), ACM, pp. 48–61.

Lessons Learned from Evaluating Eight Password Nudges in the Wild

Karen Renaud
Abertay University
k.renaud@abertay.ac.uk

Verena Zimmermann
Technische Universität Darmstadt
zimmermann@psychologie.tu-darmstadt.de

Joseph Maguire & Steve Draper
University of Glasgow
{joseph.maguire,steve.draper}@glasgow.ac.uk

Abstract

Background. The tension between security and convenience, when creating passwords, is well established. It is a tension that often leads users to create poor passwords. For security designers, three mitigation strategies exist: issuing passwords, mandating minimum strength levels or encouraging better passwords. The first strategy prompts recording, the second reuse, but the third merits further investigation. It seemed promising to explore whether users could be subtly *nudged* towards stronger passwords.

Aim. The aim of the study was to investigate the influence of visual nudges on self-chosen password length and/or strength.

Method. A university application, enabling students to check course dates and review grades, was used to support two consecutive empirical studies over the course of two academic years. In total, 497 and 776 participants, respectively, were randomly assigned either to a control or an experimental group. Whereas the control group received no intervention, the experimental groups were presented with different visual nudges on the registration page of the web application whenever passwords were created. The experimental groups' password strengths and lengths were then compared that of the control group.

Results. No impact of the visual nudges could be detected, neither in terms of password strength nor length. The ordinal score metric used to calculate password strength led to a decrease in variance and test power, so that the inability to detect an effect size does not definitively indicate that such an effect does not exist.

Conclusion. We cannot conclude that the nudges had no effect on password strength. It might well be that an actual effect was not detected due to the experimental design choices. Another possible explanation for our result is that password choice is influenced by the user's task, cognitive budget, goals and pre-existing routines. A simple visual nudge might not have the power to overcome these forces. Our lessons learned therefore recommend the use of a richer password strength quantification measure, and the acknowledgement of the user's context, in future studies.

1 Introduction

The first encounter with a new system or service, for many individuals, requires the creation of a password. This authentication approach is based on the possession of some secret shared knowledge, known only to the user and this one system.

People are asked to provide passwords so frequently, and inconveniently, that they end up choosing weak passwords, leaving themselves vulnerable to attack [30]. In effect, password choice becomes something of an obstacle to be hurdled in order to be able to satisfy legitimate goals. The primary problem is the fact that memory limitations tug people towards memorable and predictable secrets, whereas strong security mandates more effort. Strength can be achieved either by using a hard-to-remember and hard-to-guess nonsense string, or by using a long pass phrase. Both are personally more costly than a weak password.

Some believe that we should simply enforce strong passwords [15] or expire passwords regularly [18]. The problem is that neither the former nor the latter guarantee increase resistance to attack [53, 57]. Moreover, restrictive, complex password policies aimed at mandating strong passwords can conflict with users' needs, increase effort and ultimately compromise productivity and security [24, 48, 52].

The other option is to replace the password with something like a biometric or token-based authentication [5, 38]. Neither of these is perfect either. No biometric is ubiquitous and infallible [32] and tokens are expensive and easily lost or stolen.

Other alternatives are graphical passwords, mnemonic passwords or passphrases [1, 51, 28] but these have not really gained widespread acceptance and even passphrases have their flaws [29, 39].

While many focus on the password's deficiencies, it must be acknowledged that passwords also have advantages. They are easy to deploy, accessible to those with disabilities, cost-effective, preserve privacy and are easily replaced [6].

Instead of focusing myopically on the password choice event, we should contemplate password creation as one component of an entire authentication eco-system, and consider that the end user needs more support throughout the process. Horcher and Tejay [23] claim that users are poorly scaffolded during the password creation process, and that this contributes to poor password choice. Solutions that scaffold by offering dynamic feedback on password quality are designed to encourage deliberation and reflection during password creation [17]. However, such approaches have, thus far, not significantly improved the quality of passwords [12, 17, 45].

There is increasing evidence that behaviour can be influenced through surprisingly small and inexpensive interventions called "nudges" [21].

Transferring successful nudges from other areas to the authentication context was something we wanted to test to find out whether these would encourage users to create stronger passwords. The hypothesis we tested was:

> **H1:** The presence of a visual nudge will lead to longer and stronger passwords.

We carried out a longitudinal study to investigate the potential of eight different user interface nudges, displayed during password creation, calculated to influence password choice. The contributions of this paper are:

- Details of how nudges were tested in the wild, and the ethical constraints we encountered.
- Empirical evidence that the tested nudge conditions did not significantly impact password quality.
- Reflection on the results and suggested explanations for the negative finding reported by the study.

The paper concludes with a discussion of lessons learned and recommendations for future studies of this kind.

2 Background

A nudge can be considered a mechanism that guides individuals to make wiser choices without their necessarily being aware of its influence [34]. An intervention can only be considered a nudge if individuals are able easily to resist its influence [43]. A good example of a nudge is the house-fly painted on urinals in an Amsterdam airport. This nudge had the desired effect of reducing spillage, but could equally have been ignored by urinal users.

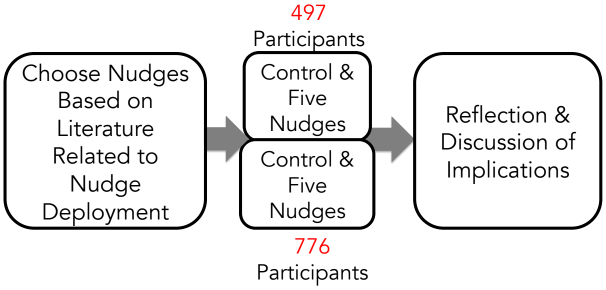

497
Participants
Choose Nudges Based on Literature Related to Nudge Deployment
Control & Five Nudges
Control & Five Nudges
Reflection & Discussion of Implications
776
Participants

Figure 1: Results Reported in this Paper

The subtle nudge approach has proved popular with western governments [44, 22], who have adopted nudges in key areas such as tax and public health [50]. A small alteration in letter text sent to individuals significantly improved tax payment rates [21]. However, such use has been criticized with the suggestion that nudges do not promote long-term behaviour change [36]. Nevertheless, this may not be an issue for use in authentication if the motivation is to promote optimal decisions at the moment of password creation.

There is an argument that people sometimes create passwords unthinkingly, basically operating using their *autopilot* (System 1) thinking, rather than deliberately engaging (System 2) level thinking to choose a good password [43]. Sunstein [41] explains that nudges can work in tandem with educational efforts by impacting System 1 thinking, with educational efforts targeting System 2 thinking, thus complementing each other.

Jeske *et al.*. [26] demonstrated such an approach when it came to nudging users to select the most secure wireless networks. They found that nudges could be effective, but that personal differences also played a role in the security decisions. Similarly, Yevseyeva *et al.* [56] experimented with nudging people towards secure wireless network selection using different variations of a prototype application. The found a combination of colour coding and the order in which the Wi-Fi networks were listed to be most effective.

Nudges have also been deployed to improve decisions surrounding privacy. Choe *et al.* [10] investigated positive and negative framing of privacy ratings to nudge individuals away from privacy-invading applications. They demonstrated that framing, as well as user ratings, had the potential to nudge individuals towards privacy-respecting applications. Similarly, Balebako *et al.* [3] suggest that nudges can support users in making more optimal decisions in privacy when it comes to location sharing. They argue that individuals, left unaided, might well make regrettable privacy decisions due to the cognitive load caused by having to consider all possible ramifications of a single privacy decision. Similarly, Almuhimedi *et al.* [2] investigated user awareness of privacy invasion by making usually invisible data sharing, visible . Almuhimedi *et al.* demonstrated that the majority were nudged to reassess their privacy permissions when data was presented.

Authentication nudge studies have delivered disappointing results so far [12, 17]. One study attempted to exploit the Decoy Effect [27]. This design involves giving users three choices: one inferior, one very expensive, and a middle-of-the-road option that designers want people to choose. The decoy study [45] offered users their own password choice, a complex hard-to-remember password and the alternative they really wanted users to choose: a long and memorable password. The relative strengths of the three passwords was displayed to influence choice. The results were disappointing [45].

Another nudge effort that has enjoyed much research attention is the password strength meter. These mechanisms provide strength feedback, either post-entry or dynamically. Mechanisms can provide colour indicators, strength indicator bars, or informative text [8].

Ur *et al.* [47] compared a number of different password strength meters and discovered that meters impacted password strength. However, they tested their meters using a Mechanical Turk survey. The fact that the created passwords carried no cost might have led to respondents formulating somewhat unrealistic passwords. Ur *et al.*'s study was an essential first step in exploring these kinds of interventions, giving us hope that nudges could be designed to work in the wild too.

Sotirakopoulos [40] attempted to influence password choice by providing dynamic feedback. No difference between a horizontal strength meter and the comparison to peer passwords emerged. Vance *et al.* [49] also reported that password strength meters only impacted password strength in conjunction with an interactive fear appeal treatment condition that included a message on the seriousness of the threat. An interactive password strength meter and a static fear appeal did not impact password strength.

Egelman *et al.* [17] did test the impact of providing password meters in the wild. They found that the meters made no observable difference to password choice, unless users perceived the account to be important. If people *do not* attribute value, then it is understandable that the password meter makes no difference to their choice.

Privacy nudges have been more successful than authentication nudges so far. Privacy choices, however, entail people having a choice between two fairly equivalent options [10, 26]. Nudging in authentication does not match this pattern of use and, in fact, initial studies on nudges in authentication have delivered mixed results, as described above. Still, nudges have been successfully deployed in other application areas, and at least two explanations for the lack of success in authentication. It might be the case that authentication is unsuited to nudging influence. On the other hand, it could be that a success authentication nudge is yet to be discovered discovered.

Much nudging in authentication has focused on password strength meters. We thus carried out a study to extend the evidence base by testing a number of visual authentication nudges. We tested nudges which focus on cognitive effects (e.g. social norms and expectation) that have rarely been tested in the authentication context.

We displayed different visual nudges during password creation events, in order to determine whether they exerted any influence over users during password creation.

3 Method

Current efforts to improve password choice focus primarily on the individual. However, situational and contextual influences could minimise the impact of individually-focused interventions [31]. Furthermore, social influence is a strong driver of compliance [11, 35]. Interventions could conceivably exploit the power of social norms to influence individual behaviours [4]. Since our target users in this study were students this context includes the University and their School. Visual nudge figures were created beforehand and displayed statically to ensure that all students saw exactly the same image. A dynamically-updated image might have confounded results because participants would then have seen different images, confounding our results. We designed one nudge for each cognitive effect we tested and that has led to positive results in other research areas.

Due to the exploratory and "in the wild" nature of this study, we decided to evaluate a range of cognitive effects with one nudge each, instead of focusing on one effect and creating several variations of nudges to exert influence in that one area. If a positive impact resulted, further exploration of the effect and variations of the nudge would be a direction for future research.

3.1 The Nudges

We conducted two studies with a similar experimental design: In each of the two studies the nudges served as an independent variable with six levels, a control group that did not receive any intervention and five different nudge conditions. From the nudges described below, nudges N1 to N5 were tested in study 1. Nudges N6 to N8 were tested in study 2 along with a replication of N2 and

N3. The dependent variables were (1) password strength measured with the strength estimator `zxcvbn.js` [54] and (2) password length (for further information see the Apparatus section). All nudges were presented to the participants on the registration page of a web-based university application that is described in the Apparatus Section below.

- **N0: Control**. The control group was presented with the standard registration page which asked users to "Choose a Password".

- **N1: Subconscious Mind.** *Testing the Priming Effect*. In authentication people are almost always prompted to provide a new password with the words: "Choose a Password". It is possible that this phrase could be partly responsible for one of the most common passwords being "password". If this admittedly subtle prime is a causative we ought to be able to influence choice by changing the word to "secret", and then see how many participants choose the password 'secret'. Thus, in our first experimental condition the phrase "*Choose a Password*" was replaced with "*Choose a Secret*".

- **N2: University Context**. *Testing the Expectation Effect*. Instead of mandating password strength requirements, the participants were shown the static graphic (Figure 2) that suggests that their password ought to be stronger than the average password chosen by other students.

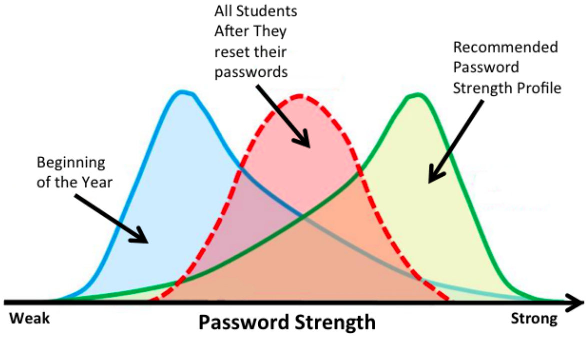


Figure 2: Expectation Effect Nudge Graph [37]

- **N3: School Context**. *Testing the Strength of In-Group Effect*. We suggested that participants identify themselves with students within their school, referred to as SoCS (Figure 3) in the graphic that was shown to them.

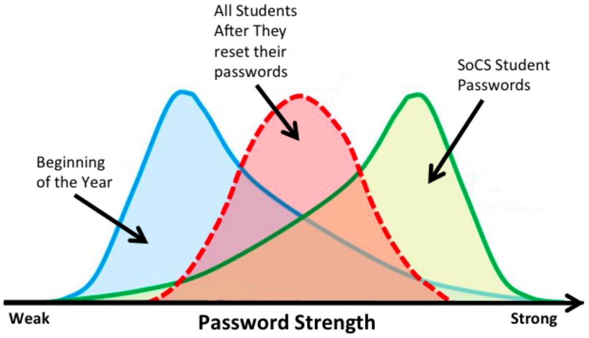


Figure 3: The In-Group Nudge Graph [9]

Some people argue that people do not know how strong their passwords are. To determine whether dynamic feedback reflecting the strength of their passwords would make a difference, we superimposed the arrow shown in Figure 4 over the images in Figures 2 and 3, giving us conditions N4 and N5. The strength feedback was based on the same strength estimator `zxcvbn.js` [54] that was used to calculate password strength in all experimental conditions (see section Apparatus for further details).

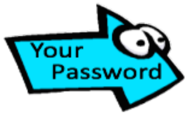


Figure 4: Strength Indicator

- **N4: University Context & Feedback**. Testing the combination of the Expectation Effect graph, with an interactive password strength meter superimposed over it. This would theoretically allow the user to see where on the x-axis their password is located, in terms of strength, as they entered it.

- **N5: School Context & Feedback**. Testing the combination of the In-Group Effect graph, with same dynamic strength feedback indicator as N4.

- **N6: Social Norm**. An image of eyes on a wall, appearing to "watch", makes people more likely to pay into an honesty box and also has the potential to reduce littering [4]. Given the impact of displayed eyes in other fields we considered it worthwhile to test whether the perception of being watched would encourage stronger passwords we displayed a pair of eyes above the password entry field.

For the final two conditions we asked the participants to reflect on the strength of their passwords to make them pause and think about the password. Due to the constraints imposed by the ethics committee, no self-report free-form text was available. Instead, the participants were asked to rate the perceived strength of their password on a scale below the Figures as shown in 2 and 3 respectively, giving us conditions seven and eight.

This was intended to drive processing up to the System 2, deliberate level, of processing, to offset the automaticity they might be subject to while choosing passwords.

- **N7: University Environment & Reflection**. This treatment displayed the same image as N2, and asked the user to rate the strength of the password he or she had just entered. The instruction referred to them as 'a student' in order to highlight their University affiliation;
- **N8: School Environment & Reflection**. This group displayed the same image as N3, in addition to asking the user to rate the strength of their password. The instruction referred to them as 'a computing science student', once again to emphasise their in-group affiliation.

Apparatus. The nudges were tested using a web-based university application where students were provided with coursework deadlines, timetable information and project allocations. The authentication scheme was based on standard alphanumeric authentication, i.e. a username and a password.

We did not enforce a password policy nor a time limit for password creation as we wanted to test the sole impact of the nudges on password creation. However, the university where the study took place generally suggests that passwords should be at least eight characters long (passphrases are recommended), include at least one non-letter and should be changed at least once a year). Access to the system was only possible with a student ID and from within the campus network. As it was not possible to install password managers on the lab machines and the use of personal laptops was not allowed, the use of password managers was largely avoided. If participants used a password manager on another device they would have to enter the stored password manually.

The website was used from October 2014 to April 2015 for Study 1 and from October 2015 to April 2016 for Study 2, thus for two consecutive academic years.

Password strength was calculated with the help of `zxcvbn.js` [54]. This in an open-source and JavaScript strength calculator that uses pattern matching and minimum entropy calculation. For this research, the score metric was used. It delivers a strength value between 0 and 4 that indicates whether the number of guesses required to break the password is less than 10^2 (score 0), 10^4 (score 1), 10^6 (score 2), 10^8 (score 3), or above (score 4) [1]. For example, the password "password" gets a rating of 0, where a password like "bootlegdrench42" is issued a rating of 4. Hence, the scores are not evenly spaced, the scale is exponential and the resulting data therefore ordinal. Password length was measured as the number of characters used for a password. For privacy and security reasons the participants' passwords were never transmitted unhashed: strength was calculated locally and the hashed password transmitted to the server.

Sample. All participants were students enrolled in technical courses, mainly specialising in Computer Science that used the web application for their studies. In Study 1, a total of 587 individuals registered to use the web application. Some students exercised their right to opt out, leaving 497 participants taking part in the study. In Study 2, 816 individuals registered to use the web application and created a password, of those 776 participants took part in the study.

Ethics. The study was conducted in agreement with the university's Ethics board. Participants were able to opt out of the experiment at enrolment, and about 15% did so. The school management would not allow us to contact the students to ask any questions because of the sensitivity and secrecy of passwords, and the fear that the students would interpret any communication as an indication that their passwords had been compromised. For privacy reasons we were not permitted to report any demographic information. We ensured that we used only public domain images during the course of this study.

Procedure. The participants were randomly assigned to the control condition or one of the experimental conditions by a script embedded in the enrolment web page.

They were informed that their actions were being logged and could be used for research purposes. They were presented with a consent form, allowing them to opt out of the experiment, but still benefit from use of the website. In all experimental conditions, the nudges were presented above the password entry field during enrolment and also during subsequent password creation events.

4 Results

4.1 Study 1

The data were first analyzed in terms of preconditions for statistical procedures such as sampling distribution and missing values. The descriptive statistics of Study 1 are listed in Table 2. The mean is reported as μ, the standard deviation as σ, the median as $\tilde{x}$ and the interquartile range as IQR. Overall, the average password strength

[1] https://blogs.dropbox.com/tech/2012/04/zxcvbn-realistic-password-strength-estimation/ (accessed 28th September 2017)

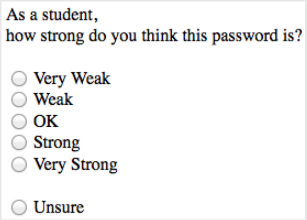
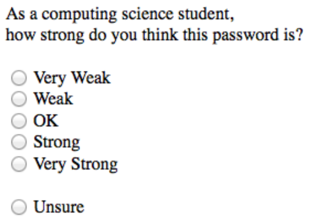
NUDGE	PROMPT	COND
Control	”Choose a Password”	N0
Framing	”Choose a Secret”	N1
Expectation	Graph in Figure 2	N2
In-Group	Graph in Figure 3	N3
Expectation & Dynamic Strength	Graph in Figure 2 + Figure 4	N4
In-Group & Dynamic Strength	Graph in Figure 3 + Figure 4	N5
Social Norms		N6
Expectation & Reflection	Graph in Figure 2 + Reflection  As a student, how strong do you think this password is? ○ Very Weak ○ Weak ○ OK ○ Strong ○ Very Strong ○ Unsure	N7
In-Group & Reflection	Graph in Figure 3 + Reflection  As a computing science student, how strong do you think this password is? ○ Very Weak ○ Weak ○ OK ○ Strong ○ Very Strong ○ Unsure	N8

Table 1: Tested Nudges (N = Nudge Condition)

was rated with $\widetilde{x}$ = 1 and IQR1 = 0, IQR3 = 3. The distribution of the password strength scores is depicted in Figure 5. The average password length was μ = 9.59 (σ = 3.25) and $\widetilde{x}$ = 9. The shortest password comprised 3, the longest 32 characters.

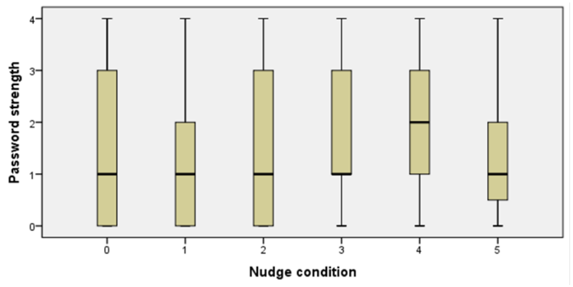


Figure 5: Password strength Study 1

Due to a non-normal sampling distribution and the password strength being measured on an ordinal scale, Mann-Whitney-U tests were conducted to compare each of the five nudge conditions with the control group. The tests were run for both password strength and length using the Benjamini-Hochberg procedure for the correction of p-values. The effect size was calculated using Cliff’s Delta [13, 14] which does not make assumptions about the underlying data distribution.

Password strength in the Priming group (N1, $\widetilde{x}$ = 1) did not differ significantly from the control group (N0, $\widetilde{x}$ = 1), U = 3419.00, z = -.351, p = .726, Cliff’s Delta = .03 [-.14, .2]. We counted two uses of the word “secret” as password in this group. However, none of the other participants, who were primed with the prompt “Provide a password” used the word ‘password’, so there is no evidence of a strong priming effect.

Likewise, there was no significant difference between the control group and the conditions In-Group Effect

		Strength Estimation					Length				
	Subjects	$\widetilde{x}$	IQR1	IQR3	min	max	μ	σ	$\widetilde{x}$	min	max
N0	82	1	0	3	0	4	9.46	3.83	8.00	4	32
N1	86	1	0	2	0	4	8.91	2.72	8.50	3	17
N2	83	1	0	3	0	4	9.95	3.51	9.00	6	24
N3	81	1	1	3	0	4	10.33	3.57	9.00	6	22
N4	82	2	1	3	0	4	9.76	2.53	9.00	6	17
N5	83	1	0	2	0	4	9.17	3.01	8.00	6	21

Table 2: Descriptive Statistics of user-generated passwords in Study 1 (μ = mean, σ = standard deviation, $\widetilde{x}$ = median, IQR = Interquartile range).

(N3, $\widetilde{x}$ = 1), U = 2955.5, z = -1.251, p = .211, Cliff's Delta = -.11 [-.28, .06], and In-Group effect with feedback (N5, $\widetilde{x}$ = 1), U = 3272.5, z = -.439, p = .661, Cliff's Delta = -.04 [-.21, .13]. Finally, also the comparison between the password strength of the control group and the Expectation Effect with feedback group (N4, $\widetilde{x}$ = 2) yielded an insignificant result due to the Benjamini-Hochberg adapted p-value threshold, U = 2708.00, z = -2.207, p = .027, Cliff's Delta = -.19 [-.36, -.02]. The same was true for the similar condition N2 without feedback ($\widetilde{x}$ = 1), U = 3080.00, z =-1.084, p = .278, Cliff's Delta = -.09 [-.26, .08]. The effect sizes are graphically depicted in Figure 6.

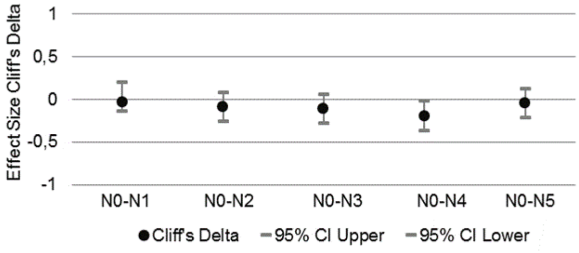


Figure 6: Effect sizes of the password strength comparisons

Password length, among others (such as use of different types of characters or of upper and lower cases), can be one factor contributing to stronger passwords. However, in line with the findings on password strength no significant effect on password length could be proven.

4.2 Study 2

The data analysis for Study 2 followed a similar approach to the one for Study 1. The descriptive statistics of Study 2 can be found in Table 3. Overall, the average password strength was rated with $\widetilde{x}$ = 2, IQR1 = 0 and IQR3 = 3. The distribution of the password strength scores is shown in Figure 7. The average password length was μ = 10.02 (σ = 2.57) and $\widetilde{x}$ = 9. The shortest password comprised 4, the longest 25 characters. Similar to Study 1, Kolmogorow-Smirnow tests and a visual inspection revealed deviations from a normal distribution leading to the use of nonparametric Mann-Whitney-U tests. From the original N=776 data sets, 39 had to be excluded due to technical problems with the java script strength estimator.

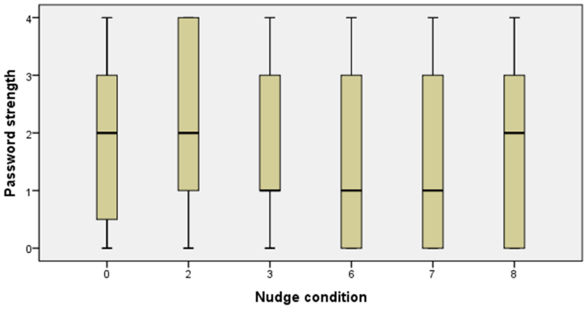


Figure 7: Password strength Study 1

Again, the control group was tested against the five experimental groups N2, N3, N6, N7 and N8 in pairwise comparisons using nonparametric Mann-Whitney-U tests and the Benjamini-Hochberg procedure for p-value correction. However, the experimental groups did not differ significantly from the control group, neither in terms of password strength nor length.

		Strength Estimation					Length				
	Subjects	$\widetilde{x}$	IQR1	IQR3	min	max	μ	σ	$\widetilde{x}$	min	max
N0	124	2	0.25	3	0	4	10.13	2.70	10.00	6	24
N2	124	2	1	4	0	4	10.23	2.56	10.00	6	19
N3	120	1	1	3	0	4	10.06	2.73	9.00	4	25
N6	124	2	0	3	0	4	9.80	2.42	9.00	6	16
N7	121	1	0	3	0	4	9.88	2.24	9.00	6	15
N8	124	1	0	3	0	4	10.02	2.77	9.00	5	17

Table 3: Descriptive Statistics of user-generated passwords in Study 2 (μ = mean, σ = standard deviation, $\widetilde{x}$ = median, IQR = Interquartile range).

4.3 Hypothesis

Based on our findings we conclude that **H1** is not supported. The presence of the visual nudges we tested did not lead to longer and stronger passwords.

5 Discussion & Reflection

Research designs strive to maximize three criteria when collecting evidence: *generalizability, precision*, and *realism*. Since it is impossible to maximize all of these, all research designs exhibit deficiencies in one or more of these dimensions [33].

For example, survey research is generalizable whereas lab experiments are more precise, and field experiments (and case studies) are realistic while being less precise due to low controllability of confounding factors. Researchers who utilize laboratory experiments to study security behaviors can control the environment and fix a number of research variables, but realism suffers because this setting only mimics reality. Field experiments are far more realistic, but are undeniably less precise. Surveys perform poorly in terms of realism and precision.

The best research projects will probably combine the findings of surveys, lab experiments and field studies in order to offset the deficiencies of individual methods. A number of surveys have been carried out in this area [47], giving us a measure of generalizability. We contribute to the field by carrying out and reporting on our nudge-related field study, adding realism to previous findings.

After the unexpected outcome of our studies we reflected on reasons for the eight nudges seemingly making no significant impact on users' password choices. The possible explanations we considered fall into two broad categories. The first concerns potential methodological and statistical issues. The second concerns the participants: their task, aims and perceptions.

(1) Methodological considerations

The strength metric

For the purpose of our study, we decided to measure password strength with the password strength meter `zxcvbn.js`. We made this decision based on the fact that it was open-source, uses pattern matching and searches for minimum entropy. However, the score rating provided by `zxcvbn.js`, and used in our study, measures password strength on an ordinal scale with '0' indicating the number of guesses required to break the password being less than 10^2 and '4' assigned to a password requiring over 10^8 guesses. The clustering of data into 5 artificial categories, however, suppressed data variance. For example, if the number of guesses to crack a password in the control group was 1100 and that of a password in one of the experimental conditions was 9900, both passwords would be assigned a score of 2 indicating between 10^3 and 10^4 guesses required to break the password. Thus, the difference in the data would not be reflected in the score.

Although we were not aware of any alternatives when we commenced our study, there are now wrappers to run `zxcvbn.js` completely offsite. We used the open-source version of the client. To protect the participants' passwords, we did not transmit unhashed passwords — strength was calculated locally and the hashed password, together with its strength rating, transmitted to the server. The unavailability of the raw data later prevented us from calculating alternative strength estimations that might have provided a greater variance and a categorization closer to the real distribution.

The loss of information negatively affected the analysis so that it is possible that existing effects were not detected. We would therefore recommend the use of a richer classification mechanism for further studies of this type.

Non-parametric tests

Another issue is that the ordinal password strength scale required the use of a non-parametric test. In our study, the Mann-Whitney-U test was conducted for the pairwise comparisons of the experimental and control groups. (Non-parametric tests make no assumptions about the probability distributions of the measured variables, as compared to parametric tests that require normality. Non-parametric tests are indicated where the normality requirement is violated: they are more robust against outliers and use characteristics such as the median and the central tendency to describe a distribution.)

However, if the requirements of parametric tests are met, the test power of such parametric tests is, generally speaking, higher than that of non-parametric tests. Tests with a higher test power are more likely correctly to reject a null hypothesis (no difference between groups) when the alternative hypothesis (difference between groups) is true. In our case, that means that a test with a higher test power might well have detected an existing difference between the experimental and the control groups, which the non-parametric test did not reveal. To quantify that potential impact we conducted a G*Power analysis [19] to compare the test power of an non-parametric Mann-Whitney-U test to an independent t-test. We fixed $\alpha = .05$ and sample size on 80 people per group similar to the sample sizes in study 1. We then manipulated the effect size Cohen's d required by G*Power to compare the results. We found the changes in test power to be below 2% (see Table 4).

Thus, the use of non-parametric tests might have contributed to our negative findings. Still, the analysis shows that the influence of the non-parametric vs. parametric test is rather small, whereas the influence of the effect size is much bigger. In our study 1 the effect sizes were only between .03 and .19. For future studies, it would therefore be beneficial to use a study design and password strength metric that offers greater variance and supports the deployment of parametric tests.

(2) Participant Considerations

Authentication is Complex

The focus of the experiment was solely on the password choice task. The user's specific goals and needs, in the context of the task, might not have been considered sufficiently. This is especially relevant in that security tasks are often secondary rather than primary goals [55]. Depending on the context, users aim to read mails, book a hotel or check their course details and grades. Many users might consider authentication a necessary evil that has to be overcome to reach a primary goal. It is just one among many elements in the choice-making ecosystem. Thus, it might be that the nudges tested in this study are not ineffective *per se* but that they were not powerful enough in the authentication context. For future work it would therefore be important to analyse the users' choice-making ecosystem holistically before designing a simple user interface display "intervention" to nudge users towards a change in behaviour.

Password Strength Perceptions

Studies by Ur *et al.* [48, 46] found that users' perceptions of what makes a strong password differs from the actual password security. Users succumb to several misconceptions. For example, many overestimated the security benefit of including a digit compared to other characters and underestimated the decrease in security that resulted from their use of common keyboard patterns. This might be an indication that users lack the understanding of what specifically contributes to a strong password. In the context of our results this means that the nudges might not have sufficiently enhanced the users' understanding of what makes a password stronger. Thus, feedback on password strength might be promising direction for future research.

However, the success of feedback meters in the literature, that dynamically display password strength to the user and thus constitute one form of feedback, is mixed. Studies in which users were not actively prompted to consider their password reported only marginal effects, whereas in others the meters weren't even noticed by users [7]. This confirms our earlier recommendation that future studies should engage in analyzing the targeted users, their tasks and mental models in a holistic way before designing nudges. Apart from that, one could assume that nudges which not only transport the message that passwords should be secure but also offer guidance on how to achieve this, might be more effective. This assumption, however, needs to be tested.

Password Reuse

People reuse passwords across sites [16, 25], a fact relied on by hackers globally. In a recent study by Wash and Rader [52] password re-use behaviour was investigated. The authors showed that for important accounts, such as university accounts, people re-used stronger and more complex passwords as compared to less important accounts. Thus, the difference between strong, re-used passwords (in the control group), and strong "nudged" passwords (in the experimental groups) might have been too small to detect. Apart from that, our nudges were designed to target the password creation process. If par-

			Test power	
Cohen's d	Sample size	α	independent t-test	Mann-Whitney-U test
0.1	80	0.05	0.1550283	0.1516025
0.2	80	0.05	0.3499859	0.3393193
0.3	80	0.05	0.5965318	0.5796253
0.4	80	0.05	0.8089716	0.7928030
0.5	80	0.05	0.9336887	0.9238465

Table 4: Comparative analysis of test power using the G*Power software [19].

ticipants were reusing passwords they might well have ignored the nudges altogether, rendering them impotent.

Nudges & Complex Behaviours

Nudges are targeted at at making users change a default rule or behaviour. This, however, isn't an easy task. Sunstein [42] explains that there are a number of reasons for users clinging to their default password behaviours despite the presence of nudges.

1. *First*, changing default behaviour requires active choice and effort, and the option towards which the person is being nudged might be more effortful than the default option. Nudges might be more effective where people have to choose between two options that are similar in terms of effort. In authentication, however, options are seldom similar. A stronger password increases the cost for the user in terms of time and memory load. Reusing a password is *much* less effortful than coming up with a new one.

2. *Second*, departing from the default way might be perceived to be risky and only become a realistic option if people are convinced that they should indeed change their default behaviour.

3. *Third*, people are loss averse. If the default is viewed as a reference point, a change might be considered a loss of routine or long-memorized passwords.

4. *Fourth*, password choice is cognitively expensive [20], not a simple activity. If people are already depleted for some reason they are even less likely to choose a stronger password and a visual nudge is hardly going to have the power to mitigate this.

In future studies, it would be interesting to test nudges that offers a benefit in return for the extra perceived effort. One idea suggested by Seitz, Von Zezschwitz and Hussmann [45] is to reward users with a stronger password by allowing them to keep the password longer than a weak password: applying a strength-dependent aging policy. Thus, weak passwords would be easier to type and memorise but would have to be changed more frequently whereas stronger passwords are harder to type and memorise but could be kept longer.

Limitations

As described above, this study was conducted in the field with a high degree of realism. However, field studies lack the controllability of laboratory experiments, even more so in our case where the requirements of the ethics committee constrained us in terms of collecting demographic and additional information to preserve participant privacy and anonymity.

Furthermore, the use of the password strength scores that are not evenly distributed resulted in a loss of variance and a decrease in test power. Future studies should therefore consider and compare other possibilities to quantify password strength (also see Methodological Considerations and Lessons Learned sections).

Another limitation is the limited generalisability of the sample that predominantly consisted of Computing Science students. It can therefore be expected that the sample was somewhat biased towards technically-adept, young and male participants. Another limitation concerns the design of the Figures that were presented to the participants in the experimental conditions N2 University Context, N3 School Context and the related conditions N4, N5, N7 and N8. Participants received dynamic feedback on their password strength in relation to the graph in N4 and N5. In N7 and N8 they were asked to rate the perceived strength of their passwords in relation to the graph. However, the participants in N2 and N3 did not receive feedback on their passwords. There, the nudge was intended merely to create the impression that the participants' peer group passwords ought to be stronger than the average. In retrospect, it seems that this, on its own, did not have the power to impact password strength.

Lessons learned

A number of lessons were learned during the course of this research. We suggest the following implications that might be useful to security researchers conducting future studies:

1. **First**, the categorization of the password strength based on the `zxcvbn.js` metric used in our study resulted in a loss of information and variance of password strength. It also required the use of nonparametric tests that, generally speaking, have a lower test power than parametric tests. It is therefore possible that existing small effects were not detected. For future studies, it would be advisable to explore and compare other metrics, e.g., the exact number of guesses required to break a password.

2. **Second**, based on the literature, nudges seem to be more effective where choices are equal in terms of effort. Stronger passwords will undeniably require more effort both in terms of memory load and typing time and complexity.

3. **Third**, authentication nudges might not come into effect when users re-use passwords. Therefore, it would be interesting either to assess password reuse as a control variable, or to prevent users from reusing passwords, e.g. by applying an idiosyncratic password policy. In this case, the increased memory load would have to be acknowledged and compensated for in some way and such a policy might well introduce unanticipated and unwanted side effects.

4. **Fourth**, to better comprehend participants' understanding of secure password creation, we ought to conduct further studies exploring their mental models. It could also be useful to compare different user groups, such as laypersons and experts, who possess different levels of knowledge and perhaps engage in different decision-making strategies. Depending on the outcome of those studies, nudges that not only increase awareness, but also offer guidance on how to create stronger passwords, might be a more promising approach.

6 Conclusion

The research reported in this paper investigated the viability of a number of nudges in the authentication context. We manipulated the choice architecture to encourage the choice of stronger passwords. We discovered that password strength was not impacted by the visual nudges.

Having reflected on our findings, we were reminded of the complexity of the password creation event. It is influenced by so many more factors than the mere appearance of the surrounding user interface. We learned some valuable lessons during the course of this research and we conclude the paper by presenting a list of these to assist other researchers wishing to work in this area.

Acknowledgement

We thank our shepherd for his support in refining this paper for the LASER workshop.

We obtained ethical approval from College of Science and Engineering at the University of Glasgow to carry out nudge-related research using the website (Approval #300140006).

The research reported in this paper was supported by the German Federal Ministry of Education and Research (BMBF) and by the Hessian Ministry of Science and the Arts within CRISP (www.crisp-da.de/).

References

[1] ABDULLAH, M. D. H., ABDULLAH, A. H., ITHNIN, N., AND MAMMI, H. K. Towards identifying usability and security features of graphical password in knowledge based authentication technique. In *Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on* (2008), IEEE, pp. 396–403.

[2] ALMUHIMEDI, H., SCHAUB, F., SADEH, N., ADJERID, I., ACQUISTI, A., GLUCK, J., CRANOR, L. F., AND AGARWAL, Y. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, ACM, pp. 787–796.

[3] BALEBAKO, R., LEON, P. G., ALMUHIMEDI, H., KELLEY, P. G., MUGAN, J., ACQUISTI, A., CRANOR, L. F., AND SADEH, N. Nudging users towards privacy on mobile devices. In *Proc. CHI 2011 Workshop on Persuasion, Nudge, Influence and Coercion* (2011), ACM.

[4] BATESON, M., CALLOW, L., HOLMES, J. R., ROCHE, M. L. R., AND NETTLE, D. Do images of 'watching eyes' induce behaviour that is more pro-social or more normative? A field experiment on littering. *Public Library of Science One 8*, 12 (2013), e82055:1–9.

[5] BHATTACHARYYA, D., RANJAN, R., ALISHEROV, F., AND CHOI, M. Biometric authentication: A review. *International Journal of u-and e-Service, Science and Technology 2*, 3 (2009), 13–28.

[6] BONNEAU, J., HERLEY, C., VAN OORSCHOT, P. C., AND STAJANO, F. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Security and Privacy (SP), 2012 IEEE Symposium on* (2012), IEEE, pp. 553–567.

[7] BONNEAU, J., HERLEY, C., VAN OORSCHOT, P. C., AND STAJANO, F. Passwords and the evolution of imperfect authentication. *Communications of the ACM 58*, 7 (2015), 78–87.

[8] Carnavalet, X. D. C. D., and Mannan, M. A large-scale evaluation of high-impact password strength meters. *ACM Transactions on Information and System Security (TISSEC) 18*, 1 (2015), 1–32.

[9] Castano, E., Yzerbyt, V., Paladino, M.-P., and Sacchi, S. I belong, therefore, I exist: Ingroup identification, ingroup entitativity, and ingroup bias. *Personality and Social Psychology Bulletin 28*, 2 (2002), 135–143.

[10] Choe, E. K., Jung, J., Lee, B., and Fisher, K. Nudging people away from privacy-invasive mobile apps through visual framing. In *IFIP Conference on Human-Computer Interaction* (2013), Springer, pp. 74–91.

[11] Cialdini, R. B., and Trost, M. R. Social influence: Social norms, conformity and compliance. In *The handbook of social psycholog*, D. T. Gilbert, S. T. Fiske, and G. Lindzey, Eds., 4 ed. McGraw-Hill, New York, 1998, pp. 151–192.

[12] Ciampa, M. A comparison of password feedback mechanisms and their impact on password entropy. *Information Management & Computer Security 21*, 5 (2013), 344–359.

[13] Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin 114*, 3 (1993), 494–509.

[14] Cliff, N. Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research 31*, 3 (1996), 331–350.

[15] Crawford, J. Assessing the value of formal control mechanisms on strong password selection. *International Journal of Secure Software Engineering (IJSSE) 4*, 3 (2013), 1–17.

[16] Das, A., Bonneau, J., Caesar, M., Borisov, N., and Wang, X. The tangled web of password reuse. In *NDSS* (2014), vol. 14, pp. 23–26.

[17] Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., and Herley, C. Does my password go up to eleven?: the impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, 2013), ACM, pp. 2379–2388.

[18] Farcasin, M., and Chan-tin, E. Why we hate it: two surveys on pre-generated and expiring passwords in an academic setting. *Security and Communication Networks 8*, 13 (2015), 2361–2373.

[19] Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods 39*, 2 (2007), 175–191.

[20] Gross, T., Coopamootoo, K., and Al-Jabri, A. Effect of cognitive depletion on password choice. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2016)* (2016), USENIX Association, pp. 55–66.

[21] Halpern, D. *Inside the Nudge Unit: How small changes can make a big difference*. WH Allen, London, 2015.

[22] Holden, J. Memorandum to the Heads of Executive Departments and Agencies. Implementation Guidance for Executive Order 13707: Using Behavioral Science Insights to Better Serve the American People, 2015. Sept 15. Executive Office of the President. Office of Science and Technology Policy `https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights\\-better-serve-american` Accessed 19 September 2016.

[23] Horcher, A.-M., and Tejay, G. P. Building a better password: The role of cognitive load in information security training. In *International Conference on Intelligence and Security Informatics, 2009. ISI'09* (The Hague, 2009), IEEE, pp. 113–118.

[24] Inglesant, P. G., and Sasse, M. A. The true cost of unusable password policies: password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, 2010), ACM, pp. 383–392.

[25] Ives, B., Walsh, K. R., and Schneider, H. The domino effect of password reuse. *Communications of the ACM 47*, 4 (2004), 75–78.

[26] Jeske, D., Coventry, L., Briggs, P., and van Moorsel, A. Nudging whom how: It proficiency, impulse control and secure behaviour. In *Personalizing Behavior Change Technologies CHI Workshop* (Toronto, 27 April 2014), ACM.

[27] Josiam, B. M., and Hobson, J. P. Consumer choice in context: the decoy effect in travel and tourism. *Journal of Travel Research 34*, 1 (1995), 45–50.

[28] Keith, M., Shao, B., and Steinbart, P. A behavioral analysis of passphrase design and effectiveness. *Journal of the Association for Information Systems 10*, 2 (2009), 63–89.

[29] Keith, M., Shao, B., and Steinbart, P. J. The usability of passphrases for authentication: An empirical field study. *International Journal of Human-Computer Studies 65*, 1 (2007), 17–28.

[30] Kritzinger, E., and von Solms, S. H. Cyber security for home users: A new way of protection through awareness enforcement. *Computers & Security 29*, 8 (2010), 840–847.

[31] Luck, M., and d'Inverno, M. Constraining autonomy through norms. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2* (Bologna, 2002), ACM, pp. 674–681.

[32] Magnet, S. *When biometrics fail: Gender, race, and the technology of identity*. Duke University Press, Durham, USA, 2011.

[33] Mcgrath, E. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000 (2nd ed.* Morgan Kaufman, 1995, pp. 152–169.

[34] Oliver, A. Is nudge an effective public health strategy to tackle obesity? Yes. *British Medical Journal 342* (2011).

[35] Orazi, D. C., and Pizzetti, M. Revisiting fear appeals: A structural re-inquiry of the protection motivation model. *International Journal of Research in Marketing 32*, 2 (2015), 223–225.

[36] Rayner, G., and Lang, T. Is nudge an effective public health strategy to tackle obesity? No. *British Medical Journal 342* (2011), d2168:1–2.

[37] Rosenthal, R., and Jacobson, L. *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart & Winston, Wales, 1968.

[38] Sasse, M. A. Usability and trust in information systems. In *Cyber Trust & Prevention Project*. Edward Elgar, 2005.

[39] Shay, R., Kelley, P. G., Komanduri, S., Mazurek, M. L., Ur, B., Vidas, T., Bauer, L., Christin, N., and Cranor, L. F. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (2012), ACM, pp. 7–26.

[40] Sotirakopoulos, A. *Influencing user password choice through peer pressure*. PhD thesis, The University Of British Columbia (Vancouver), 2011.

[41] Sunstein, C. R. Nudges Do Not Undermine Human Agency. *Journal of Consumer Policy 38*, 3 (2015), 207–210.

[42] SUNSTEIN, C. R. Nudges that fail. *Behavioural Public Policy 1*, 1 (2017), 4–25.

[43] Thaler, R. H., and Sunstein, C. R. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, 2008.

[44] THE BEHAVIOURAL INSIGHTS TEAM. Who we are, 2014. `http://www.behaviouralinsights.co.uk/about-us/` Accessed 19 Sept, 2016.

[45] TOBIAS SEITZ, EMANUEL VON ZEZSCHWITZ, S. M., AND HUSSMANN, H. Influencing self-selected passwords through suggestions and the decoy effect. In *EuroUSEC* (Darmtsadt, 2016), Internet Society.

[46] UR, B., BEES, J., SEGRETI, S. M., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. Do users' perceptions of password security match reality? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 3748–3760.

[47] UR, B., KELLEY, P. G., KOMANDURI, S., LEE, J., MAASS, M., MAZUREK, M. L., PASSARO, T., SHAY, R., VIDAS, T., AND BAUER, L. How does your password measure up? the effect of strength meters on password creation. In *Presented as part of the 21st USENIX Security Symposium (USENIX Security 12)* (Bellevue, 2012), USENIX, pp. 65–80.

[48] UR, B., NOMA, F., BEES, J., SEGRETI, S. M., SHAY, R., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. "I Added '!'at the End to Make It Secure": Observing password creation in the lab. In *Symposium on Usable Privacy and Security (SOUPS)* (2015), pp. 123–140.

[49] VANCE, A., EARGLE, D., OUIMET, K., AND STRAUB, D. Enhancing password security through interactive fear appeals: A web-based field experiment. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (Hawai'i, 2013), IEEE, pp. 2988–2997.

[50] VERWEIJ, M., AND HOVEN, M. V. D. Nudges in public health: paternalism is paramount. *The American Journal of Bioethics 12*, 2 (2012), 16–17.

[51] WARKENTIN, M., DAVIS, K., AND BEKKERING, E. Introducing the Check-off password system (COPS): an advancement in user authentication methods and information security. *Journal of Organizational and End User Computing (JOEUC) 16*, 3 (2004), 41–58.

[52] WASH, R., RADER, E., BERMAN, R., AND WELLMER, Z. Understanding password choices: How frequently entered passwords are re-used across websites. In *Symposium on Usable Privacy and Security (SOUPS)* (2016), pp. 175–188.

[53] WEIR, M., AGGARWAL, S., COLLINS, M., AND STERN, H. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proceedings of the 17th ACM Conference on Computer and Communications Security* (2010), ACM, pp. 162–175.

[54] WHEELER, D. L. zxcvbn: Low-budget password strength estimation. In *USENIX Conference 2016* (Vancouver, August 2016), USENIX, pp. 157–173.

[55] WHITTEN, A., AND TYGAR, J. D. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *USENIX Security Symposium* (1999), vol. 348.

[56] YEVSEYEVA, I., MORISSET, C., AND VAN MOORSEL, A. Modeling and analysis of influence power for information security decisions. *Performance Evaluation 98* (2016), 36–51.

[57] ZHANG, Y., MONROSE, F., AND REITER, M. K. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proceedings of the 17th ACM Conference on Computer and Communications Security* (2010), ACM, pp. 176–186.

An Empirical Investigation of Security Fatigue
The Case of Password Choice after Solving a CAPTCHA

Kovila P.L. Coopamootoo
Newcastle University
kovila.coopamootoo@ncl.ac.uk

Thomas Groß
Newcastle University
thomas.gross@ncl.ac.uk

M. Faizal R. Pratama
University of Derby

Abstract

Background. User fatigue or overwhelm in current security tasks has been called *security fatigue* by the research community [11, 24]. However, security fatigue can also impact subsequent tasks. For example, while the CAPTCHA is a widespread security measure that aims to separate humans from bots [26], it is also known to be difficult for humans [2]. Yet, to-date it is not known how solving a CAPTCHA influences other subsequent tasks.
Aim. We investigate users' password choice after a CAPTCHA challenge.
Method. We conduct a between-subject lab experiment. Three groups of 66 participants were each asked to generate a password. Two groups were given a CAPTCHA to solve prior to password choice, the third group was not. Password strength was measured and compared across groups.
Results. We found a significant difference in password strength across conditions, with $p = .002$, corresponding to a large effect size of $f = .42$. We found that solving a text- or picture-CAPTCHA results in significantly poorer password choice than not solving a CAPTCHA.
Conclusions. We contribute a first known empirical study investigating the impact of a CAPTCHA on password choice and of designing security tasks in a sequence. It raises questions on the usability, security fatigue and overall system security achieved when password choice follows another effortful task or is paired with a security task.

1 Introduction

The CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart [26]) is an important web security measure that differentiates humans from bots. It is widely used across the web [20], including at websites with a high traffic flow such as Facebook, Twitter, LinkedIn, Reddit and various email providers. Balancing both security and usability of CAPTCHAs continues to be a challenge [30, 7]. CAPTCHAs are known to be difficult for humans [2] and to pose usability issues [29].

Security research shows that on the one hand, users are fatigued and frustrated with security [11]. Literature quotes a *compliance budget* where employees comply either when there is no extra effort or after weighing the cost and benefits of extra effort [1]. In addition, a threshold exists beyond which it gets too hard and burdensome for users to maintain security [11]. Fatigue impacts current tasks with users being desensitized to security or rejecting security [24].

On the other hand, there is a suspicion that priming moderate effort can enhance security decisions such as for password choice. For example, Groß et al. [12] found that while password strength is weakest when users are cognitively depleted, moderate effort exertion is beneficial for password strength.

When registering an account, users are often asked to solve a CAPTCHA and to choose a password. Although online account registration forms often present the CAPTCHA challenge after password choice, unclear guidance into the data entry sequence or form refresh when an incorrect CAPTCHA is entered lead to a situation where the password is chosen after solving the CAPTCHA. In addition, when the Tor Anonymizer is detected, users are systematically asked to solve a CAPTCHA before they can register. Therefore the question arises whether the effort required to solve a CAPTCHA impacts password choice. While a handful of research have studied a link between cognitive effort and password choice or password management [12, 17, 10], for empirical investigation of the impact of effort previously spent on password choice, we are only aware of Groß et al. [12].

Research Question. We investigate the main RQ *"How does solving a CAPTCHA before creating a pass-*

word influence password choice?" We reproduce experimental design components of Groß et al. [12] and use validated methods from cognitive psychology to measure effort [25], stress[22, 23] and cognitive load [13].

Contribution. This paper contributes the first empirical investigation of the impact of security fatigue (experienced in a first task) on a subsequent security task. Our findings indicate that engaging in a CAPTCHA influences password choice. Solving a text- or picture-CAPTCHA lead to weaker password than not solving a CAPTCHA with a large effect size of 0.42.

Outline. In the rest of the paper, we first provide background research, followed with the study aims. We describe a pre-study. We then follow the main study methodology and provide the results before the discussion, limitations and conclusion.

2 Background

2.1 Security Fatigue

Previous research suggests that users often perceive security as a barrier that interferes with their productivity [11]. Subsequently, the term *security fatigue* was coined to describe the threshold of acceptance beyond which it gets too burdensome for users to maintain security [11]. In this sense, security fatigue describes an interference to current tasks and an additional step to be taken. *Security fatigue* has also been used to describe users' weariness or reluctance to experience anymore of something [24]. In particular, they reported that participants are tired, turned off and overwhelmed by security. The term has been used to describe user behavior both in the workplace [11] and for the general public [24].

2.2 CAPTCHA

A CAPTCHA is a program that can generate and grade tests that most humans can pass, but that current computer programs cannot [26]. By differentiating humans from bots, they are used for security applications such as preventing bots from continuous auto-voting in online polls, -registering to email accounts or preventing dictionary attacks [26]. Text-based CAPTCHAs require users to type alphanumeric characters from a distorted image, where popular ones include reCAPTCHA [27, 21] and BaffleText [5] whereas image-recognition CAPTCHAs are based on image problems, where examples include ASIRRA [8] and reCAPTCHA [21].

2.2.1 Usability & Security

While research propose that good CAPTCHAs ought to be both usable by humans and strong in resisting adversarial attacks [29, 7], CAPTCHAs are often difficult for humans [2, 4, 29, 9]. User reports on the perception, preferences and use of CAPTCHA [9] show that only every other user solves a CAPTCHA at first try, with character distortion named as the main obstacle.

Early CAPTCHAs have been broken by object recognition algorithms [18, 4] and segmentation [30, 7]. Yan and El Ahmad [30] exploited flaws in a word-image scheme via simple attacks and found that it was easy to separate foreground text from the background and that the scheme was vulnerable to segmentation attacks and dictionary attacks when English words was used. Bursztein et al. [3] provided an enhancement to the process of attacking text CAPTCHAs and they proposed randomizing the CAPTCHA length and individual character size, creating a wave shape and collapsing or overlaid lines, for improved protection against attacks. There are also claims of breaking the latest of Google's image reCAPTCHA [21].

2.3 Text Password

Text passwords are the cheapest and most commonly used method of computer authentication. However, a large proportion of users are frustrated when forced to comply to password policies such as monthly reset [15]. Effort and tiredness to a state of cognitive depletion causes users to choose weaker passwords [12], providing an indication that effort is necessary for the creation of strong passwords.

3 Aim

We investigate the main RQ *"How does solving a CAPTCHA before creating a password influence password choice?"*.

3.1 Impact on Password Strength

Password strength varies according to cognitive state, that is whether the user is depleted or fresh [12].

Research Question 1 (RQ⊨P)**.** *How does the strength of a password chosen after solving a CAPTCHA differ from not solving a CAPTCHA?*
$H_{P,0}$: *Solving a CAPTCHA does not impact password strength*
$H_{P,1.1/P,1.2}$: *Solving a [text-CAPTCHA/picture-CAPTCHA] causes weaker passwords than in the CONT condition.*

3.2 Effort Exerted

In Section 2.2, we reviewed literature exposing the difficulties of solving CAPTCHAs. These difficulties entail user effort.

Research Question 2 (RQ⊨E). *How does the overall effort of solving a CAPTCHA and choosing a password differ from only choosing a password?*
$H_{E,1.1/E,1.2}$*: Solving a [text-CAPTCHA/picture-CAPTCHA] causes exertion of more effort than not solving any CAPTCHA.*
$H_{E,1.3}$*: Solving a text-CAPTCHA causes exertion of more effort than solving a picture-CAPTCHA.*

3.3 Performance

The text and picture-CAPTCHAs can pose difficulties that affect performance differently, for example while the distortions of text-CAPTCHAs are problematic for the user, recognition or picture quality might do so for picture-CAPTCHA.

Research Question 3 (RQ⊨F). *How does the type of CAPTCHA impact the time spent, success rate and results checking rate?*
$H_{T,1}$*: Solving a text-CAPTCHA requires more time.*
$H_{S,1}$*: Solving a text-CAPTCHA has lower success rate.*
$H_{R,1}$*: Solving a text-CAPTCHA has a higher results checking rate.*

4 Pre-Study

Before the main study reported in this paper, we designed a pre-study to compare password strength across two effort inducing conditions.

4.1 Aim

Research Question 4. *How does password choice differ between different conditions of effort, in particular across control, text-CAPTCHA and 2-digit multiplication?*

4.2 Method

4.2.1 Participants

Participants were recruited on the university campus via adverts and flyers. They were paid a time compensation of $6.5. Sample size was $N = 40$, with 17 women, mean age 26.75 years ($SD = 7.540$). 9 participants were from a computer science background.

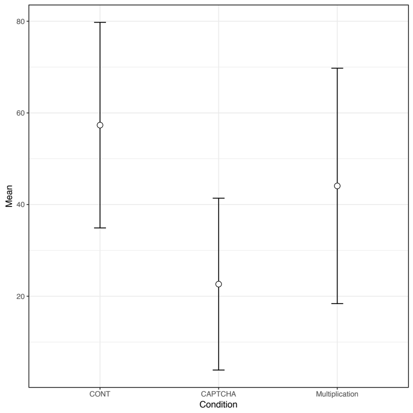


Figure 1: Confidence Intervals of the means of password strength score by condition. (Pre-Study)

4.2.2 Procedure

The procedure consisted of (a) a pre-task questionnaire for demographics, (b) a mood questionnaire, (c) a puzzle manipulation, (d) a password entry for a mock-up GMail registration, (e) a mood questionnaire and (f) a debriefing questionnaire.

We choose two puzzles: an example of the widely used text-CAPTCHA and a 2-digit multiplication. We choose a 2-digit multiplication because it is known in cognitive psychology to consume effort [16]. Solving mental multiplication problems has been shown to engage cognitive effort in particular 2-digit multiplication [14]. In this task we asked participants to solve 48 x 97 (ensuring that one of the numbers was a prime number, since prime numbers ensure shortcuts are not being used). We provide more details of the text-CAPTCHA in Section 5.2.2, which is also used for the main study.

We designed a between-subject study where participants were randomly assigned to one of the three groups. We ended up with 12 participants in the CONT, 14 in the text-CAPTCHA condition and 14 in the 2-digit multiplication condition.

4.3 Results and Discussion

Similar to Groß et al. [12], we measure password strength via password meter with NIST amendments. We computed a one-way ANOVA with the password strength score as dependent variable. There was no statistically significant effect of the experiment condition on the password strength score, $p = .074 > .05$.

Considering the intervals plot of Figure 1, we observe that the confidence intervals of the conditions CONT and CAPTCHA barely overlap, asking for further investigation. In terms of effect size, we see an effect size of Hedges' $g = 0.99$ $[0.17, 1.81]$.

5 Main Study Method

5.1 Participants

Participants were recruited on campus via adverts and flyers, and the experiment was run in a dedicated quiet zone. Participants were paid a time compensation of $6.5 for completing the experiment, which lasted between 10-15 minutes. The sample consisted of university students, $N = 66$, of which 30 were women. The mean age was 21.79 years ($SD = 3.223$). Participants were not from a computer science background, 38 were local nationals, and 63 reported undergraduate education.

5.2 Procedure

We designed a between-subject experiment, via experimental design guidelines [6]. We induce the independent variable (IV) effort, with three levels: CONT, text-CAPTCHA and picture-CAPTCHA, further described in Section 5.2.2.

The procedure consisted of (a) a pre-task questionnaire for demographics, (b) a combined short stress and mood questionnaire, (c) a CAPTCHA manipulation, (d) a password entry for a mock-up GMail registration (a reproduction of [12]), (e) a combined full stress and mood questionnaire, (f) a questionnaire for task load and (g) a debriefing questionnaire. Figure 2 depicts the experiment design.

5.2.1 Block Randomization

We ensured that each condition had equal number of participants and that participants are randomly assigned across groups. We automated a random block assignment method and we ended up with an equal number of 22 participants in each condition.

5.2.2 Manipulation Tasks

Following the review in Section 2.2, we selected the character/text-CAPTCHA and image recognition/picture-CAPTCHA as the experimental conditions. We chose these two schemes because the text-CAPTCHA has been most popular and is still used by high traffic sites such as Facebook[1] while the picture-CAPTCHA is an option for the reCAPTCHA, which is the most used CAPTCHA according to online surveys [20]. In terms of security, both schemes are also known to suffer from security flaws and have been been broken by segmentation and machine learning attacks as seen in Section 2.2.

Text-CAPTCHA We generated a CAPTCHA image using Securimage PHP CAPTCHA[2], as shown in Figure 3. Securimage distorts the code and draws random lines over the image. We used a level of perturbation of 1.75 to induce effort. 1.75 is readable yet require some effort. The number of lines on the image was set to the default 5. This CAPTCHA was also used in the pre-study provided in the Appendix.

Picture-CAPTCHA The image reCAPTCHA challenge provides a sample image and 9 candidate images. It asks the user to select images similar to the sample [21], where the correct number of images vary between 2 to 4. In our picture-CAPTCHA condition, we tweaked the image reCAPTCHA process and asked participants to count the number of times a particular image appear, here the number of cats as shown in Figure 4. We estimated that the effort spent in clicking on all occurrences would be similar as counting the number of occurrences. Participants still have to recognize particular images, yet we maintain a similar user input (text entry) as in the text-CAPTCHA condition.

5.3 Measures

5.3.1 Password Strength

Similar to [12], we use password meter Web site[3] with NIST adjustments. In addition, we evaluated the zxcvbn password strength estimator [28]. Zxcvbn provides the number of guesses, $\log_{10}$ guesses and a zxcvbn score from 0 to 4.

5.3.2 Password Strategy and Re-Use

At the debrief, we asked participants if they re-used one of their existing passwords to register to the GMail account. We also queried for password strategy employed. We report these in the Appendix.

5.3.3 Brief Mood Inventory

As [12], we use a short form of a brief mood inventory (BMI). Because we merged the stress and BMI questionnaire, instead of a 5-point Likert-type items between 1 Disagree strongly and 5 Agree strongly, we used a the 4-point Likert of the stress questionnaire with items 1 Not at all, 2 Somewhat, 3 Moderately and 4 Very much.

5.3.4 Stress and Workload

The Spielberger State Trait Anxiety Inventory is one of the most used measures of anxiety and stress in psychol-

[1] https://www.facebook.com

[2] https://www.phpcaptcha.org

[3] http://www.passwordmeter.com

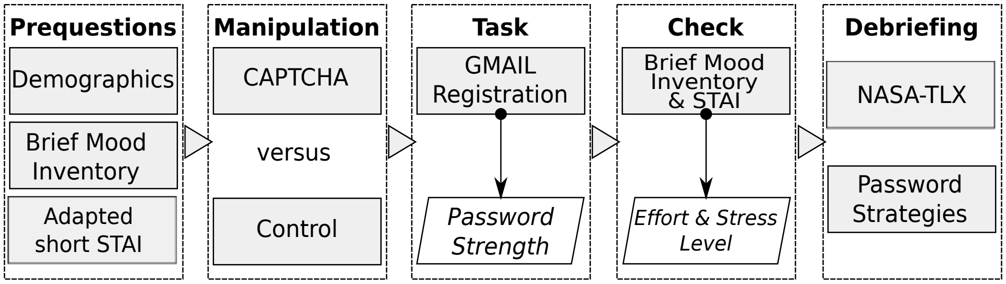


Figure 2: Overview of the experiment procedure. The experimental groups solved a CAPTCHA. The control group did not.

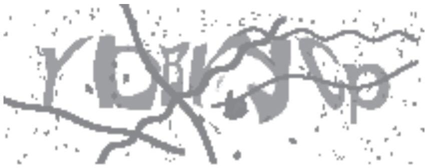


Figure 3: The text-CAPTCHA

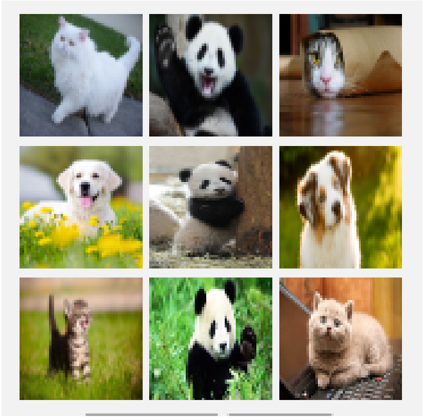


Figure 4: The picture-CAPTCHA

ogy [22, 23]. We chose the Y-1 questionnaire as measure of stress, with items towards how the participant felt in the experiment. In the post-task questionnaire, we included the full STAI.

NASA Task Load Index (NASA TLX) assesses mental workload via the dimensions of mental demand, physical demand, temporal demand, performance, effort and frustration [13].

5.3.5 Performance

Previous research recorded the time required [19] as well as the number of attempts required by participants to solve CAPTCHAs [9]. We also recorded the time taken in seconds to solve the CAPTCHAs, the final result entered by the participant and the number of times participants checked their results.

6 Results

All inferential statistics are computed at a significance level α of 5%. We estimate population parameters, such as standardized effect sizes of differences between conditions with 95% confidence intervals. A *confidence interval* is an interval estimate of a population parameter. The confidence level determines the frequency such confidence intervals would contain the population parameter if an infinite number of independent experiments were conducted.

6.1 Manipulation Check

6.1.1 Effort Exerted

We evaluate the null hypothesis $\mathsf{H}_{\mathsf{E},0}$: *Solving a CAPTCHA does not impact the effort exerted.* We calculate Diff_Tiredness (from the BMI in Section 5.3.3) as the difference in self-reported tiredness before the start of the CAPTCHA and after the registration.

A Kruskal-Wallis test showed that there was a statistically significant difference in Diff_Tiredness between the different conditions $\chi^2(2) = 12.736$, $p = .002 < .05$, with a mean rank Diff_Tiredness 26.64 for CONT, 44.00 for text-CAPTCHA and 29.86 for picture-CAPTCHA. We reject the null hypothesis $\mathsf{H}_{\mathsf{E},0}$.

We run 3 Mann-Whitney tests, with a Bonferroni-corrected significance level of $\alpha_\mathsf{B} = .0167$: (a) Diff_Tiredness was statistically significantly greater in the text-CAPTCHA condition than in the control condition, $U = 112.50$, $Z = -3.432$, $p = .001 < .0167$, $r = -.42$. This constitute a large effect size; (b) There was no significant difference in Diff_Tiredness between the picture-CAPTCHA condition and the control condition, $p = .556 > .0167$; (c) Diff_Tiredness was statistically significantly greater in the text-CAPTCHA condition than in the picture-CAPTCHA condition, $U = 140.50, Z = -2.577, p = .0100 < .0167, r = -.32$. We observe a medium to large effect size.

From these results, we conclude that the manipulation was successful in leading participants to exert more effort in the text-CAPTCHA condition than in the picture-CAPTCHA and CONT conditions.

6.1.2 Performance

We evaluate the null hypotheses $\mathsf{H}_{\mathsf{T},0}$/ $\mathsf{H}_{\mathsf{S},0}$/ $\mathsf{H}_{\mathsf{R},0}$: *A type of CAPTCHA does not impact the [time spent/success rate/results checking rate].* These DVs indicate participants' engagement and enable further evaluation of the success of the manipulation.

Time to solve CAPTCHA. We note that Levene's test for equality of variances across conditions is not significant with $p = .640 > .05$. With a two-tailed independent samples $t - test$, we find that participants in the text-based CAPTCHA condition ($M = 128.94$, $SD = 22.18$) have taken statistically significantly more completion time than participants in the picture-CAPTCHA condition ($M = 32.57, SD = 19.60$), $t(42) = 15.271$, $p < .001$. This gives an effect size of Hedges' $g = 4.52$ $[3.38, 5.64]$, a very large effect. We reject the null hypothesis $\mathsf{H}_{\mathsf{T},0}$.

Success at completing CAPTCHA. Although participants seem to have tried longer for the text-CAPTCHA, only five of them obtained a correct result; 20 did so for the picture-CAPTCHA. We run a χ^2 test, where we find a significant difference in correct results across the CAPTCHA conditions with $\chi^2(1, \text{N} = 44) = 23.21$, $p < .001$. The odds of having a correct result in the picture-CAPTCHA were 68 times higher than in the text-CAPTCHA. We reject the null hypothesis $\mathsf{H}_{\mathsf{S},0}$.

Checking Results. We counted the number of times participants checked their results. Since the number of checks failed Levene's test for equality of variance across the CAPTCHA conditions, with $p = .002 < .05$, we opt for the non-parametric Mann-Whitney test. The number of checks was significantly larger in the text-CAPTCHA condition than in the picture condition, $U = 19.5$, $Z = -5.365$, $p = .000 < .005$, $r = -.66$. This refers to a large effect size. We reject the null hypothesis $\mathsf{H}_{\mathsf{R},0}$.

6.2 Stress and Workload

We investigate the overall STAI score and the difference between the five pre/post STAI items and find no significant difference across the experimental conditions. We investigate NASA-TLX's across the dimensions of mental demand, physical demand, temporal Demand, performance, effort, frustration and the overall TLX_Score. We find no significant difference across conditions. We believe participants rated the last task only, that is the GMAIL registration.

6.3 Impact on Password Strength

We evaluate the null hypothesis $\mathsf{H}_{\mathsf{P},0}$: *Solving a CAPTCHA does not impact password strength* across both password strength measures.

6.3.1 Passwordmeter

The distribution of the Passwordmeter password strength score is measured on interval level and is not significantly different from a normal distribution for each condition. Saphiro-Wilk for (a) CONT, $D(22) = .976$, $p = .848 > .05$, (b) text-CAPTCHA, $D(22) = .967$, $p = .641 > .05$, (c) picture-CAPTCHA, $D(22) = .962$, $p = .534 > .05$. Levene's test for the homogeneity of variances show that the variances were not significantly unequal across conditions, $F(2,63) = 0.638$, $p = .532 > .05$.

We computed an one-way ANOVA with the password strength score as dependent variable. There was a statistically significant effect of the experiment condition on the password strength score, $F(2,63) = 6.716$, $p = .002 < .05$. We measure the effect size in Cohen's $f = .42$ from ($\eta^2 = .176$ $[0.043, 0.296]$) and Cohen's $\omega^2 = 0.148$. This constitutes a large effect. We provide the descriptive statistics in Table 1 and the means/interval plot in Figure 5. We reject the null hypothesis $\mathsf{H}_{\mathsf{P},0}$.

As post-hoc test, we conducted a Tukey HSD reporting that the password strength was statistically significantly lower in the text-CAPTCHA condition ($M = 31.05$, $SD = 29.16$) than in the control condition ($M = 67.68$, $SD = 37.02$) with $p = .002 < .05$. We have an effect size in Hedges' $g = 1.08$ $[0.44, 1.71]$.

Furthermore, the password strength in the picture-CAPTCHA condition ($M = 42.36$, $SD = 35.18$) was statistically significantly lower than in the control condition, $p = .042 < .05$. That is at an effect size in Hedges' $g = 0.69$ $[0.08, 1.29]$.

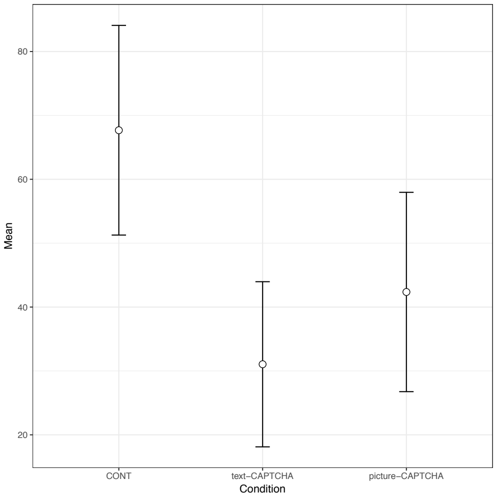


Figure 5: Passwordmeter

Figure 6: 95% Confidence Intervals on means of password strength scores by condition. (Main Experiment)

6.3.2 Zxcvbn

The distribution of the zxcvbn $\log_{10}$ guesses is measured on interval level and is not significantly different from a normal distribution for each condition. Saphiro-Wilk for (a) CONT: $D(22) = .184$, $p = .053 > .05$, (b) text-CAPTCHA: $D(22) = .148$, $p = .148 > .05$, (c) picture-CAPTCHA: $D(22) = .121$, $p = .538 > .05$. We also computed Levene's test for the homogeneity of variances. For the zxcvbn $\log_{10}$, the variances were not significantly unequal: for CAPTCHA and control conditions, $F(2,63) = 1.072$, $p = .349 > .05$.

We computed a one-way ANOVA with the zxcvbn $\log_{10}$ guesses as dependent variable. There was a statistically significant effect of the experiment condition on the zxcvbn $\log_{10}$ guesses, $F(2,63) = 4.665$, $p = .013 < .05$. We measure the effect size in Cohen's $f = .36$ from ($\eta^2 = .130$ $[0.016, 0.244]$) and Cohen's $\omega^2 = 0.1$. This constitutes a medium to large effect size. We provide the descriptive statistics in Table 2. Based on zxcvbn, we would equally reject the null hypothesis $H_{P,0}$.

As a post-hoc test, we computed Tukey HSD, reporting that password strength was statistically significantly lower in the text-CAPTCHA condition ($M = 6.38$, $SD = 2.84$) than in the control condition ($M = 8.66$, $SD = 2.11$), $p = .010 < .05$. We have an effect size measured in Hedges $g = 0.89$ $[0.27, 1.51]$. There was no significant difference between the picture-based CAPTCHA condition and the control condition.

Because the normality of the data was borderline, we computed a Kruskal-Wallis test on the zxcvbn $\log_{10}$ guesses as well, which showed that there was a statistically significant difference in the zxcvbn $\log_{10}$ guesses between the different conditions $\chi^2(2) = 10.340$, $p = .006 < .05$, with a mean rank zxcvbn score of 42.55 for CONT, of 23.95 for text-CAPTCHA and of 34.00 for picture-CAPTCHA.

In addition, zxcvbn also provides an ordinal score ranging from 0 to 4. We computed a Kruskal-Wallis test which showed that there was a statistically significant difference in the zxcvbn score between the different conditions $\chi^2(2) = 9.251$, $p = .010 < .05$, with a mean rank zxcvbn score of 40.86 for CONT, of 24.27 for text-CAPTCHA and of 35.36 for picture-CAPTCHA.

7 Discussion

Our findings that solving a CAPTCHA prior to choosing a password impacts the password strength, has wide implications because of the impact on authentication security. We note that the more effortful text-CAPTCHA led to weaker passwords. However although the effort spent (via Diff_Tiredness) was not significantly different between the picture-CAPTCHA and the control, there was still a difference in password strength between the two conditions.

While Groß et al.'s [12] showed that a combination of tasks specifically designed in psychology to cognitively deplete users (the white bear, an impulse control and the Stroop test), resulted in users choosing weak passwords, this research shows that even common security measures such as the CAPTCHA challenge has a detrimental effect on password strength.

In addition, while system designers often create account registration forms (such as Facebook, Reddit, Wikipedia) with a CAPTCHA challenge *after* password choice rather than before, our research informs future design decisions of the positioning of CAPTCHAs. Our findings indicate that we should clearly guide the sequence of user input for usability and not to put the password strength at risk. Design recommendations such as positioning the CAPTCHA on a separate page after password choice is likely to be beneficial for security. We also observe that CAPTCHAs are often deployed as a gateway to access Web sites at all, either when frequent requests from the originating IP address were observed or when the use of the TOR Anonymizer was detected. Consequently, in these cases users are systematically exposed to CAPTCHAs before they could register and choose a password.

Furthermore, apart from considering individual and subsequent effects on a security task, it is also important to consider the overall, combined cognitive ef-

Table 1: Descriptive statistics of password strength via password meter by condition.

Condition	N	Mean	Std. Dev.	Std. Error	95% CI		Min	Max
					LL	UL		
CONT	22	67.68	37.02	7.89	51.27	84.09	-8	151
text-CAPTCHA	22	31.05	29.16	6.21	18.12	43.97	-16	103
picture-CAPTCHA	22	42.36	35.18	7.50	26.76	57.96	-17	102
Total	66	47.03	36.82	4.50	37.98	56.08	-17	151

Table 2: Descriptive statistics of password strength via zxcvbn $\log_{10}$ guesses by condition.

Condition	N	Mean	Std. Dev.	Std. Error	95% CI		Min	Max
					LL	UL		
CONT	22	8.66	2.11	0.45	7.72	9.59	4.77	14.97
text-CAPTCHA	22	6.38	2.84	0.61	5.12	7.64	0.95	13.41
picture-CAPTCHA	22	7.76	2.46	0.52	6.67	8.85	2.86	13.34
Total	66	7.60	2.62	0.32	6.95	8.24	0.95	14.97

fort of different tasks. In particular whether they lead to the user rejecting security overall. So far past research has only looked at security fatigue of the current task [11, 24]. Therefore, our findings raise several questions (a) How does designing security tasks in sequence impact (i) usability, (ii) rejection of security and security fatigue, and consequently (iii) the overall security achieved? (b) Does a sequence of security tasks induce a weak link? (c) What combined effort can the user bear? (d) What combination of security tasks is within the user's cognitive effort capacity?

7.1 Ethics

We followed the ethical guidelines at our University to run both the pre-study and the main study. We did not induce more effort than is reasonable in daily life. The participants were informed of the approximate length of the studies, were guided through a consent form and were informed that they could cease participation at any time. Participants were rewarded with a compensation of $6.5 for their time.

Participants' data are kept securely under lock and key and on machines with hard disk encryption. The personal identifiable data of participants was separated from the experiment data and the experiment data anonymized.

7.2 Limitations

Ecological Validity. Although requiring a more controlled setup, we chose lab studies as a first step because it is believed in password research that such studies offer better data quality. We used the same GMail mockup as previous studies [12] which is identical to the GMail account registration page.

Our findings pertain to the chosen manipulations. We chose the text-recognition CAPTCHA, known to be widely used and an adapted picture-recognition CAPTCHA, which often comes up from the widespread reCAPTCHA. Further experiments can be conducted on other CAPTCHA schemes and sequence in security tasks.

Sample Size and Power The study fulfills recommendations have a power of at least $1-\beta = 80\%$ against an effect size of Cohen's $f = 0.5$ in the omnibus ANOVA. With the given sample size of $N = 66$, an effect of Cohen's $f = 0.4$ could still be detected at 80% power. We note that the ANOVA on zxcvbn slightly fell below that mark.

Given the sample size investigated, the parameter estimation on the means and effect sizes in differences is not especially tight. Further research with larger samples could tighten the confidence intervals on the population parameters.

Sampling Bias. Our sample was from university student population, hence an educated sample with a mean age of 21.79 in the main study. Although we found that password choice was weaker in text-CAPTCHA than in the control condition for both the pre-study and the main study, for generalizability, this study can easily be reproduced on a stratified sample. A larger sample size can also support other statistical analyses such as regressions.

Post Questionnaires. In the post-stress and the cognitive workload questionnaires, we found that participants evaluated the GMail registration only rather than the CAPTCHA and registration. Future experiments evaluating stress and workload combined across a sequence of tasks would benefit from making clearer to participants about the task being evaluated. It might also be beneficial to consider a small stress and workload evaluation in between the sequential tasks.

Our findings of security fatigue are focused on the security tasks chosen and the sequence in which they were designed. For example it might be useful to also compare the impact of setting a password on performance at solving a CAPTCHA or other security tasks.

We did not include subjective evaluation of the CAPTCHAs. Future studies can also benefit from additional measures such as self-report/subjective participant feedback on solving the CAPTCHA and also from the combination of CAPTCHA and password.

8 Conclusions

We provide a first empirical study evaluating security fatigue in relation to sequential security tasks. We find that password choice following a CAPTCHA lead to poorer passwords than without the CAPTCHA. While our findings impact design practice and research on individual security tasks together with their pairing with other tasks, they also have wide implications for the overall security of systems.

References

[1] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 workshop on New security paradigms*, pages 47–58. ACM, 2009.

[2] E. Bursztein, S. Bethard, C. Fabry, J. C. Mitchell, and D. Jurafsky. How good are humans at solving captchas? a large scale evaluation. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 399–413. IEEE, 2010.

[3] E. Bursztein, M. Martin, and J. Mitchell. Text-based captcha strengths and weaknesses. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 125–138. ACM, 2011.

[4] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski. Designing human friendly human interaction proofs (hips). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 711–720. ACM, 2005.

[5] M. Chew and H. S. Baird. Baffletext: A human interactive proof. In *Electronic Imaging 2003*, pages 305–316. International Society for Optics and Photonics, 2003.

[6] K. P. Coopamootoo and T. Groß. Evidence-based methods for privacy and identity management. In *Privacy and Identity Management. Facing up to Next Steps*, pages 105–121. Springer, 2016.

[7] A. S. El Ahmad, J. Yan, and L. Marshall. The robustness of a new captcha. In *Proceedings of the Third European Workshop on System Security*, pages 36–41. ACM, 2010.

[8] J. Elson, J. R. Douceur, J. Howell, and J. Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. In *ACM Conference on Computer and Communications Security*, volume 7, pages 366–374. Citeseer, 2007.

[9] C. A. Fidas, A. G. Voyiatzis, and N. M. Avouris. On the necessity of user-friendly captcha. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2623–2626. ACM, 2011.

[10] D. Florêncio, C. Herley, and P. C. Van Oorschot. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts. In *Usenix Security*, pages 575–590, 2014.

[11] S. Furnell and K.-L. Thomson. Recognising and addressing 'security fatigue'. *Computer Fraud & Security*, 2009(11):7–11, 2009.

[12] T. Groß, K. Coopamootoo, and A. Al-Jabri. Effect of cognitive depletion on password choice. *Learning from Authoritative Security Experiment Results (LASER'16)(July 2016), S. Peisert, Ed*, 2016.

[13] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[14] E. H. Hess and J. M. Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

[15] P. Hoonakker, N. Bornoe, and P. Carayon. Password authentication from a human factors perspective. In *Proc. Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 459–463. SAGE Publications, 2009.

[16] D. Kahneman. *Thinking fast and slow*. Farrar, Strauss, 2011.

[17] V. Kothari, J. Blythe, S. W. Smith, and R. Koppel. Measuring the security impacts of password policies using cognitive behavioral agent-based modeling. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*, page 13. ACM, 2015.

[18] G. Mori and J. Malik. Recognizing objects in adversarial clutter: Breaking a visual captcha. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003.

[19] G. Reynaga, S. Chiasson, and P. C. van Oorschot. Exploring the usability of captchas on smartphones: Comparisons and recommendations. In *NDSS Workshop on Usable Security USEC*, 2015.

[20] A. Rogers and G. Brewer. Statistics for websites using captcha technologies, 2017.

[21] S. Sivakorn, J. Polakis, and A. D. Keromytis. I'm not a human: Breaking the google recaptcha. *Black Hat,(i)*, pages 1–12, 2016.

[22] C. D. Spielberger, R. L. Gorsuch, and R. E. Lushene. Manual for the state-trait anxiety inventory. 1970.

[23] C. D. Spielberger, R. L. Gorsuch, R. E. Lushene, P. Vagg, and G. Jacobs. Stai manual for the state-trait anxiety inventory. palo alto, 1970.

[24] B. Stanton, M. F. Theofanos, S. S. Prettyman, and S. Furman. Security fatigue. *IT Professional*, 18(5):26–32, 2016.

[25] D. M. Tice, R. F. Baumeister, D. Shmueli, and M. Muraven. Restoring the self: Positive affect helps improve self-regulation following ego depletion. *Journal of Experimental Social Psychology*, 43(3):379–384, 2007.

[26] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 294–311. Springer, 2003.

[27] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[28] D. L. Wheeler. zxcvbn: Low-budget password strength estimation. In *Proc. USENIX Security*, 2016.

[29] J. Yan and A. S. El Ahmad. Usability of captchas or usability issues in captcha design. In *Proceedings of the 4th symposium on Usable privacy and security*, pages 44–52. ACM, 2008.

[30] J. Yan and A. S. El Ahmad. Captcha security: A case study. *IEEE Security & Privacy*, 7(4), 2009.

A Appendix

A.1 Password Re-Use

We asked participants whether they registered the account via a password they currently use for any services. In the control condition as well as in the picture-CAPTCHA condition, 22.7% of the participants re-used an existing password. 36.3% of the participants re-used an existing password in the text-CAPTCHA condition. This difference is not statistically significant. While a social media password was most commonly used, none of the participants re-used a banking or retail password. Table 3 provides a detailed view of password type re-used.

Table 3: Re-Use Context

Count	Service
1	email
4	social-media
2	education
2	mobile
1	social media/mobile/education
2	social media/email/mobile/education

A.2 Password Strategies

We coded participants' password strategies. Figure 7 depicts the strategies across conditions while Sections A.2.1 to A.2.6 provide the qualitative details. There was no significant difference across groups.

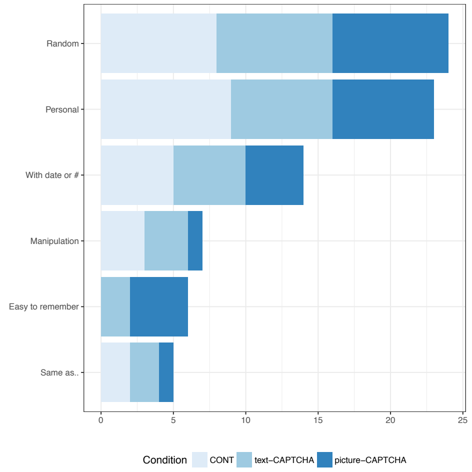


Figure 7: Password strategy across conditions

A.2.1 Random

36% of the participants did not have a strategy, described it as random or specifically said that they used a random thought.

8 participants stated that they did not have a strategy, for example, P7 expressed *“I did not have one”* or P21 in *“nothing really”*. 9 participants used a random password, such as expressed by P6 in *“I put random information together”*, by P56 in *”Trying to be funny”* or by P62 in *”random words and letters”*. 7 participants expressed a random thought, such as by P20 in *“Whatever comes to mind”*, by P29 in *“What comes to mind with different signs”* or by P58 in *“Randomly Thought of* [sic]*”*. We found that 8 participants in each condition used a random strategy.

A.2.2 Personal

36% of the participants chose a password related to their preference or something personal to them.

10 participants chose a password linked with their preferences, for example P22 expressed *“Favourite Sport with mix of capslock inbetween* [sic]*”*, or P22 *”Favourite Football player”* whereas 15 participant created a password with personal meaning, such as expressed by P26 *“Daddy's name + random letter* [sic]*”*, P30 expressed *“City where i was born.*[sic]*”* or P40 in *”My Dog's full name and the year we got him”*. We found that 9 participants created a password from a preference or with personal meaning in the control, 7 in the text-CAPTCHA and 8 picture-CAPTCHA, where 1 participant's strategy included both a preference and something personal.

A.2.3 Manipulation

We found that only 7 participants had a strategy involving complexity combinations, changing characters to numbers or the equivalent in another language, for example as expressed by P13 *”make a strong password with capital letters, small letters and numbers”* or P48 in *”a bit creative with changing the i to 1”*. 3 participants with this strategy were from the control condition, 3 in the text-CAPTCHA condition and 1 in the picture-CAPTCHA.

A.2.4 Same as ...

Only 4 participants described a re-use strategy, for example as expressed by P3 *”Same as always”* or P47 *”Same as Username”*. We found that there was 2 participants employing this strategy in the control and 1 in each of the text and picture-CAPTCHA conditions.

A.2.5 Easy to remember

Only 6 participants reported that they created an easy to remember password, for example as expressed by P4 *”just used two easy words without spaces between which is easy to remember for me”* or P15 *”making the password unreal and easy to remember”*. 2 participants in the text and 4 in picture-CAPTCHA reported this strategy compared to none in the control condition.

A.2.6 With date or number

21% of participants created a password that was combined with numbers or dates, for example P18 reported using *”My Initials and current year”* and P27 *”Favourite colour and 100”*. 5 participants employed this strategy in both the control and text-CAPTCHA conditions and 4 in the picture-CAPTCHA condition.

Dead on Arrival: Recovering from Fatal Flaws in Email Encryption Tools

Juan Ramón Ponce Mauriés[1], *Kat Krol*[2,‡],
Simon Parkin[1], *Ruba Abu-Salma*[1], *and M. Angela Sasse*[1]
[1] *University College London,*
{juan.mauries.15, s.parkin, ruba.abu-salma.13, a.sasse}@ucl.ac.uk
[2] *University of Cambridge, kat.krol@cl.cam.ac.uk*

Abstract

Background. Since Whitten and Tygar's seminal study of PGP 5.0 in 1999, there have been continuing efforts to produce email encryption tools for adoption by a wider user base, where these efforts vary in how well they consider the usability and utility needs of prospective users.
Aim. We conducted a study aiming to assess the user experience of two open-source encryption software tools – Enigmail and Mailvelope.
Method. We carried out a three-part user study (installation, home use, and debrief) with two groups of users using either Enigmail or Mailvelope. Users had access to help during installation (installation guide and experimenter with domain-specific knowledge), and were set a primary task of organising a mock flash mob using encrypted emails in the course of a week.
Results. Participants struggled to install the tools – they would not have been able to complete installation without help. Even with help, setup time was around 40 minutes. Participants using Mailvelope failed to encrypt their initial emails due to usability problems. Participants said they were unlikely to continue using the tools after the study, indicating that their creators must also consider utility.
Conclusions. Through our mixed study approach, we conclude that Mailvelope and Enigmail had too many software quality and usability issues to be adopted by mainstream users. Methodologically, the study made us rethink the role of the experimenter as that of a helper assisting novice users with setting up a demanding technology.

1 Introduction

Usability issues have been regularly cited as a barrier to the adoption of email encryption [23] since Whitten and Tygar's seminal paper "Why Johnny Can't Encrypt" [24]. The paper received the USENIX Security Test of Time Award in 2015, which might be interpreted to mean that this state of affairs persists. Recent research [17] reports that users are increasingly learning about security threats from various sources, such that they may be more receptive to adopting email encryption tools than ever before. There has been increasing effort to provide end-to-end encryption and eliminate barriers to adoption, such as key distribution [24, 23, 20, 6].

The motivation behind the study described here was to observe and analyse novice users' first encounter with such tools: 18 years after Johnny, how easy is it to configure and use an encrypted email client?

We know that users prefer to use email encryption tools which integrate with email systems they are already using [19]. Thus, we chose to study two current open-source, integrated PGP email encryption tools – Enigmail and Mailvelope. We observed users across three stages of activity, within a group-based study: an installation group session, home use over a week with assigned group communication tasks, and a debrief group session. Ten participants completed the study, divided into two groups of four and six participants using Enigmail and Mailvelope respectively. The approach was validated by findings showing that barriers were encountered across all phases of the study for both tools, in many places requiring the assistance of a knowledgeable experimenter to complete the various stages. This raises questions about the role of a knowledgeable expert in the process of learning to use a complex piece of software and overcoming barriers to effective use, where the experimenter may need to take on this duty.

2 Background

Lack of usability has been demonstrated to hamper both the adoption and actual security of email encryption. Whitten and Tygar [24] explored whether PGP 5.0 could

[‡] The study was conducted while the author was at University College London (UCL).

be used by the general public to effectively secure emails. The authors employed two evaluation methods: (1) a hybrid of a cognitive walk-through and heuristic evaluation, and (2) a lab-based user study. Problems were identified in the user interface design which introduced security risks – most lab participants were incapable of using the PGP software securely. It was concluded that making security usable requires the development of domain-specific user interface design principles and techniques.

Garfinkel and Miller [10] performed a user study of the CoPilot email client and Key Continuity Management (KCM), where KCM automates key generation, key management, and message-signing. The authors concluded that KCM and CoPilot improved usability by managing encryption tasks on behalf of users. In contrast, Ruoti et al. [20] suggested that designers should focus on manual encryption to provide transparency and engender trust in encrypted tools. Subsequent studies (e.g., [11, 9, 21, 23]) have followed the findings of Whitten and Tygar's original work, for instance, through studies of secure communications in two-way radios [7], and opportunistic email encryption [8].

Recent studies of encryption have explored socio-technical factors. Gaw et al. [12] interviewed employees in an organisation, finding alongside usability a range of social factors influence adoption of encrypted email, such as the perceived importance of specific messages and the perceived line between secrecy and paranoia. Renaud et al. [18] explored adoption factors across several dimensions, such as awareness of privacy risks and motivation to protect against violation of emails. User interviews captured mental models of email security, identifying adoption challenges, such as incomplete threat models and lack of understanding of email architecture. Ruoti et al. [19] conducted lab-based studies with pairs of novices cooperating to send an encrypted email with a range of email tools, finding that lack of transparency impacted trust, and that the availability of effective tutorials was critical.

Here, we present a novel approach to studying use of encrypted email tools – a combination of lab-based setup with groups of participants using their own computers, home use of encrypted email to perform a shared task, and debrief in a lab setting to measure perception of the tools. This allows us to explore where barriers can emerge during the process of adopting and acclimatising to encrypted email.

3 Method

Our study aimed to compare characteristics of Enigmail and Mailvelope, to understand the facilitators and obstacles behind adoption of encrypted email solutions. We chose Mailvelope and Enigmail as they are end-to-end encrypted, open-source, and available free of charge. While Enigmail is a stand-alone extension to the Thunderbird email client, Mailvelope is an integrated solution, as either a Chrome extension or Firefox add-on.

3.1 Design

We conducted a three-part study with one group of participants installing and using Enigmail alongside Thunderbird, while the other group using Mailvelope. Participants used their own laptops during the study, as follows:

- **Lab-based setup.** Participants were interviewed about their email-related habits, and asked to install, configure, and begin using their assigned tool.
- **Home use of encrypted email.** Participants were given a task to complete outside of the lab setting, organising a mock flash mob campaign via encrypted email over one week. Participants sent emails to each other to agree on the location and music for the mock event, and to confirm the location with the experimenter. They also sent emails to a new member of the group (another researcher).
- **Lab-based feedback session.** Participants discussed their experience of Enigmail or Mailvelope.

Participants were asked to bring their own laptops to the study, to preserve ecological validity [13, 14]. They were provided with printed copies of the installation guides for either Thunderbird and Enigmail or Mailvelope. Crucially, the experimenter was available to assist participants – rather than presume to lead them – during the setup phase, and was contactable during the home-use phase. Participants were asked to note when they completed specific tasks on another sheet: (1) installing Thunderbird (only for the Enigmail group), (2) installing the Enigmail extension for Thunderbird *or* the Mailvelope extension on Firefox or Chrome, (3) configuring the extension (generating a private and public key pair), (4) sharing public keys with other group members, and (5) sending an encrypted email to the study coordinator.

At the lab-based debrief session, participants completed System Usability Scale (SUS) [5] forms for both Enigmail and Mailvelope. The SUS questionnaire consists of ten statements, where users indicate how strongly they agree with each statement on a five-point Likert scale. At the end of the final session, participants received £30 for their participation.

3.2 Participants

Participants were recruited through a research participant pool at University College London. It is a participant pool where members of the general public can register and sign up for research studies. Prospective participants completed a pre-screening questionnaire to indicate occupation, age, gender, whether they had previously used an email client, and if they had any experience with email encryption tools.

Overall, 52 individuals completed the pre-screening questionnaire. Two groups were formed, with six participants each, so as to be resilient to unanticipated no-shows. Those with a background in computer science were excluded to favour non-technical users. Two of the invited Enigmail participants did not attend on the day of the lab-based setup session. The final sample was as follows: the Enigmail group had four participants, two females and two males. Their mean age was 32.7 ($SD = 20.2$, range: 23–45). The Mailvelope group consisted of four females and two males, with a mean age of 39.6 ($SD = 9.1$, range: 24–76).

3.3 Procedure

Upon arrival, participants were asked to read the information sheet and sign a consent form. The first group was tasked with installing Mozilla Thunderbird and the encryption extension Enigmail. The second group was assigned the browser extension Mailvelope that works with Firefox or Chrome web browsers. An explanation of the tasks to be completed was given, but users were not briefed on the specific goal of the study until the end of the final session one week later.

3.4 Role of the experimenter

We initially conceived the role of the experimenter to be that of a session facilitator, asking participants about their experiences with the tools, and eliciting their mental models of how encryption works. As Enigmail and Mailvelope are targeted towards mainstream users, we provided participants with official setup guides published by the developers of the tools.

At the design stage of the study, we did not envisage the role of the experimenter to be an instructor telling participants how to set up the tools. However, the pilot session we conducted before the main study sessions made us change this element of the study design. In the pilot study, we used a convenience sample consisting of colleagues who mostly had a computer science background. They were asked to perform the exact same tasks as our participants. The pilot session of the setup lab-session took in excess of 1.5 hours, where despite the sessions being full of discussion about the instructions, the pilot participants struggled with the installation process to such a degree that it was necessary for the experimenter, a domain-knowledge expert, to guide them through the process to successful installation and use. As a result, the experimenter was briefed to not actively lead participants through the setup steps, but to respond to requests for help from participants if they arose during the session(s).

3.5 Research ethics

The study was conducted after having been approved by UCL's Research Ethics Committee (approval number: 9423/001). The research was also registered with the UK Data Protection Act 1998 (Z6364106/2016/07/11). We did not collect any personally identifiable information. We temporarily stored demographics and contact detail information to be able to select participants and invite them to the study. This information and the recordings made during the group sessions were securely disposed of at the end of the study.

4 Results

4.1 Task completion and times

The average task completion times are shown in Figure 1. The average completion time for all tasks was 48.1 minutes for the Enigmail group, and 40.4 minutes for the Mailvelope group. Task times are self-reported, so values may not be precisely accurate, but are indicative of the time it took for each group to complete the tasks assigned to them. The majority of participants in both groups reached and completed the final task. However, it can be seen that even with (minimal) assistance from the knowledgeable experimenter it can take novices in the region of half an hour to set up and test encrypted email. Average task times for Enigmail are shown alongside notable participant quotes in Figure 3, and for Mailvelope in Figure 4 (see Appendix).

All Enigmail participants completed the four mock campaign tasks successfully. In the Mailvelope group, one out of six participants was unable to complete the third and fourth setup task (e.g., importing a new public key from a new participant and sending encrypted email to this person). This participant, P4-M,[1] downloaded the attachment correctly but imported an incomplete block of text as part of the public key. Participant P2-M was unable to complete task four due to a broken laptop.

4.1.1 SUS

A SUS score can range from 0 (poor) to 100 (excellent). The average score for Enigmail was 63.1 (range: 57.5–77.5, $SD = 9.7$), and for Mailvelope was 50.8 (range:

[1] Participants are referred to as P*X*-E for those in the Enigmail group, and P*X*-M in the Mailvelope group.

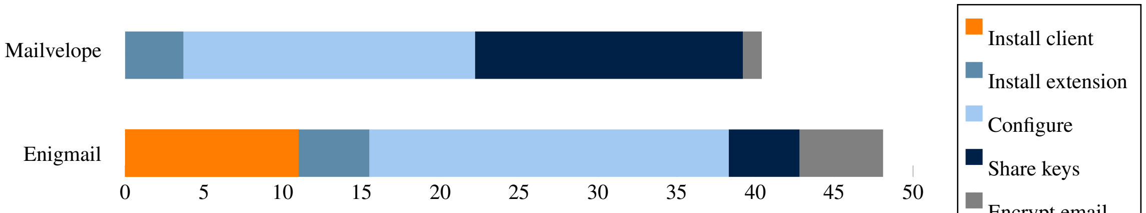


Figure 1: Task times in minutes for Enigmail and Mailvelope.

27.5–70, $SD = 19.4$). This result means that Enigmail achieved "Good Usability", whereas Mailvelope achieved "OK Usability". An unpaired t-test showed that these differences were not statistically significant ($p = 0.28$), possibly due to a small sample size.

4.2 Qualitative results

The audio-recordings of the sessions were transcribed, and the transcripts were analysed using thematic analysis [4]. The analysis identified the following themes.

4.2.1 Sharing sensitive information

Participants generally considered personally identifying information to be sensitive (e.g., when shopping online or entering passport details for flights). They felt that disclosure of this information could expose them to the risk of identity theft or leakage of, for instance, online banking details.

All participants expressed that they had needed to share sensitive information at some point. Diverse means were mentioned; two participants had shared sensitive information via regular email, a further two via the telephone, and three via messaging applications such as WhatsApp or Facebook Messenger. Two participants stressed that they, as users, have to trust the service provider, or otherwise not use the service at all:

> *"I mean... you basically have to put your trust in it, otherwise you just don't use the email or you don't use the messenger service, you know?"* (P5-M)

Participants spoke of unintended recipients who might access their emails. All Mailvelope participants agreed with P4-M's sentiment: *"Well, I think [...] Gmail, it's checked every time we use it and all of our data is known to them."* Participant P3-M argued that their emails would not be a target for malicious parties: *"We are not... important enough for somebody to hack my personal email... we are not Hillary Clinton!"*

4.2.2 Encryption

All participants had previously heard of "encryption", but did not report having used a dedicated email encryption tool. Participant P4-E had, however, previously tried to install Mailvelope a few months prior to the study:

> *"I tried to install Mailvelope, yeah, but only got half-way through 'cause I really couldn't understand how to do the rest of it... "*

Participant P2-E noted having *"seen people use PGP and stuff"* without having used it personally, despite having *"technical friends"* who encrypt their emails. In response, P3-M explained the mechanism behind encryption as follows:

> *"It kind of converts the entire message into some kind of codes and then you send to the recipient in the form of code and then something happens... I don't know what happens... "*

There was a consensus amongst participants that encryption did something to the original message that prevented an unintended person from reading the message.

Participants also commented on recent news, airing concerns about anonymous browsing and government involvement. P1-E commented: *"Recently the government was trying to block... something they were trying, they didn't want the encryption because obviously they want access to your emails... "* They further elaborated that using encryption might draw attention: *"If all our communications are being monitored, wouldn't having encryption make you a suspect of some suspicious activity instead?"*

4.2.3 Installation and configuration

Participants from both groups agreed that the installation of the extensions was straightforward (including Thunderbird for the Enigmail group). For P1-M, installing the Mailvelope extension was perhaps too seamless:

> *"There was no way of knowing if we had done it or not. It would have been good if there'd*

> *been a bar across the top saying or showing how much of it was installed, or saying it was installed because I wasn't absolutely sure if it was finished... "*

All participants agreed that configuration of the extensions was complicated. For Enigmail, the experimenter had to intervene because there was a bug in the setup wizard. When the setup wizard tried to download the GnuPG component required by Enigmail to do the cryptographic work, a progress bar was shown with the progress of this download. However, the download and installation did not actually start, and no error message or warning message was displayed.

The Mailvelope group complained that after installation, the steps required to configure the extension were unclear. They found locating the button to open the options menu was frustrating since they did not know what to look for. Participant P2-M commented: *"It was a bit complex, I had to ask many times, it was complicated... "* Enigmail users similarly complained that the process was convoluted, difficult to follow, and that it was hard to completely understand all available options, and then decide which one to choose. P4-E elaborated:

> *"... When you get all of the boxes I'm like "Oh my god! Which one do I do – this one or this one?" And that's where I start to struggle because I don't understand the technical language."*

All participants completed the steps up until key exchange without incident. Those in the Mailvelope group were frustrated at being unable to share public keys. The key-generation setup wizard had an option to automatically upload a public key to Mailvelope's key servers, but this process did not work – even when the option was selected, the keys were not uploaded. Participants instead had to copy and paste the key or download it as a file to manually share it with others.

Experimenter intervention was necessary to explain the manual process needed to effectively exchange keys. P4-M was evidently frustrated: *"It's too complicated, it's too much!"* All participants agreed that this step was the worst, as it was unclear what to do intuitively or from the official guide. P5-M: *"Finding the keys, importing them, that was pretty difficult!"*. Participants' mental models of encryption did not relate the use of two keys:

> *"I didn't understand the need for keys, this is all new to me... I can use email, but I don't know why we need a key... so I would have given up, I think!"* (P1-M)

4.2.4 Thunderbird and Enigmail

The Enigmail group generally did not like the encryption experience. When asked if any changes would make the tool better, the focus was on the setup process. P3-E saw too many steps in the installation and configuration process:

> *"I was thinking it should be built into Thunderbird, just using one piece of software, so just basically the install is like: "Where [do] you want to install it?" and then: "Do you need to set up keys?" or whatever."*

P4-E made comparisons to the use of other applications:

> *"It needs to be literally as easy as installing some of the other apps, you know, that you can just download and have encryption that way."*

When considering the design of the Thunderbird interface, P1-E commented that:

> *"I didn't really like the interface of Thunderbird, I thought it was a little bit more clunky, umm, it had very old-school interface."*

Participant P4-E said that even though she liked the idea of encryption, the whole process of getting it to work was too complicated. She attributed it to her age, after hearing about encryption, she had genuine privacy concerns:

> *"Because it's there I would use it, but It's too complicated, maybe because I'm 45 and maybe it's the younger generation of people who put their whole lives on the Internet, you know, and privacy, the idea of privacy is changing... and I... even though I haven't got any sensitive information really, it's just about protecting my own privacy. It's just like getting letters in the post, you wouldn't necessarily just leave your letters laying around for people to read... "*

P4-E explained usability was necessary for adoption by all users:

> *"If I say to some of my friends or even my elderly parents: "Hey! That's encrypted e-mail!", it's just not going to happen and it's not like I really understand it. It needs to be literally as easy as installing some of the other apps, you know, that you can just download and have encryption that way. For me, it has to get to that point really for general consumption... "*

Participants commented that once the applications had been configured, the interface in fact simplified the use of encrypted email as well as public key sharing. They all noticed the warning messages when an email was going to be sent unencrypted. They also said that sharing their public key was easy and convenient because they only had to click one button.

All those in the Enigmail group did, however, say that they would likely remove it from their laptops after the study. P3-E explained:

> *"I'll reinstall it if I have specific reasons like someone sends me an encoded message or I need to send someone something, but it's taking a lot of space."*

4.2.5 Mailvelope

Once through the process of exchanging keys, Mailvelope users felt that the rest was easy to do. P3-M and P2-M commented, respectively, that *"I think it was fairly simple to use after that and yeah!... I can see myself using this with people that I email often..."* and that *"it was kind of cool to learn that it was that easy, to be able to encrypt an email... I didn't realise that you could just add something to your Gmail... you know, an add-on and do it that easily..."*

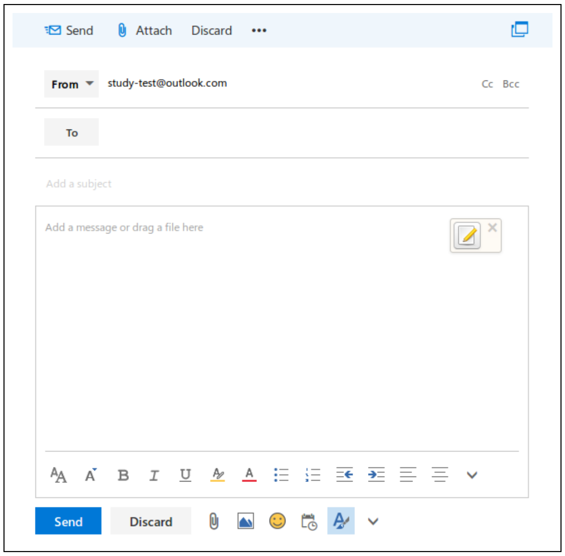
All participants felt confident using the system after a few days completing tasks, and wanted to share their comments. P1-M: *"I didn't know that just adding an extension you could do all that... encrypting and decrypting..."* P6-M struggled to complete tasks for the first few days of home use, having forgotten the passphrase for their private key. They were upset about missing the tasks:

> *"So I tried all the password permutations, so I was so confused... I still wonder why it is... I used something easy to remember... After several days, I said "Oh my goodness!" I had to tell you I had forgotten..."*

Once asked to repeat the process of generating new keys, they were excited to exchange the new public key, where *"that was one thing that I managed to do and I feel quite proud about that!"*

There were some comments as to how to improve Mailvelope's interface and the process of encrypting emails. Three participants reported that the button to activate encryption was not obvious, leaving them prone to sending unencrypted email (see Figure 2 for a screenshot depicting the encryption button). P3-M explained:

> *"Perhaps something more prominent than just that tiny button, because I did it a couple of times, I was writing the text until I realised."*



From study-test@outlook.com

Figure 2: A screenshot of the user interface for Mailvelope displaying the encryption button on the right.

P6-M expressed a concern that the tool did not warn them when they tried to send their public key, and instead attached the private key:

> *"It's just not safe, I mean, they should definitely send a warning message saying "Do you really want to send your private key...?" or something. Yeah... I sent my private key, it should at least warn once. There are so many times when you do something and it's like "Are you sure?" and for the private key it just sends..."*

All those in the group agreed that despite interface issues, Mailvelope was easy to use once they were familiar with the process. Some members of the group mentioned that they would try to use the tool with friends and family. P4-M explained:

> *"I'll keep it but to be honest, I doubt I'll use it... I just don't email sensitive information with people... that often..."*

4.2.6 Interoperability: network effects

In the final session, in both groups when discussing their possible future use of the tools, participants raised concerns that their contacts would need to install these tools as well. While it is true that their contacts would need to install a PGP-based client, the participants in both groups thought it would need to be the exact same one that they had. They were surprised when we explained to them that any PGP-client would be able to exchange encrypted messages with another PGP-client. It was an interesting mental model that could have been influenced by messaging applications for smartphones that generally do not offer interoperability. Research has shown that the adoption of such messaging apps may be influenced by network effects [1].

5 Discussion and conclusions

Participants in both groups were familiar with using email clients, both in the browser and as standalone applications. They were also aware of encryption, and had a basic understanding of what it did to messages, where learning about security technologies from popular news is not uncommon [17]. Participants simply reported that they would not use email to share sensitive information, having found other ways to share such information that were felt as being more secure, and voicing a lack of trust in the medium (in line with studies of pairs of novices using encryption tools [19]). Both products integrate with existing solutions. However, Mailvelope integrated with a browser, permitting users to continue to use existing email clients that they were familiar with. Where explicit/visible encryption is seen as necessary, the effort may lie in paving a way for these features to be integrated into existing popular platforms, and to emphasise interoperability between tools [2].

Participants used the tools on their own laptops. Integration with existing applications was highlighted as an advantage of Mailvelope, although encryption tools were compared to email clients that participants were familiar with, such as Gmail. If an encryption tool appears alien, it compounds the challenge of learning how to operate it effectively.

Both tools had bugs; downloading the GnuPG component required by Enigmail and automatically uploading a public key to Mailvelope's key servers did not work. Participants had to do both manually after being instructed by the experimenter. Mailvelope's option to encrypt was not immediately obvious as previously shown by Schochlow et al. [22], where prevention of errors is a fundamental precursor to providing usable interfaces [15]. Effective user interaction with encryption tools still lies in following basic interface design principles, and there were specific hurdles with each tool.

Ideally, the experimenter has an observatory role in a study like this, but because of the shortcomings of the technologies, they had to step out of this role and take on a more active approach of responding to participants' questions. Without an informed expert present, many participants reported that they would not have continued trying to use the tool(s) in reality. One flaw can be enough to dissuade potential users. However, with guidance, the setup was completed for all participants in both groups. Results suggest that guided habituation of encryption tools can overcome hurdles in the comprehension of encryption. This may be a useful approach for practical use of encrypted email. However, for security user studies, employing researchers who act strictly as experimenters and without domain knowledge has its own advantages [14]. Having a knowledgeable expert close by can be a natural way of learning how to use a new technology [16], where this study has also been an opportunity to observe how having a *helper* available to provide assistance can overcome obstacles which have a known – albeit complicated and demanding – solution.

Adoption barriers appeared across all three stages of our study and for both tools. Practitioners and researchers may continue to study emerging encrypted email solutions to progressively identify isolated barriers to adoption. However, security software developers continue to rely on an intuitive sense of what constitutes usability [3]. If we want any chance of promoting adoption, basic software quality and usability need to be delivered first and foremost. Furthermore, developers also need to draw on usability and design expertise: if the tools are seen as "retro", and do not meet user expectations, we can hardly expect them to be adopted.

Acknowledgements

We would like to thank all LASER attendees and anonymous reviewers for their feedback. Kat Krol's work was supported by a Secure Usability Fellowship from the Open Technology Fund and Simply Secure. Ruba Abu-Salma's work was supported by a SUDS (Supporting Usability and Design in Security) Fellowship from the Open Technology Fund and Simply Secure. We would like to thank Ania Piotrowska, Ingolf Becker and Seb Aebischer for their help in editing this paper.

References

[1] ABU-SALMA, R., KROL, K., PARKIN, S., KOH, V., KWAN, K., MAHBOOB, J., TRABOULSI, Z., AND SASSE, M. A. The Security Blanket of the Chat World: An Analytic Evaluation and a User Study of Telegram. In *European Workshop on Usable Security (EuroUSEC)* (2017).

[2] ABU-SALMA, R., SASSE, M. A., BONNEAU, J., DANILOVA, A., NAIAKSHINA, A., AND SMITH, M. Obstacles to the Adoption of Secure Communication Tools. In *IEEE Symposium on Security and Privacy (S&P)* (2017).

[3] BECKER, I., PARKIN, S., AND SASSE, M. A. Combining Qualitative Coding and Sentiment

Analysis: Deconstructing Perceptions of Usable Security in Organisations. In *Learning from Authoritative Security Experiment Results (LASER) Workshop* (2016).

[4] BRAUN, V., AND CLARKE, V. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology 3*, 2 (2006), 77–101.

[5] BROOKE, J., ET AL. SUS – A Quick and Dirty Usability Scale. *Usability Evaluation in Industry 189*, 194 (1996), 4–7.

[6] CHOE, Y. R., RUOTI, S., ANDERSEN, J., HENDERSHOT, T., ZAPPALA, D., AND SEAMONS, K. There's Hope for Johnny: Automatic vs. Manual Encryption. Tech. rep., Sandia National Laboratories (SNL-CA), Livermore, CA, USA, 2015.

[7] CLARK, S., GOODSPEED, T., METZGER, P., WASSERMAN, Z., XU, K., AND BLAZE, M. Why (Special Agent) Johnny (Still) Can't Encrypt: A Security Analysis of the APCO Project 25 Two-Way Radio System. In *USENIX Security Symposium* (2011), pp. 8–12.

[8] GARFINKEL, S. L. Enabling E-mail Confidentiality through the Use of Opportunistic Encryption. In *Annual National Conference on Digital Government Research* (2003), Digital Government Society of North America, pp. 1–4.

[9] GARFINKEL, S. L., MARGRAVE, D., SCHILLER, J. I., NORDLANDER, E., AND MILLER, R. C. How to Make Secure Email Easier to Use. In *Conference on Human Factors in Computing Systems (CHI)* (2005), pp. 701–710.

[10] GARFINKEL, S. L., AND MILLER, R. C. Johnny 2: A User Test of Key Continuity Management with S/MIME and Outlook Express. In *ACM Proceedings of the Symposium on Usable Privacy and Security (SOUPS)* (2005), pp. 13–24.

[11] GARFINKEL, S. L., SCHILLER, J. I., NORDLANDER, E., MARGRAVE, D., AND MILLER, R. C. Views, Reactions and Impact of Digitally-Signed Mail in E-commerce. In *Financial Cryptography and Data Security*. Springer, 2005, pp. 188–202.

[12] GAW, S., FELTEN, E. W., AND FERNANDEZ-KELLY, P. Secrecy, Flagging, and Paranoia: Adoption Criteria in Encrypted E-mail. In *Conference on Human Factors in Computing Systems (CHI)* (2006), pp. 591–600.

[13] KROL, K., MOROZ, M., AND SASSE, M. A. Don't Work. Can't Work? Why It's Time to Rethink Security Warnings. In *International Conference on Risk and Security of Internet and Systems (CRiSIS)* (2012), IEEE, pp. 1–8.

[14] KROL, K., SPRING, J. M., PARKIN, S., AND SASSE, M. A. Towards Robust Experimental Design for User Studies in Security and Privacy. In *Learning from Authoritative Security Experiment Results (LASER) Workshop* (2016).

[15] MOLICH, R., AND NIELSEN, J. Improving a Human-Computer Dialogue. *Communications of the ACM 33*, 3 (1990), 338–348.

[16] POOLE, E. S., CHETTY, M., MORGAN, T., GRINTER, R. E., AND EDWARDS, W. K. Computer Help at Home: Methods and Motivations for Informal Technical Support. In *ACM Conference on Human Factors in Computing Systems (CHI)* (2009), pp. 739–748.

[17] RADER, E., AND WASH, R. Identifying Patterns in Informal Sources of Security Information. *Journal of Cybersecurity 1*, 1 (2015), 121–144.

[18] RENAUD, K., VOLKAMER, M., AND RENKEMA-PADMOS, A. Why Doesn't Jane Protect Her Privacy? In *Privacy Enhancing Technologies* (2014), Springer, pp. 244–262.

[19] RUOTI, S., ANDERSEN, J., HEIDBRINK, S., O'NEILL, M., VAZIRIPOUR, E., WU, J., ZAPPALA, D., AND SEAMONS, K. "We're on the Same Page": A Usability Study of Secure Email Using Pairs of Novice Users. In *ACM Conference on Human Factors and Computing Systems (CHI)* (2016), pp. 4298–4308.

[20] RUOTI, S., KIM, N., BURGON, B., VAN DER HORST, T., AND SEAMONS, K. Confused Johnny: When Automatic Encryption Leads to Confusion and Mistakes. In *ACM Symposium on Usable Privacy and Security (SOUPS)* (2013).

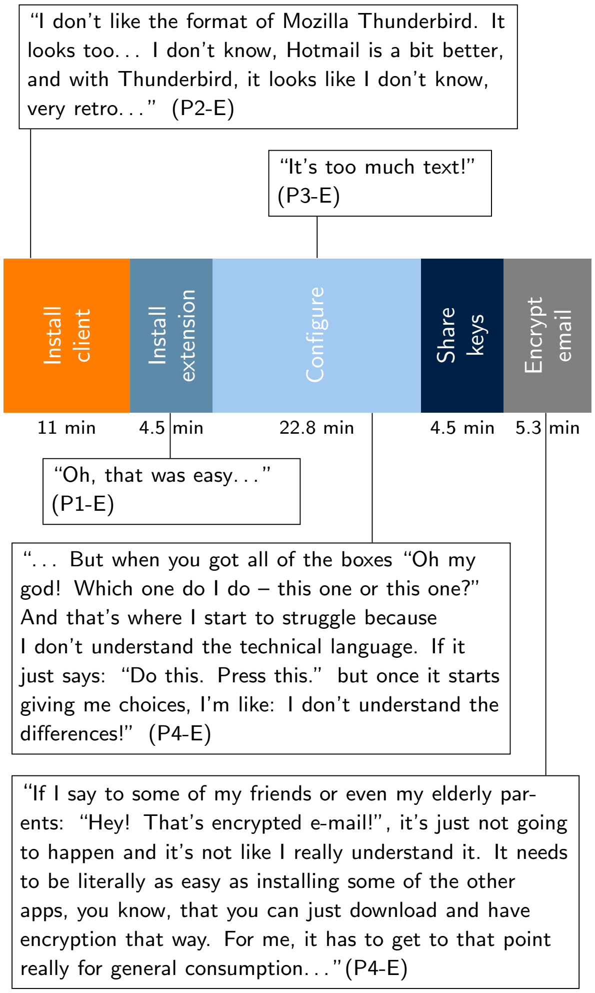
[21] RYAN, J. F., AND REID, B. L. Usable Encryption Enabled by AJAX. In *IEEE International Conference on Networking and Services (ICNS)* (2006), pp. 116–116.

[22] SCHOCHLOW, V., NEUMANN, S., BRAUN, K., AND VOLKAMER, M. Bewertung der GMX/Mailvelope-Ende-zu-Ende-Verschlüsselung. *Datenschutz und Datensicherheit – DuD 40*, 5 (2016), 295–299.

[23] SHENG, S., BRODERICK, L., KORANDA, C. A., AND HYLAND, J. J. Why Johnny Still Can't Encrypt: Evaluating the Usability of Email Encryption Software. In *ACM Symposium on Usable Privacy and Security (SOUPS)* (2006), pp. 3–4.

[24] WHITTEN, A., AND TYGAR, J. D. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *USENIX Security Symposium* (1999).

Appendix



"I don't like the format of Mozilla Thunderbird. It looks too. . . I don't know, Hotmail is a bit better, and with Thunderbird, it looks like I don't know, very retro. . . " (P2-E)

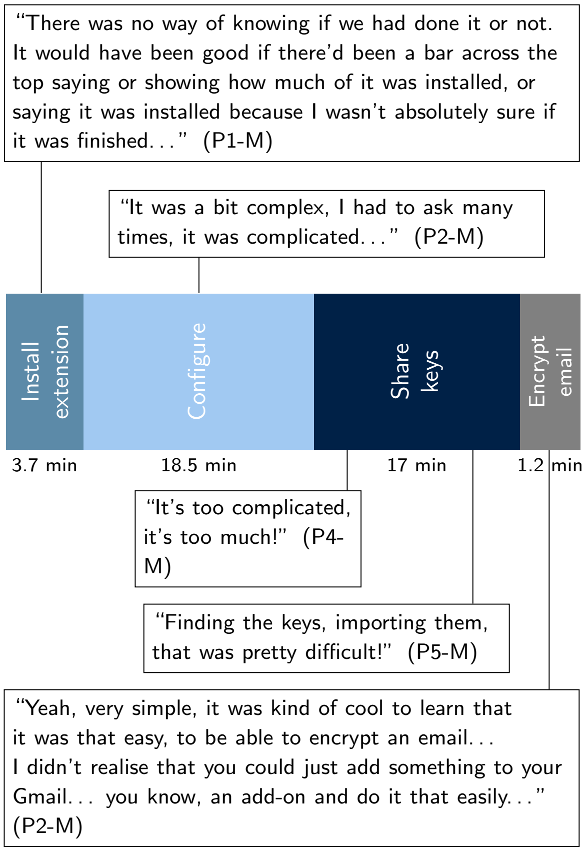
"It's too much text!" (P3-E)

"Oh, that was easy. . . " (P1-E)

". . . But when you got all of the boxes "Oh my god! Which one do I do – this one or this one?" And that's where I start to struggle because I don't understand the technical language. If it just says: "Do this. Press this." but once it starts giving me choices, I'm like: I don't understand the differences!" (P4-E)

"If I say to some of my friends or even my elderly parents: "Hey! That's encrypted e-mail!", it's just not going to happen and it's not like I really understand it. It needs to be literally as easy as installing some of the other apps, you know, that you can just download and have encryption that way. For me, it has to get to that point really for general consumption. . . "(P4-E)

Figure 3: The user journey of setting up Enigmail. The graph shows timings for each step of the setup process with notable participant quotes.



"There was no way of knowing if we had done it or not. It would have been good if there'd been a bar across the top saying or showing how much of it was installed, or saying it was installed because I wasn't absolutely sure if it was finished. . . " (P1-M)

"It was a bit complex, I had to ask many times, it was complicated. . . " (P2-M)

"It's too complicated, it's too much!" (P4-M)

"Finding the keys, importing them, that was pretty difficult!" (P5-M)

"Yeah, very simple, it was kind of cool to learn that it was that easy, to be able to encrypt an email. . . I didn't realise that you could just add something to your Gmail. . . you know, an add-on and do it that easily. . . " (P2-M)

Figure 4: The user journey of setting up Mailvelope. The graph shows timings for each step of the setup process with notable participant quotes.

The Impacts of Representational Fluency on Cognitive Processing of Cryptography Concepts

Joe Beckman, *Purdue University* Sumra Bari, *Purdue University*
Yingjie Chen, Ph.D., *Purdue University* Melissa Dark, Ph.D., *Purdue University*
Baijian Yang, *Purdue University*

Abstract

fMRI presents a new measurement tool for the measurement of cognitive processing. fMRI analysis has been used in neuroscience to determine where cognitive processing takes place when people are exposed to environmental stimuli and has been used to determine where students and experts process basic mathematical functions. This research sought to understand where cryptography was processed in the brain, how representational translation impacts cognitive processing, and how instruction focused on teaching representational fluency in cryptography concepts impacts cognitive processing of cryptography. Subjects were given a multiple-choice pretest, instructed during the semester in the concepts of interest to this research, given a multiple-choice post-test, then subjected to the fMRI scan while prompted to process these concepts. Results of the study show that cryptography is processed in areas indicative of the representational forms in which they were presented, as well as engaging the executive processing areas of the brain. For example, cryptography presented visually was processed in the brain in similar areas as other concepts presented visually, but also engaged the areas of the brain that organize and process complex concepts. However, the research team did not find significant results related to the cognitive processing of translating among representations, nor did we find significant changes in cognitive processing of cryptography for topics in which the focus of instruction was teaching representational fluency. Pre and post test results showed subject performed better on concepts instructed using representational fluency against concepts instructed without a focus on representational fluency, but the difference was not significant at α=.05.

1. Introduction

Cybersecurity is considered a top priority by the US government to defend its virtual borders. A shortage of qualified IT security professionals has long been a problem nationally and internationally [13, 15, 17]. Furthermore, the workforce shortfall is widening. According to a 2015 workforce study, 62% of respondents stated that their organizations have too few information security professionals compared to the 56% in 2013 [17].

Cybersecurity education has been and continues to be a primary focus for fortifying the workforce. The implications are many and include: the need for more students to become aware of and interested in cybersecurity; the need for a higher proportion of the students who are interested in cybersecurity to convert to a declared cybersecurity major in college; and the need to retain students in that major to boost graduation numbers so that more enter the workforce. However, quantity is not the only challenge in cybersecurity workforce development. It is equally, if not more, important that the workforce have the breadth and depth of skills needed to perform in the workforce. Cybersecurity education needs breadth that covers both technical and nontechnical skills spanning computer science, computer engineering, information systems, psychology, business and management, and many other related disciplines [3]. According to [7] we "have a shortage of the highly technically skilled people required to operate and support systems we have already deployed, we also face an even more desperate shortage of people who can design secure systems, write safe computer code, and create the ever more sophisticated tools needed to prevent, detect, mitigate, and reconstitute systems after an attack" [7]. [9] and [16] also emphasize that cybersecurity experts need deep technical skills coupled with capabilities to recognize and respond to complex and emergent behavior, mastery in using abstractions and principles, assessing risk and handling uncertainty, problem-solving, and reasoning; coupled with facility in adversarial thinking.

It is a challenge to educate cybersecruity graduates to assure that they: 1) have broad and deep technical skills, 2) are facile in abstraction, problem-solving, reasoning, and adversarial thinking, and 3) able to learn and perform in this complex and emergent domain. Teaching cyberscurity requires the educator to present the abstract concept to students in a crystal clear way, and to extend the abstract concept to practice to let the students learn the knowledge in context.

Given the newness of the field, cybersecurity's pedagogical "best practices" have not yet been adequately investigated [19]. In the past 10-15 years, articles focused on teaching practice have increased. For example, [4] discusses challenge based learning methodology to improve learning via a multidisciplinary approach which encourages students to collaborate with their peers, ask questions, develop a deeper understanding of the subject and take actions in

solving real-world challenges. [19] proposed a multifaceted hierarchical education framework to teach cybersecurity with the desired level of breath and depth [19]. [15] presents a unique teaching collaborative among 13 universities that intends to teach students agile research and development skills in cyberecurity. While there has been considerable growth in the investigation and reporting on cybersecurity teaching, we find that there is little to no substantive work on cybersecurity learning and thinking.

This work is grounded in cognitive theory and investigates students' mental models in one knowledge area of cybersecurity, i.e., cryptography. We developed Model-Eliciting Activities (MEAs), investigated students' representational fluency and the relationship of students' development of schema and changes in their cognitive processing and control when encountering cryptography concepts. In this paper, we report on students'mental models using functional magnetic resonnance imaging (fMRI) analysis of student's brain activities while solving complex security problems, as well as learning data from classroom tests.

The paper is organized as follows. Related work is reviewed in Section 2 and research methods are in Section 3. Results are presented and discussed in Section 4 and 5, respectively. Section 6 includes discussion and future work.

2. Previous Work

2.1 The Importance of Conceptual Understanding of Cryptography in Cybersecurity

Cryptography is an important subject in cybersecurity. And while cryptography is important for everyone in the field to understand, it can be an especially challenging subject to learn. The domain includes several key concepts, such as symmetric key cryptography, asymmetric key cryptography, types of ciphers, cryptanalysis and attacks, hashing, digital signatures, etc. Each of these concepts is comprised of sub concepts, which build with other sub concepts to form conceptual understanding of a key concept. Furthermore, the conceptual understanding of these concepts and sub concepts requires mathematical, language, and analytic thinking. Both breadth and depth of cryptography knowledge must be considered.

Conceptual understanding is defined as the abstract mental representation of given phenomena. Conceptual understanding occurs in the mind and the mind continuously (re)forms mental representations. The veracity of learners' conceptual understanding is the fidelity of the conceptual understanding to the external world. If conceptual understanding matters, then conceptual learning is where we need to start.

2.2 Cognitive Theory, Conceptual Learning and Measurement Thereof

Cognitive theories of conceptual learning are grounded in Piaget's work on logical mental frameworks (also called schemas and mental models) as structures in the brain that organize information and interactions among information. Interacting with new information, according to Piaget, modifies these schema, which is learning [12]. Conceptual learning is the acquisition of information about concepts and their interactions, and the ongoing modification about the body of conceptual knowledge as new concepts and their interactions are encountered [10].

Correct categorization involves making links to prior knowledge and so may require adjustment or correction of prior knowledge. Assimilation theory presented in [1] contrasts rote learning (temporary acquisition of disorganized or poorly understood isolated or arbitrarily related concepts) with meaningful learning (long-term acquisition of organized, interrelated concepts into existing cognitive structures). Conceptual learning is the process of identifying and correctly categorizing concepts such that they can later be used to make predictions or decisions [2, 11].

[10] has shown that providing learners with instruction in representational fluency can build conceptual understanding. Representations are the different forms in which a concept, principle, or phenomenon can be expressed and communicated. Common representations include graphic, pictorial, verbal, mathematical, and concrete. Each representation presents the phenomenon it is intended to describe in a different mode. Deep(er) understanding of the given concept requires understanding of and among various representations. Beyond comprehending representations, even deeper understanding means being fluent in shifting back and forth among the variety of relevant representations.

The concept of fluency is often associated with the ability to express oneself in the spoken and written word, and to move effortlessly (automatically) between the two representations. A person who is fluent in a language has this ability; they can translate from English to Chinese and back, and from written to spoken word and back (where written may be in English and spoken in Chinese).

The idea of fluency has been extended to other fields such as physics, chemistry, engineering, and mathematics. For example, a study by [8] on experts and novices found that physics problem solvers who are fluent in their use of different representations can easily translate between them, and can assess the usefulness of a particular representation in different situations. Similarly, [16] found that when learners develop multiple representations they are better able to transfer knowledge to new domains with increased cognitive flexibility.

Representational fluency in the STEM fields can include: a) visualizing and conceptualizing transformation processes abstractly; b) understanding systems that do not exhibit any physical manifestations of their functions; c) transforming physical sensory data to symbolic representations and vice versa; d) quantifying qualitative data, e) qualifying quantitative data; f) working with patterns; g) working with continuously changing qualities and trends; and h) transferring principles appropriately from one situation to the next [5]. Regardless what the transformation, representational fluency connotes continuous adaptation and flexibility of the conceptual model, and the ability to perform with facility, adeptness, and expertise. Representational fluency is an important aspect of deep conceptual understanding that has been shown to promote transfer of learning and the development of "expertise".

[18] advocates for the role of neuroscience in the study of mental models. The "mental frameworks" theorized by Piaget in [10] would require activity in the brain [18]. As learners' mental schema change to incorporate new information derived from experiences, brain function in the learners' brains changes. That is, learning changes the structures of the brain.

Advances in neuroscience offer researchers new tools, such as fMRI, to measure brain activity. To date, fMRI has been used in studies of cognitive processing of mathematics. [12] sought to understand what areas of the brain are involved in mathematical computation while [10] built on [12] by using fMRI to measure changes in cognitive processing after instructing students in multiplication in one and two digit numbers. These studies are examples of how neuroscience is being used to understand cognitive processing, so that later it can be applied to evaluate the impacts of instruction on learning.

Our study seeks to understand where cryptography is processed in the brain as a basis for understanding what instructional methods maximize cryptography learning in students.

3. Methods

3.1 Research Questions

This exploratory study first investigated where in the brain cryptography is processed. Second, we investigated the impact of representational form on cognitive processing. More specifically, we investigated whether cognitive processing increased when students were asked to translate cryptography concepts between representational forms (language to math, math to graphical, etc.) in comparison to cognitive processing of concepts using the same representational form (language to language, math to math, etc.). Third we investigated whether teaching cryptography using multiple representations changed how and/or where cryptography concepts were processed in the brain in comparison to instruction that was not focused on generating representational fluency.

The research team used fMRI scans of students to answer the research questions. In order to investigate impact of teaching using multiple representations, learners were taught five cryptography topics using multiple representations, and four topics were taught using single representations to convey concepts. Data gathered from learners' classroom performance were used in support of the fMRI analysis, as discussed below.

3.2 fMRI Component

3.2.1 Variables and Operationalization

3.2.1.1 Independent Variables

As a descriptive question, determining where cryptography concepts are processed in the brain did not have an independent variable. When considering whether translation between representational forms in the context of cryptography impacted cognitive processing, the research team defined a binary variable, Representational Translation. Either the students had to make a translation between representations, or they did not. We implemented this variable as questions that the students were asked to answer while under fMRI scanning. Students were required to make a Representational Translation when, as shown in Figure 1, Representation 1 and Representation 2 were presented in different representational forms.

Questions asked of students during fMRI were generated from material that was taught using both the representational fluency-focused instructional method, as well as the method that did not focus on the use of multiple representations in instruction. Instructional Method was defined as the independent variable in terms of our third research question regarding the impact of instruction focused on representational fluency on cognitive processing of cryptography concepts.

3.2.1.2 Dependent Variables

For our research questions, the dependent variable was Cognitive Processing of Cryptography Concepts, which illustrates where in the brain and with what intensity cryptography is processed. The variable was analyzed in different ways based on the question asked, but was implemented by comparing different periods of activity in the fMRI scan based on the question being considered against brain activity measured as the subject observed the crosshair pattern following each question as shown in Figure 1.

3.2.2 Population and Sample

Nine out of the 12 students from a graduate-level, semester-long network security course participated in fMRI scans.

3.2.3 Setting

Scans were administrated at the University MRI Facility using 3T GE Discovery MR750 and a 32-channel brain array (Nova Medical). Scans consisted of a high-resolution (1mm isotropic) T1-weighted anatomical scan for registration and tissue segmentation purposes and six functional scans (TR/TE=1500/28msec; flip angle=72°; 35 slices at 3.5mm; field of view (FOV)= 24 cm and matrix= 64x64). Each functional run focused on one topic and consisted of nine yes/no matching questions (nine blocks) using three different representational forms. The functional runs were presented in random order. The subjects were able to see the questions inside the scanner through fiber optic goggles (NordicNeuroLab; Bergen, Norway) and responded with their answers through a four-button keypad. Subject's responses were directly transmitted to a computer for storage. Each block began with 15 seconds of crosshair display during which subjects were instructed to relax and focus on the display. The subjects were then presented with a question in one of the representational form for nine seconds, the ISI was of 1.5 seconds and then they were presented with another slide consisting of a question in the same or different representation form for nine seconds. After the second representation subjects then had nine seconds to decide if both the representations (R1 and R2) presented the same concept or not and answered yes or no by pressing one of the designated buttons on the keypad.

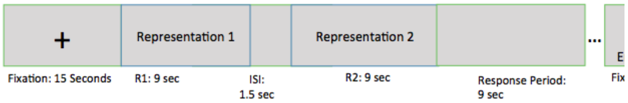


Figure 1: The protocol of a block.

3.2.4 Data Analysis

fMRI scans were processed with an in-house MATLAB code adapted from *afni_proc.py*, using AFNI and FSL. The pipeline consisted of brain extraction, outlier detection, de-spiking, slice timing correction, motion correction, alignment to the T1-weighted anatomical scan, tissue segmentation into gray matter, white matter and cerebral spinal fluid (CSF), and spatial smoothing within each tissue type (isotropic Gaussian filter with Full Width Half Maximum (FWHM) of 4mm). Anatomical and fMRI scans of all subjects were aligned to a standard template (skull stripped $1mm^3$ ICBM152) so that brain activation patterns from different subjects could be grouped together for analysis. Data were motion corrected using three motion parameters (three translational and three rotational for each x-y-z axes) and their derivatives as regressors in General Linear Model (GLM). Block regressors were used for each of the nine transitions and crosshairs in GLM.

Brain activations obtained from crosshair slides were treated as Baseline activations. Brain activations for each representation were obtained by comparing the $\boldsymbol{\beta_{Representation}}$ versus $\boldsymbol{\beta_{Baseline}}$ obtained from GLM of all subjects and all runs, using paired voxel-wise 3D t-tests followed by voxel-wise False Discovery Rate (FDR) correction. Adjacent voxels with p_{FDR}<0.05 and cluster size greater than 100 voxels were considered as significant brain activations against the baseline and are as shown in the figures.

fMRI data gathered was analyzed differently for dependent variable, Cognitive Processing of Cryptography Concepts, based on which question was being investigated. When investigating where in the brain cryptography concepts are processed, activation patterns were gathered during the presentation of the first representation of each question. Activation present during the resting period following the question (noted as the second crosshair pattern in Figure 1) was subtracted from activation patterns noted during Representation 1. Data were separated based on the representation presented in Representation 1 in Figure 1, then the data were aggregated for all student participants (n=9) by representation n=18 per student), for a total of 162 individual data points per representation.

Evaluating whether translation between representations within questions impacted cognitive processing of cryptography, the period of time during the presentation of Representation 2 and the Response Period (as shown in Figure 1) was used to gather cognitive processing data and activation noted during the second crosshair pattern was subtracted from the gathered cognitive processing data. Data were grouped by the independent variable Representational Translation and aggregated for all students. In this case, three questions per topic did not require Representational Translation. So, the total number of data points for non-translation was n=18. Six per topic did require translation for a total translation n=36. Each student answered questions on the same six topics.

Finally, cognitive processing data were gathered and analyzed by the Instructional Method independent variable. In this case, the same data gathering process was used as for analysis of Representational Translation, except that the data were grouped by the instructional method in which the topic was taught. In terms of this comparison, each student was given questions from three topics that were taught using the treatment instructional method that focused on representational fluency and three topics that were taught with the control instructional methodology. This analysis consisted of nine questions over three topics aggregated for nine students, or n=243. However, the research team delimited these comparisons by comparing only questions with the same structure to each other. For example, the fMRI results for all questions on a topic that required the subject to translate a concept from language to math (or vice versa) were aggregated to determine cognitive processing of cryptography concepts during that translation process. Therefore, the effective n=27 (three translations per topic, nine subjects in total) treatment data points and n=54 control data points.

3.3. Classroom Component

3.3.1 Research Question

Classroom data were used only in support of analysis of the fMRI results produced from this study; therefore, the research questions are the same as those discussed as part of the fMRI component earlier.

3.3.2 Variables and Operationalization

3.3.2.1 Independent Variable

The independent variable in this experiment was the method of instruction. Instructional methods were assigned by the researchers to the following topics taught in class: Zero-Knowledge Proof (ZKP), Pohlig-Hellman Ciphers (PH), Rivest Shamir Adleman Cryptosystems (RSA), Digital Cash (DC), and Public Key Infrastructure (PKI). All other content taught during the semester was taught using two representational forms not focused on representational fluency.

3.3.2.2 Dependent Variable

The dependent variable was students' pre to post-test learning gain. Learning gain was determined by normalizing students' points scored on the pre and post- tests into a percentage interval variable, subtracting the pretest score from the post-test score, and averaging the differences of the twelve students for each question. Pre to post-test score differences were aggregated by instructional method and compared using a t-test.

3.3.3 Populations and Samples

Twelve of twelve students from a graduate-level, advanced network security course offered in the Spring 2017 semester at a large university in the Midwestern United States consented to allow their pre and post-test exam scores to be used in this research.

3.3.4 Setting

Data for this experiment were gathered in one section of a graduate-level advanced networking course at a large public university in the Midwestern United States. The course was not a required course. The control topics were taught using a combination of lectures delivered by projecting slides containing individual representational forms (language, graphics, or math) to deliver concepts to learners. Instruction of the treatment topics taught: Zero-Knowledge Proof (ZKP), Pohlig-Hellman Ciphers (PH), Rivest Shamir Adleman Cryptosystems (RSA), Digital Cash (DC), and Public Key Infrastructure (PKI) using activities consisting of multiple representations and focused on representational fluency. No other aspects of the instruction or scored evaluation of the students in the classroom differed between the control and treatment groups. Student performance was evaluated using a pretest and post-test, which also served as the students' final exam.

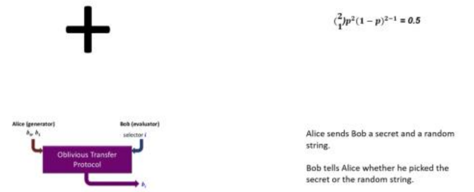
3.3.2 Population

The population from which subjects were drawn for this experiment consisted of all students enrolled in the University's graduate advanced network security course offered by the college of Technology in the Spring of 2017. Enrolled students were predominantly 18-24-year-old. Because the experiment required subjects to consent to the use of their scores on course homework, projects, and exams, those students who gave their signed consent to release their scores to the research team comprised the sample in each class section. All 12 students in the course consented to allow use of their classroom scores in this study.

4. Results

4.1 Brain Location: Cognitive Processing of Cryptography

In order to answer the question, "Where in the brain are cryptography concepts processed?", the research team analyzed blood oxygen level data (BOLD) of participants, representing brain activity, taken during the fMRI while the participants were processing cryptography questions. Measurement of blood flow to the bran, the measurement on which fMRI is based, serves as a proxy for changes in brain activity. Increased blood flow to an area of the brain indicates increased brain activity, cognitive processing, where decreased activity is signaled by reduced blood flow to areas of the brain. In this research, questions were presented using graphical, language, and mathematics representations as shown in Figure 1, which generated distinct patterns of brain activation, so we address the research question by representation.



$\binom{2}{1}p^2(1-p)^{2-1} = 0.5$

Alice sends Bob a secret and a random string.

Bob tells Alice whether he picked the secret or the random string.

Figure 1: Representations used in fMRI questions (clockwise top to bottom): crosshair pattern, mathematic, language, graphical

This analysis used brain activation detected during the presentation of the first of two slides in each question, and the resting crosshair pattern following the question as illustrated in Figure 2.

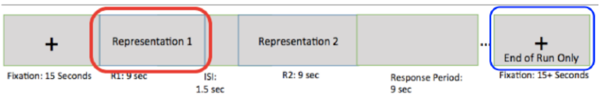


Figure 2: Activation Comparisons by Representation

The research team aggregated the BOLD signal data for all questions by the type of the first representation, that is math, graphical, or language, across the nine student participants. Cryptography concepts presented using a mathematical representation with mathemtics produced BOLD activation patterns shown in Figure 3.

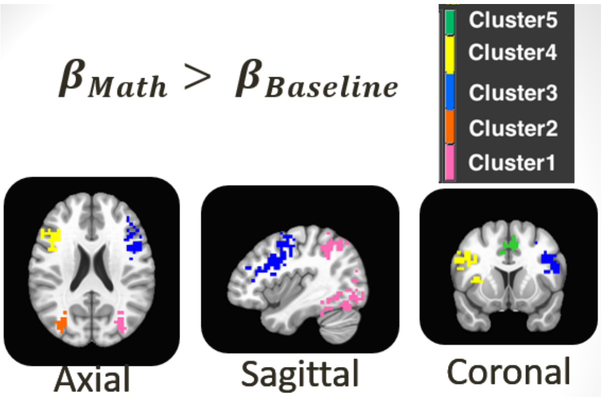


Figure 3: Brain Activation of Cryptography Concepts Presented using Matematical Representations

Five clusters of activation were noted at a significance level of α=0.05. Corresponding Broadmann Areas and usages are listed in Table 1 below.

Cluster	Broadmann Area	Gyrus	Usage
1	Left 39	Left Middle Temporal	Accessing word meaning
2	Right 39	Right Middle Temporal	Accessing word meaning
3	Left 9	Left Inferior Frontal	Representation of numbers
4	Right 44	Right Inferior Frontal	Executive processing
5	Left 9	Left Medial Frontal	Executive processing

Table 1: Math Processing Areas of Activation

Cryptography concepts presented using English language stimuli activated five areas of the brain, which are shown in Figure 4 and detailed in Table 2.

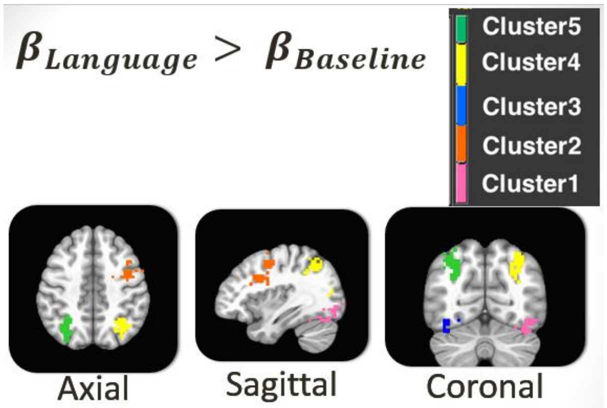


Figure 4: Brain Activation of Cryptography Concepts Presented using English Language Representations

Five clusters of activation were noted at a significance level of α=0.05. Corresponding Broadmann Areas and usages are listed in Table 2 below.

Cluster	Broadmann Area	Gyrus	Usage
1	Left 17	Left Inferior Occipital	Visual Processing
2	Left 44	Left Inferior Frontal	Executive Language Processing
3	Right 37	Right Lingual Gyrus	Visual Processing
4	Left 3	Left Inferior Parietal	Somatosensory Processing
5	Right 1,2	Right Superior Parietal	Somatosensory Processing

Table 2: Language Processing Areas of Activation

Graphical representations of cryptography concepts produced two areas of brain activation. These areas are shown in Figure 5 and decribed in Table 3 below.

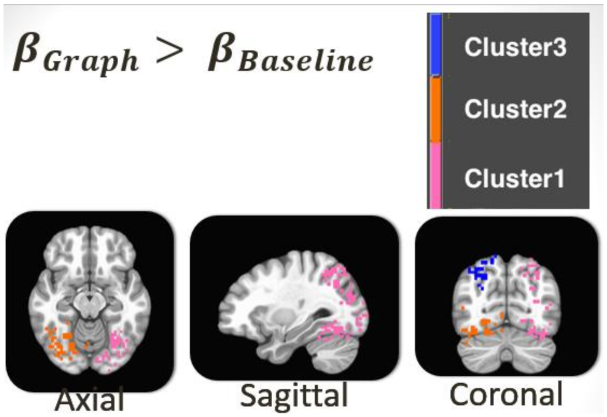


Figure 5: Brain Activation of Cryptography Concepts Presented using Graphical Representations

Three clusters of activation were noted at a significance level of α=0.05. Corresponding Broadmann Areas and usages are listed in Table 3 below.

Cluster	Broadmann Area	Gyrus	Usage
1	Left 30	Left Middle Occipital	Visual Processing
2	Right 7	Right Superior Parietal	Facial Stimuli
3	Right 37	Right Lingual	Visual and Letter Processing

Table 3: Graphica Processing Areas of Activation

4.2 Brain Activation in Cryptography Processing During Translation of Representational Forms

The research team compared students' cognitive processing on cryptography questions in which they were forced to make a translation between representational forms in order to answer the question against cognitive processing activity on questions in which no such translation was necessary. We had hypothesized, based on Thomas, Wilson, Corballis, Lim, and Yoon (2010), that questions requiring such a translation would produce more intense cognitive activity in similar brain regions than those that did not require representational translation. Our comparison of brain activation in this study did not support this hypothesis. Only brain activation patterns in the Language to Language questions, the Language to Math, and the Math to Graphical analyses showed significant activation beyond baseline. Figures 6, 7, and 8, respectively, show the brain areas of significant activation in this comparison.

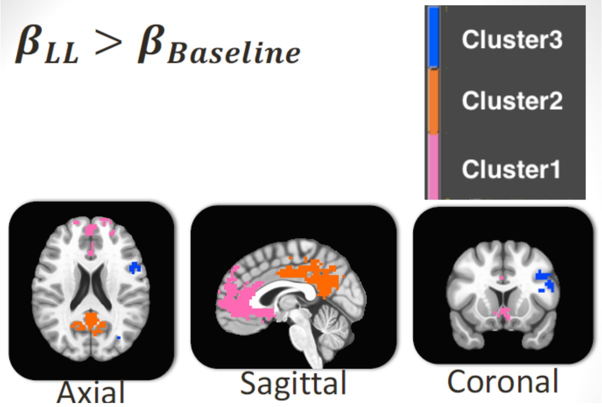


Figure 6: Brain Activation for Language to Language Comparisons of Cryptography Concepts

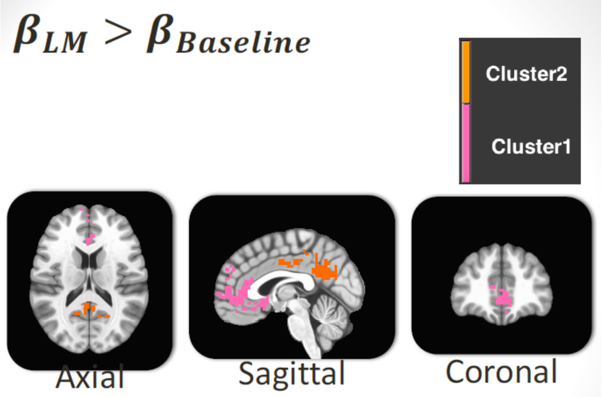


Figure 7: Brain Activation for Language to Math Comparisons of Cryptography Concepts

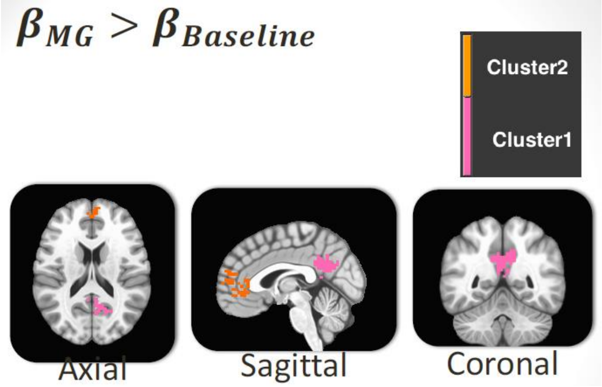


Figure 8: Brain Activation for Math to Graphical Comparisons of Cryptography Concepts

4.3 Cryptography Learning by Instructional Method

In support of the fMRI brain activation data comparison between topics instructed using MEA and focused on representational fluency and those taught using traditional lecture-based instruction, the research team also compared learning gains using the course pre-test and post-test. We hypothesized that teaching cryptography concepts using representational fluency would produce different patterns of cognitive activation compared against topics taught without a focus on representational fluency. For this comparison, pre and post-test scores were aggregated from the twelve students in the class, and across all topics that were instructed using MEA and compared to those that were instructed using the traditional method of instruction not focused on representational fluency. This analysis showed an average learning gain of 10.83% on topics instructed using MEA and 3.56% on topics instructed using methods not focused on representational fluency. These learning gains are not significant at α=0.05 (t=1.19, p=0.24). Comparing pre-test scores based on instructional method indicated a similar level of knowledge, on average, of material that would be instructed using MEA (μ: 0.57, σ: 0.23) versus topics that would be instructed using other methods (μ: 0.57, σ: 0.25)

5. Discussion and Conclusion

The purpose of this work is to design and evaluate if and how representational fluencies are related to cognitive learning. The team specifically examined the following research questions:

1) Where does the cryptography occur in the brain?

The fMRI scan analyses showed that cryptography concepts, if represented using different formats, i.e. language, graph, and math notations, activate different parts of the brain. The results were statistically significant, even when the sample size was merely 9. The activation maps also echo similar distributions as the previous study on math and physics concepts. This may suggest that from cognitive perspective, cryptography is fundamentally not very different from math and physics. Or put it differently, brain activations are directly related to the form of the representations rather than the underlying complex cryptography concepts or algorithms.

2) Where do the transitions of different representations occur in the brain?

Among nine possible representation fluencies, the research team discovered that three of them are statistically significant. They are from language to language, from language to math, and from math to graph. This suggest for the group of students that participated fMRI scans, longer and stronger brain activities were recorded when students were asked to translate the same concept from language to language, from language to math, and from math to graph. Interestingly, the translation from math to language and the translation from graph to math were not shown the same statistical significance. If the research results are reliable, it can be inferred that representation translations are uni-directional. That is the brain reacts differently when translating language to math than translating math to language. If we further assume stronger or longer brain activations are related to more difficult tasks, then it may suggest the three transitions that showed statistically significance might be the ones that students having trouble with.

3) How does representational fluency impact the classroom learning results?

The classroom learning results showed an average of 10.38% gains between pretests and posttests when the instructional methods were delivered using MEAs that were specifically designed to train students on the representational fluencies. In contrast, the gains were merely 3.56% when conventional instructional methods were adopted in the classroom. However, the p value of the paired t-test was 0.26: too large for the research team to declare the findings are statistically significant. There were two major reasons accounted for this "non-significance". The first was due to small sample size of 12. They second was the very high average pretest scores. More specifically, students averaged 56.8% ($\sigma = 23.4\%$) on topics to be instructed using MEA and 57.2% ($\sigma = 25.3\%$) on topics to be taught not using MEA. Students participated in this study were all graduate students and may possess strong prior knowledge of cryptography. If high levels of prior knowledge contributed to the relatively high pretest scores, the large standard deviations in both pre and posttest scores indicate that very different levels of prior knowledge were present among the students (posttest σ, MEA: 23.8%, non-MEA: 24.2%). Further study is needed with a bigger sample size, and preferably at undergraduate level to fully understand the impact the representational fluencies on the classroom learning results. Adding more questions to the pre and posttests at each Bloom level of learning would add clarity to how instruction impacted understanding of the cryptography concepts being researched.

6. Future Work

The results of this study present several avenues for future research. Given the limitations of this experiment, future work could validate our findings regarding where cryptography concepts are processed in the brain. Our failure to find significant results relating to cognitive processing activation during representational translations or cognitive processing related to representational fluency leave these areas open for additional research. In particular, it is possible that different types of classroom instruction or classroom measures of that instruction could also be performed in order to evaluate the effects on cognitive processing and learning. With a cognitive processing baseline set in this work for processing of cryptography, many aspects of learning can be compared against these baselines toward the goal of increasing cryptography learning in information security students.

References

[1] Ausubel, D. P., Novak, J. D., & Hanesian, H. (1978). Educational Psychology: A Cognitive View, 2nd edn (New York: Holt, Rinehart and Winston). *Reprinted (1986). New York: Warbel and Peck.*

[2] Brown, A. L., Cocking, R. R., & Bransford, J. D. (2000). How people learn. JD Bransford (Ed.).

[3] Burley, D. L. (2014). Cybersecurity education, part 1. *ACM Inroads*, *5*(1), 41–41.

[4] Cheung, R. S., Cohen, J. P., Lo, H. Z., & Elia, F. (2011). Challenge based learning in cybersecurity education. In *Proceedings of the 2011 International Conference on Security & Management* (Vol. 1).

[5] Dark, M. J. (2003). A models and modeling perspective on skills for the high performance workplace. Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching, 279-293.

[6] Delazer, M., Domahs, F., Bartha, L., Brenneis, C., Lochy, A., Trieb, T., & Benke, T. (2003). Learning complex arithmetic—an fMRI study. Cognitive Brain Research, 18(1), 76-88.

[7] Evans, K., & Reeder, F. (2010). *A Human Capital Crisis in Cybersecurity: Technical Proficiency Matters*. CSIS.

[8] Hsu, L., Brewe, E., Foster, T. M., & Harper, K. A. (2004). Resource letter RPS-1: Research in problem solving. American Journal of Physics, 72(9), 1147-1156.

[9] McGettrick, A., Cassel, L. N., Dark, M., Hawthorne, E. K., & Impagliazzo, J. (2014). Toward curricular guidelines for cybersecurity. In *Proceedings of the 45th ACM technical symposium on Computer science education* (pp. 81–82). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2538990

[10] Moore, T. J., Miller, R. L., Lesh, R. A., Stohlmann, M. S., & Kim, Y. R. (2013). Modeling in engineering: The role of representational fluency in students' conceptual understanding. *Journal of Engineering Education*, *102*(1), 141-178.

[11] Özdemir, G., & Clark, D. B. (2007). An Overview of Concep-tual Change Theories. Eurasia Journal of Mathematics, Sci-ence & Technology Education, 3(4).

[12] Piaget, J. (1964). Part I: Cognitive development in children: Piaget development and learning. *Journal of research in science teaching*, *2*(3), 176-186.

[13] Pierce, A. O. (2016). *Exploring the Cybersecurity Hiring Gap*. Walden University. Retrieved from http://scholarworks.waldenu.edu/dissertations/3198

[14] Rickard, T. C., Romero, S. G., Basso, G., Wharton, C., Flitman, S., & Grafman, J. (2000). The calculating brain: an fMRI study. Neuropsychologia, 38(3), 325-335.

[15] Rowe, D. C., Lunt, B. M., & Ekstrom, J. J. (2011). The role of cyber-security in information technology education. In *Proceedings of the 2011 conference on Information technology education* (pp. 113–122). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2047628

[16] Schneider, F. B. (2013). Cybersecurity education in universities. IEEE Security & Privacy, 11(4), 3-4.

[15] Sherman, A., Dark, M., Chan, A., Chong, R., Morris, T., Oliva, L., ... & Wetzel, S. (2017). The INSuRE Project: CAE-Rs Collaborate to Engage Students in Cybersecurity Research. *arXiv preprint arXiv:1703.08859*.

[16] Spiro, R. J. ea (1992). Cognitive flexibility, constructivism and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. Duffy, Thomas M. und David H. Jonassen (Hg.): Constructivism and the Technology of Instruction: A Conversation. Hillsdale, NJ, 57-75.

[17] Suby, M., & Frank Dickson. (2015). The 2015 (ISC) 2 Global Information Security Workforce Study. *Frost & Sullivan in Partnership with Booz Allen Hamilton for ISC2*.

[18] Szűcs, D., & Goswami, U. (2007). Educational neuroscience: Defining a new discipline for the study of mental representations. *Mind, Brain, and Education*, *1*(3), 114-127.

[19] Wei, W., Mann, A., Sha, K., & Yang, T. A. (2016). Design and implementation of a multi-facet hierarchical cybersecurity education framework. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)* (pp. 273–278). https://doi.org/10.1109/ISI.2016.7745488

Self-Protective Behaviors Over Public WiFi Networks

David Maimon, *University of Maryland*
Michael Becker, *University of Maryland*
Sushant Patil, *University of Maryland*
Jonathan Katz, *University of Maryland*

Abstract

The proliferation of public WiFi networks in small businesses, academic institutions, and municipalities allows users to access the Internet from various public locations. Unfortunately, the nature of these networks pose serious risks to users' security and privacy. As a result, public WiFi users are encouraged to adopt a range of self-protective behaviors to prevent their potential online victimization. This paper explores the prevalence of one such behavior---avoidance of sensitive websites---among public WiFi network users. Moreover, we investigate whether computer users' adoption of an online avoidance strategy depends on their level of uncertainty regarding the security practices of the WiFi network they login to. To answer these questions, we analyze data collected using two phases of field observations: (1) baseline assessment and (2) introduction of a private (honeypot) WiFi network. Phase one baseline data were collected using packet-sniffing of 24 public WiFi networks in the DC metropolitan area. Phase two data were obtained through introducing a honeypot WiFi network to 109 locations around the DC Metropolitan area and an implementation of a quasi-experimental one-group-post-test-only research design. Findings reveal that although most WiFi users avoid accessing banking websites using established public WiFi networks, they still use these networks to access social networks, email, and other websites that handle sensitive information. Nevertheless, when logged in to a WiFi network that has some uncertainty regarding the legitimacy and security practices of its operator, WiFi network users tend to avoid most websites that handle sensitive information.

1. Introduction

The expansion of public WiFi networks in small business (for instance coffee shops, restaurants), academic institutions, and municipalities in the USA and around the world [11,3] allows users to login to the Internet from various public locations and at all times of day. In most cases, these wireless networks are easily accessible to customers and other users, and do not require any form of user authentication or identification for using them [23]. Once logged in to these networks, public WiFi users tend to check their email accounts, access social networks, shop online, and even access their bank accounts [18]. Unfortunately, since many of the public WiFi networks are unencrypted [23] and allow for an easy distribution of malware [11], man-in-the-middle attacks [1], and hijacking connection [20], they pose series risks to their users' security and privacy.

Acknowledging these risks, the Federal Trade Commission (FTC) encourages public WiFi users to take specific precautions when using these networks. For instance, users are instructed to use encrypted WiFi networks, only enter personal identifying information on secured websites (i.e. websites that their URL address begins with https), use Virtual Private Network (VPN) connections, and avoid sending emails containing personal information (see https://www.consumer.ftc.gov). Few experts even go further to suggest that since malicious WiFi networks could be easily deployed by criminals in order to trick people to log into them [23], users should completely avoid online banking and accessing sensitive data when using a public WiFi network (even if these websites are encrypted). Unfortunately, despite the continued efforts that are being made to improve public WiFi users' awareness of these hazards and the security measures that they need to take [10], we still lack understanding of how common self-protective behaviors are among public WiFi users. Moreover, it is relatively unknown what could spark self-protective behaviors among internet users who employ WiFi hotspots.

Addressing these issues, this paper seeks to answer two key research questions; first, how established the self-protective practice of avoidance from accessing websites that handle sensitive information is among public WiFi network users? And second, does the uncertainty regarding the legitimacy of the WiFi network operator determine computer users' avoidance from accessing websites that handle sensitive information? To answer these questions, we analyze data collected using both survey and experimental research designs. The integration of two complimentary research designs allows a more thorough investigation of public WiFi users' online self-protective behaviors, as well as the context in which these behaviors are more likely to occur. We begin this paper with a brief overview of the important role of self- protective behaviors in preventing the completion of a criminal event, and situate this discussion in the context of the online environment and public WiFi users' decision-making process when accessing the network. We continue with a description of the survey methodology (phase 1) and the experimental research design (phase 2) we employed in our research. Followed by that we discuss findings from statistical analyses we performed. We conclude by considering the theoretical and policy implications of these findings.

2. Theoretical Framing

2.1 Victim Self-Protective Behaviors

Victim Self-Protective Behaviors (VSPB) occur when individual attempts to protect himself from becoming the victim of crime [2]. Broadly, criminologists differentiate between two major types of VSPB: forceful and non-forceful resistance. Forceful resistance refers to active aggressive behaviors like pushing, biting, and kicking, that are introduced by a victim directly against a perpetrator in order to prevent an act of a criminal event [21]. Non-forceful resistance, on the other hand, refers to passive resistance techniques that are used by a victim to avoid offenders, and consequently, reduce the probability of a criminal event [9]. Examples of behaviors that could be classified as non-forceful strategies include avoiding an offender, escaping, pleading and begging. Findings from past criminological research suggest that both forceful and non-forceful resistance can decrease the likelihood of sexual abuse and rape [15], domestic violence [2] and robbery [24,9] from occurring or escalating. These findings coincide with the theoretical rationale extended by two key criminological theories that aim to explain the probability of a successful criminal event to be completed: The Routine Activities Theory [5] and the Situational Crime Prevention perspective [4].

The Routine Activities Theory [5] focuses on identifying behaviors, activities and situational contexts that put would-be targets at risk for criminal victimization [19]. In their original formulation of the theory, Cohen and Felson suggested that the structure of aggregated daily routines determine the convergence in time and space of motivated offenders, suitable targets and capable guardians, and influence trends of predatory crime. For the purposes of this study, capable guardianship, or the presence of individuals capable of, and motivated to intervene on behalf of potential victims, is notably absent in the context of public WiFi. Reflecting on the relevance of VSPB in the context this theory, one may suggest that greater use of VSPB would complicate offenders' attempts to complete a criminal event and reduce its occurrence [9]. Moreover, victims' use of non-forceful resistance technique like evasion and avoidance will remove the victim from the criminogenic situation, and prevent the occurrence of a criminal event [24]. Simply put in the original context, the application of VSPBs should reduce the suitability of potential targets.

The Situational Crime Prevention perspective [4] is focused on the occurrence and development of criminal events. The underlying premise of this perspective is that criminals are rational, weighing the costs and benefits of their prospective behaviors, so successful crime prevention efforts must involve the design and manipulation of human environments to make offenders' decisions to get involved in crime less attractive [4]. Therefore, Clarke recommended the adoption of crime-specific prevention strategies (for instance, strategies targeting theft, robbery, burglary, vandalism, etc.) that fall into five categories: *increase offenders' effort, increase offenders' risks, reduce offenders' rewards, reduce provocations, and remove excuses* [7]. VSPB on its various forms are of utmost relevance in the context of this perspective since victim's resistance would increase offenders' effort to complete a criminal event and offset offenders' cost and benefit calculations [9].

Although past research has focused on the effect of VSPB on preventing offline victimization, we suspect that non-forceful resistance VSPBs are also relevant in preventing online victimization. For example, like installing a security or alarm system in someone's home to prevent burglary, installing antivirus software on one's computer is considered an effective practice for preventing malware attacks [16-17]. Similarly, while avoiding a potential neighborhood or street segment is proved to be an effective non-forceful strategy for reducing the probability of robbery [24], spending less time on untrusted or untrustworthy websites and downloading copyright protected material illegally to a computer may reduce individual likelihood to experience a wide range of cybercrimes [10]. Importantly, we believe that there is a need to differentiate between offline and online non-forceful VSPB in order to understand how these strategies reduce the probability of an online criminal event. For instance, [14] report that public WiFi users attempt to protect their privacy when working with the network by tilting or dimming their computer screens, as well as sitting with their computers angled toward the wall. These behaviors could be classified as offline non-forceful resistance strategies. In contrast, installing an antivirus package, using a secure VPN connection, and avoiding accessing and handling sensitive information while using public WiFi networks could be classified as an online non-forceful resistance strategies that reduce the probability of cybercrime from progressing.

2.2 The Current Research

Our focus in this paper is on internet users' online avoidance from accessing sensitive websites while using a public WiFi network. Specifically, we seek to determine *how common avoidance from accessing websites that handle sensitive information (banking, email, social networks and personal cloud – e.g. google drive, dropbox, etc.) among WiFi networks is.* Indeed, previous research has already investigated public WiFi users' online routines. For example, findings reported by the Identify Theft Resource Center [12] suggest that 57% of the public WiFi users they sampled logged into a work-related system like email or file sharing while using a public WiFi network and that 24% of respondents made purchase using a credit card while using the network. Similarly, [22] reports that 83% of public WiFi users use their emails, 68% use their social media accounts, 43% access work specific information, 42% shop online and 18% access banking websites while using public WiFi networks. While these reports are informative and suggest variation with respect to the type of websites that public WiFi users tend to access while employing public WiFi networks, these reports draw on problematic samples, employ questionnaires for gathering data from subjects, and fail to take into consideration the physical and temporal conditions which may influence public WiFi users' decisions to engage in these online behaviors. We suspect that a more hands on approach to assess public WiFi users' online routines with the network is to see what people are actually doing on public WiFi network by monitoring locations which host a public WiFi hotspot and observing the traffic they generate.

In addition to exploring how likely public WiFi users are to avoid accessing websites that handle sensitive information, we also explore whether uncertainty regarding the owner of a WiFi network shapes users' avoidance from accessing websites that handle sensitive information. To this end, prior psychological theory and research indicates that decision makers tend to be ambiguity-averse [8,13]. Accordingly, when forming expectations about the consequence of their possible behaviors, individuals opt for prospects with known risks as opposed to unknown risks. In line with this rationale, we believe that the introduction of ambiguous information (i.e. missing information that prevents decision makers' ability to estimate the probability of an event) regarding a WiFi network and its operator, will disrupt public WiFi users' calculations of their risks of becoming the victims of cybercrime, and will induce more cautious online behaviors in contrast to when using a network whose owner is known.

3. Data and Methods

To answer these two research questions, we collected data across two phases: (1) a baseline assessment of user behavior on extant WiFi networks, and (2) an evaluation of if, and how individuals use an unknown network that was introduced. Phase one baseline data was collected by packet-sniffing extant public WiFi networks at 24 locations in the DC metropolitan area. In phase two, we introduced our own WiFi network in 109 locations around the DC Metropolitan area and implemented a quasi-experimental one-group-post-test-only research design. Like in phase one, for the second phase, we deployed private WiFi networks (honeypots) and packet-sniffed the internet traffic on these private networks.

3.1 Public WiFi Baseline Assessment

To explore public WiFi users' online behaviors we collected public WiFi network data by launching 72 packet sniffing sessions in 24 locations across Maryland and the DC metropolitan area using the software "Wireshark". Wireshark is a network protocol analyzer that can monitor and capture network packets that have not been addressed to the host. We used "Wireshark" to collect packet data in one hour sessions at three times of day (morning, noon and evening[1]), recorded the public WiFi speed, and counted the number of devices that used the network. Since in six of the sniffing session no computer users attend the location we report data from 66 sessions. These data from the public WiFi networks were used for identifying how computer users are using public wireless networks. Specifically, we examined unencrypted WiFi traffic to determine websites visited by users and the activities with which these websites are associated (e.g., checking email, watching videos, using a P2P file-sharing service, etc.), whether or not end-to-end encryption (e.g., SSL or a VPN) is being used, and whether malware is detected on the host and/or in the inspected traffic. Importantly, in order to protect people's privacy and maintain anonymity, the collected data was aggregated across each data collection session and not linked to specific users.

In addition to network data, we also collected data from the physical environment in which the public WiFi hotspot operated. Figure 1 presents an example of data collected from a location in Washington, DC during a 1-hour sniffing session by one of our research assistants. As may be observed in the figure, once arriving at a research location our research assistants diagrammed the physical layout of the space as well as recorded information about *the number of individuals who were present in each research site, number of male, number of female, number of customers, number of employees, number of observed smartphone devices, number of laptops, percent of individuals sharing a table, and percent of people sitting in adjacent tables.*

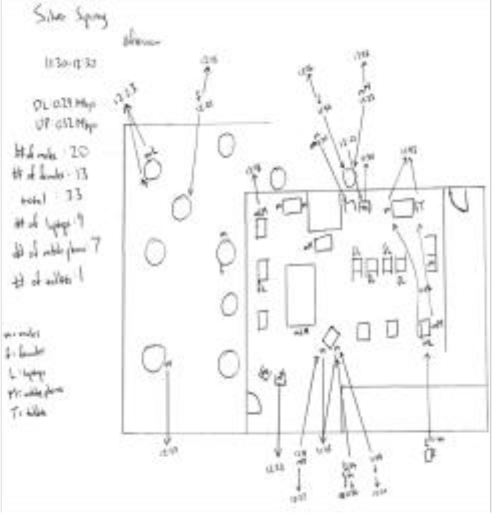


Figure 1. Observations Recorded During A One Hour Sniffing Session in Sliver Spring MD (Afternoon)

Finally, information regarding neighborhood demographic and social characteristics was downloaded from the U.S. Census website (available at www.census.gov). Specifically, we download neighborhood (census tract level) information regarding *the total population in the neighborhood, percentage of residents that are below the poverty line, percentage of residents in the community who are unemployed, percentage of households in the community that are headed by a female, percentage of residents living in the same house in the last 5 years, and percentage of foreign-born resident in the community.*

3.2 WiFi Network Honeypots

Next, to understand public WiFi users' behaviors on the network we also explored computer users' willingness to login to a WiFi network they were not familiar with (and which we owned). To do so, we introduced a new and unknown network to locations similar to those selected in phase one. In this experimental design, the characteristics and outcomes of interest were measured across both phases and thus can be compared on observable attributes.

[1] Morning sessions were defined as entirely within the hours of 8:00am and 11:00am, afternoon sessions were within the hours of 12:00pm and 2:00pm, and evening sessions were between the hours of 5:00pm and 8:00pm on weekdays.

Adopting this research design in our work, we selected 109 research sites with a wireless router of our own at three times of day (morning, noon and evening)[2] for each location. The router allowed users easy access to the Internet since it did not require login credentials (i.e. password and user names). Traffic on this network was closely monitored by a student who packet-sniffed our network using the "Wireshark" software and tools native to the router. Our goal in this was to determine the proportion of public WiFi users who are likely to roam around and look for WiFi networks to login to and use. We were also interested to understand these users' online behaviors while on the untrusted network. All in all, we observed and analyzed internet traffic on 34 of the 109 locations we visited (i.e. 31% of the research sites). Importantly, the current research does not seek to explain the variation between locations in which computer users accessed and did not access our networks. Instead the current work is focused on the type of traffic we observed on the WiFi networks we deployed. Thus, consistent with the data collected in the public WiFi baseline assessment phase, we collected information on online users' online behaviors and susceptibility to cybercrime victimization using "Wireshark". We also collected relevant information on the physical environment using observations. Finally, we downloaded information regarding neighborhood demographic and social characteristics from the U.S. Census website.

3.3 Ethical and Privacy Considerations

We have applied for an IRB approval for this project and the IRB team in the University of Maryland determined that our project does not involve human subjects, and hence does not require an IRB approval. Further, honeypot networks deployed in phase 2 of this study were clearly labeled as "private", and thus potential users knowingly trespassed on an unknown private network. In addition, we also consulted with the legal team at the University of Maryland and verified that the act of sniffing is legal in the state of Maryland. Indeed, the use of a free and public program to sniff in unsecure public networks has been ruled to be legal under the Wiretap Act (see "In re INNOVATIO IP VENTURES, LLC PATENT LITIGATION", District Court, ND Illinois 2012) and has been employed by [3] in their investigation of public WiFi networks in 20 international airports (located in 4 countries). However, in line with the University of Maryland Legal Team's recommendation, we did not initiate a sniffing session in public WiFi locations in which this activity was specifically prohibited by the network owner.

3.4 Dependent Variables

The data collected using Wireshark during our packet-sniffing sessions indicated that WiFi users employed the network for accessing wide range of websites. Indeed, we observed packet-data of advertisement, E-commerce, education, news, sport and video streaming websites. However, since our goal in this paper is focused on WiFi users' online self-protective behaviors, we observe in this work the relative number (i.e. proportion) of sniffing sessions and WiFi networks on which users accessed websites that handle sensitive information as a dependent measure. Specifically, we calculated the *proportion of packet sniffing sessions and WiFi network hotspots on which packet-data that is associated with banking, email, social network and personal cloud* websites was observed.

4. Results

4.1 How prevalent is avoidance from accessing sensitive websites among public WiFi network users?

We begin by presenting findings regarding the prevalence of packets originating from sensitive websites and observed over public WiFi locations. In the following, the unit of analysis is the location.[3] Figure 2 shows the proportion of sniffing sessions (N=66) at which banking, social network, email, and personal internet packets were observed. As indicated in the figure, banking websites packets were observed at 38% of the sniffing sessions we collected. In addition, packets originated in social network websites were observed at 86% of the sniffing sessions, packets from email accounts on 68% of the sniffing sessions, and packets from a personal cloud on 73% of the sniffing sessions.

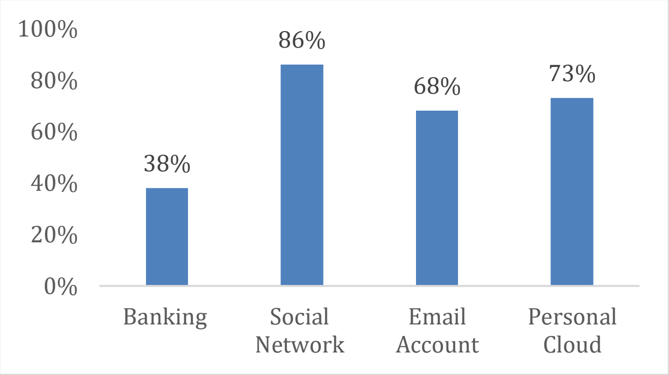


Figure 2. Internet Traffic Observed on Public WiFi Hotspots in the DC Metropolitan Area (N=24 Unique Locations).

Since we packet-sniffed the 24 locations during three times of day (morning, afternoon, and evening), we further explored whether the presence of packets from websites that handle sensitive information varies by time of day. Findings from this analysis are presented in Figure 3. As indicated in the figure, with few exceptions, the presence of packet data from banking, social network, email and personal cloud websites on public WiFi hotspot tended to be consistent throughout the day. Indeed, it appears that banking packets are less common during evening sniffing sessions than during morning and afternoon sessions, and that both email and personal cloud packets are less common on public WiFi hotspots during morning sniffing sessions than during afternoon and evening sessions. However, analyses from a chi-square test suggests that these differences are not statistically significant.

[2] With the same criteria as in phase one.

[3] These data were aggregated up from one to three hours of data collection sessions dependent upon the hours of operation for each location and most fairly represent the limitations of using DNS packet queries as an indicator of network traffic rather than presuming to measure the volume of said traffic.

These findings also suggest that public WiFi users generally do not avoid accessing websites that handle sensitive information. In fact, evidence from our packet-sniffing sessions suggests that public WiFi users access social media, email, and personal cloud accounts while using public WiFi hotspots. Still, the relatively low prevalence of locations where banking packets were observed indicates that public WiFi users may be taking steps to avoid accessing sensitive banking information from these networks.

4.2 Does Computer Users' Online Avoidance Depend on the Level of Uncertainty Regarding the WiFi Network?

Next, we explore whether ambiguity regarding the WiFi network and its owner determine users' probability of accessing sensitive websites. To answer this question, we compare the proportions of extant public WiFi hotspots on which banking, email, social media, and personal cloud website packets were observed with the proportions of honeypot WiFi networks on which similar packets were observed. Note that while the analyses performed to answer our first research question were focused on the presence of packet data on the *sniffing session (i.e each time we sniffed the network)*, we answer our second research question by investigating the presence of packet data on the *WiFi network (i.e. aggregating the three sniffing sessions we ran per each location l)*. Specifically, we employ the data collected during our initial phase of public WiFi network assessment (i.e. 24 locations) and compared it with packet data collected on our honeypot WiFi networks (i.e. 34 locations).

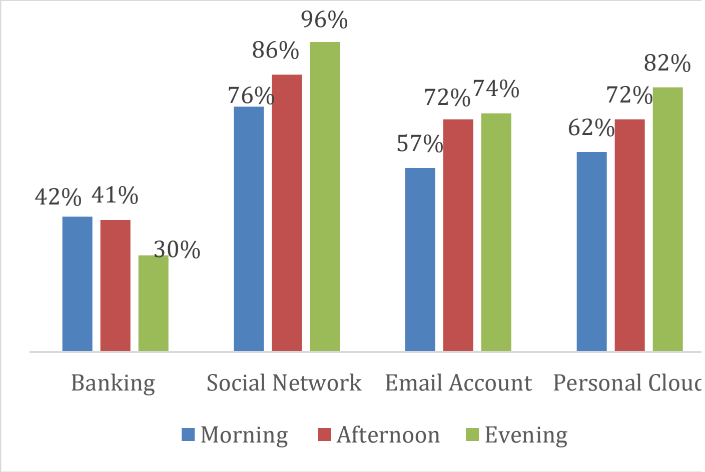


Figure 3. Internet Packets Observed During 66 Sniffing Sessions on Public WiFi Hotspots in the DC Metropolitan Area Across Three Times of Day

Before turning to answer our second research question, we first compare both the location and neighborhood level characteristics in which our research team either sniffed the public WiFi network, or deployed and observed traffic on the honeypot WiFi network that was deployed. These findings are presented in Table 1 and Table 2. As indicated in Tables 1 and 2, both the physical and social landscape and census tract characteristics of the locations we attended both in the assessment and honeypot phases of our project are very similar. In fact, the only significant difference between the contextual characteristics of the extant public WiFi hotspots and the contexts in which the honeypot WiFi networks were introduced was with respect to the number of mobile devices observed. Specifically, we observed a significantly higher number of mobile devices during the assessment of extant public WiFi network locations than in the locations where we deployed our own WiFi network. At the neighborhood level, it appears that the neighborhoods in which we deployed our WiFi networks had a significantly higher percentage of foreign-born residents than in the public locations with extant WiFi networks. Moreover, residential stability (i.e. percent living in the same house for more than 5 years) is significantly higher in the neighborhoods in which we surveyed the extant public WiFi networks than in the neighborhoods where we deployed our honeypot network.

Location Physical and Social Characteristics	Extant Public WiFi Network	Honeypot WiFi Network
	Mean (SD)	Mean (SD)
Number of people	23.47 (12.30)	21.16 (17.39)
Number of males	11.25 (5.75)	10.66 (9.75)
Number of females	10.97 (6.18)	10.50 (8.42)
Number of customers	20.93 (11.49)	18.66 (16.39)
Number of employees	2.53 (1.69)	2.49 (2.14)
Number of mobile devices (observed)	8.22 (6.64)	2.77* (3.13)
Number of Laptops (observed)	4.31 (5.03)	2.70 (6.05)
% people sharing a table	61.88 (23.94)	69.77 (43.23)
% people sitting in adjacent tables	74.16 (25.98)	77.16 (56.85)

* p<0.05 ** p<0.01

Table 1. Location Physical and Social Characteristics of Public WiFi Hotspots and Locations in which WiFi Networks Were Deployed

Next to investigation of significant differences between the physical and social landscapes and neighborhood characteristics across networks, we also test for significant differences between the presence of traffic to websites that do not require accessing sensitive information on the two types of networks. Findings from that analysis are reported in Table 3. As shown in Table 3, users of both extant public WiFi networks and the honeypot WiFi networks used the Internet for accessing educational, news, sport and video streaming websites. Moreover, packets reflecting advertisement traffic were observed on both type of networks. However, the proportion of extant public WiFi hotspot locations with packets in these five website types is significantly higher than the proportion of honeypot WiFi network locations with the same type of packets.

Finally, to answer our second question we compared the proportion of extant public WiFi locations and locations where our own WiFi networks were deployed on user access to websites that handle sensitive information. Findings from this analysis are presented in Figure 4. As indicated in the figure, banking website packets were observed on 54% of the extant public WiFi hotspots that we surveyed. In addition, packets indicative of social network website use were observed on 100% of the extant public WiFi hotspots, packets from email sites on 83% of the hotspots, and packets from a personal cloud on 87.50% of the public WiFi hotspots. In contrast, no banking, email or personal cloud packets were observed on the honeypot WiFi networks. However, in close to 68% of the locations with WiFi networks we deployed we observed packets indicative of social media website use. To test whether the proportion of extent public WiFi locations and locations where our own WiFi networks differ on the presence of packets of websites that handle sensitive information we ran a T-test for determining whether the difference between the two proportions is significant. Findings from these t-tests reveal statistically significant difference between public WiFi and honeypot WiFi for each type of packet that is originated in website that handle sensitive information. Thus, this finding suggests that internet users are more likely to avoid accessing websites that transmit sensitive data when employing WiFi networks that carry uncertainty with respect to their owners.

Neighborhood Characteristics	Public WiFi	Unfamiliar WiFi Network
	Mean (SD)	Mean (SD)
Total population	3405 (1384.24)	4213 (2781.90)
Percent poverty	14.97 (9.09)	13.92 (13.26)
Percent unemployed	5.70 (4.00)	4.43 (3.10)
Percent foreign born	13.62 (10.42)	21.34* (14.46)
Percent female headed household	25.18 (18.03)	35.11 (61.17)
Percent living in the same house for more than 5 years	77.86 (9.40)	70.06** (11.07)

* p<0.05 ** p<0.01

Table 2. Census Tract Characteristics of Extant Public WiFi Hotspots and Honeypot WiFi Deployment Locations

5. Discussion

As public WiFi use proliferates and the number and speed of hotspots continues to grow, the commensurate risk of cybercrime on these networks is likely to rise accordingly. Drawing on the VSPB perspective, we designed and collected two phases of data to assess first how individuals make use of known, albeit often unsecured, wireless networks, and second, if, and how individuals would utilize an unknown network of uncertain management and origin. First, we asked how established the VSPB of avoidance is on websites that handle sensitive information among WiFi network users. Second, we sought to consider if uncertainty regarding the provenance of the WiFi network is associated with differential adoption of this avoidance technique. Findings from our unique field study provide several insights. First, we find some support for the extension of the VSPB framework to cyber environments. Insofar as self-protective behaviors may be concerned in the physical world, it appears that when connected to a public WiFi network, in more than half of the locations that we observed, individuals did not access banking websites. This finding was somewhat attenuated when considering the traffic to Social Networks, Email, and Personal Cloud services. This suggests that while there may be a salient risk associated with accessing ones bank on public WiFi, either due to the ubiquity of, or ambivalence toward disclosure of potentially less sensitive details available on social media, in emails, and backed up on personal cloud services, this traffic may not be conceived of as concerning to users.

Packets type	Proportion of extant Public WiFi Locations with Packets Observed (n=24)	Proportion of honeypot WiFi Locations with Packets Observed (n=31)
Advertisement	.83	.65**
Education	.41	.21**
News	.70	.27**
Sport	.41	.09**
Video streaming	.67	.23**

* p<0.05 ** p<0.01

Table 3. Proportion of Extant Public WiFi and Honeypot WiFi Network Locations in the DC Metropolitan Area with Different Types of Packets

Second, we find support for the notion that the introduction of uncertainty to the source and management of WiFi networks (as on our honeypots) could serve as a deterrent for sensitive web traffic by users. Consistent with [8] and [13], individuals who chose to login to honeypot networks appeared to be more cautious in their sensitive web traffic, only accessing social media in addition to less vulnerable sites. Again, the evidence of social media traffic suggests that the inter-connected world that we live in may habituate individuals to sharing such details as are present on their public social media profiles. However, this is not to say that such ambivalence to disclosing these details is without risk. As can be seen from cases of cyber-stalking and cyber-bullying, access to an individual's social media account can be a very damaging in the wrong hands. In sum, the application of avoidance as a VSPB online, when incorporated with the use of appropriate antivirus software and safe internet behavior when on unsecured networks retains an important role in limiting victimization risk on public WiFi.

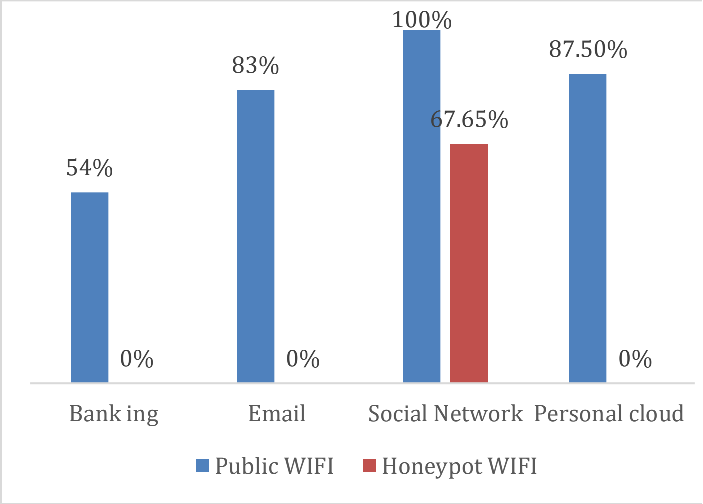


Figure 4. Internet Packets Observed on 24 Public WiFi Location and 34 Honeypot WiFi Networks

Finally, it behooves us to account for the limitations of this project. Due to the abundance of public WiFi networks, establishing a sampling frame from which to draw a representative sample of locations or networks was beyond the scope of this project. Thus, the findings presented herein are descriptive in nature and should be qualified as such. Future research should consider a means from which to obtain a census of specific types of WiFi hotspots from which to draw a more generalizable sample. Furthermore, additional characteristics of WiFi traffic and network users should be considered and controlled for in future analyses, including the base rate of traffic to given types of websites, the number of devices on networks, and duration of device network use. Additionally, while the use of Wireshark for categorizing DNS packet queries to servers represents an important first step in assessing network traffic on extant public and honeypot WiFi networks, future research should consider the use of HTTP and HTTPS packets for greater granularity of traffic data.

6. Conclusions

Avoidance from accessing websites that handle sensitive information is a type of online self-protective behavior that could be easily employed by public WiFi users to prevent their potential cybercrime victimization. While this avoidance strategy is rare among public WiFi users' in the context of social media, email, and personal cloud services, it appears to be quite common with respect to banking websites. Moreover, increasing the level of uncertainty regarding the WiFi network's legal owner and operator is associated with an increased likelihood of avoiding websites that handle sensitive information.[4]

References

[1] Aime, M. D., Calandriello, G., & Lioy, A. 2007. Dependability in wireless networks: Can we rely on WiFi?. *IEEE Security & Privacy*, *5*(1).

[2] Bachman, R., Saltzman, L. E., Thompson, M. P., & Carmody, D. C. 2002. Disentangling the effects of self-protective behaviors on the risk of injury in assaults against women. *Journal of Quantitative Criminology*, *18*(2), 135-157.

[3] Cheng, N. , Xinlei W, Wei C., Prasant M. and Aruna S. 2013. "Characterizing Privacy Leakage of Public WiFi Networks for Users on Travel." Proceeding of INFOCOM'13 IEEE.

[4] Clarke, R V. 1995. Situational crime prevention. *Crime and justice* 19: 91-150.

[5] Cohen, L.E. and Felson, M. 1979. Social Change and Crime Rate Trends: A Routine Activity Approach. American Sociological Review 44: 588-608.

[6] Consolvo, S., Jung, J., Greenstein, B., Powledge, P., Maganis, G., & Avrahami, D. 2010, September. The Wi-Fi privacy ticker: improving awareness & control of personal information exposure on Wi-Fi. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 321-330). ACM.

[7] Cornish, D. B., & Clarke, R. V. 2003. Opportunities, precipitators and criminal decisions: A reply to Wortley's critique of situational crime prevention. *Crime Prevention Studies, 16*, 41-96.

[8] Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics*, 643-669.

[9] Guerette, R. T., & Santana, S. A. 2010. Explaining victim self-protective behavior effects on crime incident outcomes: A test of opportunity theory. *Crime & Delinquency*, *56*(2), 198-226.

[10] Holt, T. J., & Bossler, A. M. 2014. An assessment of the current state of cybercrime scholarship. *Deviant Behavior*, *35*(1), 20-40.

[4] We would like to thank our reviewers for their thoughtful comments in improving the clarity of this and future work.

[11] Hu, H., Myers, S. Colizza V., &Vespignani, A. 2009. “WiFi Networks and Malware Epidemiology.” PNAS 106(5): 1318-1323.

[12] Identity Theft Research Center. 2012. Public WiFi Usage Survey. Available at: https://www.idtheftcenter.org/images/surveys_studies/PublicWiFiUsageSurvey.pdf

[13] Kahneman, D., & Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, 263-291.

[14] Klasnja, P., Consolvo, S., Jung, J., Greenstein, B. M., LeGrand, L., Powledge, P., & Wetherall, D. 2009, April. When i am on wi-fi, i am fearless: privacy concerns & practices in eeryday wi-fi use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1993-2002). ACM.

[15] Kleck, G., & Sayles, S. 1990. Rape and resistance. Social Problems, 37(2), 149-162.

[16] Lalonde Lévesque, F., Nsiempba, J., Fernandez, J. M., Chiasson, S., & Somayaji, A. 2013. A clinical study of risk factors related to malware infections. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (pp. 97-108). ACM.

[17] Lalonde Lévesque, F. L., Fernandez, J. M., & Somayaji, A. 2014. Risk prediction of malware victimization based on user behavior. In *Malicious and Unwanted Software: The Americas (MALWARE), 2014 9th International Conference on* (pp. 128-134). IEEE.

[18] Norton. 2013. 2013 Norton Report. Symantec

[19] Pratt, T.C., Holtfreter, K. & Reisig, M.D. 2010. Routine Online Activity and Internet Fraud Targeting: Extending the Generality of Routine Activity Theory. Journal of Research in Crime and Delinquency 47: 267-296.

[20] Song,, Y., Yang C., & Gu, G. 2010. Who is Peeping at Your Password at Starbucks? – To Catch an Evil Twin Access Point. In: Proc. 2010 IEEE/ IFIP International Conference on Dependable Systems and Networks (DSN).

[21] Ullman, S. E. 1997. Review and critique of empirical studies of rape avoidance. Criminal Justice and Behavior, 24(2), 177-204.

[22] Xirrus. 2016. Rolling the Dice with Public WiFi. Available at: https://www.xirrus.com/pdf/Rolling-The-Dice-With-Public-WiFi.pdf

[23] Zafft A. & Ago E. 2012. “Malicious WiFi Networks: A First Look.” 7th IEEE Workshop on Security in Communication Networks.

[24] Ziegenhagen, E. A., & Brosnan, D. 1985. Victim responses to robbery and crime control policy. Criminology, 23, 675-695.

Measuring the Success of Context-Aware Security Behaviour Surveys

Ingolf Becker, Simon Parkin and M. Angela Sasse
University College London
$\{$*i.becker, s.parkin, a.sasse*$\}$*@cs.ucl.ac.uk*

Abstract

Background. We reflect on a methodology for developing scenario-based security behaviour surveys that evolved through deployment in two large partner organisations (A & B). In each organisation, scenarios are grounded in workplace tensions between security and employees' productive tasks. These tensions are drawn from prior interviews in the organisation, rather than using established but generic questionnaires. Survey responses allow clustering of participants according to predefined groups.
Aim. We aim to establish the usefulness of framing survey questions around active security controls and problems experienced by employees, by assessing the validity of the clustering. We introduce measures for the appropriateness of the survey scenarios for each organisation and the quality of candidate answer options. We use these scores to articulate the methodological improvements between the two surveys.
Method. We develop a methodology to verify the clustering of participants, where 516 (A) and 195 (B) free-text responses are coded by two annotators. Inter-annotator metrics are adopted to identify agreement. Further, we analyse 5196 (A) and 1824 (B) appropriateness and severity scores to measure the appropriateness and quality of the questions.
Results. Participants rank questions in B as more appropriate than in A, although the variations in the severity of the answer options available to participants is higher in B than in A. We find that the scenarios presented in B are more recognisable to the participants, suggesting that the survey design has indeed improved. The annotators mostly agree strongly on their codings with Krippendorff's $\alpha > 0.7$. A number of clusterings should be questioned, although α improves for reliable questionsby 0.15 from A to B.
Conclusions. To be able to draw valid conclusions from survey responses, the train of analysis needs to be verifiable. Our approach allows us to further validate the clustering of responses by utilising free-text responses. Further, we establish the relevance and appropriateness of the scenarios for individual organisations. While much prior research draws on survey instruments from research before it, this is then often applied in a different context; in these cases adding metrics of appropriateness and severity to the survey design can ensure that results relate to the security experiences of employees.

1 Introduction

Engaging users is important to develop meaningful, effective security behaviour surveys. If studies are conducted out of context, reproduction of results is difficult [24]. Yet much of security awareness research examines individuals' abilities to internalise and enact knowledge of security risks and controls in an abstract setting. Efforts to measure security behaviour frequently assess individuals' competency in general security skills (see Section 2). Much of this research ignores the bounded effort of the individual [6, 16], and that employees in organisations have other responsibilities [3].

Here we run a validation exercise on scenario-based surveys conducted in two large organisations each with many thousands of staff. Scenarios are built on frictions between security and regular business tasks derived from prior exploratory interviews with a cross-section of employees. The core principles of the methodology underlying the two surveys are: determining attitudes toward security provisions and policy in the organisation, and; characterising how individuals act independently or with others to enact security-related behaviours. The differences between the surveys represent an evolution in survey design as lessons have been learned, where we develop measures which account for these differences and allow cross-comparison between survey deployments. We describe the framework for our scenario-based surveys in Section 3.

Here we explore the capacity to utilise additional types of questions to reflect on the survey design without further effort by the researchers. If the participants are given an opportunity to indicate the applicability of the scenarios to their environment, we can tailor the results not just to specific user groups, but also reflect on how a survey engages with diverse groups and their security needs.

From the analysis we formulate further metrics for measuring how aligned the security apparatus of an organisation is with the employees who are governed by the policies and controls that are in place. This is achieved through Likert-scale questions added to the existing questionnaire, as described in our methodology (Section 4), which serve as internal validity measures.

The surveys conducted with our two partner organisations contain questions structured in this way, and we discuss the results of our research in Section 5. We find that the appropriateness and applicability of the questions of the survey have improved from A to B. Similarly, the reliability of the clustering of the answer options has improved. Yet participants judge the answer options in B to be more severe and less balanced than in A.

This immediately available feedback allows researchers to continuously evaluate their survey design and discard unreliable questions from further analysis. In broader terms, we reflect on this approach in the discussion that follows in Section 6, and in the conclusions in Section 7.

2 Related Work

We consider scenario-based security behaviour survey research from two perspectives: Initially we examine the construction and motivation of these surveys, and in the second stage we focus on the reliability of survey analysis (given that surveys would be deployed in specific organisational contexts). Our review of related work highlights the need to situate scenarios in the participants' environment to build a reliable picture of how security provisions and workplace conditions interact.

Most of the works reviewed do not evaluate the external validity of the questionnaires applied but instead rely on additional prior work. Our methodology brings obvious benefits to survey designs in research and practice alike.

Egelman et al. [14] developed the Security Behavior Intentions Scale (SeBIS) to predict security behaviours for common controls ('awareness', 'passwords', 'updating', and 'securement'). The SeBIS survey comprises of 16 items on a 5-point Likert scale. SeBIS was deployed on several occasions through Mechanical Turk (and in one case, PhoneLab). The goal of the work was to determine if self-reported, *intended*, behaviours translated into actual behaviour. To this end, tasks were set relating to each behaviour category (such as identifying fake login pages). The authors accepted that the designed tasks were targeted and narrow in scope, but with a focus on exploring SeBIS' predictive capabilities in this limited setting. Here we use scenarios and options based in real organisational settings to establish an individual's behaviour type and attitude toward the security apparatus around them; the focus is not on predicting behaviour, but rather to capture a snapshot of how effectively security provisions are perceived to be supporting the business.

Parsons et al. [26] sought to validate a survey tool for measuring information security awareness and awareness initiatives, the Human Aspects of Information Security Questionnaire (HAIS-Q).Two studies were conducted: in one study, participants completed the HAIS-Q and were tested for security skills (in this case, identifying potential phishing links amongst a range of fabricated emails); in the second study, engagement of participants in the survey was examined by establishing the level of *non-responsivity*. Here, we similarly seek to determine whether the scenarios and response options in our surveys resonated with participants, through examination of internal measures within our situated surveys. By doing so we identify repeatable measures for measuring engagement.

Rajivan et al. [27] propose a questionnaire for capturing users' level of *security expertise*, presented as being a critical factor in how well an individual can assess risk and use available security controls. The questionnaire seeks to separate respondents across the dimensions of skills, rules, and knowledge, toward understanding how individuals apply these in different situations. Here we discuss our survey methodology as a means to not only determine how employees use the tools available to them as individuals and groups, but also how they respond to specific risks which can potentially arise in their working environment. Rajivan et al. also included free-text questions to capture additional comments from participants, where we use a similar internal mechanism in our situated scenarios so that participants can further describe security experiences from their own perspective (further informing the picture of security *on the ground*).

Karlsson et al. [20] posit that in organisations, information security compliance must be evaluated relative to employees' work tasks (and with this, competing goals and their related *values* such as productivity and efficiency). The authors speak of there being "tensions and dilemmas" where one option is preferable to others that are available. While the argument is made that situational context is critical to understanding how tensions are resolved, the authors' questionnaire however is free from any contextual settings. This does allow it to be applied to any *"white-collar individual"*, but may limit how the questionnaire captures the *"tensions and dilemmas"* that exist in a specific organisation. Here we are assessing scenarios which are grounded in prior interviews with employees, specifically to identify those regular tensions and dilemmas which occur in the workplace.

We argue that surveys that are situated in scenarios that the participants can relate to will engage them and evoke genuine responses, which can inform efforts to improve the effectiveness of security solutions in an organisation. Some research has focused on basing scenario design in literature, where Blythe [9] argues that scenarios should *"[avoid] unusual events and characters but nonetheless resonate with the respondent in a way that they are readily understood while presenting multiple solutions."* Siponen and Vance argue that research needs to be practically relevant by ensuring contextual relevance [29]. Their five suggestions focus on studying information system policy violations, but are equally transferable to other behavioural research (and related to the principles derived by Krol et al. in [23] for studying usability in security and privacy). Many examples of security behaviour research using questionnaire instruments do not consider the role of task conflicts characterised by Karlsson et al. [20]. These works instead draw on existing questions from prior research [13, 19, 30, 31, 33]. The importance of scenario-based surveys is underlined by Wash et al.'s findings that individuals do not self-report security accurately [32]: it undermines much of the traditional self-reported constructs used for inferring personal security behaviour.

Sohrabi et al. for example argue that *"the lack of information security awareness, ignorance, negligence, apathy, mischief, and resistance are the root of users' mistakes"* [30]. Their questionnaire uses nine information security constructs from prior literature to model their interactions. These questions and mappings are adopted from four previous studies [13, 19, 31, 33]. However each of these articles in turn is standing on the shoulders of giants, citing a total of 29 prior works to support their survey design, which, in turn, cite over 100 other unique articles to support the quality of their survey design. The sources span the fields of sociology, education, criminology, information systems and medical research.

Of the literature referenced by Sohrabi et al. and their references in turn, the vast majority source their constructs and survey questions from further literature. A number of articles construct their own questions. The most rigorous of those papers in the chain to do any pre-testing validation of their question design is by Bulgurcu et al. [11], who conduct two rounds of card sorting by 11 students followed by two rounds of pilot testing by 110 individuals. Huang [17] and Chang and Chuang [12] also conduct some limited pre-testing.

While it is good scientific practice to rely on constructs that have been rigorously tested in prior works, only one of the papers ([19]) discussed above cites the primary literature which validates the constructs in their original setting ([11]). Further, as throughout the literature, the questions are taken out of context of their original research premise, where the validity of the original validation should be revisited in each new context. It is understandable that a full pre-study is infeasible for every new questionnaire, yet augmenting the survey with additional questions (as described subsequently) to support the measurement of validity post-hoc would be cheap and desirable.

The data of our research is grounded in competing goals in realistic scenarios, so that (i) security managers can better understand how employees' attitude and behaviour toward security policy and controls influence the approach to problem resolution, and (ii) researchers can gain further insight into the shape of tensions between security and productive tasks in an organisational setting.

3 Background

In this section we describe the two surveys that contribute the primary data in this paper.

The first organisation to be studied we shall call Company A[1] (which has many thousands of employees). In this organisation 118 semi-structured interviews were conducted, exploring conflicts between security and business processes lasting on average 40 minutes. These interviews were analysed with thematic analysis, and form the basis of workplace-based scenarios and possible solutions informed by approaches reported by employees. Scenarios are combined to create a scenario-based survey. A similar approach is described by Blythe et al. [10] for conducting interviews around security behaviours within organisations, presenting *dilemmas* as short stories with a central character in a specific context (and informed by an organisation's security policies). A small proportion of Company A's workforce was sampled through interviews, where a wider survey would attempt to capture the prevalence of the issues identified in interviews across the wider company. A total of 1486 participants provided complete responses. Further, the respondents gave 516 additional comments at various stages of the survey, through text-entry fields provided.

A second, similarly large organisation (which we call Company B) was studied subsequently, where lessons learned from the analysis of Company A improved the organisation of the survey process. These lessons have caused the progression in groupings as described in Table 1 and discussed in Section 4. The initial attitude and behaviour types [4] were derived from Adams [1].

[1] for brevity the companies will be referred to as 'A' and 'B' for the remainder of this paper

85 interviews were conducted in Company B of similar lengths to the interviews in A, where again these were used in a similar methodology as in A to develop a larger-scale survey. A survey was conducted using scenarios built upon themes emerging from the qualitative analysis of the interviews. 641 employees responded to this survey, including 195 free text responses. While the survey results have been analysed [4, 7] here we explore how the free-text responses can indicate the success of the survey in engaging with the organisation's employees. Further analysis of the interviews that informed the survey designs can be found in [22] for Company A, [21] for Company B, and [8] for a combination of both organisations.

3.1 Employee types

In each of the two surveys we attempt to position participating employees on two dimensions. In A these are Attitude and Behaviour types, whereas in B these have evolved to Maturity levels and Behaviour types. For the definitions of the types please see Table 1. Foremost, these two dimensions can be examined individually and in combination – across age groups, business divisions, and physical locations – to target interventions which reduce friction between security and productivity in the workplace.

The attitude types in A focus on individuals' interaction with security apparatus. In B, these have evolved to a scale of Maturity Levels, which are ranked levels of individuals' interaction with the organisation's policy (such that interventions would act to improve employees' working interactions with centralised security policy and security provisions). In both A and B, the participants were also asked to assign an appropriateness score for each answer option on a 5-point Likert scale, ranging from *not acceptable at all* to *very acceptable*.

The behaviour types in A are a measure of the individuals' likelihood to trade-off security for productivity. This evolved to the more abstract concepts in B where the answers are now mapped to four distinct behaviour types as defined by Adams [1], to better represent the role of teams and organisational culture in individual security behaviours. Additionally, participants were asked to assign a severity score to each answer option of the behaviour type questions, as well as give a general indicator as to how acceptable to the business it would be for the participant not to finish the task described in each scenario.

(a) Attitude Types for A

1 Discount suspicions, cause no bother, passive,
2 Report suspicions but take no direct action,
3 Take direct action through official channels,
4 Take direct personal action against the threat.

(b) Behaviour Types for A

1 Prepared to perform insecure acts to maximise productivity,
2 Show a minor priority for work over security when the two conflict,
3 Passive, expects others to take the initiative to ensure security,
4 Tries to remain secure wherever possible.

(c) Maturity Levels for B

1 Is not engaged with security in any capacity,
2 Follows security policy only when forced to do so by external controls,
3 Understands that a policy exists and follows it by rote,
4 Has internalised the intent of the policy and adopts good security practises even when not specifically required to,
5 Champions security to others and challenges breaches in their environment.

(d) Behaviour Types for B

Individualists rely on themselves for solutions to problems,
Egalitarians rely on social or group solutions to problems,
Hierarchists rely on existing systems or technologies for solutions to problems,
Fatalists take a 'naive' approach to solving problems, feeling that their actions are not significant in creating outcomes.

Table 1: The dimensions by which survey responses are measured in Company A and Company B

3.2 Survey design

Figure 1 shows one of the scenarios in B (note that participants did not see the Individual-

Question: File Storage (Behaviour-type)

Concerned about the safety of his current work, Shamal decides to back up his data, some of which is confidential. As he uses his own laptop under the ‘bring your own device’ scheme, he usually stores all his work on his drive on the central server but he wants to have a second copy just in case something happens or he loses connectivity to the company network. He thought about using one of the common drives but none of the ones he regularly uses have sufficient space.

Individualist: Create a local copy on the hard drive of your BYOD laptop, it is the only machine you work on so you know it will be safe and this ensures you will always have access to it if needed.

Egalitarian: Use a common drive that you used for an old project and still have access to, as your credentials were never revoked. It has enough space although you do not know who manages it now.

Hierarchist: Use an online service, such as Dropbox, to store the data as it is more under your control.

Fatalist: Back your work up onto a USB stick – you have ordered an encrypted one but while you wait for it to arrive you use a personal stick you have to hand.

Figure 1: Scenario ‘File Storage’ (QFS) in B

ist/Egalitarian/Hierarchist/Fatalist labels, as defined in Table 1). In each organisation, surveys were crafted for participants based upon their department to improve relevance, see [4, 7] for more details. For the example scenario in Table 1, we found through the interviews that data availability was a predominant issue in the organisations, where many interview participants mentioned the use of security *workarounds* [5] to guarantee that they reached their business goals. Question design attempted to offer the participants a number of options which would all be regarded as equally appropriate, based upon the themes identified from the preceding interviews with employees. The participants were asked to rank the four options in order of their preferences. Additionally, participants were allowed to offer additional comments, which included the following example:

> *“Shamal needs to find out who manages the common drive now, and whether the company authorises use of Dropbox and personal USB sticks, before using any of those options.”*

As part of the work described here, two annotators coded the volunteered comments for the types (without reference to the alignment of responses to types already defined for each question). For example, the quotation above could be coded as a Hierarchist’s point of view, as the individual falls back to existing structures for solutions to the problem.

> *“This scenario could easily be avoided by providing sufficient space on the common drives.”*

Conversely, this statement has been coded as a Fatalist. The employee is frustrated that the natural solution to the problem is outside their reach.

4 Methodology

The methodology laid out in this section establishes three metrics to measure the quality of the survey design and its external validity retrospectively. The quality of the survey includes how engaged participants are in considering a scenario, and how relevant a scenario and its options are to their own experiences. If an organisation is committed to measuring how well its security provisions support the effective completion of business tasks towards identifying and removing frictions, decision-makers would have a natural interest in having a realistic picture of the current experiences of employees.

4.1 Appropriateness and applicability

For each of the answer options to Attitude and Maturity questions (see Tables 1a and 1c) participants were asked to specify the acceptability of that answer on a 5-point Likert scale ranging from “Not acceptable at all” to “Very acceptable”. There are of course biases present here, namely that given a participant’s type they may see some options as more acceptable than others. Indeed there is a statistically significant (at $p < 0.001$) correlation in our survey response data between the ranking of options and their associated behaviour types with Kendall’s τ of 0.62. Yet the ideal scenario design would leave the participants with four objectively equally acceptable options, and allow the participant to freely rank the option.Hence a high appropriateness score is desired.

Similarly, for each of the answer options to Behaviour type questions (see Tables 1b and 1d) participants were asked to specify the severity of each option as well as the *acceptability of failing to complete the task* for each scenario on a 5-point Likert scale. Again, the severity scores are statistically significantly (at $p < 0.001$) correlated with the ranking of the answers (Kendall’s $\tau = -0.20$), with less severe answers being ranked as more preferable. The severity scores of the different answers should be ranked equally by the participants (as questions are designed with no one 'right' answer), resulting in a low standard deviation throughout the questions. The ideal mean of the standard deviation of severity of options is 0, which would imply that all options given to the participants are perceived as equally severe.

The *acceptability of failing to complete the task* metric would ideally be identically distributed for all questions in order to allow for inter-question comparison. This is a metric that is difficult to establish through prior analysis. If participants think that for a scenario it is more acceptable for it not to be completed given the given consequences as in the scenario and its options, the participants do not fully commit to their choices of behaviour types, as in no scenario there is an option to do nothing (and in turn avoid side effects from the chosen solution).

4.2 Validation of ranked types by free-text responses

In the survey design for both A and B, participants are asked to rank four answer options according to their preferences. Participants are also invited to provide additional comments on the questions. We find that there are two common types of responses: those that further confirm a respondent's answer, or elicit suggestions / solutions that are not included in the question and the associated options. We code these according to the applicable mapping (Attitude or Behaviour) in each organisation as listed in Table 1, e.g., for each free text response the annotators have to choose from one of four options. While this is opportunistic (not all participants provided additional comments), we can validate the mappings by calculating inter-annotator agreement metrics as described in the following sections.

4.2.1 Inter-annotator agreement

	Coder A			
Coder B	T1	T2	T3	T4
T1	4	3	0	0
T2	0	36	2	15
T3	0	2	59	1
T4	0	0	1	32

Table 2: Confusion matrix for Question 1 in A between the coders' assignment of types to the free text responses

The calculation of the inter-annotator agreement between the two coders is straightforward. We first calculate a confusion matrix for each question (an example is shown in Table 2), and then calculate Krippendorff's chance corrected inter-annotator agreement metric α. Krippendorff's α ranges from -1 to $+1$, where 0 corresponds to chance agreement and $+1$ to perfect agreement. As the attitude types and maturity levels in Tables 1a and 1c are on a ranked scale, we weight the disagreement linearly. For the other two types described in Tables 1b and 1d the agreement is binary.

4.2.2 Validating the mapping

	Coding type			
Rank	T1	T2	T3	T4
1	2	6	84	3
2	5	26	22	10
3	2	19	12	34
4	2	43	6	34

Table 3: Confusion matrix for Question 1 in A between participants assigned ranks to the potential answers and the types assigned to the participants by the coders based on the coding of the free text responses

We validate the mapping of the survey answer options (an example is shown in Figure 1) by treating the participants as another annotator and calculating the inter-annotator metric α. However, the participants rank their options, but the coders annotate separate (but not independent) text statements. For example, a participant may provide a ranking of Type 3 $>$ Type 2 $>$ Type 4 $>$ Type 1 for a specific question, and from the coders we may see Coder X: Type 3, Coder Y: Type 2.

In this case the standard agreement table approach [15] for > 2 annotators cannot be used. Yet Krippendorff's α naturally extends to non-square weight matrices. In our case, this leads to a confusion matrix such as in Table 3. Here we tabulate the frequency that a coder has annotated a statement with a specific type with the rank that the participant gave that type. Perfect validation would therefore imply that all types chosen by the coders have rank 4; i.e. the only non-zero entries are in the bottom row of the confusion matrix. Given this matrix we can execute the calculation of Krippendorff's α with a weights matrix that treats numbers in the bottom row as perfect agreement, and linearly increases disagreement for lower ranked options.

4.2.3 Estimating confidence in α

In order to calculate the confidence in the calculated value of α we rely on α's standard deviation. As an analytic expression is not available, we bootstrap the calculation of α. In the following sections the confidence intervals are calculated using 1000-fold bootstrapping.

5 Results

In this section we present the application of the metrics defined in Section 4 to the datasets described in Section 3. There are four tables to consider in this section; Tables 4 and 5 for the analysis of the secondary coding of the free text responses, Table 6 for the analysis of appropriateness scores on attitude/maturity questions, and Table 7 for the analysis of the severity metrics on behaviour type questions.

5.1 Analysis of clustering

Table 3 is an example confusion matrix calculated based on the methodology presented in Section 4.2.2. The column headers list the four possible types assigned to the free-text responses by the coders. If a free-text response by the coders was judged to be type 1, but the participant ranked the answer corresponding to type 1 as rank 2, this would increment the number in row 1, column 2. Perfect agreement would be represented by the type assigned through coding of the free-text responses always being ranked highest (rank 4) by the participants. This would be a confusion matrix of non-zero entries in the bottom row only.

The strong disagreement between the coders and the assigned rank in question 1 can be identified by the strong mismatch in type 3: of the 124 statements assigned to type 3 by the coders, 84 were ranked least likely (rank 1) by the participants. This implies that the answer option assigned to type 3 *"Request that those with access share their (main log-in) account details and passwords with those without to allow them access to the information"*) does not match behaviour type 3 (as defined in Table 1) (*"Passive, expects others to take the initiative to ensure security"*).

Interestingly, this disagreement is not reflected in the coding of the free-text responses themselves. Table 2 shows the confusion matrix for Question 1 for the two coders. There is virtually no disagreement for types 1, 2 and 3; but some disagreement for type 4, where 15 statements assigned to type 4 by coder A were considered to be type 2 by coder B. The internal validity for the coding of free text responses for Question 1a can be accepted based on Krippendorff's α of 0.77 ± 0.00 as shown in Table 4, but we are unable to validate the mapping of answer options to types.

Tables 4 and 5 list the number of free-text responses coded and Krippendorff's α for both the validation of the mapping as well as the coders agreement.

Question	#	Mapping α	Coder's α
Q4	40	0.21 ± 0.02	0.29 ± 0.02
Q5	34	0.35 ± 0.02	0.02 ± 0.03
Q6	2	-0.33 ± 0.76	0.00 ± 0.67
Q8	29	0.30 ± 0.04	0.94 ± 0.02
Q10	37	0.23 ± 0.03	0.73 ± 0.02
Q1	155	-0.03 ± 0.01	0.77 ± 0.00
Q2	137	0.43 ± 0.01	0.91 ± 0.00
Q3	12	0.33 ± 0.08	0.38 ± 0.10
Q7	25	0.24 ± 0.03	0.13 ± 0.04
Q9	45	0.13 ± 0.02	0.76 ± 0.02

Table 4: Krippendorff's α measures for compA with 95% confidence intervals.

Question	#	Mapping α	Coder's α
QID	33	0.27 ± 0.02	0.85 ± 0.02
QCDP	53	0.31 ± 0.01	0.38 ± 0.02
QT	22	0.24 ± 0.04	0.34 ± 0.05
QSD	27	0.53 ± 0.02	0.47 ± 0.04
QRM	12	0.27 ± 0.07	0.42 ± 0.07
QVPN	23	-0.09 ± 0.03	0.37 ± 0.04
QFS	18	0.19 ± 0.05	0.46 ± 0.05
QCC	7	0.38 ± 0.13	0.75 ± 0.21

Table 5: Krippendorff's α measures for compB with 95% confidence intervals.

5.1.1 Suitable values for α

Before discussing this data further we must delineate the boundaries for which we consider Krippendorff's α to be reliable. From a statistical perspective we can conduct a t-test where the null hypothesis is $\alpha = 0$, i.e. the data is equivalent to chance. This t-test is represented in our tables through the use of 95% confidence intervals. Indeed all rows that are statistically significant at the 95% confidence interval are also significant at the 99% confidence interval. However the literature [15] is clear that primary data is only sufficiently reliable for further analysis at $\alpha > 0.667$.

It is clear that most of the coder's agreement values in A satisfy this criteria. There are a number of exceptions: *Q5*, *Q3* and *Q7*. The inter-coder agreement is not as strong in B, where only *QID* satisfies this criteria. When focusing on the validation of the mapping/clustering however, none of the scenarios satisfies this stringent criteria.

Considering the difficulty the coders have to establish agreement on the free-text responses in B, the low mapping α values are not surprising: the coding is a difficult task (given the brevity of comments and potential lack of contextual information). Yet rather than discarding the results at this stage, it may be more important to identify the scenarios which are indistinguishable from random data: scenarios *Q1* in A and *QVPN* in B. Apart from these two scenarios, our data allows the focus of further investigations and policy decisions to be guided by data with known uncertainty.

5.2 Appropriateness

		Mean		1st choice	
Question	#	mean	std	mean	std
Company A					
Q4	374	0.626	0.120	0.923	0.195
Q5	820	0.570	0.110	0.925	0.161
Q6	137	0.427	0.138	0.821	0.321
Q8	364	0.529	0.085	0.983	0.084
Q10	903	0.483	0.082	0.917	0.185
Company B					
QID	152	0.488	0.122	0.778	0.316
QCDP	456	0.508	0.108	0.893	0.220
QT	164	0.499	0.095	0.873	0.252
QSD	292	0.546	0.118	0.939	0.181

Table 6: Appropriateness scores for each attitude question in compA, the higher, the more appropriate. As each answer option is assigned an appropriateness score, the mean represents the mean appropriateness score of all answer options irrespective of that answer's ranking. The *1st choice* only considers the appropriateness assigned by the participants to their top choice.

Table 6 shows the appropriateness scores the participants have given the answer options for specific questions. The scores vary from 0 (not appropriate) to 1 (very appropriate). The mean appropriateness score is more varied in A than in B, although it is close to 0.5 for all questions, indicating that the average answer option is balanced. This is desirable as it offers participants the option to swing to both extremes as necessary. The appropriateness score given by participants to their highest ranked choice is very high, confirming the participant's stance that they view their preferred choice as most appropriate.

5.3 Severity

Table 7 compares the distribution of severity scores and *acceptability of failing the task* scores across the different scenarios and organisations. There are a number of variations: Scenarios in A are considered less acceptable to be left undone, however the standard variations of the severity scores across the different scenario options are higher. According to Table 7 the scenarios in B are therefore believed by the participants to be more applicable to their environment (particularly *QRM* and *QCC*), but the answer options are more balanced in severity in A, implying that options represent potential solutions that may be seen in everyday work in A compared to the more contrived answer options in B.

		Failing		Std of Severity	
Question	#	mean	std	mean	std
Company A					
Q1	903	0.281	0.307	0.270	0.128
Q2	893	0.270	0.296	0.239	0.123
Q3	137	0.394	0.340	0.271	0.122
Q7	291	0.458	0.393	0.296	0.144
Q9	374	0.668	0.449	0.274	0.123
Company B					
QRM	152	0.196	0.312	0.377	0.101
QVPN	152	0.439	0.370	0.323	0.120
QFS	164	0.430	0.318	0.297	0.114
QCC	292	0.240	0.410	0.182	0.163

Table 7: Acceptability of failing to complete the task (higher more acceptable) and standard deviation of severity of options for behaviour type scenarios in compA.

6 Discussion

This research supports a process of continuous improvement to organisational security, by providing measures for (i) typical workaround to regular frictions with security in the workplace (by analysing the perceived suitability of solutions derived from interviews), and (ii) how the interactions employees have with security apparatus can be designed to minimise the demand on their 'compliance budget' [6]. Employee's willingness to expend effort for the security of not only themselves but those around them can be explored by articulating embodied security cultures which may arise in any number of situations in the workplace where security controls can be applied. Both the survey results and the free-text responses can inform targeted interventions as part of incremental improvement, an approach advocated by Renaud and Goucher [28]. Unfortunately in striving for internal validity for security behaviour constructs it is easy to overlook the need to establish the applicability of the results to the real world, that is, to measure the quality of *engagement* with employees (where tensions can arise with local demands on effort and capacity). The related works discussed in Section 2 demonstrate this well.

Security managers ought to identify the non-divisible security behaviours in their own organisations, and equally deploy information security surveys that shine light on previously unseen *workaround* or *compromise* behaviours by engaging employees. To do this, available options (and ideally, additional feedback from users) must point to clear responses to security-related challenges that employees see as acceptable given the pressures they perceive in a particular situation. Where respondents imply confusion about what is being asked of them in a scenario-based survey question – or indeed, see two or more behaviours as one and the same – this implies that more can be done to clearly separate candidate behaviours. In turn, this can be achieved if security managers act to grow their understanding of how security manifests for employees who have other competing demands for their attention (see also Ashenden and Lawrence [2], Herley [16] and Parkin et al. [25]).

Our proposed survey methodology and validity measures address these challenges. This is achieved both internally (by way of inter-annotator agreement), and externally (by way of appropriateness and severity scores). We are able to highlight strengths and shortcomings in the survey design which not only inform the design of realistic scenarios by researchers, but also inform the investment in security by policy managers when designing interventions. Organisational environments are complex, and researchers cannot assume that they have a full understanding of security behaviours prior to deploying a survey. This research helps to identify these known unknowns. Security practitioners considering potential investments may do well to understand the quality of the data they base their decisions upon [18].

7 Conclusions

In this paper we have described a methodology for post-hoc assessment of the quality of situated security behaviour survey designs. We utilise free-text responses and reflective metrics to measure the surveys' external validity. We have demonstrated this approach on two surveys in two large organisations, drawing on 711 free text responses and over 7000 reflective scores in the process. This has allowed us to quantify the evolution of our scenario-based surveys through clearly-defined and repeatable metrics, and partially validate the mapping from survey responses to constructs. This knowledge will allow security managers to tailor future improvements to their organisation's security policy and behavioural interventions more accurately to the local working environment, relative to the demonstrable strengths and weaknesses of the survey design.

We strongly advice researchers designing surveys in future to include open questions that can be answered by the participant without being biased. Survey designers should not assume they know everything about the responders even if the survey is grounded in qualitative research, and continually look for ways to involve respondents to gather more context-specific information, such as by including the reflective questions described in this research.

Dataset

The participant's assigned categories and the two sets of coding for both organisations as well as the analysis code can be found at DOI `10.14324/000.ds.10038283`.

Acknowledgements

We would like to thank the LASER program committee and workshop attendees for their feedback, as well as the participating organisations. The authors are supported in part by UK EPSRC grants, no. EP/G037264/1.

References

1. Adams, J. Risk and morality: three framing devices. *Risk and morality*, 2003: 87–106.
2. Ashenden, D., and Lawrence, D. Can we sell security like soap?: a new approach to behaviour change. *Proceedings of the 2013 workshop on New security paradigms workshop*. 2013, 87–94.
3. Ashenden, D., and Lawrence, D. Security Dialogues: Building Better Relationships between Security and Business. *IEEE Security & Privacy*, 14(3), 2016: 82–87.
4. Beautement, A., Becker, I., Parkin, S., Krol, K., and Sasse, M. A. Productive Security: A Scalable Methodology for Analysing Employee Security Behaviours. 2016.
5. Beautement, A., and Sasse, A. The economics of user effort in information security. *Computer Fraud & Security*, 2009(10), 2009: 8–12.
6. Beautement, A., Sasse, M. A., and Wonham, M. The compliance budget: managing security behaviour in organisations *Proc. NSPW '08*, 47–58.
7. Becker, I., Parkin, S., and Sasse, M. A. Finding Security Champions in Blends of Organisational Culture. *Proc. USEC*. 2017, 11.
8. Beris, O., Beautement, A., and Sasse, M. A. Employee Rule Breakers, Excuse Makers and Security Champions: Mapping the risk perceptions and emotions that drive security behaviors. 2015.

9. Blythe, J. Information security in the workplace: A mixed-methods approach to understanding and improving security behaviours. PhD thesis. Northumbria University. 2015.
10. Blythe, J. M., Coventry, L. M., and Little, L. Unpacking Security Policy Compliance: The Motivators and Barriers of Employees' Security Behaviors. *SOUPS*. 2015, 103–122.
11. Bulgurcu, B., Cavusoglu, H., and Benbasat, I. Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. *MIS Quarterly: Management Information Systems*, 34, 2010: 523–548.
12. Chang, H. H., and Chuang, S.-S. Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. *Information & Management*, 48(1), 2011: 9–18.
13. Cheng, L., Li, Y., Li, W., Holm, E., and Zhai, Q. Understanding the violation of IS security policy in organizations: An integrated model based on social control and deterrence theory. *Computers & Security*, 39, 2013: 447–459.
14. Egelman, S., Harbach, M., and Peer, E. Behavior ever follows intention?: A validation of the security behavior intentions scale (SeBIS). *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, 5257–5261.
15. Gwet, K. L. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.
16. Herley, C. More Is Not the Answer. *IEEE Security Privacy*, 12(1), 2014: 14–19.
17. Huang, C.-C. Knowledge sharing and group cohesiveness on performance: An empirical study of technology R&D teams in Taiwan. *Technovation*, 29(11), 2009: 786–797.
18. Hubbard, D. W. How to measure anything: finding the value of intangibles in business. John Wiley & Sons, 2014.
19. Ifinedo, P. Information systems security policy compliance: An empirical study of the effects of socialisation, influence, and cognition. *Information & Management*, 51(1), 2014: 69–79.
20. Karlsson, F., Karlsson, M., and Åström, J. Measuring employees' compliance-the importance of value pluralism. *Information & Computer Security*, 25(3), 2017:
21. Kirlappos, I., Parkin, S., and Sasse, M. A. Shadow security as a tool for the learning organization. *ACM Computers and Society*, 45(1), 2015: 29–37.
22. Kirlappos, I., Parkin, S., and Sasse, M. A. Learning from "Shadow Security": Why understanding non-compliance provides the basis for effective security. *Proc. USEC*. 2014.
23. Krol, K., Spring, J. M., Parkin, S., and Sasse, M. A. Towards robust experimental design for user studies in security and privacy. *Learning from Authoritative Security Experiment Results (LASER) Workshop*. 2016.
24. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015: aac4716.
25. Parkin, S., Krol, K., Becker, I., and Sasse, M. A. Applying Cognitive Control Modes to Identify Security Fatigue Hotspots. *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. 2016.
26. Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A., and Zwaans, T. The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies. *Computers & Security*, 66, 2017: 40–51.
27. Rajivan, P., Moriano, P., Kelley, T., and Camp, L. J. Factors in an end user security expertise instrument. *Information & Computer Security*, 25(2), 2017: 190–205.
28. Renaud, K., and Goucher, W. The Curious Incidence of Security Breaches by Knowledgeable Employees and the Pivotal Role of Security Culture. *Human Aspects of Information Security, Privacy, and Trust*. 2014, 361–372.
29. Siponen, M., and Vance, A. Guidelines for improving the contextual relevance of field surveys: the case of information security policy violations. *Eur J Inf Syst*, 23(3), 2014: 289–305.
30. Sohrabi Safa, N., Von Solms, R., and Furnell, S. Information security policy compliance model in organizations. *Computers & Security*, 56, 2016: 70–82.
31. Tamjidyamcholo, A., Bin Baba, M. S., Shuib, N. L. M., and Rohani, V. A. Evaluation model for knowledge sharing in information security professional virtual community. *Computers & Security*, 43, 2014: 19–34.
32. Wash, R., Rader, E., and Fennell, C. Can People Self-Report Security Accurately?: Agreement Between Self-Report and Behavioral Measures. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, 2228–2232.
33. Witherspoon, C. L., Bergner, J., Cockrell, C., and Stone, D. N. Antecedents of organizational knowledge sharing: a meta-analysis and critique. *J of Knowledge Management*, 17(2), 2013: 250–277.