What breach? Measuring online awareness of security incidents by studying real-world browsing behavior

SRUTI BHAGAVATULA, Carnegie Mellon University, USA
LUJO BAUER, Carnegie Mellon University, USA
APU KAPADIA, Indiana University Bloomington, USA

Learning about real-world security incidents and data breaches can inform people how their information is vulnerable online and thus encourage safer security behavior. This paper examines 1) how often people read about security incidents online, 2) of those people, whether and to what extent they follow up with an action (e.g., trying to read more about the incident), and 3) what influences the likelihood that they will read about an incident and take some action. Our quantitative study of the real-world internet-browsing behavior of 303 participants finds a low level of awareness. Only 16% of participants visited any web page related to six widely publicized large-scale security incidents; few read about an incident even when it was likely to have affected them. We also found that more severe incidents and articles that constructively spoke about the incident were associated with more action. Our findings highlight two issues: 1) security awareness needs to be increased; and 2) current awareness is so low that expecting users to be aware and take remedial action may not be effective.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**.

Additional Key Words and Phrases: security incidents, data breaches, security awareness

ACM Reference Format:
Sruti Bhagavatula, Lujo Bauer, and Apu Kapadia. 2021. What breach? Measuring online awareness of security incidents by studying real-world browsing behavior. In *European Symposium on Usable Security 2021 (EuroUSEC '21), October 11–12, 2021, Karlsruhe, Germany.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3481357.3481517

1 INTRODUCTION

With the rise of security incidents and data breaches, security awareness is crucial for people to have the tools and know-how for keeping their computers and online data safe [44]. High-profile incidents and breaches in the past decade such as WannaCry, Heartbleed, Petya, and NotPetya have compromised over 300,000 systems worldwide [35, 46, 76]. The data compromised has ranged from passwords, names, and email addresses to credit card and social security numbers. People affected by these incidents typically need to become aware of them before they can take remedial action. More generally, awareness of the extent and effects of security incidents increases the adoption of better security practices [44, 48].

To this end, research about awareness of security incidents (completed using surveys and interviews) has found that people learn about breaches from a variety of sources and that some breaches are more likely to be discussed than others [25]. One survey found that almost half of respondents heard about a breach from a source other than the breached company [5]. People's reported willingness to take action was shown to be correlated with the source of information [94]

EuroUSEC '21, October 11–12, 2021, Karlsruhe, Germany

ACM ISBN 978-1-4503-8423-0/21/10.
https://doi.org/10.1145/3481357.3481517

and if they perceived a tangible security benefit [50]. Overall, these studies provide an important step towards understanding how people learn about and react to incidents. However, research thus far has relied largely on participants' recollection of past behavior or hypothetical situations, which is constrained by common limitations of self-reported methodologies [31, 40, 43, 80, 88, 90].

In this paper, we take a significant step toward a more detailed understanding of how people learn about and take action after incidents, specifically through online browsing. For six national-scale security incidents of potentially varying relevance to people, we use *longitudinal, real-world browsing data* to examine to what extent people may become aware of these incidents and the subsequent actions they may take (e.g., to learn more about the incident or generally about security). With the underlying goal of improving the spread of incident information through online media, we specifically study these problems in the context of *online browsing* without considering other channels by which this information may be shared. Our dataset was collected from the home computers of 303 participants between October 2014 and August 2018 and includes all URLs visited and passwords used to log onto online services from participants' home computers. As users could also read about security incidents on devices from which we do not collect data, we conducted a follow-up survey of 109 participants to validate our results (Sec. 6).

We explore two main topics: First, we examine *how often* participants read about incidents on the web and whether the likelihood of reading about incidents varies by demographics, browsing habits, or self-reported security behaviors. Second, we seek to understand *how participants came to read* about incidents, *how they reacted* to reading about them, and how different ways of finding out about incidents affect the action they take. For example, we examined whether the type of web content (e.g., news vs. social media) on which we first observed participants reading about incidents affected whether they took constructive action, such as further investigating an incident.

We found that only 16% of the 303 participants visited an incident-related web page about any of six major security incidents between 2014 and 2017. For example, only 15 of 59 likely Equifax credit-report holders read about the breach online in our dataset. These numbers remain low even after accounting for mobile browsing not captured by our dataset. Overall, we found that older and more tech-savvy participants as well as those who were more proactively aware about security [29] were more likely to read about security incidents on the internet.

Of the participants whom we observed reading about an incident, 73% subsequently visited additional web pages with information about the incident or about security and privacy in general. Reasonably, the higher the severity of the data compromised, the more likely participants were to visit related web pages. Participants' likelihood of taking action was higher if the content through which they found out about the incident had a positive sentiment; no other property of the incident-related content seemed to be associated with taking action, even though our power analyses showed we had a sufficient sample size to detect medium-sized effects.

Overall, our results suggest low awareness of, or inclination to follow up on, security incidents. The implications of these results are two-fold: first, our results suggest that people may not sufficiently engage with information about security incidents and that for those who do engage, the presentation of the information can play a role in inducing action. Second, the low rates of engagement may also indicate that increasing awareness is not the most effective avenue for keeping users safe. Further research should study other approaches to improving user security while systems that people use should take steps to keep their users safe without requiring them to maintain their own security.

We next survey related work (Sec. 2) and describe our dataset (Sec. 3). We then describe the methodology for and results of investigating *how many* and *which* people read about incidents (Sec. 4) and *how* people learn about incidents online and how this affects their actions (Sec. 5).

We also describe a follow-up study that substantiates our results using self-reported data (Sec. 6). Finally, we discuss the limitations and implications of our work (Sec. 7 and 8).

2 RELATED WORK

We survey three categories of related work: security incidents and how people perceive them or react; the dissemination of security and privacy information or advice and its influence; and methodologies for measuring security behavior.

2.1 Awareness and perceptions

Much of the existing work about security incidents studies how people interact with incidents such as data breaches (e.g., how people perceive data breaches and notifications and the risks involved) [50, 94], what influences people to take action after a breach [50, 94], and how people hear about data breaches [5, 25, 50, 94].

Recent work found, for example, that people reported to be more willing to take remediation actions if they perceived a tangible security benefit [50]. Research focused on the Equifax breach found that people reported that the source of advice about steps to take after the breach affected their willingness to take action [94]. Researchers also found that customers' spending at a retailer was significantly reduced after an announcement of a breach of the retailer's site [47]. Other researchers found that only a minority of survey participants would stop business with a company after the company suffered a breach [5] and that almost half learned about breaches from a source other than the affected company. More recent work surveyed people's reactions to notifications of password compromise. When advised or required to change their passwords, less than a third of respondents reported any intention to change their passwords [39]. Other work has shown that when a security incident involves a major social network people exhibit a variety of responses, from inaction to actively seeking out information [77].

Previous work has also studied people's general awareness of breaches and how breach information comes to people's attention and found that social media accounted for almost a third of participants' information sources [25]. More recent work found that only 26% of people were aware of a security breach even when they were affected by the breach [63].

Our work draws inspiration from previous work that examines how people come across incident information and suggests that the source of information is important for taking action. Our work is also motivated by the low self-reported awareness of breaches found in prior work. We focus on incident information on web pages and base our analyses on participants' real browsing behavior.

2.2 Information dissemination and influence

Related work also studied the mechanisms and sources from which people learn about security and privacy, often finding that social media and other web-based methods are good channels for this task. For example, prior work found that people use Twitter to complain or share opinions regarding security incidents [28] and that conversations about security and privacy drive people to share with and advise others [23, 75]. Recent qualitative work found that older adults tend to rely less on internet sources and more on social resources such as advice from friends and family [68]. Prior work has also studied themes in security and privacy advice across three different sources—news articles, web pages with security advice, and informal stories from family or peers—and found that each source presents information in a uniquely constructive way [74]. Research also found that the sources of security and privacy advice were important factors for people's digital security habits [79] (as also described in Sec. 2.1), and that the amount of advice that people reported receiving was not distributed evenly among economic classes [78]. Researchers have also looked at the ways in which presenting people with security information may help convince them to adopt good security practices. One such work proposed interfaces for filesystems that show people how others implement security [26]. Similar work found that showing people that their friends use security-enhancing features on social networks increases the uptake of these features [24].

Our work is motivated by the findings that web-based media are useful mechanisms for spreading computer security and privacy information. With a long-term of goal of sharing such information and advice more effectively, we specifically aim to understand empirically how relevant information is consumed via web browsing.

2.3 Measuring security behavior

Two approaches have been used to measure security behavior: collecting self-reported data through surveys, interviews, or controlled experiments (e.g., people's behaviors when exposed to internet attacks [69], password updating habits [41], and willingness to take remediation measures after a breach [50, 94]) and instrumenting users' computers to observe security behaviors (e.g., measuring password reuse [22, 32, 71] or the presence of malware on people's computers [34]). Since self-reported data can be prone to biases and may not be representative of the reality of peoples' security [9, 31, 40, 43, 80, 88, 90], we focus, unlike most prior research on security incidents, on empirical measurement of actual behavior.

Previous work has measured how often people update their computer systems [19, 34, 90], what security settings they use on their computers [19, 34], whether they are infected with malware [19, 34], and the presence of third-party applications [90]. Prior work has also measured how often people click on unsafe links [19, 82], their private-browsing [40] and password-reuse habits [32, 71, 90], and whether they install security-enhancing extensions [90]. We focus on security behaviors related to web usage, as we are specifically studying the use of web-based media in spreading information.

Previous work that extracted security behaviors from real data has collected data in multiple ways. One set of researchers partnered with an internet service provider that recorded all HTTP traffic of consenting participants [82]. Others asked study participants to install a tool that collected their system logs and information about the passwords they entered on web pages [90].

We leverage real-world behavioral data of home computer users (dataset described next, in Sec. 3) that was collected through instrumenting participants' operating systems and browsers.

3 DATA COLLECTION AND DATASET

Data collection. We obtained data collected as part of the Security Behavior Observatory (SBO) project, a longitudinal study of the security behaviors of Windows computer users [33] from October 2014 to July 2019. Data collected by the SBO includes information about system events and browser-related data such as browsing history. To collect this information, participants' home computers were instrumented with software that collects data via system-level processes and browser extensions. Data related to passwords entered into web pages was collected starting January 2017 and only in the Google Chrome and Mozilla Firefox browsers.

The SBO and its use for our work were approved by the ethics review board at Carnegie Mellon University. Data collected by the SBO has been used to study, for example, private-browsing habits [40], people's ability to detect phishing attacks [19], people's maintenance of their systems for security [19, 34], and password reuse habits [13, 71]. The SBO dataset contains data about a broad range of people across multiple demographics (described in Sec. 4.2).

Our study is based on the following longitudinal datasets collected by the browser extensions.

Browsing history: The browsing data we analyze spans a subset of the whole SBO dataset from October 2014 to July 2018, encompassing 303 participants who were active in the study at the time of when at least one of several security incidents was publicly announced (see Sec. 4). Participants enrolled in the SBO study on different dates and for different durations. The average duration for which the 303 participants were enrolled was 505 days. This dataset includes information about every URL visited in the web browser, along with page titles and timestamps.

Password data: This dataset spans from January 2017 to August 2018 and includes 233 of the 303 participants. The data includes information about every entry made into a password field in a web page, as determined by a browser extension, including: a salted one-way hash of the password and the URL of the form in which the password was submitted. We filtered this dataset to exclude passwords entered during failed login attempts or entered by a user other than the main computer user by replicating the filtering process used by prior work [13].

The browsing data was retrieved from participants' main computers. We assessed the accuracy of our results in the context of participants' overall browsing across multiple devices through a follow-up study of 109 SBO participants (see Sec. 6) which appeared to support our main findings. We further discuss the limitations of this dataset in Sec. 7.

4 WHO READS ABOUT SECURITY INCIDENTS

We examine how many and which people visit security-incident-related web pages and what factors are associated with their likelihood of doing so. We focus on selected security incidents (Sec. 4.1.1) and model participants by their demographics and technical backgrounds, self-reported security intentions, and internet browsing behavior (Sec. 4.1.2). We report on the relationship between these features and the likelihood that participants visited pages related to security incidents (Sec. 4.2).

4.1 Methodology

4.1.1 Identifying who reads about security incidents. We examined six security incidents that occurred between 2013 and 2017 [8, 56, 92] for which we expect most people to have read about *at least one*. We selected incidents that 1) were large-scale incidents (not affecting only a local population), 2) spanned a variety of incident types (from personal financial data losses and company document leaks to cyber attacks on home computers), 3) are well-known, and 4) were represented in our browsing history dataset. We selected well-known incidents because people's awareness and engagement are likely stronger and easier to observe for such incidents. In particular, we studied:

- **Equifax breach:** September 2017 breach of the credit reporting site that compromised the personal information of almost 150 million customers [16].
- **Uber breach:** Late 2016 breach that compromised the personal information of 57 million Uber users [58].
- **Ashley Madison breach:** Data breach on the affair-centric dating site in July 2015 and compromised around 33 million users' private information [60].
- **Panama Papers:** April 2016 breach of 11.5 million files from the database of the world's fourth largest offshore law firm, Mossack Fonseca [11, 45].
- **WannaCry:** Ransomware attack in May 2015 that initially affected over 70, 000 computers in 99 countries [12, 76].
- **Yahoo! breaches:** Two breaches: one in late 2014 affecting over 500 million user accounts and another in 2013 affecting over 1 billion user accounts [38, 73]. It was later revealed that all user accounts were compromised [57].

Each incident we studied may have been relevant to users in different ways. They could have been affected by it, they could be users of the compromised service and may want to be more cautious in the future, or they could learn about general security and privacy dangers. For example, although Panama Papers may not be directly relevant to most users, we included it because awareness about it could indirectly encourage users to be cautious about their own private records (e.g., medical records) and maybe be selective in trusting institutions with their data.

To study who reads about these incidents, we studied the 303 SBO participants who were active in the study before the incident became public and for three months after.

For each incident, we identified participants who visited an incident-related page (henceforth, we may call this *reading about an incident*). This page visit could have been the first exposure to the incident or an attempt to learn more about the incident online. Since we seek to study how often people actually signal their intent to learn more about an incident rather than simply "hearing about it," we did not consider a participant to have read about an incident if they may have seen it on some web page (e.g., social media) but did not click on the article.

To determine whether a participant read about an incident, we performed a keyword search over the URLs and titles of all the pages in their browsing history. For each incident, we manually selected a set of keywords that we believed would identify web pages that focus on that incident. For example, we searched for various combinations of "Yahoo" and one of the following: "compromise", "attack", "breach", and "hack". To confirm that our keyword lists were inclusive enough and that our identification process was robust, we also performed multiple Google searches using a variety of search terms to find web pages about the incidents and then confirmed that each of the top 100 Google search results about each incident would be identified by our keyword lists. We then manually verified that each page visit that matched a keyword actually corresponded to a page about the incident. For example, a page on `yahoo.com` with the path containing the word "hack" referring to a page about life hacks would not be considered an incident-related page.

Equifax and Yahoo! users. To provide further context for our observations of how many participants read about an incident, we observed, for people who were likely to have been *affected* by an incident, how many of them read about the incident as part of our analysis. Equifax and Yahoo! are the two breaches for which we were able to relatively accurately estimate how many participants were *actually affected* by examining whether they logged in to certain web sites. In both cases, the number of affected people was all or almost all users or consumers [16, 42, 57].

We determined which participants were likely to have had an Equifax credit report by observing who had entered a password on `equifax.com` or on one of the popular credit-report sites that report Equifax scores [20] (`identityforce.com`, `identityguard.com`, `annualcreditreport.com`, `creditsesame.com`, `creditkarma.com`, and `quizzle.com`) before Sep. 7, 2017, when the breach became public. While most Americans were likely to have been affected [16, 42] regardless of whether they had an account with a credit-reporting site, for this analysis, we considered this set of participants that were *very likely* to have been affected according to the above criteria.

Similarly, we determined which participants had a Yahoo! account by searching for participants who had entered a password on the `yahoo.com` domain before each of February 15, 2017—when the breach had first become public—and October 3, 2017—the time of the second breach announcement.

4.1.2 Studying which people read about incidents. After determining which participants read about an incident, we studied what participant characteristics correlated with visiting pages related to the security incidents. We modeled participants and their behavior using three feature sets and then performed a logistic regression for each feature set, where the binary outcome variable in each regression indicates whether a participant read about an incident.

Feature set 1: Demographic characteristics. Based prior work that showed that demographics were correlated with how people share security and privacy news and their comfort with uses of breached data [25, 50], we hypothesized that there could be a relationship between demographics and whether participants read about a security incident. Our first feature set contains demographic information: age, gender, income, highest education level, student status, whether the participant's primary profession involves programming, and whether the participant knows at least one programming language. The last two serve as measures of technical savviness.

Feature set 2: Self-reported security intentions. Prior work found self-reported security intentions (as measured by the SeBIS scale [29]) to be correlated with how people heard about and shared security and privacy news [25]. Hence, our second feature set is comprised of four continuous feature values of the SeBIS scale (values in $[1, 5]$), which participants optionally filled out upon enrollment in the SBO. The four values represent the extent to which participants 1) secure their devices, 2) generate strong and varied passwords across accounts, 3) demonstrate proactive awareness of security issues or safety of websites and links, and 4) update the software on their computers.

Feature set 3: Participants' observed internet behavior. We hypothesized that the types of web pages people browse are correlated with their likelihood to encounter information about a security incident (e.g., people who browse more technology-related news articles are more likely to come across web pages about security incidents). To test this hypothesis, we examined the kinds of topics of web pages that participants typically visited and the amount of their web browsing that involved visiting web pages on technical topics. We describe each of these next.

Characterization of browsing behavior: We used a topic-modeling algorithm to generate a set of topics that categorized participants' browsing. To generate the set of topics (which was the same for all participants), we looked up the category of the domain of every web-page visit in Alexa Web Information Services (AWIS) [2], resulting in a multiset (i.e., bag) of words. We performed topic modeling using the Non-negative Matrix Factorization (NMF) algorithm [59] on that multiset, which identified two topics as the most coherent for modeling participants' browsing (by determining the number of topics with the highest intra-topic cosine similarities [86]). One topic appeared to correspond to browsing that was professional or work-related; the other to browsing that was leisure- or entertainment-related. (See App. A for more details.) NMF also outputs a value that describes how much of each participant's web browsing matches each topic. We use these values as the features that characterize participants' browsing behavior.

Amount of technical content browsed: We also characterized people's browsing by how many of the web pages they visited were technical or technology-related. We again used NMF to build a topic model with two topics: one representing technology-related content and the other comprising all other content. This topic model was computed over the content of web pages visited by the participants. We considered a web page to be technical or technology-related if the AWIS category for the web page's domain contained the word "technical" or "technology". We downloaded the content of web pages in a sample of each participant's browsing history using the newspaper library [70] and used NMF to learn two topics based on two documents: the content of downloaded web pages with a technical AWIS category and the downloaded content of all other web-pages. We then applied this topic model trained for two topics on the multiset of tokens of downloaded content for each participant's browsing sample. Similarly to the characterization of browsing behavior, the NMF algorithm outputs, for each topic and participant, a weight for the topic within the sample of the participant's browsing history. We characterize the amount of technical content a participant browses by the weight corresponding to technical content. (See App. B for more detail.)

4.2 Results

The 303 participants we studied spanned a broad range of demographics (see Table 3 in App. C). We were surprised to discover that only 48 of the 303 (16%) read about[1] *any* of the six incidents[2]. In three additional instances, participants searched for incident-related keywords but did not visit any of the search results. Table 1 shows how many participants visited a page about each incident. This was computed based on browsing history from participants' home computers, which our confirmatory study suggests accounts for the majority of participants' browsing (see Sec. 6).

We also examined a subset of participants that we hypothesized were particularly likely to have been affected by the Equifax or Yahoo! breaches (see Sec. 4.1.1). Table 2 shows that very few likely affected participants for both incidents read about each incident. For example, of the 59 participants with likely Equifax reports, only 15 read about the incident in our dataset. We substantiate these low numbers through our follow-up study (Sec. 6).

Incident	*# users*	*% users*
Equifax	26	54%
Yahoo!	6	13%
Uber	4	8%
Ashley Madison	6	13%
WannaCry	14	29%
Panama Papers	10	21%

Table 1. Number of participants who read about each security incident; some read about multiple incidents.

We analyzed the relationship between the binary outcome of whether a participant read about any of the incidents and each of the three feature sets described in Sec. 4.1.1 through three logistic regression models (see App. E for model assumptions) using a significance level of 0.05.

First, we computed a model exploring the effect of demographic characteristics over the 303 participants (Table 4 in App. D). We found that participants' ages and whether they knew a programming language were significant factors. Specifically, older and more technology-savvy participants were more likely to read about incidents. The odds of reading about an incident increased by 3.216× ($p = 0.017$) if the participant was 50 years old or older and the odds of reading about an incident increased by 3.917× ($p = 0.002$) if a participant knew a programming language.

Incident	*# users likely affected*	*% (#) users that read*
Equifax	59	25% (15)
Yahoo!	48	2% (1)

Table 2. Number of participants likely affected by the Equifax or Yahoo! breaches and the percentage of those who read about the incident.

Our second model examines the relationship between whether participants read about an incident and their self-reported SeBIS scale values (Table 5 in App. D). This model was computed over 247 participants who provided SeBIS data to the SBO at enrollment time. Only one of the four SeBIS scale values, which we modeled by its Z-score for easier interpretation, was statistically significant: the odds of reading about an incident increased by a factor of 1.594 ($p < 0.01$) for each standard deviation increase in a participant's proactive awareness score.

Our third model examines the relationship between reading about an incident and participants' internet browsing behavior (i.e., browsing topics and amount of technical browsing; see Sec. 4.1.2). This model (see Table 6 in App. D for table results) was computed over 302 participants who had enough browsing data from which a sample sufficient for computing the technical browsing descriptor could be drawn (see App. B). Of the factors examined by this model, only the amount of technical or technology-related browsing was a significant factor (again modeled by its Z-score).

[1] We say that these participants "read about the incident," even though we cannot confirm they understood the content of the pages they visited.

[2] While not all participants may be interested in *every* incident, the incidents we study were chosen so that the majority of participants was *affected by* one or more incidents.

The odds of reading about an incident were increased by a factor of 3.315 ($p < 0.01$) for every standard deviation increase in the technical browsing score.

We built a fourth logistic regression model in which the features were the four significant features from the above three regression models. Table 7 in App. D shows the results of this model in which all features were again found to be significant and to increase the likelihood of reading about an incident. For example, with this model, participants over the age of 50 were 4.021× more likely to read about an incident than their younger counterparts.

4.3 Summary of findings

Overall, participants who were older, more proactively aware about computer security, and who were more technology-inclined were more likely to come across information about security incidents online. This indicates a potential imbalance in the dissemination of important security information.

5 HOW PEOPLE LEARN ABOUT INCIDENTS AND TAKE ACTION

We now study how the 48 participants who read about security incidents came to visit incident-related web pages and what behavior they exhibited in response. We first explain how we characterize reading about (discovery) and taking action after an incident (Sec. 5.1). We then examine how the characteristics of discovery or of the incident relate to participants’ reactions (Sec. 5.2).

5.1 Methodology

We defined features that characterize the process of discovery of web pages about incidents (Sec. 5.1.1) and we characterized participants’ actions after discovery (Sec. 5.1.2).

5.1.1 Learning about incidents. We examined the *browsing trajectories* (sequences of page visits that surround the visit of an incident-related web page) of each participant for each incident. We then measured the characteristics of the page visits that were part of the trajectory before the visit to an incident-related web page, and, separately, the characteristics of page visits after the first visit to the incident-related web page. We analyzed how the actions people take—as observed by examining the part of the trajectory after first visiting the incident-related web page—were related to characteristics of the incident and of the browsing path up to reading about the incident, participants’ demographics, and browsing behavior.

We constructed browsing trajectories as follows: we first identified each participant’s visits (if any) to web pages related to any of the incidents from Sec. 4.1.1. For each *first occurrence*—the first visit to any incident-related page about a specific incident—we defined a trajectory to be composed of the 20 page visits immediately preceding this first visit, the actual first visit, and the 20 page visits that immediately followed. In this manner, we constructed one trajectory per incident for each participant who visited any page about that incident.

To study how people read about incidents through browsing and their subsequent actions, we defined and analytically examined several features:

Precursor web page type (precursor_type) This feature describes the type of web page on which the participant clicked on a link that took them to the first occurrence of a page about the incident. This is commonly called the “referrer” page; we call it the precursor page because we identified these pages manually instead of via referrer headers, which are often not available. To create this feature, we manually categorized all the precursor pages as follows:

- **Social media:** A social media site (e.g., Facebook) page or home page.
- **Message boards:** A message forum such as 4chan message boards or Reddit.
- **News page:** A web page on a news website.
- **Purposeful:** Search engine results about the incident.

- **General browsing:** The page did not fall into one of the above categories but contained a link to the first-occurrence page (e.g., a stackexchange page with a sidebar link to an incident-related page).
- **Unknown:** No pages in the trajectory preceding the first occurrence of an incident-related page appeared to have a link to that page (e.g., the participant entered the URL manually or clicked on a link in an external tool).

Whether the precursor page was a home page (precursor_is_homepage) This feature captures whether the precursor page was a home page or whether the participant had to have browsed more deeply into a website before encountering the precursor page. We examined this feature to determine whether the link to the first-occurrence page was easily visible to anyone (i.e., on a home page) or would be seen only by some visitors to that site (i.e., those who navigated to a specific section).

First-occurrence page type (1st_occur_type) This feature categorizes the first-occurrence page according to whether it was specifically about the incident, and, if so, whether it was descriptive or prescriptive. We used the newspaper library [70] to extract the main content of each page. We then manually examined the content and developed the following three categories, using which we then classified each page: 1) general information about this specific incident; 2) advice about this specific incident; and 3) not specifically about this incident, but mentions it.

First-occurrence page sentiment (1st_occur_sentiment) Inspired by research on how the sentiment of social media posts influences the poster's followers [10], we hypothesized that people's reactions to web pages about incidents might be related to the sentiment of the pages: in particular, that a positive sentiment might correlate with more constructive action. We computed the sentiment for the main content of each first-occurrence page (collected as described above) using the NLTK Vader library [37]. The library returned a score in $[-1, 1]$; lower values indicated more negative sentiment, higher more positive sentiment, and zero indicated neutral sentiment. The feature we defined consists of three categories depending on this score: "positive," "neutral," and "negative". Scores greater than or equal to 0.1 fell into the "positive" category; scores less than or equal to -0.1 fell into the "negative" category; and all other scores were classified into the "neutral" category.

Incidents previously read about (incident_num) We hypothesized that people react to incidents differently depending on how many incidents they have come across through web browsing. Hence, for each incident that a participant read about, we counted the number of trajectories previously constructed for this participant for other incidents and exposed this as a feature in our analyses.

Type of data compromised (data_compromised) This feature represents the type of data compromised in the incident. We broadly grouped the data types and incidents as follows: **PII:** names, phone numbers, partial credit card numbers, email or physical addresses (Ashley Madison[3], Uber); **PII++:** PII with credit card information or social security numbers (Equifax); **passwords** (Yahoo!); and **miscellaneous** (WannaCry, Panama Papers).

5.1.2 Actions after reading about incidents.

We manually examined the 20 page visits in each trajectory immediately following the first occurrence as well as any visits to incident-related web pages *after* the first visit. We called one of these page visits an *action* taken in response to reading about the incident if it fell into at least one of the following categories:

- **Educating themselves about the incident:** e.g., reading additional articles about how the incident occurred, who was responsible, or implications of the incident.

[3] We categorized the Ashley Madison breach as only including PII and not passwords since passwords were cracked and leaked months after the original leak became public.

- **Educating themselves about general security:** e.g., reading articles about how to secure their network or whether using personal emails for work is safe.
- **Taking action to make themselves more secure:** e.g., attempting to freeze their credit reports or reading "what you need to do" articles.

We then counted the number of actions after reading about an incident. We use this raw count of actions as the outcome variable in our analyses (see Sec. 5.2.2). For example, if a participant visited two more pages that discussed the incident as well as one page with a "what you need to do" article, this would count as having taken three actions after reading about the incident.

So as to treat incidents uniformly, we did not consider actions tailored to any specific incident (e.g., changing passwords after a password breach).

5.2 Results

In Sec. 5.2.1, we describe how participants came to read about incidents and their subsequent actions. In Sec. 5.2.2, we analyze the amount of action participants took in relation to the type of incident, how they came across incident-related pages, participant characteristics, and their browsing behaviors.

5.2.1 Descriptive results. Using the methodology described in Sec. 5.1.1, we identified 66 distinct trajectories across the 48 participants (out of 303) who visited an incident-related web page. About twice as many trajectories described a participant reading about a PII++ breach (26) than a PII breach (10), and about four times as many described a PII++ breach than a password breach (6).

The types of web pages that led participants to visit the first incident-related page (precursor_-type) were relatively evenly distributed across social media (11), message boards (9), news pages (13), searching for the incident (9), and general browsing (10). For 14 trajectories we could not identify the precursor page. For the other 52, approximately half (24) the precursor pages were home pages (precursor_is_homepage); the others (28) were pages deeper in a website.

When we categorized the first incident-related page visits according to their content (1st_occur_-type), we found that 10 were advice articles, 35 were pages with general information about the incident, and 21 had content related to the incident (e.g., a story about a woman's identity stolen 15 times after the Equifax breach) without specific information about the incident. The sentiments of the first-visited incident-related pages (1st_occur_sentiment) were slightly positively skewed, with 31 trajectories categorized as "positive," 20 as "negative," and 15 as "neutral".

Most participants (71%) visited pages about only one of the six incidents. Only 10 visited pages about two incidents, two about three incidents, and two visited pages about four distinct incidents.

Most participants (73%) who visited an incident-related web page afterward took at least one action (i.e., visited another page about the incident, a page about security in general, or a page describing how to react to an incident). The mean number of actions taken across the 66 trajectories was 3 with a standard deviation of 7.19, the median 1, and the maximum 48. Figure 1 shows how the number of actions is distributed across the trajectories.

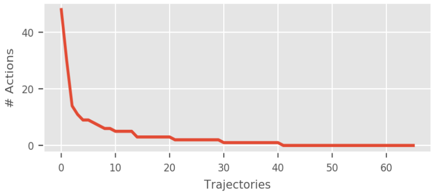


Fig. 1. Number of actions taken per trajectory.

5.2.2 Relating actions to features. We now examine the relationships between how much action participants took and four feature groups: features relating to the trajectory, the participant's demographics, the type of incident, and the participant's internet browsing behavior.

Three of the analyses are over the 48 participants (and 66 trajectories) who read about an incident. The analysis of participants' reactions relative to what led them to visit an incident-related web page is over 52 trajectories, since we removed trajectories for which we could not determine what led the participant to visit an incident-related page (i.e., precursor_type was "Unknown"; see Sec. 5.1.1).

Actions in relation to trajectories. The first analysis studies the number of actions participants took related to the following features of the browsing trajectories: precursor_type, precursor_is_homepage, 1st_occur_type, incident_num, and 1st_occur_sentiment.

We modeled this relationship by a quantile regression model, a non-parametric linear model suitable when the assumptions of linear regression are not satisfied [53]. In particular, we built quantile regression models to predict the conditional 0.25, 0.5, 0.75, and 0.9 quantiles of the number of actions outcome [18]. In this analysis, each trajectory (not each participant) is one data item and we used the raw action count as the outcome variable. We modeled each of the categorical features (with n levels) with $n - 1$ indicator variables compared to a baseline level.

The only factor that was correlated with the number of actions participants took after visiting an incident-related page was the sentiment of the content of this first visited page. This feature was only significant for the trajectories with many actions in the 0.9 quantile (90th percentile), where the number of actions in the 0.9 quantile increased by seven if the article's sentiment was positive instead of negative. Table 8 in App. F shows the results of the regression model for the 0.9 quantile.

Actions in relation to participant demographics. The second analysis examines the number of actions participants took relative to their demographics via a linear regression (see App. G for model assumptions). Table 9 in App. F shows the detailed results. If a participant read pages about multiple incidents with multiple trajectories, we averaged their actions across the trajectories.

No participant descriptors were statistically significant in relation to the actions participants took. We conducted a power analysis following previous work [7, 21, 30, 62, 67, 87] aiming for an experimental power of 0.8, a p-value (α) of 0.05, and a medium effect size. We calculated that a minimum sample size of 36 was necessary to see medium-sized effects with our model, a criterion which our sample of 48 participants exceeded. This suggests that if our model did not show a factor to be statistically significant, that factor was likely to not have had a "medium" or greater effect.

Actions in relation to the type of incident. We examine the relationship between the number of actions taken per trajectory and the type of data compromised (data_compromised) using the Kruskal-Wallis one-way test of variance [55].

The amount of action differed significantly between categories of the data compromised ($\chi^2 = 19.843$, $df = 3$, $p < 0.001$). To understand which groups were statistically different from each other and in what direction, we conducted a post-hoc analysis with pairwise comparisons using Dunn's test between each group, applying the Bonferroni correction for each comparison [14, 27].

Participants took an average of 5.35 actions after a PII++-compromised incident, 3.04 after a miscellaneous incident, but only 0.5 and 0.3 after reading about a passwords or PII incident, respectively. The greater number of actions taken for PII++ was significantly higher than for passwords ($Z = 3.002, p = 0.02$) or just PII ($Z = 3.85, p < 0.001$).

Actions in relation to participants' browsing behavior. Finally, we tested for relationships between the number of actions people took and the types of pages they visited on the web and the amount of technical browsing (as described in Sec. 4.1.2). The linear regression model, though suitable for this analysis (see App. G for model assumptions), did not reveal any statistically significant relationships, although, as before, a power analysis showed that we had sufficient power to see medium-sized or larger effects. Table 10 in App. F shows the results of this model.

5.3 Summary of findings

In summary, participants came across pages about security incidents through a variety of media in similar proportions. Most of the times participants came across the page about the incident after browsing deeper through a website, suggesting that such pages about incidents are not easily accessible (e.g., from a homepage). Participants were likely to read more about the incident or take an action when the page exhibited a positive sentiment but no other features were correlated with taking action, implying that the lack of action was nearly universal across our dataset.

6 CONFIRMING DATASET VALIDITY

Our findings (Sec. 4–5) are based on the browsing activity collected from one home computer of each participant. However, participants could have read about incidents or taken action on other devices, data about which is not captured in our dataset.

To shed light on how representative our dataset is of participants' overall browsing behavior, we collected additional self-reported data. We conducted a survey of 109 SBO participants who were active in May 2019, in which we asked about their familiarity with, and any actions after, several security incidents, as well as about how much web browsing they perform on which devices (see App. H). This study was approved by the review board at Carnegie Mellon University. The survey took between one and five minutes and participants received $5. Many participants in our main dataset were not active when we conducted this follow-up survey and vice versa; 84 of the 109 survey participants were in our original SBO dataset. Hence, we use this survey as *a measure of the self-reported behavior of SBO participants in general*, rather than of individuals who were in both datasets. (Results for the 109 participants described in this section are consistent with results computed over the overlapping 84 participants only.) Since the 109 participants may not have been active in the SBO around at least one of the incidents we studied, we did not ask participants about each of those incidents. Instead, we asked about a variety of events and security incidents, and additionally about incidents with a wide impact.

Participants reported that they conducted, on average, 59% of their web browsing on their SBO computers. As the amount of browsing on desktop and laptop computers may have decreased over time in favor of browsing on mobile devices [64], this 59% is likely a lower bound. Participants earlier in the study likely performed more of their overall browsing on their SBO computers.

We also asked participants how often (on a 5-point Likert scale [89]) they read about security incidents on (1) their SBO computer or (2) any other devices. We found no significant difference between the two distributions (Kolmogorov-Smirnov [54]: $D = 0.056, p = 0.997$). We also found no significant indication that the distribution of browsing on SBO computers vs. other devices varied by participant age (Spearman's correlation test: $S = 180990, p = 0.155$). Finally, to gauge the accuracy of the self-reported data, we asked how familiar participants were (on a 5-point Likert scale) with five security incidents, four non-incident-related events, and one fictitious security incident (an Airbnb social security number breach). 8% indicated moderate or extreme familiarity with the fake Airbnb breach, suggesting that the self-reported results may exaggerate actual familiarity.

Our results from Sec. 4 indicate that 25% of the participants in our main dataset who were likely affected by the Equifax breach read about the breach, a surprisingly small percentage. If we assume that this percentage is computed based on 59% of all browsing, then the actual percentage of people who read about the breach—if our data included 100% of all browsing—could be as high as 42%, which is still low considering the significance of the breach. However, when asked what action they took following the Equifax breach, 41 of the 86 survey respondents who indicated at least slight familiarity with the breach (48%) responded that they read about the breach online or visited the Equifax website, with the majority of the rest answering "didn't do much/didn't do anything". Five of the 41 respondents additionally replied that they "can't remember" and/or "didn't do much/didn't do anything," implying that the actual number of participants who read about the incident online or visited the website might be even lower than reported.

Similarly, our results from Sec. 4 indicate that 2% of participants in our main dataset who had a compromised Yahoo! password [57] read about the breach online. Self-reported data also suggests low awareness: when asked about reading and reacting to the Yahoo! breach, only nine of the 72 participants who indicated at least slight familiarity with the breach (13%) answered that they read about the breach online. Two respondents answered "can't remember" and/or "didn't do much/didn't do anything" in addition to reading online, again indicating that a lower number of participants than self-reported may have actually read about the incident.

Overall, our results suggest that the browsing data that we used for our analyses (Sec. 4–5) covers the majority (with a lower bound of approximately 59%) of the browsing performed by the participants. While the additional browsing participants performed on non-SBO devices may dilute some of our findings, the self-reported data supports the big picture: a surprisingly small subset of users may read about incidents and try to learn more about the incidents.

7 LIMITATIONS

The SBO data includes rich browsing and password data that is typically infeasible to obtain, at the cost of a limited participant pool and concerns about generalizability. We believe it is the big picture revealed by our results that matters—that very few people seem to engage with information about security incidents—rather than the specific percentages involved. As SBO browsing histories may not be representative of all the browsing users do, we conducted the confirmatory study, which supports the high-level findings based on SBO data: participants were rarely familiar with or read about major security incidents regardless of the devices on which they browsed the internet.

Since browsing history was represented via URLs and page titles, we could not include analyses that depended on the content of dynamic pages (e.g., social network pages). We also could not distinguish between content that participants consumed and content they loaded but did not read. Finally, since participants might have opened multiple pages in parallel and click on links in pages opened earlier, we could not always accurately determine the precursors to the first incident-related pages participants visited. In practice we found only a few instances where this was a problem.

The data we analyzed was collected only from Windows computer users. Windows is the dominant OS for personal computers [17], but users of non-Windows operating systems might exhibit behaviors different from the behaviors of the participants in our dataset.

Although data from SBO participants has been used for several security- or privacy-related studies [19, 34, 40, 71], the SBO participants may be biased towards less privacy- and security-aware people, given the nature of the SBO data collection infrastructure.

Finally, the subsets of participants we used in specific analyses were of sufficient size to make uncovering medium-sized effects likely, but not so large as to reliably discover small-sized effects.

8 DISCUSSION

Our dataset allows a comprehensive view of the actual browsing behaviors of 303 participants across 44 months. Although our sample is small, our results substantiated by our confirmatory study suggest low security engagement in general.

Our findings show that for the people in our dataset, the consumption of security and privacy incident information was not as prevalent in people's online activities as was ideal. Even when information was presented and consumed, participants often did not attempt to learn more about the incident or show further interest in reading about it. On the one hand, further research is needed to study how this information can be disseminated more widely and be studied in the context of a general population. To elicit and encourage interest, websites should better highlight problems, the implications and risks, and suggestions for staying secure or maintaining privacy in light of the incident [91, 93, 95]. On the other hand, further research is needed to understand how companies can keep their users safe without requiring them to have security awareness and take action.

Improving dissemination of security incident information. Our results highlight the challenge of increasing awareness of security incidents. Although the Equifax breach affected more than half of adult US residents, and hence, likely the majority of the participants, only 25% of likely affected participants visited an article related to the breach. Without adequate awareness of such incidents, people are unlikely to understand the importance of safe security behavior or that the implications of the incident may be relevant to them even if they were not directly affected [48]. We confirmed through a Google search that each of the three most popular news sites [4] published multiple articles describing each of the incidents during the three months after each became public. That is, there appeared to be sufficient publicity about each incident. Since these incidents were highly publicised, perhaps "security fatigue" made people reluctant to learn more about them [85]. Additionally, while it is expected that people will be more engaged when they have more at stake (as supported by our findings that the most severe incidents were associated with more action), challenges remain with improving their engagement in the context of everyday services [83].

Recent work examined what kinds of "stories" are more likely to be shared or to affect people's security behaviors [75]. We found that several incident-related articles participants first discovered were stories about the impacts of the incident, and not about the incident itself. Additionally, articles with a more positive sentiment were associated with people exhibiting security-enhancing behaviors in the form of trying to learn more about the incident. For instance, posts with a high positive sentiment score had titles such as "The One Move to Make After Equifax Breach" or "'WannaCry' on Linux systems: How do you protect yourself?", which suggest that the articles contain constructive advice and information. On the other hand, an example of an article title with a strong negative sentiment was "The next ransomware attack will be worse than WannaCry," which seems largely about warning people of the perils of the incident. News organizations reporting on security incidents could encourage action by presenting constructive advice. In general, these organizations could benefit from research on what kinds of stories and content are most likely to influence security behaviors and the further sharing of such content.

Demographic factors related to dissemination and action. Most of the demographic factors that we examined were not significantly associated with the likelihood to come across or act on incident-related articles. However, older and more technology-savvy participants were significantly more likely to have read about incidents. The latter may be because information about incidents is disseminated more towards technical audiences, perhaps because of the challenges of disseminating incident information (which may be seen as more technical) on non-technical outlets. Prior work found that security information is disseminated unevenly based on socio-economic status [78], which could also be linked to technology savviness. Another potential explanation is that technology-savvy people are more receptive to such information, and so it remains an open challenge to convince less technology-savvy people about the importance of security incidents and effectively communicate online risks to them. Recent work found that video communication can raise the saliency of risk and might be more effective at reaching less technology-savvy populations [36]. Thus, in addition to exploring what types of "stories" are more effective, further work is needed to explore the medium of delivering such stories for different populations.

Improving security without increasing awareness. Given the low engagement with security information that we observed, relying on users' awareness to ensure their security may not be viable.

Our analysis yielded several negative results, both in terms of what was correlated with coming across incident information and with following up on an incident. The negative results suggest that there may be a deeper issue in terms of how concerned about incidents people are in the first place. Reading articles about incidents may be too high of a burden if users do not have interest in the topics. Placing the burden of awareness on users also raises additional issues, such as requiring users to differentiate between correct and incorrect information they encounter online. For users to do this effectively, they need to already be educated about security practices.

Instead of relying on users to become aware of and proactively seek help after incidents, affected organizations could inform their users of an incident directly, with instructions for how to perform any needed remediation. While breach notification is a legal requirement in some countries, such as the US and the EU [1, 3], some companies that suffered a breach or incident that we studied apparently did not notify their customers. For example, Yahoo! did not notify their customers about their data breaches [65]. Companies could additionally provide detailed guidance on what consumers need to do to protect their data and accounts beyond on the affected site (e.g., identifying theft insurance, changing reused passwords on other domains). When possible, companies could also take immediate steps to protect their users on their behalf after a security incident. After password breaches, for example, companies can force a reset on all passwords. For widespread cyberattacks, a patch can be automatically deployed to all computers on the affected platforms. To help users stay secure in general, system designers should consider from the start how to remove the responsibility of security decision-making from users [6, 81].

9 CONCLUSION

Using the real browsing histories of 303 participants over four years, we measured how often participants read web pages about security incidents, what actions they took after reading, and what factors were associated with how likely they were to read about an incident or take action. We found that only a small minority (16%) of participants visited an incident-related web page about at least one of six large-scale security incidents. Furthermore, few participants who read about an incident showed further interest in the incident or took some action by reading more about it.

Our results highlight the challenges of increasing awareness of security incidents. Even when an incident was highly publicized and participants were likely to have been affected, few showed engagement with or awareness of the incident, e.g., only 25% read about the Equifax breach. Without adequate awareness, it is unlikely that people will act to improve their security. We found the low rate of discovery, and of constructive action after discovery, to be nearly universal across participants. Participants who were older, who exhibited a higher proactive security awareness, or who had an affinity for technology were more likely to read about incidents. Other factors, including the remaining demographics that we explored, had no impact. When reading web pages that spoke about the incident in a constructive way, participants were more likely to try and learn more about the incident or take action. However, no other factors correlated with taking action. Even though our results are based on a relatively small population, they highlight the need for wide and effective dissemination of security incident information and advice and for exploring alternative avenues to ensure user safety that do not rely on user awareness and education.

ACKNOWLEDGEMENTS

This work was supported in part by the Carnegie Mellon University CyLab Security and Privacy Institute. Parts of the dataset we used were created through work supported by the National Security Agency under Award No. H9823018D0008. We would also like to thank Jeremy Thomas and Sarah Pearman for help with working with the SBO data.

REFERENCES

[1] 2018. Art. 33 GDPR. Notification of a personal data breach to the supervisory authority. *General Data Protection Regulation (GDPR)* (Mar 2018). https://gdpr-info.eu/art-33-gdpr/

[2] 2018. AWS | Alexa Web Information Service - Traffic metrics for any website. https://aws.amazon.com/awis/

[3] 2021. Security breach notification laws. *National Conference of State Legislatures* (2021). https://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx

[4] 2021. Top 15 most popular news websites: March 2021. *eBizMBA | The eBusiness Guide* (2021). http://www.ebizmba.com/articles/news-websites

[5] Lillian Ablon, Paul Heaton, Diana Catherine Lavery, and Sasha Romanosky. 2016. *Consumer attitudes toward data breach notifications and loss of personal information.* Rand Corporation.

[6] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* (1999).

[7] Elham Al Qahtani, Mohamed Shehab, and Abrar Aljohani. 2018. The effectiveness of fear appeals in increasing smartphone locking behavior among Saudi Arabians. In *Symposium on Usable Privacy and Security (SOUPS).*

[8] Taylor Armerding. 2018. The 17 biggest data breaches of the 21st century. https://www.csoonline.com/article/2130877/data-breach/the-biggest-data-breaches-of-the-21st-century.html

[9] Farzaneh Asgharpour, Debin Liu, and L. Jean Camp. 2007. Mental models of security risks. In *Financial Cryptography and Data Security (FC).*

[10] Younggue Bae and Hongchul Lee. 2011. A sentiment analysis of audiences on twitter: who is the positive or negative audience of popular twitterers? In *International Conference on Hybrid Information Technology (ICHIT).*

[11] BBC News 2016. Panama Papers Q & A: What is the scandal about? https://www.bbc.com/news/world-35954224

[12] BBC News 2017. Cyber-attack: Europol says it was unprecedented in scale. https://www.bbc.com/news/world-europe-39907965

[13] Sruti Bhagavatula, Lujo Bauer, and Apu Kapadia. 2020. (How) Do people change their passwords after a breach? *arXiv preprint arXiv:2010.09853* (2020).

[14] J Martin Bland and Douglas G Altman. 1995. Multiple significance tests: the Bonferroni method. *The BMJ* (1995).

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JML)* (2003).

[16] Donna Borak and Kathryn Vasel. 2018. The Equifax hack could be worse than we thought. *CNNMoney* (2018). https://money.cnn.com/2018/02/09/pf/equifax-hack-senate-disclosure/index.html

[17] Ed Bott. 2013. Latest OS share data shows Windows still dominating in PCs. *ZDNet* (2013). https://www.zdnet.com/article/latest-os-share-data-shows-windows-still-dominating-in-pcs/

[18] Kevin J. Boudreau, Nicola Lacetera, and Karim R. Lakhani. 2011. Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science* (2011).

[19] Casey Canfield, Alex Davis, Baruch Fischhoff, Alain Forget, Sarah Pearman, and Jeremy Thomas. 2017. Replication: Challenges in using data logs to validate phishing detection ability metrics. In *Symposium on Usable Privacy and Security (SOUPS).*

[20] Heather Catalano. 2018. Best free credit report site of 2018. *The Simple Dollar* (2018). https://www.thesimpledollar.com/best-free-credit-report-site/

[21] Stephane Champely. 2018. *pwr: Basic functions for power analysis.* https://CRAN.R-project.org/package=pwr R package version 1.2-2.

[22] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. 2014. The tangled web of password reuse. In *Network and Distributed System Security Symposium (NDSS).*

[23] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A. Dabbish, and Jason I. Hong. 2014. The effect of social influence on security sensitivity. In *Symposium on Usable Privacy and Security (SOUPS).*

[24] Sauvik Das, Adam D. I. Kramer, Laura A. Dabbish, and Jason I. Hong. 2014. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *ACM Conference on Computer and Communications Security (CCS).*

[25] Sauvik Das, Joanne Lo, Laura Dabbish, and Jason I. Hong. 2018. Breaking! A typology of security and privacy news and how it's shared. In *ACM Conference on Human Factors in Computing Systems (CHI).*

[26] Paul DiGioia and Paul Dourish. 2005. Social navigation as a model for usable security. In *Symposium on Usable Privacy and Security (SOUPS).*

[27] Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association (JASA)* (1961).

[28] Paul Dunphy, Vasilis Vlachokyriakos, Anja Thieme, James Nicholson, John C. McCarthy, and Patrick Olivier. 2015. Social media as a resource for understanding security experiences: A qualitative analysis of #password tweets. In *Symposium on Usable Privacy and Security (SOUPS).*

[29] Serge Egelman and Eyal Peer. 2015. Scaling the security wall: Developing a security behavior intentions scale (SeBIS). In *ACM Conference on Human Factors in Computing Systems (CHI).*

[30] Paul D. Ellis. 2010. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Cambridge University Press.

[31] Xitao Fan, Brent C. Miller, Kyung-Eun Park, Bryan W. Winward, Mathew Christensen, Harold D. Grotevant, and Robert H. Tai. 2006. An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods* (2006).

[32] Dinei Florencio and Cormac Herley. 2007. A large-scale study of web password habits. In *The Web Conference.*

[33] Alain Forget, Saranga Komanduri, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, and Rahul Telang. 2014. Building the security behavior observatory: an infrastructure for long-term monitoring of client machines. In *Symposium on Hot Topics in the Science of Security (HotSoS).*

[34] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. 2016. Do or do not, there is no try: User engagement may not improve security outcomes. In *Symposium on Usable Privacy and Security (SOUPS).*

[35] Josh Fruhlinger. 2019. The 6 biggest ransomware attacks of the last 5 years. *CSO* (2019). https://www.csoonline.com/article/3212260/the-5-biggest-ransomware-attacks-of-the-last-5-years.html

[36] Vaibhav Garg, L. Jean Camp, Katherine Connelly, and Lesa Lorenzen-Huber. 2012. Risk communication design: Video vs. text. In *Privacy Enhancing Technologies Symposium (PETS).*

[37] CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Web and Social Media (ICWSM).*

[38] Vindu Goel and Nicole Perlroth. 2016. Yahoo says 1 billion user accounts were hacked. *The New York Times* (Dec 2016). https://www.nytimes.com/2016/12/14/technology/yahoo-hack.html

[39] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. 2018. What was that site doing with my Facebook password?: Designing password-reuse notifications. In *ACM Conference on Computer and Communications Security (CCS).*

[40] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. 2018. Away from prying eyes: analyzing usage and understanding of private browsing. In *Symposium on Usable Privacy and Security (SOUPS).*

[41] Hana Habib, Pardis Emami Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2018. User behaviors and attitudes under password expiration policies. In *Symposium on Usable Privacy and Security (SOUPS).*

[42] Robert Hackett. 2017. Equifax underestimated by 2.5 million the number of potential breach victims. *Fortune* (2017). http://fortune.com/2017/10/02/equifax-credit-breach-total/

[43] Ameya Hanamsagar, Simon S. Woo, Chris Kanich, and Jelena Mirkovic. 2018. Leveraging semantic transformation to investigate password habits and their causes. In *ACM Conference on Human Factors in Computing Systems (CHI).*

[44] Bartlomiej Hanus and Yu Andy Wu. 2016. Impact of users' security awareness on desktop security behavior: A protection motivation theory perspective. *Information Systems Management (ISM)* (2016).

[45] Luke Harding. 2016. What are the Panama Papers? A guide to history's biggest data leak. *The Guardian* (2016). https://www.theguardian.com/news/2016/apr/03/what-you-need-to-know-about-the-panama-papers

[46] Roy Horev. 2018. The Top 7 vulnerabilities of the decade. *Vulcan* (2018). https://blog.vulcan.io/top-7-vulnerabilities

[47] Ramkumar Janakiraman, Joon Ho Lim, and Rishika Rishika. 2018. The effect of a data breach announcement on customer behavior: Evidence from a multichannel retailer. *Journal of Marketing (JM)* (2018).

[48] Henri Barki Janine L. Spears. 2010. User participation in information systems security risk management. *Management Information Systems Quarterly (MIS Quarterly)* (2010).

[49] Bin Ju, Mincao Ye, Yuntao Qian, Rong Ni, and Chenxi Zhu. 2014. Modeling behaviors of browsing and buying for Alidata discovery using joint non-negative matrix factorization. In *International Conference on Computational Intelligence and Security (CIS).*

[50] Sowmya Karunakaran, Kurt Thomas, Elie Bursztein, and Oxana Comanescu. 2018. Data breaches: User comprehension, expectations, and concerns with handling exposed data. In *Symposium on Usable Privacy and Security (SOUPS).*

[51] Kassambara, Eva, Visitor, and Tomer Mann. 2018. Linear regression assumptions and diagnostics in R: Essentials. *STHDA* (Mar 2018). http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

[52] Kassambara and Michael U. 2018. Logistic regression assumptions and diagnostics in R. *STHDA* (Mar 2018). http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

[53] Roger Koenker and Gilbert Bassett Jr. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society* (1978).

[54] Andrey Kolmogorov. 1933. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* (1933).

[55] William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association (JASA)* (1952).

[56] Selena Larson. 2017. 10 biggest hacks of 2017. *CNNMoney* (2017). https://money.cnn.com/2017/12/18/technology/biggest-cyberattacks-of-the-year/index.html

[57] Selena Larson. 2017. Every single Yahoo account was hacked. *CNNMoney* (2017). https://money.cnn.com/2017/10/03/technology/business/yahoo-breach-3-billion-accounts/index.html

[58] Selena Larson. 2017. Uber's massive hack: What we know. *CNNMoney* (2017). https://money.cnn.com/2017/11/22/technology/uber-hack-consequences-cover-up/index.html

[59] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* (1999).

[60] Nate Lord. 2017. A timeline of the Ashley Madison hack. *Digital Guardian* (2017). https://digitalguardian.com/blog/timeline-ashley-madison-hack

[61] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval.*

[62] Arunesh Mathur and Marshini Chetty. 2017. Impact of user characteristics on attitudes towards automatic mobile application updates. In *Symposium on Usable Privacy and Security (SOUPS).*

[63] Peter Mayer, Yixin Zou, Florian Schaub, and Adam J. Aviv. 2021. "Now I'm a bit angry:" Individuals' awareness, perception, and responses to data breaches that affected them. In *USENIX Security Symposium.*

[64] Mary Meeker. 2018. Internet trends report 2018. *Kleiner Perkins* (2018). https://www.kleinerperkins.com/perspectives/internet-trends-report-2018/

[65] Renae Merle. 2019. Yahoo fined $35 million for failing to disclose cyber breach. *The Washington Post* (2019). https://www.washingtonpost.com/news/business/wp/2018/04/24/yahoo-fined-35-million-for-failing-to-disclose-cyber-breach/

[66] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[67] James Nicholson, Lynne Coventry, and Pam Briggs. 2017. Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection. In *Symposium on Usable Privacy and Security (SOUPS).*

[68] James Nicholson, Lynne Coventry, and Pamela Briggs. 2019. If it's important it will be a headline: Cybersecurity information seeking in older adults. In *ACM Conference on Human Factors in Computing Systems (CHI).*

[69] Kaan Onarlioglu, Utku Ozan Yilmaz, Engin Kirda, and Davide Balzarotti. 2012. Insights into User Behavior in Dealing with Internet Attacks. In *Network and Distributed System Security Symposium (NDSS).*

[70] Lucas Ou-Yang. 2013. Newspaper3k: Article scraping & curation: http://newspaper.readthedocs.io/en/latest/. *Python Software Foundation* (2013). http://newspaper.readthedocs.io/en/latest/

[71] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. 2017. Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *ACM Conference on Computer and Communications Security (CCS).*

[72] Marco Pennacchiotti and Siva Gurumurthy. 2011. Investigating topic models for social media user recommendation. In *The Web Conference.*

[73] Nicole Perlroth. 2016. Yahoo Says Hackers Stole Data on 500 Million Users in 2014. *The New York Times* (2016). https://www.nytimes.com/2016/09/23/technology/yahoo-hackers.html

[74] Emilee Rader and Rick Wash. 2015. Identifying patterns in informal sources of security information. *Journal of Cybersecurity* (2015).

[75] Emilee Rader, Rick Wash, and Brandon Brooks. 2012. Stories as informal lessons about security. In *Symposium on Usable Privacy and Security (SOUPS).*

[76] Steve Ranger. 2017. Ransomware attack: The clean-up continues after WannaCry chaos. *ZDNet* (2017). https://www.zdnet.com/article/ransomware-attack-the-clean-up-continues-after-wannacry-chaos/

[77] Elissa M Redmiles. 2019. "Should I Worry?" A Cross-Cultural Examination of Account Security Incident Response. In *IEEE Symposium on Security and Privacy (SP).*

[78] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2016. How I Learned to be Secure: a Census-Representative Survey of Security Advice Sources and Behavior. In *ACM Conference on Computer and Communications Security (CCS).*

[79] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. 2016. I Think They're Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *IEEE Symposium on Security and Privacy (SP).*

[80] Robert Rosenman, Vidhura Tennekoon, and Laura G. Hill. 2011. Measuring bias in self-reported data. *International Journal of Behavioural and Healthcare Research* (2011).

[81] Martina Angela Sasse and Ivan Flechais. 2005. Usable security: Why do we need it? How do we get it? *O'Reilly Media, Inc.* (2005).

[82] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. 2018. Predicting Impending Exposure to Malicious Content from User Behavior. In *ACM Conference on Computer and Communications Security*

(CCS).

[83] Ruth Shillair, Shelia R. Cotten, Hsin-Yi Sandy Tsai, Saleem Alhabash, Robert LaRose, and Nora J. Rifon. 2015. Online safety begins with you and me: Convincing Internet users to protect themselves. *Computers in Human Behavior* (2015).

[84] Atsushi Shimada, Fumiya Okubo, and Hiroaki Ogata. 2016. Browsing-Pattern Mining from e-Book Logs with Non-negative Matrix Factorization.. In *Educational Data Mining (EDM).*

[85] Brian Stanton, Mary F. Theofanos, Sandra Spickard Prettyman, and Susanne Furman. 2016. Security fatigue. *IT Professional* (2016).

[86] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).*

[87] Lynn Tsai, Primal Wijesekera, Joel Reardon, Irwin Reyes, Serge Egelman, David Wagner, Nathan Good, and Jung-Wei Chen. 2017. Turtle Guard: Helping Android users apply contextual privacy preferences. In *Symposium on Usable Privacy and Security (SOUPS).*

[88] Blase Ur, Jonathan Bees, Sean M Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Do Users' Perceptions of Password Security Match Reality? In *ACM Conference on Human Factors in Computing Systems (CHI).*

[89] Wade M Vagias. 2006. Likert-type scale response anchors. *Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management, Clemson University* (2006).

[90] Rick Wash, Emilee Rader, and Chris Fennell. 2017. Can People Self-Report Security Accurately?: Agreement Between Self-Report and Behavioral Measures. In *ACM Conference on Human Factors in Computing Systems (CHI).*

[91] Jane K. Winn. 2009. Are better security breach notification laws possible. *Berkeley tech. LJ* (2009).

[92] Kim Zetter. 2015. The Year's 11 Biggest Hacks, From Ashley Madison to OPM. *Wired* (2015). https://www.wired.com/2015/12/the-years-11-biggest-hacks-from-ashley-madison-to-opm/

[93] Yixin Zou, Shawn Danino, Kaiwen Sun, and Florian Schaub. 2019. You 'might' be affected: An Empirical Analysis of Readability and Usability Issues in Data Breach Notifications. In *ACM Conference on Human Factors in Computing Systems (CHI).*

[94] Yixin Zou, Abraham H. Mhaidli, Austin McCall, and Florian Schaub. 2018. "I've Got Nothing to Lose": Consumers' Risk Perceptions and Protective Actions after the Equifax Data Breach. In *Symposium on Usable Privacy and Security (SOUPS).*

[95] Yixin Zou and Florian Schaub. 2019. Beyond mandatory: Making data breach notifications useful for consumers. *IEEE Security & Privacy* (2019).

A APPENDIX

A Characterizing browsing behavior

To identify topics that characterize participants' browsing, we applied the Non-negative Matrix Factorization (NMF) topic-modeling algorithm [59] to participants' browsing histories. NMF has been used in prior work for mining browsing behavior patterns [49, 84].

To build a topic model, we created one document per participant, each consisting of the tokens of the Alexa Web Information Services (AWIS) categories of the participant's web-page visits. For example, the category for `google.com` is `Top/Computers/internet/Searching/Search_-Engines/Google`.

For each participant, we tokenized the AWIS categories of the domain of each page visit and discarded the "Top" token to create a multiset of tokenized categories. If a domain appeared multiple times in a participant's browsing history, the tokenized AWIS categories appeared an equal number of times. We then computed the term frequency-inverse document frequency (TF-IDF) score [61] for each token to produce the document-token matrix to be used as input by the NMF algorithm.

We applied the NMF topic modeling algorithm to this matrix. We varied the number of topics from two to 10 and identified the optimal number of topics by observing when the most-frequently occurring tokens in a topic were on average most similar to each other [72, 86]. To determine this similarity, we computed the average of the pairwise cosine similarities of the top 20 tokens within each topic, using a Word2Vec model [66] trained on the same documents used to train the topic model, and then averaged these average similarities. The average within-topic cosine similarity was highest when the number of topics was two.

We examined the top 20 tokens in each topic to determine the themes of the topics. The words in one topic seemed to represent more leisure-oriented browsing ("Social_Networking", "Shopping", etc.) and the other professional-oriented ("Education", "Business", "E-mail", etc.). We used the weights of topics in the resulting document-topic matrix computed over each participant's AWIS multiset as two features.

We also experimented with Latent Dirichlet Allocation (LDA) [15] but we observed the resulting topic clusters to not be as coherent as the ones derived with NMF.

B Characterizing of browsing behavior of technical or technology-related content

To identify how much of a participant's browsing was technology-related, we again used the NMF topic modeling algorithm to build a topic model that had two topics: 1) technical or technology-related content and 2) all other content.

We trained this model using a 1% sample of each of the participants' browsing histories and Alexa categories described in App. A. Here the input only contains two documents: one for the technical webpages and one for the non-technical pages. We built each document to be a representation of the content that is categorized as technical or non-technical by its domain's AWIS category, i.e., the technical document contains content about web pages that have the word "Technology" or "Technical" in its AWIS category and the non-technical document contains all other content. We downloaded the content of each web page in the sample with the newspaper library [70], tokenized each page's content, and concatenated the tokens from all technical web pages to construct the technical document and from all other web pages to construct the non-technical document (from a sample of web pages of the same size as the set of pages in the technical category). When computing the TF-IDF scores for tokens in each document, we only included tokens that appeared in one document but not the other. This way, we could construct topics with tokens that were unique to either the technical or non-technical category.

After training the two-topic topic model, we determined the index of the column in the resulting document-topic matrix that corresponded to the technical document and therefore, to the technical topic. We applied the trained model on the sample of each participant's browsing history (a multiset of tokens of the page content of web pages with a defined AWIS category). The model computed two weights for each participant, of which we used the weight of the technical topic as the feature characterizing the amount of a participant's browsing related to technical content.

C Demographics

Table 3 describes the demographic distribution of the 303 participants whose data we studied.

D Additional information on who read about incidents

Tables 4, 5, and 6 show the results of the logistic regression models exploring associations between the likelihood of coming across incident information and the three feature sets describing demographics, self-reported security intentions on the SeBIS scale, and internet browsing behavior.

Table 7 shows the results of the logistic regression model (see App. E for model assumptions) exploring the relationship between the four significant features from the aforementioned three models and the outcome of whether a participant came across an incident.

E Logistic regression assumptions

Figure 2 shows the approximately linear relationships between each continuous feature and the log-odds of the outcome for the model studying the effects of the SeBIS feature set [52]. Similarly for the model studying browsing behavior, Figure 3 shows the approximately linear relationships between the continuous browsing features and the log-odds of the outcome. The linearity assumption for

Category	***Distribution***
Age	Range: 20 to 83 Mean age: 36 Median age: 29 Standard deviation: 16.28
Gender	Female: 59% Male: 41% Did not provide: <0.5%
Education	High-school: 8% Associates degree or some other college: 29% Bachelor's degree: 40% Advanced degree: 23%
Income	≤ 50k: 50% 50k-100k: 24% 100k-200k: 10% ≥ 200k: 2% Did not provide: 14%
Is_student	Students: 48%

Table 3. Demographic distribution of the 303 participants.

	baseline	***coef.***	***exp(coef.)***	***std.err.***	***z***	***p***
(Intercept)		**-2.293**	**0.101**	**0.492**	**-4.661**	**<0.01**
age: ≥ 50	**<50**	**1.168**	**3.216**	**0.491**	**2.380**	**0.017**
gender: male	female	0.011	1.011	0.346	0.032	0.975
gender: not provided	female	-11.663	<0.01	882.744	-0.013	0.989
education: ≥ ugrad	<ugrad	-0.080	0.923	0.349	-0.229	0.819
income: >$50k	<$50k	-0.637	0.529	0.378	-1.684	0.092
income: declined to answer	<$50k	-0.530	0.588	0.544	-0.975	0.329
knows_prog_lang: yes	**no**	**1.365**	**3.917**	**0.441**	**3.093**	**<0.01**
is_programmer: yes	no	0.090	1.094	0.423	0.212	0.832
is_student: yes	no	-0.075	0.927	0.480	-0.157	0.875

Table 4. Logistic regression model describing the relationship between whether a participant read about an incident and characteristics of the participant including their demographics. "Ugrad" denotes that the participant indicated achieving a Bachelor's degree.

the demographics model was met by default due to all the features being categorical. The other two logistic regression assumptions, i.e., lack of influential observations and lack of multicollinearity, were satisfied for all models.

Table 4 depicts linearity assumptions for the logistic regression model studying the outcome of whether participants read about an incident and all four significant features from the first three models. There were no influential observations or multicollinearity.

F Additional information on actions after reading about incidents

Table 8 shows the results of the quantile regression model for the 0.9 quantile exploring the relationship between browsing trajectories and the number of actions participants took after

	coef.	*exp(coef.)*	*std.err.*	*t*	*p*
(Intercept)	**-2.828**	**0.059**	**1.344**	**-2.105**	**0.035**
device_securement	0.044	1.045	0.151	0.290	0.772
password_generalization	0.386	1.471	0.345	1.120	0.263
Z(proactive_awareness)	**0.467**	**1.594**	**0.169**	**2.768**	**<0.01**
updating	0.167	1.182	0.200	0.836	0.403

Table 5. Logistic regression model describing the relationship between whether a participant read about an incident and the SeBIS scale values they provided. The proactive_awareness feature was represented by its Z-score.

	coef.	*exp(coef.)*	*std.err.*	*t*	*p*
(Intercept)	**-4.773**	**0.008**	**1.760**	**-2.712**	**<0.01**
Z(browsing_technical)	**1.199**	**3.315**	**0.464**	**2.582**	**<0.01**
sq(browsing_leisure+1)	1.519	4.566	1.253	1.212	0.226
browsing_professional	4.859	1.289	3.187	1.525	0.127

Table 6. Logistic regression model describing the relationship between whether a participant read about an incident and characteristics of their internet browsing behavior. The browsing_leisure feature was applied a square transformation to meet the linearity assumptions of logistic regression and browsing_technical was represented by its Z-score.

	baseline	*coef.*	*exp(coef.)*	*std.err.*	*z*	*p*
(Intercept)		-5.655	0.004	1.108	-5.104	<0.01
age: ≥ 50	<50	**1.391**	**4.021**	**0.463**	**3.005**	**<0.01**
knows_prog_lang: yes	**no**	**1.201**	**3.322**	**0.414**	**2.898**	**<0.01**
sebis_proactive_awareness		**0.911**	**2.487**	**0.366**	**2.489**	**0.013**
browsing_technical		**1.465**	**4.326**	**0.458**	**3.196**	**<0.01**

Table 7. Logistic regression model describing the relationship between whether a participant read about an incident and the four significant features from Tables 4, 5, and 6.

reading about an incident. Tables 9 and 10 show the results of the linear regression models studying the number of actions in association with participant browsing behavior and demographics.

G Linear regression assumptions

Figure 5 contains the plots showing that the linear regression model studying the number of actions in relation to browsing behavior approximately satisfies the assumptions of linearity, normality of residuals, and homogeneity of variance [51]. Figure 6 contains similar plots for the linear regression model for the number of actions in relation to participant demographics. For both models, the outcome variable was log-transformed and features were transformed as necessary to meet the model assumptions.

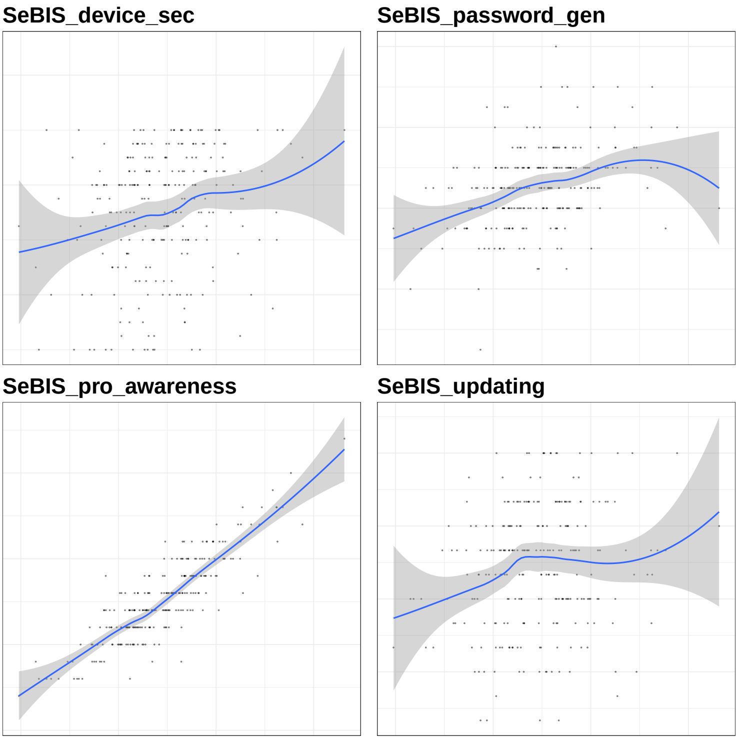


Fig. 2. Scatterplots depicting approximate linearity between the four continuous SeBIS features and the log-odds of the outcome for the SeBIS model.

	baseline	*coef.*	*std.err.*	*t*	*p*
(Intercept)		2.000	3.600	0.556	0.581
precursor_is_homepage: yes	no	-1.000	1.886	-0.530	0.599
1st_occur_type: info	advice	5.000	3.152	1.586	0.120
1st_occur_type: related	advice	1.000	3.403	0.294	0.770
incident_num: not first	first	3.000	2.035	1.474	0.147
1st_occur_sentiment: pos	**neg**	**7.000**	**2.485**	**2.816**	**<0.01**
1st_occur_sentiment: neu	neg	-1.000	2.875	-0.348	0.730

Table 8. Quantile regression model for the 0.9 quantile of the relationship between actions participants took and the trajectories that led them to reading about incidents. We grouped the values of incident_num into two buckets: “first” and “not first,” where the second bucket means that there was at least one incident previously read about. The model excludes precursor_type due to its correlation with precursor_is_homepage.

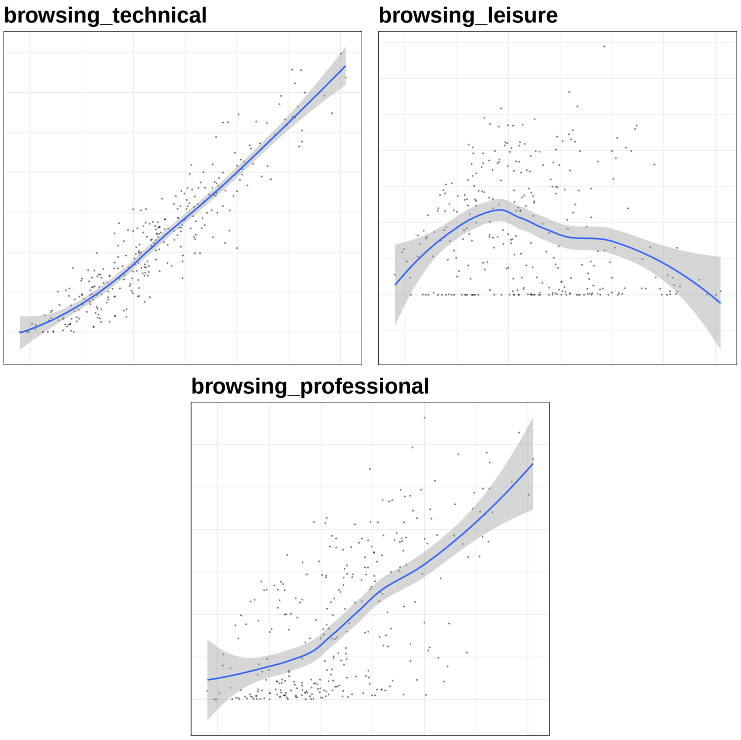


Fig. 3. Scatterplots depicting approximate linearity between the three continuous browsing features and the log-odds of the outcome for the browsing behavior model.

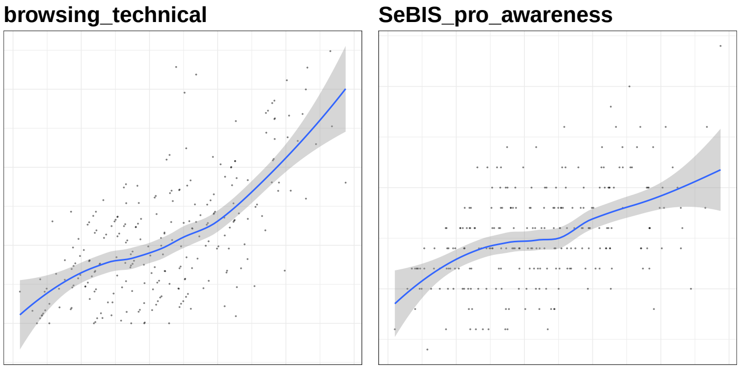


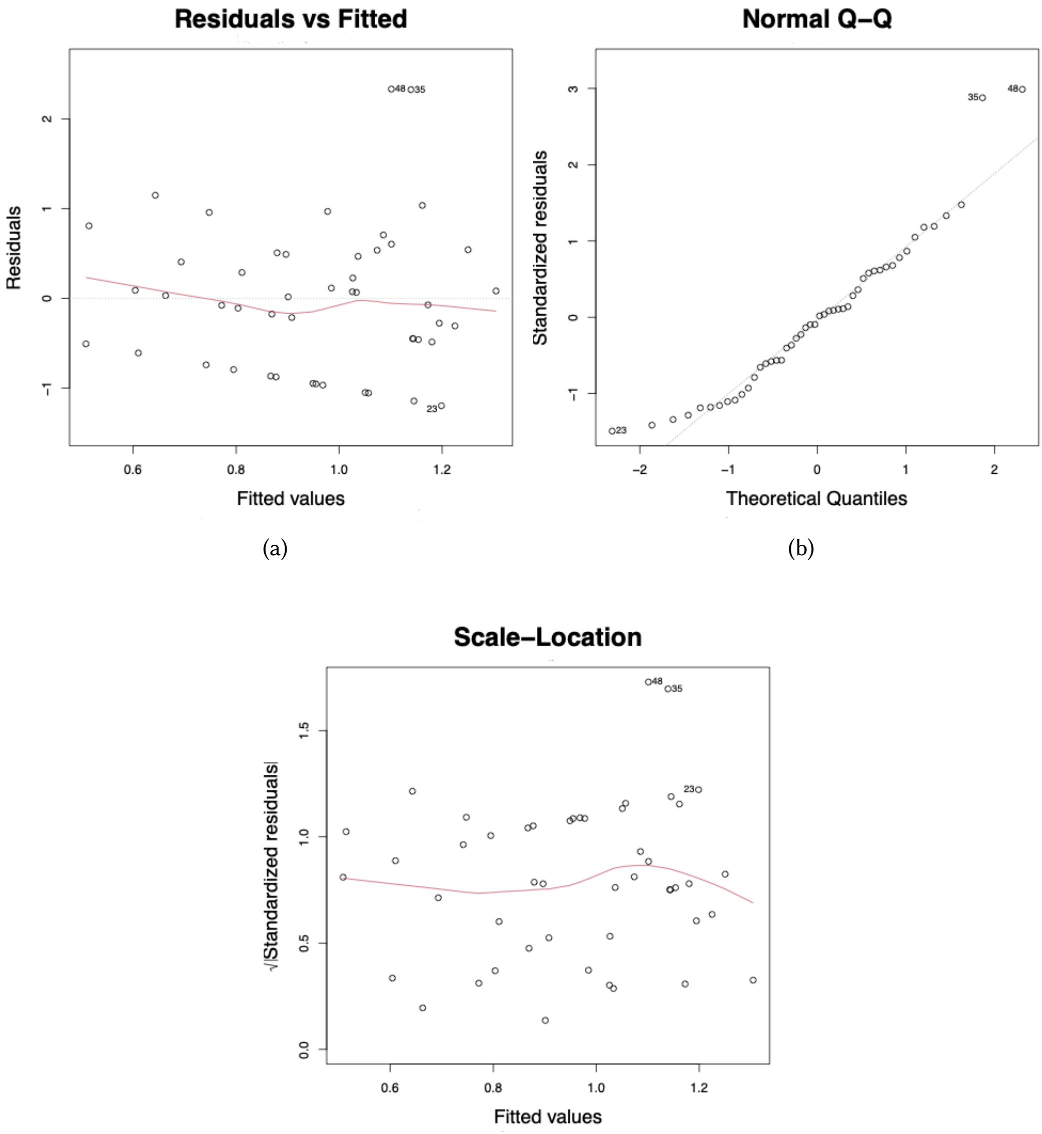
Fig. 4. Scatterplots depicting approximate linearity between the continuous features and the log-odds of the outcome for the model studying the four significant features from the three feature sets.

	baseline	***coef.***	***std.err.***	***t***	***p***
(Intercept)		0.782	0.533	1.467	0.150
(1/age)		6.932	19.085	0.363	0.718
gender: male	female	-0.271	0.299	-0.906	0.371
knows_prog_lang: yes	no	0.143	0.301	0.477	0.636
is_programmer: yes	no	-0.536	0.366	-1.465	0.151
is_student: yes	no	-0.028	0.439	-0.063	0.950
education: ≥ ugrad	<ugrad	0.507	0.284	1.785	0.082
income: >$50k	<$50k	-0.424	0.291	-1.453	0.154
income: declined to answer	<$50k	-0.142	0.420	-0.337	0.738

Table 9. Linear regression model of the relationship between actions participants took and their demographics. The outcome variable is log(actions taken) + 1. The age feature was transformed to its reciprocal for the model to meet the assumptions of linear regression.

	coef.	***std.err.***	***t***	***p***
(Intercept)	**1.304**	**0.419**	**3.111**	**<0.01**
browsing_technical	0.159	0.324	0.491	0.626
browsing_leisure	-3.413	2.190	-1.558	0.126
browsing_professional	-0.671	1.399	-0.479	0.634

Table 10. Linear regression of the relationship between actions participants took and characteristics of their internet browsing behavior. The outcome variable is log(actions taken) + 1.

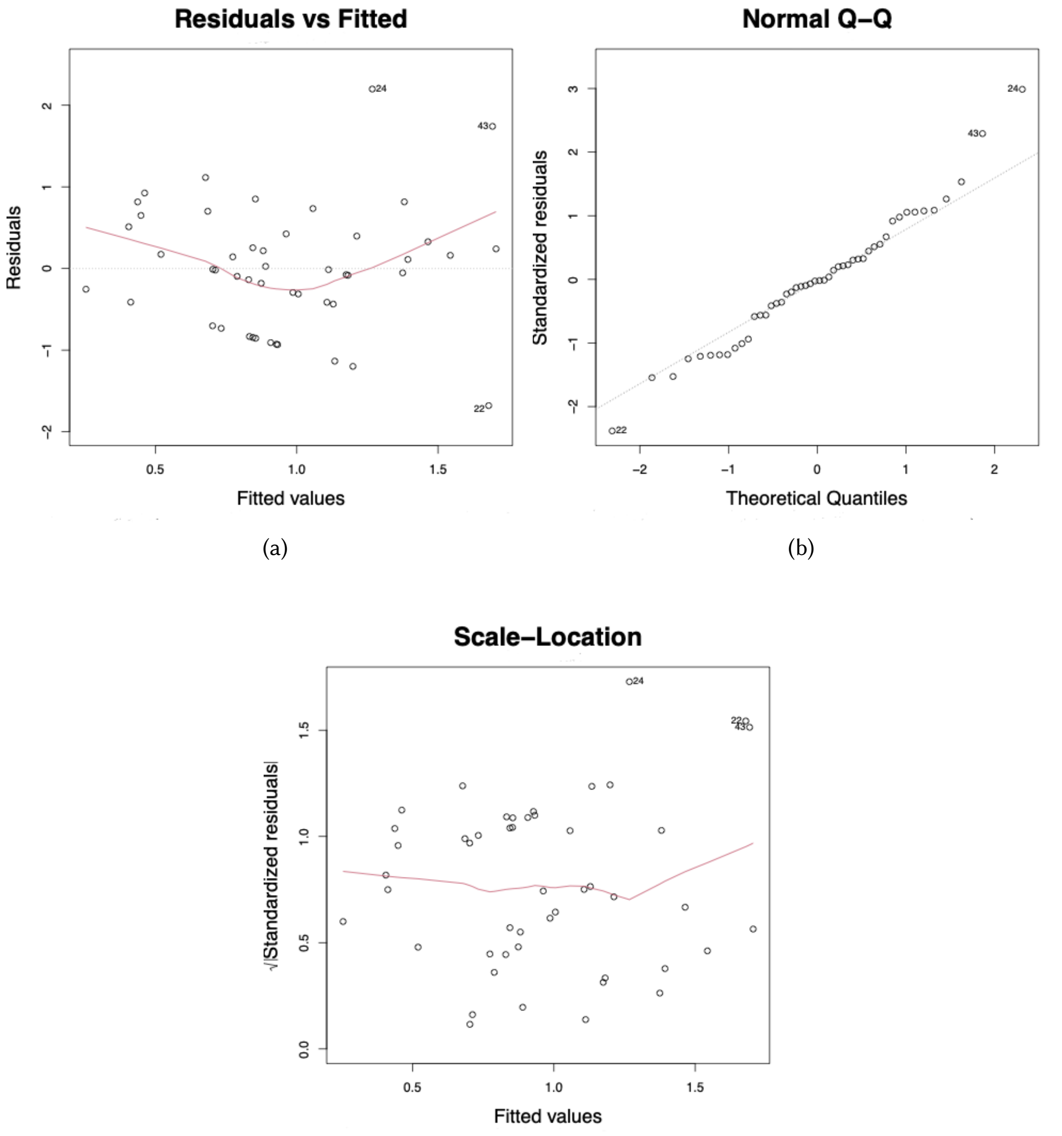


(a)

(b)

(c)

Fig. 5. Scatterplots depicting approximate linearity, normality of residuals, and homogeneity of variance for the model studying actions in relation to browsing behavior. The linearity plot (a) shows an approximately horizontal distribution of the points scattered around the red line with no particular pattern. The normality of residuals plot (b) shows that most of the points approximately fall along the diagonal line. The homogeneity of variance plot (c) shows an approximately horizontal line with the points evenly scattered around it.



(a)

(b)

(c)

Fig. 6. Scatterplots depicting approximate linearity, normality of residuals, and homogeneity of variance for the model studying actions in relation to demographics. The linearity plot (a) shows an approximately horizontal distribution of the points scattered around the red line with no particular pattern. The normality of residuals plot (b) shows that most of the points approximately fall along the diagonal line. The homogeneity of variance plot (c) shows an approximately horizontal line with the points evenly scattered around it.

H Confirmatory study survey

The following survey contains questions about your computer usage and other behaviors. In some questions, we are specifically asking about the computer on which you have installed the SBO software, which we refer to as "**SBO computer**" throughout.

(1) For each of the following events, please indicate whether you are familiar with the event. [*1=Not at all familiar, 2=Slightly familiar, 3=Somewhat familiar, 4=Moderately familiar, 5=Extremely familiar*]
   (a) Hurricane Katrina
   (b) Yahoo! passwords breach
   (c) Airbnb social security number breach
   (d) Russia meddling in the 2016 presidential elections
   (e) Equifax data breach
   (f) The 2018 Royal wedding
   (g) WannaCry ransomware attack
   (h) Panama papers leak
   (i) 2018 Soccer World Cup

(2) (If answer to 1.e >= Somewhat familiar) Was your personal information leaked during the **Equifax data breach** (i.e., was your data stolen)? [*Yes, No, Not sure*]

(3) (If answer to 1.e >= Slightly familiar) Did you take any of the following actions following the **Equifax data breach**? (check as many as apply)
   (a) Can't remember
   (b) Didn't do much/didn't do anything
   (c) Read more about it online
   (d) Read more about it somewhere else
   (e) Visited the Equifax website
   (f) Called Equifax
   (g) Informed myself about the breach in another way
   (h) Froze my credit report
   (i) Other: _________

(4) When you **read about the Equifax data breach online** or **visited the Equifax site**, did you do so on your SBO computer or on another device (i.e., any other laptop/desktop/mobile/tablet)? [*Yes, No, Not sure*]
   (a) On your SBO computer
   (b) On another device

(5) (If answer to 1.b >= Somewhat familiar) Was your password stolen in the **Yahoo! data breach**? [*Yes, No, Not sure*]

(6) (If answer to 1.b >= Slightly familiar) Did you take any of the following actions following the **Yahoo! data breach**? (check as many as apply)
   (a) Can't remember
   (b) Didn't do much/didn't do anything
   (c) Read more about it online
   (d) Read more about it somewhere else
   (e) Informed myself about the breach in another way
   (f) Changed my Yahoo! password
   (g) Other: _________

(7) (If answer to 1.g >= Slightly familiar) Did you take any of the following actions following the **WannaCry attack**?

(a) Can't remember
(b) Didn't do much/didn't do anything
(c) Read more about it online
(d) Read more about it somewhere else
(e) Informed myself about the breach in another way
(f) Paid ransom
(g) Downloaded software patch
(h) Other: _________

(8) Have you ever been affected by some data breach or computer attack ***other than*** the Equifax breach, the Yahoo! passwords breach, or WannaCry? [*Yes*, *No*, *Not sure*]

(9) In general when you **read web pages** (e.g., news articles, links you clicked on) **about data breaches** (e.g., Equifax, Yahoo! passwords breach, Ashley Madison breach, Target credit card data breach), how often do you read them on your SBO computer or on another device (i.e., any other laptop/desktop/mobile/tablet)? [*Never*, *Rarely*, *Sometimes*, *Often*, *Always*]
(a) On your SBO computer
(b) On another device

(10) More generally, over ***all*** of your web browsing, what percentage of it do you on your SBO computer vs. on any other device (i.e., any other laptop/desktop/mobile/tablet)? [*0%/25%/50%/75%/ 100% on your SBO computer*]