Mental Models of Domain Names and URLs

Richard Roberts, Daniela Lulli, Abolee Raut, Kelsey R. Fulton, and Dave Levin
University of Maryland

Abstract

Many domain impersonation strategies have been observed in the wild. Phishers utilize typos, uncommon top-level domains, Unicode characters, subdomains, and extra tokens to make domains that are visually similar to high-value target domains like banks, email providers, or shopping websites. Quantitative research tells us that many users fall for these impersonation attempts, but it is not clear *why* users have a difficult time comprehending how domains and URLs are structured. Even engaged users are often unable to identify examples of domain impersonation when primed to look for them. In this poster and abstract, we describe the design of a semi-structured interview to obtain qualitative data on the mental models that users have of URL structure, domain name ownership, and web browsers. We describe the outcome of a pilot of our interview study on five subjects, and discuss our plans for running the study at full scale.

1 Introduction & Background

The web's public key infrastructure (PKI) is tasked with ensuring that users can determine with whom they are communicating online. After a series of successful cryptographic checks, browsers commonly show a lock icon to indicate that the connection is "secure." However, it is ultimately up to the users to evaluate whether the secure connection is to whom they *intended* to connect. As a result, the web's PKI tacitly depends on users' ability to identify if the domain name they are connecting to is the domain they intended to connect to.

Unfortunately, this makes the domain name and URL potent tools in a phisher's arsenal. A phisher can take advantage of gaps or misunderstandings in a user's mental model of URLs to convince them that they are on a target website, when they are actually on an attacker-controlled impersonation. For instance, irs.gov-login.pw looks to some users as if it is a website for the US Internal Revenue Service, but in reality "irs" is merely a subdomain of the true website, gov-login.pw.

The PKI does not help root out such impersonations, as it is strictly tasked with detecting exact matching forgeries, not with identifying potential *misperceptions* of names. As a result, a phisher can easily obtain a cryptographic certificate because, to the PKI, the attacker is exactly who it claims to be (irs.gov-login.pw)—which is not always what the attacker *appears* to users to be [11]. In other words, phishers can trick users while remaining fully compliant with PKI protocols.

Many forms of domain impersonation have been defined and measured in prior literature, including typosquatting [1,10,12,14] (faceboook.com), combosquatting [7] (facebookaccount.com), homographs [2–5,8] (faceb00k.com), and target embedding [11] (facebook.com-login.pw). As we transition to an HTTPS-everywhere web, attackers are increasingly obtaining TLS certificates for impersonating domains [11]. Many users consult URLs in the address bar when evaluating the website they are on [6,9,13], and those that do not can be educated to consult the URL and lock icon as a means to protect themselves online [9,16]. While UI changes have been explored to draw users' attention to URLs and the address bar [13,15,16], even engaged users can still have difficulty detecting fake URLs [11].

Our work complements these works by establishing *why* users fall for impersonating domains. We hypothesize that users' mental models of URLs and domains diverge from how URLs and domains are used in practice, and attackers take advantage of these gaps to fool users. We have designed a semi-structured interview protocol to obtain qualitative data on how users think URLs work, what the different pieces of a URL mean, and how URLs interact with browser security indicators like the lock icon. From these interviews, we can create a taxonomy of mental models that users have, and identify problematic models which lead to dangerous misconceptions. We have piloted our interview script on five subjects to test the efficacy of our questions, and gain preliminary insight into users' mental models of domains.

2 Methodology

We ran a pilot of our interview study to ensure that our questions were appropriately surfacing subjects' beliefs, and to determine if our interview design was appropriately covering all the topics we intended. The only difference between how we ran this pilot and how we will run our full study is that the pilot is on a smaller scale with only five subjects.

2.1 Recruitment

Advertisement Participants were recruited with an advertisement on Craigslist. Our interviews were initially designed to be conducted in-person[1], so we recruited subjects in the Washington D.C. area. The advertisement directed potential participants to take a pre-screening demographic survey hosted by Qualtrics. The number of responses to our pre-screening survey was capped at 50, as we were looking to recruit 5 subjects to participate in our pilot interviews. 50 subjects allowed us to recruit a diverse cohort to interview, as well as have alternate options if anyone who agreed to participate backed out unexpectedly.

ID	Gender	Age	Educ.	Ethn.
P1	F	18–29	B	W
P2	M	50–59	B	B
P3	F	40–49	B	AHP
P4	F	40–49	B	HL
P5	M	30–39	MD	AHP

Gender: F - Female, M - Male
Education: B - Bachelor's degree, MD - Master's Degree
Ethnicity: AHP - Asian/Native Hawaiian/Pacific Islander, B - Black/African American, HL - Hispanic/Latino, W - White/Caucasian

Table 1: Demographic information for the 5 subjects who participated in the pilot of our interview study.

Participant Selection We selected a pool of 5 participants from the 50 who responded to our Craigslist pre-screening survey. Those participants were invited to participate in an hour-long interview study. Study participants were given a $30 digital gift card to Amazon.com as compensation for their time. Table 1 shows the demographic makeup of the subjects selected to participate in our interviews. While we aimed to recruit a diverse sample along as many demographics as possible for our pilot, all of our subjects had at least a Bachelor's degree as their highest level of education completed. We will ensure we recruit a more diverse population with respect to education in the full-scale study.

2.2 Interview Protocol

Our interview was designed to take approximately one hour, over a Zoom video call. At certain parts of the interview, subjects were asked to share their screen with the interviewer; at other times, the interviewer shared their screen with the subject. The interviews were recorded, but we only stored the audio of the interviews for analysis. We designed a semi-structured interview script to allow researchers to ask participants follow-up questions on ideas related to, but not explicitly outlined in, the prepared interview questions. Section 3 details the design of our interview protocol's questions.

[1]Due to COVID-19, the interviews were redesigned to take place online.

2.3 Ethics

This pilot study has been approved by the University of Maryland Institutional Review Board. Informed consent was collected from all subjects who responded to our screening survey. For subjects selected to participate in the interviews, informed consent was again obtained prior to the interview.

3 Interview Questions & Preliminary Results

In this section, we describe the 9 high-level blocks that make up our interview protocol, including questions we asked subjects and how we anticipate those questions will help us understand their mental models around URLs and domain names. Our pilot lacks the rigorous qualitative analysis that we plan to use in our full study (see Section 4 for more details), but we present preliminary findings below each block. These insights will be used to tool our questions prior to running the full study, and hint at some of the themes we expect to see in common mental models.

3.1 Navigation Tasks

Browsers and the PKI writ large expect users to check the address bar to verify the identities of the websites they visit. To understand how users engage with domains and the address bar, we opened the interview by asking participants to share their screens with the interviewer and perform three tasks: (1) navigate to youtube.com, (2) find a list of the top 10 most expensive movies of all time, and (3) log into an email account they own [2]. During and after the tasks, subjects were asked questions about how they determined what web page they were currently on, how they chose a reliable source when searching for information, and whether they took any extra precautions when entering their password. Providing the users with tasks to complete during the interview grounded their answers in a concrete experience, rather than relying on recall of their habits from memory.

Pilot Results Some users refer to the address bar as a "search bar" and associate it with conducting web searches rather than inputting a destination domain. When asked "how do you know what website you are currently on?" all subjects referred to page content (especially website logos) but only some explicitly said the domain name matched expectation. All 4 subjects who typed in the password to their email account (one had their password auto-loaded by their browser) said they felt comfortable entering it as the log-in page looked like how they "expected it to." Only one of those four said they verified the domain and looked for the browser's lock icon.

[2]Subjects were instructed to use their browser's private browsing mode to ensure they were not already logged into any accounts.

3.2 Security Hygiene

This section was designed to understand what steps subjects take to protect themselves online, how they learned those behaviors, and how confident they are in their ability to protect themselves. We also ask subjects if there are any protection methods they have heard about but do not use themselves, and why. Finally we ask subjects if their web browsing behavior differs between mobile devices and laptop/desktop devices, or between personal and work devices.

Pilot Results All subjects were moderately to very confident in their ability to stay safe online, despite citing very different and non-overlapping security behaviors (antivirus software, common sense, firewalls, etc.). In the full study, we will probe subjects for thoughts on specific techniques, e.g. "what does your firewall do to protect you?" The two youngest subjects (P1 and P5) mentioned that "growing up with computers" improved their confidence, and subjects P1 and P4 said their confidence came from the fact that they have never had computer security problems in the past.

3.3 Domain Presentation and Registration

Next we ask subjects about how domains/URLs are used. We ask subjects to describe the purpose of a browser's address bar, and what happens if they enter a URL into the address bar. Here we also ask questions about domain registration: how do companies make or obtain their URL, what does that process look like, and what would happen if someone attempted to register a domain that another party already owned? This section closes by asking if there is anything prohibited from being included in a domain name (words, symbols, or other characters), and what would happen if someone attempted to break any conventions they believe exist.

Pilot Results Most subjects believed that a domain name somehow indicates what server your computer should connect two, but one said they believed that entering domain names into the address bar was more similar to conducting a web search. All subjects believed in an entity that functions as a domain registrar that sells domains, though some could not articulate who this entity was. Subjects generally felt that most things were allowed to be present in a domain name (except possibly some symbols), and that nobody can purchase a domain currently owned by somebody else. P4 was uncertain if you could own a domain that included hateful content, and noted a possible conflict with "freedom of speech."

3.4 Domain Comprehension

In each of the next four sections, the interview shares their screen with the subject and shows them a series of URLs. In the first section, the subjects are shown 9 URLs with varying pieces missing, represented by a blank line. They are asked what purpose the piece represented by the blank line serves, and provide some examples of what they would expect to appear there. We asked subjects about the TLD, e2LD[3], subdomain, path, and query parameters. We show them a URL with both the path and a subdomain represented by blank lines, and ask them what the difference between them is (if any). We also ask about the functional difference between two different subdomains. Finally ask them about domains that look like "google.com-____" and "google.com.____", two patterns one may not expect to encounter in practice but are commonly used in domain impersonation attacks.

Pilot Results Subjects viewed the TLD as a descriptor for the "type" of entity that owned the website: "company" for .com and "organization" (sometimes non-profit) for .org. The users showed confusion in describing the difference between a subdomain and the URL path; subjects felt they both served the same purpose of describing what specific page you were accessing on a website. P2 claimed to have never seen a subdomain before. Subjects were generally familiar with query parameters, describing them as either "random codes" related to encryption, or somehow tied to the results of a search query. When shown "google.com-____", subjects felt this could be "another internal page" (P1), "something international, or maybe phishing" (P3), "a search but for something you don't want" (P4), or functionally equivalent to the slash in "google.com/" (P2).

3.5 Brand Identification

Subjects were asked to identify the name of the fictitious company that owned each of the next 15 URLs, such as "hyglyph.org" or "login.zestpond.com". Typically, the company name would be the same as the domain's e2LD. In many impersonation attacks, an attacker will use a domain designed to trick the user into believing that the e2LD is somewhere else: "crumptury.com-login.secure". If a subject accepts that "Crumptury" owns this domain without hesitation or confusion, then their mental model of domain structure may conflict with how domains work in practice.

Pilot Results Subjects generally showed confidence identifying the company name even in the mock impersonating examples. Yet, all five subjects identified "crumptury" in the example above without hesitation, some even saying the page was likely a secure page for logging into an account. P3 and P4 said they look for the word after "www" or before "com" to identify a URL's owner; this is problematic because attackers can mimic those tokens in almost any part of a URL.

3.6 Domain Component Comparison

We showed subjects 6 pairs of URLs, and asked them to draw lines between the corresponding parts of each URL

[3]The e2LD, or "effective second-level domain" is the token preceding a domain's TLD, such as "google" in "google.co.uk".

pair. For the example pair "google.com" and "yahoo.com", we would expect subjects to draw a line from "com" to "com" (the TLDs) and from "google" to "yahoo" (the e2LDs). Of note were questions asking if the subdomains "www" and "login" represent the same functional unit in a domain, whether they consider "login" as a subdomain equivalent to "login" appearing in a URL path, and how subjects choose to draw lines when presented with a benign domain ("google.com/login.pw") and a target embedding domain ("google.com-login.pw").

Pilot Results This section may require redesign to stress that we are asking subjects to pair blocks based on how they technically work and not whether they semantically represent the same thing. All five subjects drew every line connecting pieces of the target embedding domains, seeing them as identical to one another.

3.7 Free-Form Domain Responses

Finally, we showed subjects 14 domains and asked them to share any thoughts they had for each domain: whether they recognize it, what they might expect to see when visiting it, and if anything stands out as interesting or odd. Included in the domains were websites with different TLDs than they are commonly associated with, typosquatting domains, combosquatting domains, target embedding domains, a homograph impersonating domain and its Punycode counterpart, alongside other atypical domains. We use each domain to further our understanding of the subject's mental model of URLs by observing how confident the subject is in their evaluation, and whether their evaluation is compatible with a correct technical understanding of URLs.

Pilot Results All participants caught the typosquatting domain "faceboook.com" and commented that it was likely not safe. All subjects were tricked by other impersonating domains, but expressed varying levels of confidence. Also, some subjects expressed hesitation or confusion at domains that were unorthodox but functionally benign ("results-www.wikipedia.org").

3.8 Lock Icon & Impersonation

In the last section, we asked subjects questions about the browser's lock icon, and impersonation. We want to determine how subjects interact with the lock icon, and whether they understand that the lock indicates session security without necessarily establishing the nebulous "safety" of the underlying web page. To build up subjects' mental models about website impersonation, we ask subjects to place themselves in the shoes of a hacker, and describe the process, resource-intensiveness, and possible harm of impersonating a web page. Finally, we ask the subject if it is possible for a malicious, impersonating URL to appear next to a lock icon in their browser, and why they feel that way.

Pilot Results Subjects had very different understandings of the purpose of the lock icon. P1 felt that it indicates a page is secure, but only sees it on a small number of specific websites. P2 and P4 implied that the lock should not appear next to nefarious domains, but nefarious parties may mimic the lock or have it appear temporarily "before they are caught." P3 expressed privacy consciousness throughout the interview, and believed the lock icon was related to the type of content a website collects from its visitors; a website without a lock icon "would not be allowed" to collect personal data. Only P5 correctly identified that the lock icon communicates session security, but they still claimed that it was not possible for the lock to appear on a website that was maliciously pretending to be another website.

3.9 Conclusion

The interview concluded by asking participants if they had any additional ideas on what would help them stay safe online, and if they had any other thoughts or questions related to the topics discussed that were not expressed in a previous answer.

4 Plans for Full-Scale Study

Recruitment Recruitment and advertising for the full study will remain identical to recruitment for our pilot. We intend to interview subjects until we reach saturation in responses (that is, new themes stop emerging with subsequent interviews). We cannot definitively say how many subjects we will ultimately interview, but we estimate interviewing 20-25 subjects for the full study. As mentioned in Section 2.1, we will ensure we recruit a more diverse population in the full study with respect to education than we achieved in our pilot.

Analysis After interviewing all participants in the full study, we will transcribe the audio interviews into text. From there, we will have two research members analyze the data with iterative open coding. The coders will incrementally develop a codebook as they code the transcripts. We will establish coding agreement by calculating Krippendorff's α. Once agreement is reached, we will establish higher level mental models from the low-level codes using iterative axial coding.

Interview Protocol Changes Overall we feel that our interview protocol does a sufficient job at exposing subjects' beliefs about URLs. In some cases (see Section 3) we plan to change the phrasing of questions, and the specific URLs shown to subjects, for clarity and completeness. While we do not anticipate major changes to our interview protocol, we welcome the SOUPS community's feedback and recommendations for additional questions or topics to explore.

References

[1] Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *Network and Distributed System Security Symposium (NDSS)*, 2015.

[2] Yahia Elsayed and Ahmed Shosha. Large Scale Detection of IDN Domain Name Masquerading. In *APWG Symposium on Electronic Crime Research (eCrime)*, 2018.

[3] Anthony Y. Fu, Xiaotie Deng, Liu Wenyin, and Greg Little. The methodology and an application to fight against unicode attacks. In *Symposium on Usable Privacy and Security (SOUPS)*, 2005.

[4] Evgeniy Gabrilovich and Alex Gontmakher. The homograph attack. *Communications of the ACM*, 45(2), 2002.

[5] Tobias Holgers, David E. Watson, and Steven D. Gribble. Cutting through the confusion: A measurement study of homograph attacks. In *USENIX Annual Technical Conference*, 2006.

[6] Markus Jakobsson, Alex Tsow, Ankur Shah, Eli Blevis, and Youn-Kyung Lim. What Instills Trust? A Qualitative Study of Phishing. In *Usable Security (USEC)*, 2007.

[7] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, Yizheng Chen, Roza Romero-Gómez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. Hiding in plain sight: A longitudinal study of combosquatting abuse. In *ACM Conference on Computer and Communications Security (CCS)*, 2017.

[8] Chris Larsen. Bad guys using internationalized domain names (IDNs). https://www.symantec.com/connect/blogs/bad-guys-using-internationalized-domain-names-idns, 2009.

[9] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does Domain Highlighting Help People Identify Phishing Sites? In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2011.

[10] Tyler Moore and Benjamin Edelman. Measuring the perpetrators and funders of typosquatting. In *Financial Cryptography (FC)*, 2010.

[11] Richard Roberts, Yaelle Goldschlag, Rachel Walter, Taejoong Chung, Alan Mislove, and Dave Levin. You Are Who You Appear to Be: A Longitudinal Study of Domain Impersonation in TLS Certificates. In *ACM Conference on Computer and Communications Security (CCS)*, 2019.

[12] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. The long "taile" of typosquatting domain names. In *USENIX Security Symposium*, 2014.

[13] Christopher Thompson, Martin Shelton, Emily Stark, Maximilian Walker, Emily Schechter, and Adrienne Porter Felt. The web's identity crisis: Understanding the effectiveness of website identity indicators. In *USENIX Security Symposium*, 2019.

[14] Yi-Min Wang, Doug Beck, Jeffrey Wang, Chad Verbowski, and Brad Daniels. Strider typo-patrol: Discovery and analysis of systematic typo-squatting. In *USENIX Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, 2006.

[15] Min Wu, Robert C Miller, and Simson L Garfinkel. Do security toolbars actually prevent phishing attacks? In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2006.

[16] Aiping Xiong, Robert W Proctor, Weining Yang, and Ninghui Li. Is Domain Highlighting Actually Helpful in Identifying Phishing Web Pages? *Human Factors*, 59(4), 2017.