Eyes on URLs: Relating Visual Behavior to Safety Decisions

Niveta Ramkumar
nr1063@wildcats.unh.edu
Univ. of New Hampshire
Durham, NH, USA

Vijay Kothari
vijayk@cs.dartmouth.edu
Dartmouth College
Hanover, NH, USA

Caitlin Mills
caitlin.mills@unh.edu
Univ. of New Hampshire
Durham, NH, USA

Ross Koppel
rkoppel@sas.upenn.edu
Univ. of Pennsylvania
Philadelphia, PA, USA

Jim Blythe
blythe@isi.edu
Information Sciences Inst.
Marina del Rey, CA, USA

Sean Smith
sws@cs.dartmouth.edu
Dartmouth College
Hanover, NH, USA

Andrew L. Kun
andrew.kun@unh.edu
Univ. of New Hampshire
Durham, NH, USA

ABSTRACT

Individual and organizational computer security rests on how people interpret and use the security information they are presented. One challenge is determining whether a given URL is safe or not. This paper explores the visual behaviors that users employ to gauge URL safety. We conducted a user study on 20 participants wherein participants classified URLs as safe or unsafe while wearing an eye tracker that recorded eye gaze (where they look) and pupil dilation (a proxy for cognitive effort). Among other things, our findings suggest that: users have a cap on the amount of cognitive resources they are willing to expend on vetting a URL; they tend to believe that the presence of www in the domain name indicates that the URL is safe; and they do not carefully parse the URL beyond what they perceive as the domain name.

CCS CONCEPTS
• Security and privacy → Social engineering attacks; Social aspects of security and privacy: Usability in security and privacy; • Human-centered computing → User studies; • Social and professional topics → Spoofing attacks.

KEYWORDS
usable security, phishing, user study, eye tracking, cognitive psychology, reading

ACM Reference Format:

1 INTRODUCTION

As people surf the web, check their email, and do other computer-related tasks, they interact with web addresses or Uniform Resource Locators (URLs) [Wikipedia contributors 2019c]. Unfortunately, URLs do not only serve legitimate content; bad actors may use URLs under their control to conduct attacks, e.g., to serve malware or steal credentials by masquerading as a legitimate service. Thus, users must be vigilant. Trusting an unsafe URL could present a security threat to the individual or their organization. Yet users don’t want to ignore safe URLs either. This problem is compounded by user misperceptions of URL syntax, the sheer time required to vet URLs, and some practices of legitimate services (e.g., use of URL redirectors). These factors make it very difficult for users to vet URLs. Consequently, many attacks rely on the victim unwittingly clicking on a malicious URL.

From a security standpoint, it is critical to safeguard users from malicious websites. And so, numerous solutions have been developed. Some companies specialize in security training for users (e.g., [KnowBe4 2019; Proofpoint 2019b]). Others focus on limiting user exposure to unsafe URLs: Products and services like Microsoft Office 365 APT Safelinks [Microsoft 2019] and Proofpoint URLDefense [Proofpoint 2019a] check for malicious content served by URLs before allowing users to visit them. Some browsers similarly warn the user when they detect unsafe URLs (e.g., [Mozilla 2019]). There is also abundant research on why users fall for URL-based phishing attacks (e.g., [Dhamija et al. 2006; Hong et al. 2013]), on training techniques (e.g., [Kumaraguru et al. 2009; Miyamoto et al. 2014; Wen et al. 2019]), and on defenses (e.g., [Fette et al. 2007; Maurer et al. 2011]), as well as other foci. However, to the best of our knowledge, this is the first study that solely focuses on understanding users’ visual attention as they process URLs. Studying users’ visual attention while processing URLs allows us to determine why certain attacks succeed, to measure the influence of URL characteristics on visual processing and cognition, and to determine the efficacy of countermeasures.

The work presented here serves as a first step toward developing a descriptive model of the relationship between URL characteristics and user visual behavior. We conducted a user study where users were asked to classify URLs as safe or unsafe while wearing an eye tracker. One key finding is that participants spent more time on processing URLs as URL length increased but only up to a point. Another is that participants relied more upon the authority component of URLs than any other component.
2 RELATED WORK

2.1 Eye Tracking and Reading

Eye tracking is considered to be a window into users’ cognitive states [Konig et al. 2016; Reichle et al. 2012]. It has been employed to assess cognitive load [Palinko et al. 2010; Pappusetty et al. 2017; Pomplun and Sunkara 2003; Zagermann et al. 2016], reading strategies [Beymer et al. 2008; Hyon et al. 2002; 2003; Rayner et al. 2006], and design implications [Bergstrom and Schall 2014; Goldberg et al. 2002]. We study users’ eyes as they process URLs.

Users assess the safety of a URL by reading. The amount of visual attention given while reading reflects moment-to-moment cognitive processing [Rayner 1998; Zagermann et al. 2016]. Researchers have sought to examine the relationships between reading and eye movements by using measures like fixations, saccades, regressions, and backtracks [Beymer and Russell 2005; Sibert et al. 2000]. Fixations are pauses in eye movements during which new information is acquired. Research has shown that users fixate longer while reading when “the processing load is greater” [Just and Carpenter 1980].

Reading and scanning text differs with respect to fixations and word skipping [Rayner and Fischer 1996]. When and where someone looks next while reading is influenced by the reader’s ongoing mental processing [Rayner and Fischer 1996]. Six commonly used eye-tracking measures are: fixation count, fixation count on various areas of interest (AOIs), proportion of time spent on each AOI, average fixation duration, fixation rate (fixation count/second), and gaze duration mean on each AOI [Lai et al. 2013]. We used all of these measures, as well as pupil dilation and backtrack fixation count.

2.2 Pupil Dilation and Cognitive Load

As users read and evaluate URLs, they use cognitive resources. A common measure of cognitive load is pupil dilation [Kun et al. 2013; Palinko et al. 2010; Poole and Ball 2006]. When users face challenging tasks, their pupils dilate on the order of 0.1 to 0.5 mm [Beatty 1982; Pfleging et al. 2016]. This task-evoked pupillary response (TEPR) indicates the cognitive load of the task. However, pupil dilation is also influenced by other factors like the amount of light entering the pupil (pupillary light reflex) [Palinko and Kun 2012; Pfleging et al. 2016] and one’s emotional state [Bradley et al. 2008; Stanners et al. 1979; Xu et al. 2011]. To reduce these effects, we conducted the experiment in a windowless light-controlled room.

2.3 Neutral Mood Induction

Mood can affect a person’s ability to comprehend text and their judgment [Bohn-Gettler and Rapp 2011; Forgas 1989]. Mood induction is used to understand and reduce the effect of mood [Mills et al. 2019]. Watching a film or a story is one of the most effective mood induction techniques [Westermann et al. 1996]. To reduce the effect of mood and improve replicability, we had participants watch a video chosen to induce a neutral mood.

2.4 URL Security and Phishing

Phishing is the act of masquerading as a legitimate entity to gather sensitive user information [Wikipedia contributors 2019]. Adversaries often use URL obfuscation to carry out phishing attacks. In fact, URL security is primarily studied in relation to phishing.

Researchers have studied the efficacy of different phishing techniques and demographic factors affecting phishing susceptibility [Dhamija et al. 2006; Downs et al. 2007; Hong et al. 2013; Sheng et al. 2010; Wu et al. 2006]; the impact of psychological manipulation on phishing susceptibility [Goel et al. 2017]; and the effect of communication medium on phishing susceptibility [Benenson et al. 2017; Benenson et al. 2014]. Phishing and URL obfuscation techniques have been categorized, e.g., [Althobaiti et al. 2019; Drake et al. 2004; Ollmann 2004]. However, there are also (ostensibly) legitimate reasons to obfuscate or otherwise break user expectations of where URLs go, e.g., URL redirection [Wikipedia contributors 2019d], tracking links [Cyphers et al. 2018]. Researchers have developed and compared phishing training approaches and educational materials [Arachchilage et al. 2016; Kumara Pragur et al. 2007; Sheng et al. 2010, 2007; Stockhard et al. 2016; Wen et al. 2019]. Companies even provide security training [KnowBe4 2019; PhishingBox 2019; PhishLabs 2019; Proofpoint 2019b; SANS 2019a,b].

Many defenses have also been pursued. Researchers have: compared browser indicators and warnings [Dhamija et al. 2006; Egelman et al. 2008]; developed ways to effectively convey security information [Maurer et al. 2011; Schechter et al. 2007]; and studied ML approaches for email filtering and URL classification [Aldomani et al. 2013; Bergholz et al. 2010; Blum et al. 2010; Fette et al. 2007]. Browsers [Mozilla 2019] and search engines [Whittaker et al. 2010] use blacklists and other techniques to protect users. Some products vet URLs in emails before allowing user access, e.g., [Microsoft 2019; Proofpoint 2019a,b]. However, these defenses are not always foolproof, e.g., [Nathaniel 2017].

Recently, there has been growing interest in using eye trackers for usable security. An eye-tracking based system was developed to train users to look at the status bar [Aiyamotuo et al. 2014]. Another study involved participants classifying websites, not just URLs, while wearing an eye tracker to examine how users gauge website legitimacy and evaluate security indicators [Alsharnoubi et al. 2015]. Our study is similar in spirit. However, we exclusively focus on how users visually process URLs. This narrow focus lets us dissect URLs into smaller components and examine how people process them. We seek to understand which parts people pay attention to, when people give up, and how their eyes process different URLs, amongst other things.

2.5 A Brief Introduction to URL Structure

A uniform resource locator (URL) is a string of characters that specifies the location of a web resource and how to access it [Berners-Lee et al. 1994, 1998]. Every URL in our corpus uses the https scheme.

Researchers have made several observations about URLs: they are used for meerous reasons to obfuscate or otherwise break user expectations of where URLs go, e.g., URL redirection [Wikipedia contributors 2019d], tracking links [Cyphers et al. 2018]. Researchers have developed and compared phishing training approaches and educational materials [Arachchilage et al. 2016; Kumara Pragur et al. 2007; Sheng et al. 2010, 2007; Stockhard et al. 2016; Wen et al. 2019]. Companies even provide security training [KnowBe4 2019; PhishingBox 2019; PhishLabs 2019; Proofpoint 2019b; SANS 2019a,b].

Many defenses have also been pursued. Researchers have: compared browser indicators and warnings [Dhamija et al. 2006; Egelman et al. 2008]; developed ways to effectively convey security information [Maurer et al. 2011; Schechter et al. 2007]; and studied ML approaches for email filtering and URL classification [Aldomani et al. 2013; Bergholz et al. 2010; Blum et al. 2010; Fette et al. 2007]. Browsers [Mozilla 2019] and search engines [Whittaker et al. 2010] use blacklists and other techniques to protect users. Some products vet URLs in emails before allowing user access, e.g., [Microsoft 2019; Proofpoint 2019a,b]. However, these defenses are not always foolproof, e.g., [Nathaniel 2017].

Recently, there has been growing interest in using eye trackers for usable security. An eye-tracking based system was developed to train users to look at the status bar [Aiyamotuo et al. 2014]. Another study involved participants classifying websites, not just URLs, while wearing an eye tracker to examine how users gauge website legitimacy and evaluate security indicators [Alsharnoubi et al. 2015]. Our study is similar in spirit. However, we exclusively focus on how users visually process URLs. This narrow focus lets us dissect URLs into smaller components and examine how people process them. We seek to understand which parts people pay attention to, when people give up, and how their eyes process different URLs, amongst other things.

2.5 A Brief Introduction to URL Structure

A uniform resource locator (URL) is a string of characters that specifies the location of a web resource and how to access it [Berners-Lee et al. 1994, 1998]. The original URL specification details URL structure [Berners-Lee et al. 1994]. Here, we present the bare essentials of URL structure at an appropriate level of granularity to understand our work.¹

Each URL in our corpus has the form:

```
<scheme>://<authority>/<rest>
```

The scheme component [Berners-Lee et al. 1998, 1994; WHATWG 2019] corresponds to the scheme name, which specifies how to interpret the text following the colon. Common schemes are http, ftp, and file. Every URL in our corpus uses the https scheme.

¹A more thorough treatment of URLs can be found in URL and URI specifications and standards [Berners-Lee et al. 1998, 1994; WHATWG 2019].
We created hypotheses to examine how users visually process URLs which may be empty; it may also include queries, fragments, and
and how URL features affect this processing:

3.1 Hypotheses
We created a URL corpus comprising 64 URLs partitioned into 8 categories.² Categories are defined by features corresponding to (1) safety, (2) complexity, (3) a leading www in the authority component, and (4) the attack type for unsafe URLs. The corpus contains 8 URLs for each of the 8 categories. To reduce variability and maintain uniformity between categories, every URL uses https as the scheme component and com as the top-level domain.

The categories are defined by the following 4 features:

4.1.1 Safety: URLs that are safe use domain names associated with popular services within the USA, such as Facebook. We selected the fully qualified domain names used in these URLs primarily from the top 1,000 US websites in the Quantcast Top One Million list³, although we consulted other lists as well. For the subset that were complex and included rest components, we chose the rest components by searching for legitimate content served by these domain names.

URLs that are unsafe have fully qualified domain names that, at the time of corpus construction, were eligible for purchase, did not have a domain name server record, or were spoofed websites. While many URLs with the unsafe feature were not actually unsafe to visit, it is exceedingly unlikely that participants would be knowledgeable about the status of the URLs tagged as unsafe, and, if an adversary wished to acquire the corresponding domains, they could do so. This decision allowed for greater control over the corpus.

² Materials used in this study can be found at https://drive.google.com/drive/folders/1ZNMLxXMxOU4R2n6a-6d7nMxx4Qvdy9y4
³ https://www.quantcast.com/top-sites

Table 1: Disaggregation of a URL into its three components.

<table>
<thead>
<tr>
<th>scheme</th>
<th>delims</th>
<th>authority</th>
<th>rest</th>
</tr>
</thead>
<tbody>
<tr>
<td>https</td>
<td>://</td>
<td><a href="http://www.google.com">www.google.com</a></td>
<td>/forms/about/</td>
</tr>
</tbody>
</table>

The authority component specifies a subset of the host, port, username, and password [Berners-Lee et al. 1998; WHATWG 2019]. For URLs in our corpus, the authority component has either the form host or user@host where host represents the host and user represents the username. In this study, the host is always a fully qualified domain name (e.g., www.wikipedia.org) - “a sequence of domain labels separated by ’.’” [Berners-Lee et al. 1994]. The last domain label is the top-level domain. For URLs in our corpus, the authority component comprises everything following the leading https:// until either the next /, if present, or the end of the line.

We call the last component rest, a catch-all term that is not borrowed from any specification or standard. It captures everything following the authority component. The rest component includes the path [Berners-Lee et al. 1998, 1994; WHATWG 2019], which may be empty; it may also include queries, fragments, and accompanying delimiters [Berners-Lee et al. 1998, 1994; WHATWG 2019]. For every URL in our corpus, if the rest component is non-empty, it includes a path that "identifies" the resource within the scope of [the] scheme and authority" [Berners-Lee et al. 1998], it begins at the first / character following the authority component, and it is the last part of the URL. Table 1 provides an example of a URL disaggregation into these three components. Please note the formatting style used for these components. Later, we define areas of interest of the same names but different formatting styles.

4 METHOD

4.1 URL Corpus and Classification
We created a URL corpus comprising 64 URLs partitioned into 8 categories.² Categories are defined by features corresponding to (1) safety, (2) complexity, (3) a leading www in the authority component, and (4) the attack type for unsafe URLs. The corpus contains 8 URLs for each of the 8 categories. To reduce variability and maintain uniformity between categories, every URL uses https as the scheme component and com as the top-level domain.

The categories are defined by the following 4 features:

Figure 1: The left side of the figure is a processed frame from the eye tracker video (This is not the same as what the participant sees). The red cursor indicates gaze position and the four colored boxes represent four AOIs: the scheme AOI (red), the authority AOI (green), the rest AOI (blue), and the response AOI (yellow). The right side is an image of a participant performing the task wearing the eye tracker.

H1: Total time spent on the scheme per character is less than that of the authority and rest components.

H2: For URLs that have an authority component of form user@host where user ends with .com, participants spend signifi-

H3: There exists a URL length threshold over which increasing URL length does not result in more time being spent on processing URLs.
We discuss the participant selection, the GUI, data collection, data processing, and data analysis.

We conducted a within-subject experiment that was approved by the Institutional Review Board (IRB). Each of the 20 participants were shown the 64 URLs from the corpus over two sessions. The task was to classify each URL as safe or unsafe. Participants completed the task by viewing one URL at a time and clicking a button on the GUI to indicate whether they believed the URL was safe. After signing the consent form, the participant was given a brief neutral mood induction video to control for the effects of mood induction video. They then saw an introduction to the study and the user interface. Following the URL classification task, the participant filled in a questionnaire comprising: demographic questions; questions pertaining to security knowledge and behaviors, especially regarding URLs and phishing; and questions to help assess experimental validity.

### 4.3 Data Collection, Processing, & Analysis

We discuss the participant selection, the GUI, data collection, data processing, and data analysis:

#### 4.3.1 Participants:

We collected data from 20 participants (3 female, mean age = 22.68, SD = 2.65). All participants were students who participated in the user study as part of their coursework. We discarded data from 4 participants due to technical issues with the data extraction from the eye tracker. Hence, we report on the data from 16 participants (2 female).

#### 4.3.2 User Interface:

The application was created using GUIs in MATLAB. It was presented to participants on a 24” monitor with a resolution of 1920×1200. Each URL image was created using bold monospace font [Wikipedia contributors 2019a] of size 64. The screen was made up of two panes. The first included the URL image, which was scaled and displayed on screen over 2-7 lines with a full line having approximate height of 20mm and width of 280mm. The second pane included the question “Is the web address safe to visit?”, accompanied by two response buttons that read “Safe” and “Unsafe” (see Figure 1). Four markers were embedded in the application to identify the surface plane to mark various AOIs during post-processing of the eye-tracking data. Times of clicks and corresponding classifications/responses captured via button clicks were also recorded.

#### 4.3.3 Eye Tracking:

We used the head-mounted Dikablis eye tracker to collect gaze positions. It contains three cameras: two eye cameras sampling the eye at 60 Hz and a scene camera sampling at 30 Hz. Gaze positions are computed from the pupil movements and mapped onto the video from the scene camera. Establishing a mathematical mapping between the features of eye and the target being looked at is referred to as calibration. We used the four-point operator-controlled calibration method [Nyström et al. 2013].

#### 4.3.4 Post-task questionnaire:

Following the URL classification task, the participant filled in a questionnaire comprising: demographics questions; questions pertaining to security knowledge and behaviors, especially regarding URLs and phishing; and questions to help assess experimental validity.

#### 4.3.5 Data Analysis:

We used MATLAB for post-processing the eye-tracking data. We used JMP Pro 14 and R for statistical analysis. The Shapiro-Wilk test indicated that all of our data were non-normally distributed, thus we used non-parametric tests (Kruskal-Wallis test and Wilcoxon test) for analysis.

#### 4.4 Procedure:

After signing the consent form, the participant was given a brief introduction to the study and the user interface. They then saw a short neutral mood induction video to control for the effects of mood induction video.
We examined five AOIs. Figure 1 captures the first four AOIs, when the user clicks on a button to classify it. This is a proxy for AOIs that correspond to the URL. We now present the five AOIs.

Table 3: Disaggregation of a URL in accordance with the first three AOIs. This differs from Table 1 in that the scheme AOI includes the “://” following the scheme.

<table>
<thead>
<tr>
<th>scheme AOI</th>
<th>authority AOI</th>
<th>rest AOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>https://</td>
<td><a href="http://www.google.com">www.google.com</a></td>
<td>/forms/about/</td>
</tr>
</tbody>
</table>

mood. They then filled in a pre-task questionnaire to assess their mood [Schaefer et al. 2010], wore the eye tracker, and completed a practice trial to familiarize themselves with the task and the GUI.

Before calibration, we adjusted a nose pin and head band to reduce the movement of the eye tracker during the study; we did not use a chin rest. Next, we focused the eye and scene cameras and calibrated the eye tracker using the four-point operator-controlled calibration method. The participant then classified URLs for the first session and took a break. The calibration procedure was then repeated and the participant classified URLs for the second session. Last, they filled in the post-task questionnaire. The distance between the screen and the participant was kept at about 0.6 meters.

4.5 Measures

4.5.1 Mood: Each participant’s mood was assessed along six emotional states: awake, pleasant, angry, fearful, happy, and sad [Mills et al. 2019]. The assessment used a 10-point scale, where 1 indicated that the participant’s mood was not associated with the given emotional state, and 10 indicated that it was highly associated.

4.5.2 Score: The score represents the number of correctly classified URLs within a set with no penalty for incorrect classification.

4.5.3 Total Time Spent: The total time spent on classifying a URL is the time (seconds) from the presentation of the URL to the time when the user clicks on a button to classify it. This is a proxy for the cumulative effort and engagement in classifying the URL.

4.5.4 Time Spent on Areas of Interest: Using the UTC timestamps of each data point recorded by the eye tracker, we computed the percentage dwell time on five AOIs (Areas of Interest). These measures express the distribution of users’ visual attention and help us understand which URL components users use to gauge URL safety. We examined five AOIs. Figure 1 captures the first four AOIs and Table 3 gives a disaggregation of a URL in accordance with the AOIs that correspond to the URL. We now present the five AOIs:

- The scheme AOI captures the scheme component and the delimiters immediately following it. As every URL in our corpus uses the https as the scheme, this AOI always corresponds to the leading https:// in the URL.
- The authority AOI captures the authority component. For classes C1 through C7, the authority component is a fully qualified domain name, e.g., www.google.com is the authority component of https://www.google.com. For class C8, the authority component has form user@host, e.g., as in www.google.com@evil.com. To test HS, the authority AOI was further split into two smaller AOIs, the user AOI and the host AOI corresponding to the user and host components.
- The rest AOI captures the rest component.
- The response AOI captures the response portion of the screen containing the “Safe” and “Unsafe” buttons.

4.5.6 Normalized Pupil Area: The eye tracker records raw pupil area of both eyes in pixels. We used the right eye pupil area. We used the Hampel identifier technique to remove outliers [Foroughi et al. 2017; Pearson et al. 2016]. Due to the non-uniform sampling rate, we interpolated the data to obtain a uniform sampling frequency of 60 Hz [Pfleging et al. 2016]. Then, we normalized the data to compare it between participants.

4.5.7 Accounting for Length Differences in URLs: URLs may differ in the number of characters in their scheme, authority, and rest components. Thus, for the corresponding AOIs, we calculated the time spent per character (total time spent on AOI divided by number of characters in AOI) and the fixation count per character (total number of fixations occurring on AOI divided by total number of characters in AOI). For the overall comparison, we computed overall time spent per character (total time spent/total URL length), overall fixation count per character (total fixation count/total URL length), and backtrack fixation count as a function of URL length (total backtrack fixations/total URL length).

5 RESULTS

5.1 Mood Induction Measures

On average participants were awake (ranking of M=7.50, SD=1.59), felt relatively pleasant (M=7.69, SD=1.40), and were mildly happy (M=6.75, SD=1.44). They did not feel angry (M=1.81, SD=0.83), fearful (M=1.56, SD=1.09), or sad (M=1.50, SD=0.82).

5.2 Scores

The average score was 40.44 out of 64. From the post-task questionnaire, we were able to identify whether the participants knew of the services associated with the safe URLs. Table 4 indicates the probabilities of participants correctly classifying the URL given that

Table 4: Probabilities of correctly classifying safe URLs given the participant knew of the service.

| Probabilities       | P[correct|known] | P[correct|unknown] |
|---------------------|------------|---------------|
| C1 (simple, www)    | 0.92       | 0.63          |
| C2 (simple, non-www)| 0.83       | 0.19          |
| C3 (complex, www)   | 0.76       | 0.5           |
| C4 (complex, non-www)| 0.58     | 0.46          |

- The last AOI captured visual targets other than the previous four areas of interest.

4.5.5 Fixations and Backtracking Fixations: Fixating is the act of maintaining one’s gaze at a particular target for a certain duration of time. It represents the time where new information is gathered [Ramkumar et al. 2019]. We extracted fixations of 100ms or more following prior research guidelines [Irwin and Zelinsky 2002; Munn et al. 2008; Salvucci and Goldberg 2000].

Backtracking is the process of revisiting information that was previously processed or skipped [Bruneau et al. 2002]. It usually occurs to re-establish previously processed information or it signifies a cognitive interest in an area with respect to the given task [Burton and Daneman 2007]. We measured the backtrack fixation count, i.e., the number of fixations involving backtracking.

4.5.6 Probabilities

We applied the following prior research guidelines [Irwin and Zelinsky 2002; Munn et al. 2008; Salvucci and Goldberg 2000]. We extracted fixations of 100ms or more following prior research guidelines [Irwin and Zelinsky 2002; Munn et al. 2008; Salvucci and Goldberg 2000].

We used the Hampel identifier technique to remove outliers [Foroughi et al. 2017; Pearson et al. 2016]. Due to the non-uniform sampling rate, we interpolated the data to obtain a uniform sampling frequency of 60 Hz [Pfleging et al. 2016]. Then, we normalized the data to compare it between participants.


classification.

- The scheme AOI captures the scheme component and the delimiters immediately following it. As every URL in our corpus uses the https as the scheme, this AOI always corresponds to the leading https:// in the URL.

- The authority AOI captures the authority component. For classes C1 through C7, the authority component is a fully qualified domain name, e.g., www.google.com is the authority component of https://www.google.com. For class C8, the authority component has form user@host, e.g., as in www.google.com@evil.com. To test HS, the authority AOI was further split into two smaller AOIs, the user AOI and the host AOI corresponding to the user and host components.

- The rest AOI captures the rest component.

- The response AOI captures the response portion of the screen containing the “Safe” and “Unsafe” buttons.

We applied the following prior research guidelines [Irwin and Zelinsky 2002; Munn et al. 2008; Salvucci and Goldberg 2000]. We extracted fixations of 100ms or more following prior research guidelines [Irwin and Zelinsky 2002; Munn et al. 2008; Salvucci and Goldberg 2000].

We used the Hampel identifier technique to remove outliers [Foroughi et al. 2017; Pearson et al. 2016]. Due to the non-uniform sampling rate, we interpolated the data to obtain a uniform sampling frequency of 60 Hz [Pfleging et al. 2016]. Then, we normalized the data to compare it between participants.

- The last AOI captured visual targets other than the previous four areas of interest.
they knew the service. The Kruskal-Wallis test showed no significant difference between the four categories of safe URLs (C1-C4) in terms of the participant knowing the services associated with the domain names $[X^2(3)=6.9674, p=0.0729]$.

### 5.3 Overview of Eye-Tracking Results

Table 2 presents some key results. The overall distribution of visual attention on the AOIs is shown in Figure 6. Using Kruskal-Wallis test, we found that the time spent per character was significantly different between the three AOIs corresponding to the URL $[X^2(2)=30.4152, p<0.0001]$. Post hoc analysis indicated time spent per character on the authority AOI was significantly higher than that of the scheme AOI and that of the rest AOI. The fixation count per character was significantly different between the three AOIs $[X^2(2)=23.9356, p<0.0001]$. Post hoc analysis indicated that fixation count per character on the rest AOI was significantly lower than the other two. However, we found no evidence that fixation duration was significantly different between the three AOIs $[X^2(2)=3.1692, p=0.0516]$.

The Kruskal-Wallis test indicated a significant difference in normalized pupil area $[X^2(2)=8.7532, p=0.0126]$. Post hoc analysis indicated a lower pupil area for the scheme AOI relative to other AOIs, suggesting less cognitive effort was expended on the scheme AOI.

### 5.4 Complexity

We saw a significant difference in overall time spent (seconds) processing between complex and simple URLs $[\text{Wilcoxon test: } Z=3.4865, p=0.0005]$. More time was spent on complex URLs ($M=7.26, SD=2.41$) compared to simple URLs ($M=4.58, SD=1.35$). This can also be seen pictorially in Figure 4. Wilcoxon test indicated significant differences in overall time spent per character $[Z=8.9998, p<0.0001]$, overall fixation count per character $[Z=6.4883, p<0.0001]$, and backtrack fixation count as a function of URL length $[Z=4.4399, p<0.0001]$.

People spent less time per character on complex URLs ($M=0.06, SD=0.01$) than simple URLs ($M=0.13, SD=0.04$). Figure 2 shows the time spent per character decreases as URL length increases. But the backtrack fixation count was higher on complex URLs ($M=3.68, SD=2.44$) relative to simple ones ($M=2.08, SD=1.18$). We found no significant difference in the score between complex ($M=4.76, SD=2.10$) and simple URLs ($M=5.34, SD=2.51$). Examining complex URLs of different lengths tells a more nuanced story. Figure 5 suggests a peak in time spent per character that occurs near 100 characters. We observed similar trends with fixation count per character and backtrack fixation count as a function of URL length for complex URLs.

### 5.5 Existence of www

We compared safe URLs that have authority components that begin with www (C1&C3) to those that do not (C2&C4). Wilcoxon test results indicated a significant difference in time spent per character on the authority AOI between www URLs ($M=0.16, SD=0.04$) and non-www URLs ($M=0.21, SD=0.04$); $[Z=4.2094, p<0.0001]$. Also, there was a significant difference in the fixation count per character on the authority AOI between www URLs ($M=0.24, SD=0.09$) and non-www URLs ($M=0.34, SD=0.12$); $[\text{Wilcoxon test: } Z=3.2292, p=0.0012]$. The score obtained (maximum score: 8) was also significantly different between www URLs ($M=6.50, SD=1.48$) and non-www URLs ($M=4.09, SD=1.90$); $[\text{Wilcoxon test: } Z=4.7020, p<0.001]$. The score indicates that people spent less time per character on www URLs than non-www URLs.
5.6 User@Host Attack Type vs. Regular URLs

To examine user visual attention for the user@host URLs (C8), we considered two special AOs at a finer granularity than the authority AOI: the user AOI and host AOI. We compared measurements on these two AOs for the user@host URLs (C8) to those for the authority AOI for safe URLs of similar structure (C3). Using the Kruskal-Wallis test we found a significant difference on time spent per character between the authority AOI of C3, the user AOI of C8, and the host AOI of C8 [χ²(2)=32.1735, p=0.0001]. A significant difference was also observed with fixation count per character [Kruskal-Wallis test: χ²(2)=11.3323, p=0.0035]. Post hoc analysis indicated that both sets of measurements for the host AOI for C8 were lower than those of the user AOI for C8 and the authority AOI for C3; the measurements between the user AOI for C8 were comparable to those of the authority AOI for C3. These results suggest that users process the user AOI of C8 and the authority AOI of C3 similarly. Also, there was a significant difference in the score between the user@host attack type (M=3.37, SD=2.41) and safe URLs of similar structure (M=5.81, SD=1.51); [Wilcoxon test: Z=6.9176, p=0.0035].

6 DISCUSSION

First, participant responses to the pre-task questionnaire following the mood induction video [Schaefer et al. 2010] indicate they were awake and in a neutral mood. Responses to the post-task questionnaire reveal that participants did not fatigue, and, on average, correctly identified the safety of about 40 of the 64 URLs (63%). Let us now turn to a detailed discussion of the results.

6.1 URL Processing & Classification Factors

6.1.1 URL Length: The overall time spent on classifying simple (and shorter) URLs (C1, C2, C5, C6) was less than the total time spent on classifying complex (and longer) URLs (C3, C4, C7, C8). This weakly supports H1, though follow-up work must be done to disentangle length from other complexity factors.

For complex URLs, we found URL length negatively correlated with time spent per character and fixation count per character. This supports H2.

We did not observe a correlation between URL length and score. Also, while Figure 4 suggests participants spent more time parsing URLs as URL length increases, Figure 2 suggests time spent per character decreases as we increase URL length. Moreover, the positive correlation between URL length and time spent seems to cease at a point, which supports H3. Specifically, Figure 5 suggests that at a threshold of approximately 100 characters, time spent stops increasing as we increase URL length. Similar trends were observed with fixation count per character and backtrack fixation count per character. We also observed no statistical difference between time spent on complex URLs under 100 characters and those above. One interpretation is captured by a notion similar to that of the compliance budget proposed by Beaumetem et al. [Beaumetem et al. 2008]: the user may only expend a finite budget of resources (here, time is a proxy for expended resources) to classify a URL, and, if the resources required to fully process a URL exceeds this budget, the user will not expend them. While the peculiarities of where that threshold is may depend on factors other than just URL length, we expect this notion of a finite budget applies more generally.

6.1.2 AOIs: We examine the influence of the AOIs:

- **Scheme AOI:** The decrease in the pupil area for the scheme AOI indicates reduced cognitive attention. Previous work found the frequency with which a user encounters a word affects the fixation duration and processing of that word [Rayner and Duffy 1986]. Users usually spend less time on frequently encountered words. Most legitimate websites use https nowadays, which is also used in each of the 64 URLs in our corpus. This explains the decrease in cognitive load for the scheme AOI. We observed a statistically significant difference in time spent per character between the scheme AOI and the rest AOI. Therefore, we do not have evidence to support H4.

- **Authority AOI:** The results indicate the time spent per character on the authority AOI is significantly higher than that of other AOIs. Time spent and fixation count per character on the authority AOI suggests users find www at the beginning of the domain name to be a strong indicator of URL safety.

- **Rest AOI:** Reduced fixation count while reading is characteristic of scanning text [Rayner and Fischer 1996]. The fixation count per character for the rest AOI is significantly
lower than it is for other AOIs, which suggests participants scanned the rest AOI.

6.1.3 Attack Types: Participants classified positive, unsafe URLs (C5) correctly 55% of the time and they classified negative, unsafe URLs (C6) correctly 74% of the time. This suggests people are more inclined to trust URLs that use positive words or phrases, even if they have no familiarity with the domain name. Table 4 shows that participants, on average, correctly classified the URLs 77% of the time, given that they had heard of the associated services.

Results suggest users visually process the user component of URLs with the user@host attack type (C8) similar to how they process the authority of URLs without a user component. In general, the fixation count per character was low for the rest component relative to both the scheme and authority components. For C8, we observed a reduced fixation count per character and time spent per character on the host component, which suggests participants perceived the host component as part of the rest component. Visual evidence suggests participants misidentified the user component as the host for URLs in C8. Of the unsafe URL categories, participants scored worst on C8. Participants spent significantly more time per character on the user component than the host component for C8, in support of HS.

We expect classification accuracies observed in this study are upper bounds on what users achieve in practice without additional safeguards in place. Sophisticated attacks that use URL features participants do not know about will likely be more effective. We also expect that attacks that use obfuscation in the rest component - or what users perceive as the rest component - are more likely to succeed given that participants spent less time on the rest component than the authority component in our study.

6.2 Improving Security in Practice

The study suggests a sort of ceiling effect: as URL length increases, participants spent more time vetting the URL until it capped out at around 100 characters. It also provides visual evidence of user misperceptions regarding URL structure. These insights into how users process and perceive URLs suggest concrete steps and best practices for services to improve the perceived security - and, we argue, the actual security - associated with the URLs they serve. For example, from a purely technical standpoint, there is no intrinsic security benefit to serving a URL that is short, has a domain name that begins with www, and has few special characters. But if those URLs match users’ safety expectations, users would be better at classifying both safe URLs served by the service and unsafe, obfuscated URLs served by adversaries.

Some unsafe URLs from our corpus were classified as safe because they exploited uncommon URL features that users rarely encounter in practice with legitimate services. Ironically, this makes such URLs easy for a computer to classify as risky. Surprisingly, we found that some web browsers offer no user protection against such URLs, even though simple-to-write parsers could easily detect them. This provides an opportunity to improve security at minimal cost.

Last, our findings can improve the quality of security awareness training programs. Our study identifies various misperceptions held by users. It also provides concrete evidence of where users look as they process URLs. This study’s methods and data may help in assessing, comparing, and improving training modules that aim to help users correctly identify URLs.

7 LIMITATIONS

Several considerations may have affected study generalizability: Participants were predominantly male college students pursuing electrical engineering degrees. To ensure the eye tracker accurately picked up on AOIs, we used a large font and displayed URLs over multiple lines. URLs were presented in isolation; contextual factors (e.g., the device on which a URL is displayed, the application on which a URL is viewed, or beliefs regarding who sent it) may affect visual behaviors and responses. Also, repeatedly asking participants whether URLs were safe likely sensitized them to phishing attacks.

However, we took precautions to minimize unintended effects. We conducted pilot runs to ensure the interface was clear and user fatigue was minimized. We used the post-experiment questionnaire to evaluate experimental validity. And we used a neutral-mood-inducing video to reduce variability in mood.

The available indicators provide some evidence of the study’s validity. The average participant score of 63% is within the ballpark of similar studies, e.g., [Dhamija et al. 2006; Sheng et al. 2010]. Post-task survey responses indicate most participants took the task seriously, exercised equal or only slightly more caution than they would in practice, and were not fatigued. Though we did not observe significant bias, we believe any bias would be in the direction of more caution and would be unlikely to invalidate our security recommendations as problems during the classification task would also be at play in the real world. We also note that applications and interfaces in the wild may vary regarding font properties so there is no one-size-fits-all approach for conducting such studies.

Last, the URLs may have had features we could not identify that affected participants’ visual behaviors and responses. We attempted to mitigate these concerns by including eight URLs per category, but further work is needed. Also, we only considered a few flavors of URL-based attacks. Notably, no attacks made use of the rest component, which may have affected participants’ visual behaviors.

8 CONCLUSION AND FUTURE WORK

Eye tracking provides a window to examine security behavior. This paper is a first step toward developing a model that captures how users visually process, derive meaning from, and operationalize URL security information to gauge URL safety. We conducted a user study in which participants saw URLs and then classified them while wearing an eye tracker. The findings suggest that participants relied on poor security indicators such as presence of www to gauge URL legitimacy, that they spent more time and cognitive resources to vet longer URLs but only up to a point, and that, for the unsafe, user@host URLs, participants perceived the user component to be the host component. In future work, we plan to study other contextual factors such as mood, additional flavors of URL obfuscation, and the effectiveness of training the user.

ACKNOWLEDGMENTS

The work of Ramkumar and Kun was supported in part by NSF grant OISE 1658594.