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

Sruti Bhagavatula
Carnegie Mellon University
srutib@cmu.edu

Lujo Bauer
Carnegie Mellon University
lbauer@cmu.edu

Apu Kapadia
Indiana University Bloomington
kapadia@indiana.edu

***Abstract*—**
Awareness of security and privacy risks is important for developing good security habits. Learning about real-world security incidents and data breaches can alert people to the ways in which their information is vulnerable online, thus playing a significant role in encouraging safe security behavior. This paper studies how often people read about security incidents online by quantitatively examining real-world internet-browsing data from 303 participants.

Our findings present a bleak view of awareness of security incidents. Only 16% of participants visited any web pages related to six widely publicized large-scale security incidents; few participants read about an incident even when an incident was likely to have affected them (e.g., the Equifax breach almost universally affected people with Equifax credit reports). While our findings suggest that increased security awareness may be necessary, it highlights that current awareness is so low that burdening users with the responsibility of awareness and action may not be effective. We conclude with a discussion of how user safety can be achieved without relying on user awareness.

***Index Terms*—data breaches, security, privacy, awareness**

I. 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, accounts, and data safe [26]. High-profile incidents and breaches in the past decade such as WannaCry, Heartbleed, Petya, and NotPetya have compromised over 300,000 systems worldwide [21], [28], [47]. The data compromised has ranged from passwords to credit card numbers and social security numbers. In addition to the need for affected people to become aware of incidents and take action, it is important for people to generally be aware of the extent and effects of security incidents and as a result people are more likely to implement better security practices [29], [26].

To this end, research about awareness of security incidents, relying on surveys and interviews, has found that people learn about breaches from a variety of sources and that some breaches are more likely to be talked about than others [14]. One survey found that almost half of the respondents heard about a breach from a source other than the breached company [1]. Overall, these studies provide an important step towards understanding how people learn about security incidents. However, research thus far has relied largely on participants' recollection of past behavior or hypothetical situations, and so is constrained by common limitations of self-reported methodologies [53], [23], [55], [50], [18], [25].

In this paper, we take a significant step towards a more detailed understanding of how often people learn about incidents, specifically through online browsing. For a set of six national-scale security incidents of potentially varying relevance to people, we use *longitudinal, real-world browsing data* to examine to what extent people become aware of these incidents. Keeping in mind an underlying goal of improving the spread of incident information through online media, we specifically study this problem in the context of *online browsing*, without considering other channels through 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, among other types of data, all URLs visited and the salted hashes of passwords used to log onto online services from participants' home computers. We further conducted a follow-up survey of 109 participants asking about their device usage and knowledge of security incidents to confirm that our longitudinal measurements contain enough data to support our conclusions. The results of this follow-up survey suggest that the browsing data in our dataset represent more than half of participants' overall browsing (59%). We discuss the limitations of our data set in Sec. VI.

We examine *how often* people read about incidents on the web and whether the likelihood of reading about incidents is associated with demographics, browsing habits, or self-reported security behaviors. 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. In particular, only 15 of 59 likely Equifax credit-report holders read about the breach online in our dataset. Furthermore, these numbers remain alarmingly low even after accounting (through our confirmatory survey) for mobile browsing not captured in our dataset. Overall, we found that older and more tech-savvy participants were more likely to read about security incidents on the internet, as were participants with higher self-reported proactive awareness about their security [17] and participants who browse more technology-related web pages.

Overall, our results suggest remarkably low awareness of security incidents. The implications of these results are two-fold:

first, our results suggest that people do not sufficiently engage with information about security incidents. Much work remains both to help people become aware of security incidents and to help guide them towards improved security hygiene. Second, the low rates of engagement may also indicate that increasing awareness is not the most effective avenue for keeping users safe. In particular, systems that people use should take steps to keep their users safe without requiring them to be aware of or maintain their own security.

II. Related work

Existing work about security awareness and the dissemination of security information, including about security incidents, studied 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, previous work has 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 their participants' information sources [14]. Prior work also found that people use Twitter to complain or share opinions regarding security incidents [16] and that conversations about security and privacy drive people to share with and advise others [46], [12]. 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 [41]. Research also found that the sources of security and privacy advice were important factors for people's digital security habits [49] and that the amount of advice that people reported receiving was not distributed evenly among economic classes [48]. Researchers have also looked at the ways in which presenting people with security information may help convince them to adopt better security practices [15], [13], [3].

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

III. Data collection and dataset

We obtained data collected as part of the Security Behavior Observatory (SBO) project, a longitudinal study of the security behaviors of Windows computer users [19], [20], [43] from October 2014 to July 2019. Data collected by the SBO includes information about system configuration, system events, operating system updates, installed software, and browser-related data such as browsing history, browser settings, and the presence of browser extensions. 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 is approved by the ethics review boards at our institution and the SBO's home institution. SBO data has been used to study, for example, private-browsing habits [23], people's ability to detect phishing attacks [10] and password reuse habits [43], [6]. The SBO dataset contains data about a broad range of people across multiple demographics. We describe these demographics further in Sec. IV-B.

Our study is based on longitudinal data collected by the browser extensions. In particular, we use the following two sets of data.

Browsing history: The browsing data we analyze spans a subset of the whole SBO dataset from October 2014 to June 2018, encompasses 505 participants, and covers participants' browsing on their main computers using Google Chrome, Mozilla Firefox, and Internet Explorer. We study a subset of 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. IV). 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 data about 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 filter this dataset to exclude passwords used during failed login attempts or entered by a user other than the main computer user by replicating the filtering process used by prior work that examined passwords collected through the SBO [6].

We discuss the limitations of this dataset in Sec. VI.

IV. Who reads about security incidents

Here we examine how many and which people visit security-incident-related web pages, as well as what factors are associated with their likelihood of visiting such a page.

A. Methodology

1) Identifying who reads about security incidents: We examine six security incidents that occurred between 2013 and 2017 [33], [2], [56] that are significant enough that we would expect most people to have read about *at least one* incident. We selected these security incidents because they 1) were large-scale incidents (not affecting only a local population), 2) spanned a variety of incident types from personal financial data losses to company document leaks to cyber attacks on home computers, and 3) were represented in our browsing history dataset. We study the following incidents:

- **Equifax breach:** September 2017 breach of the credit reporting site that compromised the personal information of almost 150 million customers [8].
- **Uber hack:** Late 2016 breach that compromised the personal information of 57 million Uber users [35].
- **Ashley Madison breach:** Data breach on the affair-centric dating site in July 2015 and compromised around 33 million users' private information [37].

- **Panama Papers:** April 2016 breach of 11.5 million files from the database of the world's fourth largest offshore law firm, Mossack Fonseca [27], [4].
- **WannaCry:** Ransomware attack in May 2015 that initially affected over 70,000 computers across 99 countries [5], [47].
- **Yahoo! breaches:** Two breaches: one in late 2014 affecting over 500 million user accounts and another in 2013 affecting over 1 billion user accounts [45], [22]. It was later revealed that all user accounts were hacked [34].

Each incident we study may be relevant to users in different ways, e.g., 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 security and privacy dangers in a broader context. 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 the safety of 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 focus on the participants who were active in the study before the incident became public and for three months after.

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", "hack". To confirm that our keyword lists were inclusive and robust enough, 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 of the users or consumers of the respective organizations [34], [24], [8].

We determined people who were likely to have had an Equifax credit report by those who entered passwords on credit-card reporting sites that reported on Equifax credit ratings. To determine who was likely to have an Equifax credit report (and were hence likely to be affected by the breach [8]), we searched for participants who had entered a password on `identityforce.com`, `identityguard.com`, `annualcreditreport.com`, `creditsesame.com`, `creditkarma.com`, `quizzle.com`, or `equifax.com` before September 7, 2017, which is when the breach became public. We picked these seven domains because six were on a list of six popular credit-report sites reporting Equifax scores [11] and one was the Equifax homepage itself. While most Americans were likely to have been affected [24], [8] regardless of whether they had an account with a credit-reporting site, we considered this set of participants *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 February 15, 2017—when the breach had first become public. We repeated this search for participants who had a Yahoo! account before the second breach announcement, October 3, 2017.

Since we had access to passwords only for Chrome and Firefox users, these are the only participants that we can judge were affected by the Yahoo! or the Equifax breach.

2) Studying which people read about incidents: After determining which participants read about at least one of the incidents, we study what characteristics of participants are correlated with them visiting pages related to the security incidents. We model participants and their behavior using three distinct feature sets and then perform a logistic regression for each feature set, where the outcome variable in each regression is a binary variable indicating whether a participant read about an incident.

Feature set 1: Demographic characteristics Based on findings from prior work showing that demographics were correlated with how people share security and privacy news and their comfort with uses of breached data [14], [31], we hypothesized that certain demographics would also be correlated with whether they read about a security incident. Therefore, the first feature set contains demographic information about each participant: age, gender, income, highest education level, whether the participant is a student, whether the participant's primary profession involves programming, and whether the participant knows at least one programming language.

Feature set 2: Self-reported security intentions Prior work found self-reported security intentions (as measured by the SeBIS scale [17]) to be correlated with how people heard about and shared security and privacy news [14]. Hence, our second feature set comprises the four continuous feature values of the SeBIS scale [17], which participants optionally filled out upon enrollment in the SBO. The four values represent the extent to which participants a) secure their devices, b) generate strong and varied passwords across accounts, c) demonstrate proactive awareness of security issues or safety of websites and links, and d) update the software on their computers.

Feature set 3: Participants' observed internet behavior We hypothesized that the types of web pages people browse would be correlated with people's likelihood to encounter information about a security incident. For example, we hypothesized that

people who browse more technology-related news articles may be more likely to come across web pages about security incidents. To test these hypotheses, we examined two types of internet behaviors: 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 categorize 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, resulting in a multiset (i.e., bag) of words. We performed topic modeling using the Non-negative Matrix Factorization (NMF) algorithm [36] 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 [52]). One topic appeared to correspond to browsing that was professional or work-related; the other topic related to browsing that was leisure- or entertainment-related. (See App. A for more details.) The output of NMF also includes, for each topic and participant, a value that describes how much of that participant's web browsing matches the topic. Hence, each participant's browsing behavior is characterized with two features (corresponding to two topics).

Amount of technical content browsed: We further characterize people's browsing according to how many of the web pages they visit are technology-related. We again use the NMF algorithm to build a topic model with two distinct topics: one topic covering technology-related content and the other comprising all other types of content. This topic model is computed over the content of web pages visited by the participants. We consider a web page to be technology-related if the AWIS category for the domain of the web page contains the word "technical" or "technology"; otherwise we consider it not technology-related. For a sample of each participant's browsing history, we download the content of the web pages in the sample using the newspaper library [42] and train a topic modeling algorithm 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 with a non-technical AWIS category. We then apply this topic model trained for two topics on the multiset of tokens of downloaded content for each participant's browsing sample. Similarly to when characterizing 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 details.)

B. Results

Our filtered set of participants includes 303 participants who were active in the study around the time of at least one incident announcement. The participants span a broad range of demographics (see Table V in App. C for details).

Incident	*# participants*
Equifax	26
Yahoo!	6
Uber	4
Ashley Madison	6
WannaCry	14
Panama Papers	10

TABLE I: Number of participants who read about each security incident; some read about multiple incidents.

We were surprised to discover that only 48 of the 303 (16%) visited a web page that discussed *any* of the six security incidents.[1] In three additional instances participants searched for incident-related keywords but did not visit any of the search results or other incident-related pages. Table I shows how many participants visited a page about each incident.

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. IV-A1). Of the 59 participants who entered passwords on credit-card reporting sites that reported on Equifax credit ratings, 15 (25%) read about[2] the Equifax breach. In contrast, of the 48 participants who entered a Yahoo! password in the relevant time periods—and hence were definitely affected by the breach [34]—only one (2%) read about the breach.

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. IV-A1 by computing three logistic regression models. When interpreting results, we used a significance level of 0.05.

First, we computed a model exploring the effect of demographic characteristics over the 303 participants (Table II). We found that participants' ages and whether they know a programming language were significant factors. Specifically, older and more technology-savvy participants were more likely to read about incidents. The effect of age was only marginal—the odds of reading about an incident increased by $1.003\times$ ($p = 0.02$) for each additional year of age—but the odds of reading about an incident increased by $1.149\times$ ($p = 0.002$) if a participant knew a programming language.

Our second model examines the relationship between whether participants read about an incident and their self-reported SeBIS scale values (Table III). This model was computed over 247 participants who provided SeBIS data to the SBO at the time of enrollment. Only one of the four SeBIS scale values was statistically significant, which we model by its Z-score for easier interpretation; the odds of reading about an incident were increased by a factor of 1.078 ($p = 0.05$) for each standard deviation increase in the SeBIS proactive awareness score of a participant (a value in $[0, 1]$).

[1] While not all participants may be interested in *every* incident, the incidents we study were chosen so that the majority of participants found one or more incidents *relevant to* them in some way.

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

	baseline	*coef.*	*exp(coef.)*	*std.err.*	*t*	*p*
(Intercept)		-0.070	0.933	0.074	-0.953	0.341
age		**0.003**	**1.003**	**0.001**	**2.311**	**0.021**
gender: male	female	0.025	1.026	0.035	0.715	0.475
Education: $\geq$ ugrad	<ugrad	0.026	1.027	0.035	0.748	0.455
Income: >$25k	<$25k	0.002	1.002	0.040	0.053	0.957
Income: declined to answer	<$25k	-0.031	0.969	0.056	-0.562	0.574
knows_prog_lang: yes	**no**	**0.139**	**1.149**	**0.045**	**3.084**	**0.002**
is_programmer: yes	no	-0.028	0.972	0.049	-0.583	0.560
is_student: yes	no	0.050	1.052	0.047	1.078	0.282

TABLE II: Logistic regression model describing the relationship between whether a participant learned about a breach and characteristics of the participant including their demographics. "ugrad" denotes that the participant indicated achieving a Bachelor's degree.

	coef.	*exp(coef.)*	*std.err.*	*t*	*p*
(Intercept)	-0.180	0.835	0.178	-1.013	0.312
Device Securement	0.000	1.000	0.018	0.009	0.993
Password Generalization	0.020	1.021	0.039	0.535	0.593
Proactive Awareness	**0.075**	**1.078**	**0.038**	**1.990**	**0.047**
Updating	0.002	1.023	0.023	0.969	0.333

TABLE III: Logistic regression model describing the relationship between whether a participant learned about a breach and the SeBIS scale values they provided.

Our third model examines the relationship between whether participants read about an incident and their internet browsing behavior (i.e., browsing topics and amount of technical browsing; see Sec. IV-A2). This model 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). The odds of reading about an incident are increased by a factor of 0.026 ($p < 0.001$) for every standard deviation increase in the technical browsing score. Table IV contains the results of this logistic regression model.

C. Summary of findings

Overall, participants who were older, exhibited a higher proactive awareness about computer security, and those who were more technology-inclined were more likely to come across security incident information online. This finding indicates an imbalance in the dissemination of important security information to different user demographics and populations.

V. Confirming dataset validity

Our findings are based on the browsing activity collected from each participant's main computer. 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 to supplement our main dataset. We conducted a survey (see App. D) of 109 SBO participants who were active in May 2019, in which we asked about their familiarity with, and any reactions in response to, several security incidents, as well as about how much web browsing they perform on which devices. This study was approved by our institution's review board and the review board of the SBO's home institution. The survey took between one and five minutes and participants were compensated with $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 focusing on specific individuals who were in both datasets. However, our survey results for the 109 participants described below are consistent with the results computed over the overlapping 84 participants only. Since the 109 participants may or may not have been active in the SBO around at least one of the incidents we studied, we did not ask survey participants about each of the incidents. Instead, we asked about a variety of events including a few security incidents and asked follow-up questions about incidents with a wide impact.

	coef.	*exp(coef.)*	*std.err.*	*t*	*p*
(Intercept)	0.182	0.156	1.200	1.172	0.242
Browsing: leisure	-0.423	0.423	0.655	-0.999	0.319
Browsing: professional	-0.387	0.286	0.679	-1.353	0.177
Z(Browsing: technical)	**0.087**	**0.026**	**1.090**	**3.377**	**<0.001**
Overall pwd reuse	0.264	0.200	1.303	1.321	0.188
Common pwd reuse	-0.206	0.148	0.8137	-1.392	0.165

TABLE IV: Logistic regression model describing the relationship between whether a participant learned about a breach and characteristics of their internet browsing behavior.

When asked to report the fraction of browsing they performed on their SBO computers, participants indicated that they used them, on average, for 59% of their web browsing. As the amount of browsing on desktop and laptop computers may have decreased over time in favor of browsing on mobile devices [39], this 59% is likely a lower bound; participants earlier in the study likely performed a higher fraction of their overall browsing on their SBO computers.

We also asked participants how often (on a 5-point Likert scale [54]) they read about security incidents on (1) their SBO computer and (2) on any other devices. We found no statistically significant difference between the two distributions (Kolmogorov-Smirnov [32]: $D = 0.056, p = 0.997$). We also examined whether the distribution of browsing on SBO computers vs. other devices varied by participant age, but found no significant indication that it did (Spearman's correlation test: $S = 180990, p = 0.155$). Finally, to gauge the accuracy of the self-reported data, we asked participants how familiar they 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% of respondents indicated moderate or extreme familiarity with the fake Airbnb breach, suggesting that the self-reported results may slightly exaggerate actual familiarity with incidents.

Our results from Sec. IV indicate that 25% of the participants in our main dataset who were likely to have an Equifax credit report (and therefore, likely to have been affected) read about the Equifax 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 27%, which is still low considering the significance of the breach. When asked what action they took following the Equifax breach, 41 of the 86 participants who indicated at least slight familiarity with the breach (48%) responded that they read about the breach online and/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. IV indicate that 2% of participants in our main dataset who had a compromised Yahoo! password [34] read about the breach online. Self-reported data also suggests very 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 use for our analyses covers the majority (with a lower bound of approximately 59%) of the browsing performed by our participants. While the additional browsing participants performed on non-SBO devices may dilute some of our findings about how often people read about incidents, the self-reported data supports the big picture: a surprisingly small subset of users reads about incidents.

VI. Limitations

Although our work provides valuable insights into how often and which people come across security incidents, it is subject to a few limitations, including those due to the nature of the data collection.

Our dataset contains data about (relatively) few participants due to the difficulty of recruiting participants to the SBO. However, the SBO data represents a tradeoff: it offers 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 that our results reveal that matters—that a very small fraction of people seem to engage with information about security incidents—rather than the specific percentages involved. Similarly, people's desktop browsing history may not be representative of all the browsing they perform. Hence the confirmatory study, which suggests that the high-level results of the original SBO analysis hold: participants were rarely familiar with or read about major security incidents regardless of the devices on which they browsed the internet.

Despite the richness of our dataset, it does not capture information about other sources from which people became aware of incidents or participants' reasons for their behavior. However, the goal of our study is to study consumption of incident information through online browsing based on empirical measurements of browsing behavior, therefore, our analyses do not rely on the information that is not captured.

Since browsing history was represented via URLs and page titles, we could not include analyses that depended on the content of pages that are dynamic (e.g., social network or news pages with endless scrolling). We also could not distinguish between content that participants actually consumed and content they loaded but did not read.

The data we analyzed is collected only from Windows computer users. Users of non-Windows operating systems might exhibit behaviors different from the behaviors of the participants in our dataset. However, as Windows is the dominant OS for personal computers [9], we do not believe this limitation is likely to fundamentally affect our findings.

Although data from SBO participants has been used for several security- or privacy-related studies [43], [23], [10], [20], the SBO participants may be biased towards less privacy- and security-aware people, given the nature of the SBO data collection infrastructure.

VII. Discussion and Conclusions

Using the actual browsing histories of 303 participants over four years, we measured how often people read web pages about security incidents and the factors associated with how likely they are to read about an incident. Our findings are bleak: Only a small minority (16%) of participants visited a web page about any of six large-scale security incidents.

Our results highlight the challenges of increasing awareness of security incidents and of disseminating information about them. Even when an incident was highly publicized and participants were likely to have been affected, few showed engagement with or awareness of the incident through the webpages they browsed, e.g., only 25% of likely Equifax credit report holders read a webpage about the Equifax breach online. Without adequate awareness, it is unlikely that people will act to improve their security.

On the other hand, our analysis yielded a number of negative results in terms of what is correlated with coming across incident information, suggesting that there may be a deeper issue in terms of how concerned people are about incidents in the first place. Further research is needed to identify more effective ways for people to stay safe and secure beyond increased awareness of and engagement with security advice. For example, companies may be able to take immediate steps to protect their users after a security incident without requiring awareness on the users' part. After password breaches, for example, companies can force a reset on all passwords. For widespread cyberattacks on specific platforms, 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.

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] L. Ablon, P. Heaton, D. C. Lavery, and S. Romanosky. *Consumer attitudes toward data breach notifications and loss of personal information*. Rand Corporation, 2016.

[2] T. Armerding. The 17 biggest data breaches of the 21st century. *CSO Online*, 2018.

[3] F. Asgharpour, D. Liu, and L. J. Camp. Mental models of security risks. In *FC*, 2007.

[4] Panama papers Q & A: What is the scandal about? *BBC News*, 2016.

[5] Cyber-attack: Europol says it was unprecedented in scale. *BBC News*, 2017.

[6] S. Bhagavatula, L. Bauer, and A. Kapadia. (How) Do people change their passwords after a breach? In *ConPro*, 2020.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JML*, 2003.

[8] D. Borak and K. Vasel. The Equifax hack could be worse than we thought. *CNNMoney*, 2018.

[9] E. Bott. Latest OS share data shows Windows still dominating in PCs. *ZDNet*, 2013.

[10] C. Canfield, A. Davis, B. Fischhoff, A. Forget, S. Pearman, and J. Thomas. Replication: Challenges in using data logs to validate phishing detection ability metrics. In *SOUPS*, 2017.

[11] H. Catalano. Best free credit report site of 2018. *The Simple Dollar*, 2018.

[12] S. Das, T. H. Kim, L. A. Dabbish, and J. I. Hong. The effect of social influence on security sensitivity. In *SOUPS*, 2014.

[13] S. Das, A. D. I. Kramer, L. A. Dabbish, and J. I. Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *CCS*, 2014.

[14] S. Das, J. Lo, L. Dabbish, and J. I. Hong. Breaking! A typology of security and privacy news and how it's shared. In *CHI*, 2018.

[15] P. DiGioia and P. Dourish. Social navigation as a model for usable security. In *SOUPS*, 2005.

[16] P. Dunphy, V. Vlachokyriakos, A. Thieme, J. Nicholson, J. C. McCarthy, and P. Olivier. Social media as a resource for understanding security experiences: A qualitative analysis of #password tweets. In *SOUPS*, 2015.

[17] S. Egelman and E. Peer. Scaling the security wall: Developing a security behavior intentions scale (SeBIS). In *CHI*, 2015.

[18] X. Fan, B. C. Miller, K.-E. Park, B. W. Winward, M. Christensen, H. D. Grotevant, and R. H. Tai. An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, 2006.

[19] A. Forget, S. Komanduri, A. Acquisti, N. Christin, L. F. Cranor, and R. Telang. Building the security behavior observatory: an infrastructure for long-term monitoring of client machines. In *HotSoS*, 2014.

[20] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. Telang. Do or do not, there is no try: user engagement may not improve security outcomes. In *SOUPS*, 2016.

[21] J. Fruhlinger. The 6 biggest ransomware attacks of the last 5 years. *CSO*, 2019.

[22] V. Goel and N. Perlroth. Yahoo says 1 billion user accounts were hacked. *The New York Times*, Dec 2016.

[23] H. Habib, J. Colnago, V. Gopalakrishnan, S. Pearman, J. Thomas, A. Acquisti, N. Christin, and L. F. Cranor. Away from prying eyes: analyzing usage and understanding of private browsing. In *SOUPS*, 2018.

[24] R. Hackett. Equifax underestimated by 2.5 million the number of potential breach victims. *Fortune*, 2017.

[25] A. Hanamsagar, S. S. Woo, C. Kanich, and J. Mirkovic. Leveraging semantic transformation to investigate password habits and their causes. In *CHI*, 2018.

[26] B. Hanus and Y. A. Wu. Impact of users' security awareness on desktop security behavior: A protection motivation theory perspective. *ISM*, 2016.

[27] L. Harding. What are the panama papers? A guide to history's biggest data leak. *The Guardian*, 2016.

[28] R. Horev. The top 7 vulnerabilities of the decade. *Vulcan*, 2018.

[29] H. B. Janine L. Spears. User participation in information systems security risk management. *MIS Quarterly*, 2010.

[30] B. Ju, M. Ye, Y. Qian, R. Ni, and C. Zhu. Modeling behaviors of browsing and buying for alidata discovery using joint non-negative matrix factorization. In *CISIS*, 2014.

[31] S. Karunakaran, K. Thomas, E. Bursztein, and O. Comanescu. Data breaches: User comprehension, expectations, and concerns with handling exposed data. In *SOUPS*, 2018.

[32] A. Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 1933.

[33] S. Larson. 10 biggest hacks of 2017. *CNNMoney*, 2017.

[34] S. Larson. Every single Yahoo account was hacked. *CNNMoney*, 2017.

[35] S. Larson. Uber's massive hack: What we know. *CNNMoney*, 2017.

[36] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[37] N. Lord. A timeline of the Ashley Madison hack. *Digital Guardian*, 2017.

[38] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. 2008.

[39] M. Meeker. Internet trends report 2018. 2018.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[41] J. Nicholson, L. Coventry, and P. Briggs. If it's important it will be a headline: Cybersecurity information seeking in older adults. In *CHI*, 2019.

[42] L. Ou-Yang. Newspaper3k: Article scraping & curation: http://newspaper.readthedocs.io/en/latest/. *Python Software Foundation*, 2013.

[43] S. Pearman, J. Thomas, P. E. Naeini, H. Habib, L. Bauer, N. Christin, L. F. Cranor, S. Egelman, and A. Forget. Let's go in for a closer look: Observing passwords in their natural habitat. In *CCS*, 2017.

[44] M. Pennacchiotti and S. Gurumurthy. Investigating topic models for social media user recommendation. In *WWW*, 2011.

[45] N. Perlroth. Yahoo says hackers stole data on 500 million users in 2014. *The New York Times*, 2016.

[46] E. J. Rader, R. Wash, and B. Brooks. Stories as informal lessons about security. In *Symposium On Usable Privacy and Security, SOUPS*, 2012.

[47] S. Ranger. Ransomware attack: The clean-up continues after WannaCry chaos. *ZDNet*, 2017.

[48] E. M. Redmiles, S. Kross, and M. L. Mazurek. How I learned to be secure: a census-representative survey of security advice sources and behavior. In *CCS*, 2016.

[49] E. M. Redmiles, A. R. Malone, and M. L. Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *S&P*, 2016.

[50] R. Rosenman, V. Tennekoon, and L. G. Hill. Measuring bias in self-reported data. *International journal of behavioural & healthcare research*, 2011.

[51] A. Shimada, F. Okubo, and H. Ogata. Browsing-pattern mining from e-book logs with non-negative matrix factorization. In *EDM*, 2016.

[52] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring topic coherence over many models and many topics. In *EMNLP-CoNLL*, 2012.

[53] B. Ur, J. Bees, S. M. Segreti, L. Bauer, N. Christin, and L. F. Cranor. Do users' perceptions of password security match reality? In *CHI*, 2016.

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

[55] R. Wash, E. Rader, and C. Fennell. Can people self-report security accurately?: Agreement between self-report and behavioral measures. In *CHI*, 2017.

[56] K. Zetter. The year's 11 biggest hacks, from Ashley Madison to OPM. *Wired*, 2015.

APPENDIX

A. Characterizing browsing behavior

To identify topics that characterize participants' browsing, we apply the Non-negative Matrix Factorization (NMF) topic-modeling algorithm [36] to participants' browsing histories. NMF has been used in prior work for mining browsing behavior patterns [51], [30].

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 webpage 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 [38] 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 [52], [44]. 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 [40] 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.). As mentioned above, 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) [7] 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 is technology-related, we again used the NMF topic modeling algorithm to build a topic model that has 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 [42], 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 include tokens that appear in one document but not the other. This way we are able to construct topics with tokens that are unique to either the technical or non-technical category.

After training the two-topic topic model, we determine the index of the column in the resulting document-topic matrix that corresponds to the technical document and therefore to the technical topic. We apply 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 computes two weights for each participant, of which we use the weight of the technical topic as the feature characterizing the amount of a participant's browsing related to technical content.

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	$\leq$ 50k: 50% 50k-100k: 24% 100k-200k: 10% $\geq$ 200k: 2% Did not provide: 14%
Is_student	Students: 48%

TABLE V: Demographic distribution of the 303 participants.

C. Demographics

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

D. 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*]