Where is the Digital Divide? A Survey of Security, Privacy, and Socioeconomics

Elissa M. Redmiles
University of Maryland
eredmiles@cs.umd.edu

Sean Kross
Johns Hopkins University
kross@jhu.edu

Michelle L. Mazurek
University of Maryland
mmazurek@cs.umd.edu

ABSTRACT

The behavior of the least-secure user can influence security and privacy outcomes for everyone else. Thus, it is important to understand the factors that influence the security and privacy of a broad variety of people. Prior work has suggested that users with differing socioeconomic status (SES) may behave differently; however, no research has examined how SES, advice sources, and resources relate to the security and privacy incidents users report. To address this question, we analyze a 3,000 respondent, census-representative telephone survey. We find that, contrary to prior assumptions, people with lower educational attainment report equal or fewer incidents as more educated people, and that users' experiences are significantly correlated with their advice sources, regardless of SES or resources.

ACM Classification Keywords

H.1.2 User/Machine Systems: Human factors

Author Keywords

usable security, digital divide, computer science education

INTRODUCTION AND BACKGROUND

Security and privacy are intrinsically collective behaviors—one person who clicks on a malicious email may spread that email to their entire network. As such, it is important to understand how differing knowledge, habits, and priorities can shape security behavior and associated experiences across a broad range of people.

Previous research has established the existence of a *digital divide*: an access, skill, and knowledge gap in digital literacy between lower- and higher-socioeconomic status (SES) populations [11, 12, 14, 28, 32, 33]. However, the bulk of research on this topic has not directly addressed security and privacy. Some researchers have theorized that low-income users "do not value their data as highly," that low-SES users may experience discrimination as a result of differing privacy norms, or may be unable to pay for the increasingly monetized privilege of privacy, leading to worse outcomes [9, 18, 19]. Empirically, we know that variation in SES, along with differences in skills and advice sources that may be partially correlated with SES, are associated with differences in users' security and privacy beliefs and behaviors [10, 13, 23–25, 27, 31, 34]. No prior work, however, has examined how users' SES, advice sources, and resources relate to each other and to the security and privacy experiences that users report.

To better understand this intersection of demographics, information, and experiences—particularly with respect to low-SES users—we analyzed data from a probabilistic, census-representative, telephone survey of 3,000 U.S. respondents, which we received through a data grant from Data&Society. The survey over-sampled from the low-SES population. The thoroughly pretested survey queried respondents' security and privacy experiences, including becoming the victim of a scam, having your identity stolen, having an email or social media account compromised, losing a job or other opportunity as a result of something posted online, and having someone post something about you online without consent; other questions examined respondents' advice sources, available internet resources, and demographics. The relationships identified in our analysis, and the prevalence of the experiences reported by respondents, are accurate within 2.7% of their true values in the entire U.S. population.

In line with prior work, we find that less educated users have different sources of security advice than more educated users. Contrary to prior assumptions, however, we find that low-education users report equal or lower prevalence of negative security and privacy incidents as compared to higher-education users, that there is no relationship between prevalence of reported incidents and income, and that users' reported experiences are related to their advice sources, regardless of their SES or resources. These findings have important implications for how we develop, distribute, and evaluate security and privacy advice, as well as how we think about the digital divide from a security and privacy perspective.

METHODOLOGY

To examine the relationship between SES, advice sources, resources, and self-reported security and privacy incidents, we modeled the results of a 3,000-respondent telephone survey using binary logistic regression. Our Institutional Review Board (IRB) determined that our analysis of existing data did not constitute human subjects research. Below, we discuss the dataset and survey development process, sampling procedure, details of our statistical analysis, and limitations of our work.

CHI 2017, May 6-11, 2017, Denver, CO, USA.

http://dx.doi.org/10.1145/3025453.3025673

Dataset

The dataset was collected by Princeton Survey Research Associates International (PSRAI) for Data&Society via a computer-assisted-telephone-interview (CATI), random digit dial (RDD) census-representative survey of 3,000 respondents from November 18 to December 23, 2015. We received this dataset through a Data Grant from Data&Society (funded by the Digital Trust Foundation). [1]

The survey was developed by a senior researcher at Data&Society with the intent of releasing the data for analysis regarding the impact of SES on security and privacy. She assembled the survey both by authoring and pre-testing new items and by leveraging a number of pre-tested questions from surveys conducted by Pew and Reason-Rupe [3, 5–7]. The survey asks questions regarding respondents' security and privacy experiences including their advice sources, their prior negative experiences, and the resources available to them, as well as standard demographic questions. The order in which the questions were asked was randomized to prevent order bias [21]. Additionally, demographic questions were placed at the end of the questionnaire to minimize sensitivity and bias, as per expert recommendations and best practices [29].

Prior to deployment, the questionnaire was pretested with a small number of respondents. These interviews were monitored by PSRAI and conducted by experienced interviewers to ensure that respondents understood the questions.

The survey was administered via CATI by professionally trained interviewers in both English and Spanish. Calls were made throughout the day, on multiple days to both landline and cell phones to maximize the chance of connecting with different respondents. Every person in the United States had a non-zero chance of being selected for the survey. [2] This was a probabilistic survey, the dataset was weighted to be representative of the U.S. population, and the findings we report are accurate within 2.7% of the true prevalence in the population. A full outline of the survey items, weighting methodology, and analysis code can be found at go.umd.edu/2124.

Analysis

We built two sets of binary logistic regression models in our analysis, using the survey R library to incorporate the survey weights [22]. The first set of models was used to predict the odds of an individual reporting having experienced a security or privacy incident: one model included respondents' advice sources, another included SES status, and the third combined both. The second set of models predicted the likelihood that respondents with different SES used particular advice sources. We chose a simple grouped model rather than five individual models for ease of interpretability.

To reduce the chance of overfitting our data, we deliberately chose parsimonious models with input factors based on prior work [26]. To further prevent over fitting, we performed 5-fold cross validation in line with commonly used classification and regression practices [15]. We calculated the Akaike Information Criterion (AIC) [8] across five folds for each model, and we found that the AIC values for each fold were within an average of 3% of each other. For each model, we present the outcome variable, including factors, log-adjusted regression coefficients (*odds ratios*), 95% confidence intervals (moderated by the survey design effect [20]), and p-values.

[1] The survey development and deployment portion of this study was approved by Chesapeake IRB [4].

[2] Those who did not have a telephone were contacted via mail and, if interested, were provided with a phone to use for the survey.

Limitations

Self-reported surveys have several common limitations, chiefly related to under- and over-reporting, which may be caused by satisficing (selecting the first satisfactory answer without thinking deeply) [17], recall bias (misremembering experiences), desirability bias (selecting a socially desirable rather than honest answer), and the potential for questions to be misinterpreted. These were mitigated by using thorough question-development and pre-testing processes and by interviewers reminding respondents to answer thoroughly and honestly. The survey was brief, minimizing respondent fatigue.

This survey measures only whether respondents have ever used certain advice sources or had certain negative experiences. As a result, we cannot determine how often a particular advice source was consulted or how many negative experiences a respondent had; nor can we determine whether an advice source was consulted before or after any negative event. Thus, we report our findings together with several hypothetical explanations and suggest that future work should investigate these relationships further. In addition, we did not conduct a controlled experiment, and thus these results should not be interpreted as implying causality.

RESULTS

Below, we describe the survey sample and the factors that relate to users' security and privacy experiences.

Sample

Our unweighted sample was nearly representative of the U.S. population with respect to gender, age, education, geographic region, number of adults in the household, population density, household phone usage, and race/ethnicity. The weighted sample is fully representative of the population, such that the 95% confidence interval for this survey is 2.7 points. This confidence interval is calculated based on the survey design effect, which represents the loss in statistical efficiency that results from a disproportionate sample design and systematic non-response. Table 1 compares a subset of the demographics of our weighted and unweighted sample to the 2013 American Community Survey [2]. Further, the prevalence of negative experiences in our data is in line with prior work. [3]

Security and Privacy Incidents

All internet-using respondents were asked questions regarding negative security and privacy incidents that they had experienced, such as "Have you ever had important personal information stolen, such as your Social Security Number, your credit card, or bank account information?" We find that 49% of all respondents in the weighted data reported at least one of

[3] See go.umd.edu/2124 for a comparison with Pew 2013 data [1].

Metric	Unweighted	Weighted	Census
Male	52.4%	48.7%	48.2%
Female	47.6%	51.3%	51.8%
Caucasian	58.1%	62.8%	65.8%
Hispanic	18.6%	15.6%	15%
African American	14.0%	11.8%	11.5%
Other	6.7%	7.4%	7.6%
LT H.S.	12.8%	12.6%	13.3%
H.S. grad	27.4%	27.8%	28.0%
Some college	24.0%	30.0%	31.0%
B.S. or above	34.6%	28.7%	27.7%
18-29 years	16.3%	20.1%	20.9%
30-49 years	24.6%	32.6%	34.7%
50-64 years	28.8%	25.4%	26.0%
65+ years	27.0%	18.6%	18.4%
<$20k	20%	*NA*	32%
$20k-$40k	21%	*NA*	19%
$40k-$75k	18%	*NA*	18%
$75k-$100k	10%	*NA*	11%
$100k-$150k	8%	*NA*	12%
$150k+	7%	*NA*	8%

Table 1. Sample demographics, percentages may not add to 100% due to non-response. Income was the unweighted variable of interest.

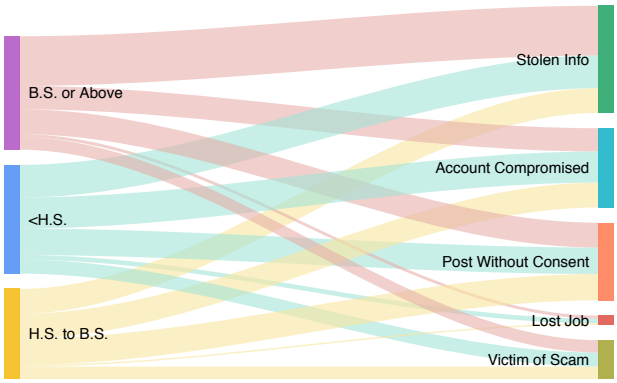

Figure 1. Prevalence of those who reported a negative experience by education level. Interactive diagram: jsfiddle.net/5orqbkp4/3/.

the negative experiences shown in Figure 1. To determine how these reported incidents relate to respondents' SES, advice sources, and resources, we utilized binary logistic regression models to predict a participant's likelihood of reporting one or more of these experiences. We created three models, detailed below. The results of all three models are presented in Table 2.

First, to understand how SES related to respondents' reported security and privacy incidents, we modeled these incidents as a function only of SES factors. In this model, we find that education is the only factor significantly related to a respondent's likelihood of reporting a negative experience (income was not correlated). Surprisingly, we find that those with lower levels of education—less than a high school diploma, and less than a bachelor's—are 60% and 35% *less* likely, respectively, to report at least one of the five negative experiences (Table 2). While 53% of those in the weighted dataset who hold a bachelor's or above reported a negative experience, only 47% of those who had less than a bachelor's reported such an incident.

Model	Factor	OR	CI	*p*-value
SES	<H.S.	0.40	[0.2, 0.81]	0.01*
& Resources Only	H.S. to B.S.	0.65	[0.44, 0.97]	0.03*
	<$20K	1.09	[0.69, 1.74]	0.71
	$20-$40K	1.20	[0.79, 1.84]	0.39
	R: Cell only	0.74	[0.52, 1.06]	0.11
	R: Home internet	1.94	[0.87, 4.32]	0.1
Advice	A: Friend	1.85	[1.25, 2.73]	< 0.01*
Only	A: Website	1.92	[1.15, 3.21]	0.01*
	A: Coworker	1.59	[0.98, 2.58]	0.06
	A: Gov. Website	1.59	[0.87, 2.88]	0.13
	A: Librarian	1.73	[0.75, 4.02]	0.2
	A: Teacher	0.95	[0.45, 2.01]	0.9
Advice	A: Friend	1.84	[1.24, 2.72]	< 0.01*
& SES	A: Website	1.76	[1.06, 2.94]	0.03*
	A: Coworker	1.53	[0.95, 2.46]	0.08
	A: Gov. Website	1.52	[0.85, 2.74]	0.16
	A: Librarian	1.88	[0.82, 4.31]	0.14
	A: Teacher	0.92	[0.44, 1.96]	0.83
	<$20K	1.09	[0.68, 1.76]	0.72
	$20-$40K	1.19	[0.77, 1.83]	0.44
	<H.S.	0.53	[0.25, 1.09]	0.09
	H.S. to B.S.	0.75	[0.5, 1.13]	0.17
	R: Mostly cell	0.77	[0.53, 1.12]	0.17
	R: Home internet	1.73	[0.77, 3.88]	0.18

Table 2. Regression results for three different models of reporting at least one negative experience (binary). 'A' and 'R' indicated boolean advice sources and resources, respectively. "Mostly cell" indicates primary internet access via mobile, and "Home internet" means internet at home. Baseline for the categorical household income factor is >$40K; baseline for education is a bachelor's or above. OR is the odds ratio between the given factor and the baseline; CI is the 95% confidence interval.

Figure 1 illustrates that negative experiences were unevenly distributed across educational groups; 32% of those holding a bachelor's or above reported having information stolen, compared to 20% of those with less than a bachelor's. There are several potential explanations for this finding, which should be explored in future work: less-educated users may be targeted less frequently for scams or identity theft, they may have more difficulty recognizing or recalling negative events, or they may have protective skills or resources not measured in this survey.

Prior work suggests that advice sources are related to users' security behaviors, and therefore potentially to their security experiences [24–26]. Thus, our second model evaluates whether the likelihood of reporting at least one negative incident is significantly related to advice sources. We find that respondents who take advice from friends and websites are 85% and 92% more likely to report at least one negative experience, respectively. Of those who took advice from friends, 49% reported a negative experience, compared to 25% of those who took advice from a co-worker, 21% from a non-governmental website, 14% from a government website, and 8% from a teacher or librarian. This may indicate that respondents more often seek advice from certain sources after a negative experience, that librarians and teachers give particularly good advice, or that respondents are receiving detrimental or difficult to interpret advice from friends, coworkers, and websites. .

Finally, we wanted to understand whether advice and SES were both related to the security and privacy incidents that users report, or whether if we controlled for both variables, only one would remain significant. We therefore constructed

Factor	OR	CI	*p*-value
<H.S.	0.01	[0, 0.06]	< 0.01*
H.S. to B.S.	0.49	[0.31, 0.79]	< 0.01*
<$20K	0.86	[0.42, 1.73]	0.66
$20-$40K	0.66	[0.36, 1.22]	0.19
R: Cell only	0.69	[0.4, 1.18]	0.17
R: Home Internet	1.62	[0.5, 5.24]	0.42

Table 3. Regression results for website advice source model. Tables for all advice sources: go.umd.edu/2124.

a third model containing both advice and SES as explanatory factors. We find that only advice sources are significant factors. Using a likelihood ratio test [30], we find that this combined model has a goodness of fit significantly better than the SES-only model (X^2=45.09, $p < 0.001$, $df = 1164$) and not significantly different from the advice-only model (X^2=7.33 $p = 0.29$, $df = 1164$). This suggests that users' negative experiences relate to their advice sources, regardless of SES.

Advice, SES, and Resources

While we find that privacy and security incidents are related to respondents' advice sources, rather than their SES, we were curious to determine whether our prior finding—that there is an SES gap in advice sources [25]—held true in our sample. To do so, we constructed logistic regression models with users' advice sources as outcome variables and SES factors as inputs.

Our results show that users' advice sources are related to their level of education; Figure 2 provides an overview of respondents' reported advice sources organized by education. We find that users who hold less than a high school diploma are 99% less likely to report a coworker as an advice source, and those who hold less than a bachelor's degree, but who completed high school, are 51% less likely (Table 3). Similarly, those who held a high school diploma were 50% less likely to report coworkers and those with under a high school education were 73% less likely to report using government websites. Perhaps surprisingly, there was no significant difference in the SES or resources of respondents who reported taking advice from librarians, friends, and teachers. Overall, these results confirm our prior findings.

We hypothesize that these findings relate to less-educated users having different job roles, possessing relatively fewer internet skills [14], and distrusting websites that provide general advice without a clear source [26]. We also hypothesize that advice from websites may be more difficult to read and interpret than advice from other sources. Of note, there was no relationship between available internet resources and advice sources, implying that accessibility of advice related to devices and internet access may not be a problem.

DISCUSSION AND FUTURE WORK

In this study, we used logistic regression to investigate how users' advice sources, SES, and resources relate to the security and privacy experiences they report. Our findings, accurate within 2.7% of the true prevalence in the U.S. population, suggest first that advice is significantly correlated with security and privacy incidents; second, that reported incidents are not directly tied to SES but to a divide in where users of differing

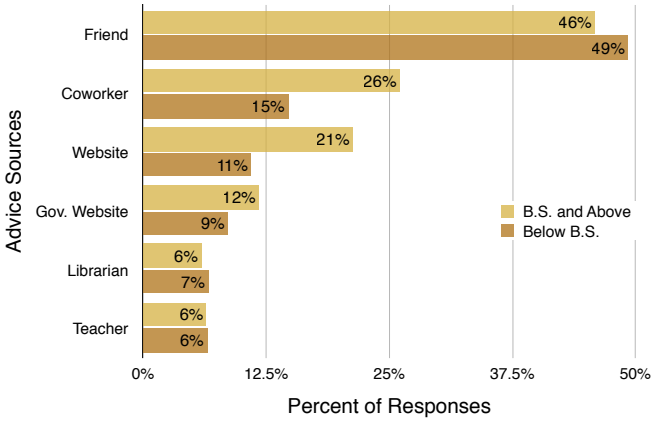

Figure 2. Respondents' advice sources by education (weighted).

educational attainment seek advice; and third, that less educated users report equal or fewer security and privacy incidents as more educated users. Below, we place our results in context and provide suggestions for future work.

Advice Matters

Our findings show a clear relationship between respondents' security and privacy experiences and advice sources. The direction of this relationship is unclear: do people receive bad advice that leads to worse experiences, or do they wait to seek advice until after a negative experience? We hypothesize some of both. In either case, however, this finding confirms that the current advice ecosystem is not working, and should be reevaluated. Future work should revisit what makes bad advice bad—outdated or incorrect content, poor presentation, a lack of readability, belief in the talisman of useless advice [16], or some combination—and look for ways to remove it or replace it with better advice. We should also evaluate the utility of new channels, such as public service announcements, for helping users find good advice.

Redefining the Digital Divide

Researchers have previously identified differences in skills, resources, and advice sources between lower- and higher-SES users. Our findings challenge the assumption that these differences lead to worse security and privacy outcomes for low-SES users. In particular, we find that income and available resources, such as in-home internet, have no impact on reported incidents. In line with prior work [25], we find that less educated users rely on less traditionally authoritative advice sources, such as friends and family. Unexpectedly, this difference is correlated with reporting slightly fewer negative incidents; further study is required to understand the causes of this result. There may be valuable lessons to learn from how less-educated users transmit security and privacy skills.

ACKNOWLEDGEMENTS

Our thanks to Cody Buntain, Tamara Clegg, Jon Froehlich, Mary Madden, and Brenna McNally. This work was supported by Data&Society and by Maryland Procurement Office contract no. H98230-14-C-013.

REFERENCES

1. 2013. Anonymity Omnibus Dataset. (2013). http://www.pewinternet.org/datasets/july-2013-anonymity-omnibus/
2. 2014. American Community Survey 5-Year Estimates. (2014). http://www.census.gov/programs-surveys/acs/news/data-releases/2014/release.html
3. 2015. National Cybersecurity Alliance. (2015). https://staysafeonline.org/
4. 2016. Chesapeake IRB. (2016). https://www.chesapeakeirb.com/
5. 2016a. Pew American Trends Panel. (2016). http://www.pewresearch.org/methodology/u-s-survey-research/american-trends-panel/
6. 2016b. Pew Internet and American Life Project. (2016). http://www.pewinternet.org/
7. 2016. Reason-Rupe Surveys. (2016). http://reason.com/poll
8. H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* (1974). DOI: http://dx.doi.org/10.1109/TAC.1974.1100705
9. boyd D., Levy K., and Arwick A. 2013. The Networked Nature of Algorithmic Discrimination. Data and Discrimination: Collected Essays. (2013). http://newamerica.org/downloads/OTI-Data-an-Discrimination-FINAL-small.pdf
10. C. Ciampa. 2013. A comparison of password feedback mechanisms and their impact on password entropy. *Information Management & Computer Security* (2013). DOI:http://dx.doi.org/10.1108/IMCS-12-2012-0072
11. E. Hargittai. 2002. Second-Level Digital Divide: Mapping Differences in People's Online Skills. *First Monday* (2002). http://arxiv.org/abs/cs.CY/0109068
12. E. Hargittai. 2003. *The Digital Divide and What to Do About It*. http://www.eszter.com/research/pubs/hargittai-digitaldivide.pdf
13. E. Hargittai. 2007. Whose Space? Differences Among Users and Non-Users of Social Network Sites. *Journal of Computer-Mediated Communication* 13, 1 (2007), 276–297. DOI: http://dx.doi.org/10.1111/j.1083-6101.2007.00396.x
14. E. Hargittai and Y.P. Hsieh. 2012. Succinct Survey Measures of Web-Use Skills. *Soc. Sci. Comput. Rev.* (2012). DOI: http://dx.doi.org/10.1177/0894439310397146
15. T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer New York. http://www.springer.com/us/book/9780387848570
16. C. Herley. 2016. The Unfalsifiability of Security Claims. (2016). https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/herley
17. A. L. Holbrook, M. C. Green, and J. A. Krosnick. 2003. Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias. *Public Opinion Quarterly* (2003). http://poq.oxfordjournals.org/cgi/citmgr?gca=pubopq; 67/1/79
18. A. E. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. 2012. The Psychology of Security for the Home Computer User.. In *IEEE S&P*. http://www.ieee-security.org/TC/SP2012/papers/4681a209.pdf
19. J. Jerome. 2013. Buying and Selling Privacy: Big Data's Different Burdens and Benefits. *Stanford Law Review* (2013). http://www.stanfordlawreview.org/online/privacy-and-big-data/buying-and-selling-privacy
20. L. Kish. 1965. *Survey sampling*. http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471109495.html
21. J. A. Krosnick. 2010. *Handbook of Survey Research*. http://www.sciencedirect.com/science/book/9780125982269
22. T. Lumley. 2016. R 'survey': Analysis of Complex Survey Samples. (2016). https://cran.r-project.org/web/packages/survey/survey.pdf
23. M. Micheli. 2016. Social networking sites and low-income teenagers: between opportunity and inequality. *Information, Communication & Society* (2016). DOI: http://dx.doi.org/10.1080/1369118X.2016.1139614
24. E. Rader, R. Wash, and B. Brooks. 2012. Stories As Informal Lessons About Security. In *SOUPS*. DOI: http://dx.doi.org/10.1145/2335356.2335364
25. E.M. Redmiles, S. Kross, and M. L. Mazurek. 2016a. How I Learned to be Secure: a Census-Representative Survey of Security Advice Sources and Behavior. In *CCS*. http://dl.acm.org/citation.cfm?id=2978307
26. E.M. Redmiles, A. R. Malone, and M. L. Mazurek. 2016b. I Think They're Trying to Tell Me Something: Advice Sources and Selection in Digital Security. In *IEEE S&P*. http://www.ieee-security.org/TC/SP2016/papers/0824a272.pdf
27. E.M. Redmiles, S. Silverstein, W. Bai, and M. L. Mazurek. 2016c. More Skilled Internet Users Behave (A Little) More Securely. In *SOUPS*. https://www.usenix.org/sites/default/files/SOUPS16poster20-redmiles.pdf
28. R. E. Rice. 2006. Influences, usage, and outcomes of Internet health information searching: Multivariate results from the Pew surveys. *International J. Medical Informatics* (2006). DOI: http://dx.doi.org/10.1016/j.ijmedinf.2005.07.032
29. N.C. Schaeffer and S. Presser. 2003. The Science of Asking Questions. *Annual Review of Sociology* (2003). DOI: http://dx.doi.org/10.1146/annurev.soc.29.110702.110112

30. A.J. Scott and J.N.K. Rao. 1984. On Chi-squared Tests For Multiway Contingency Tables with Proportions Estimated From Survey Data. *Annals of Statistics* (1984). https://www.jstor.org/stable/2241033

31. S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. 2010. Who Falls for Phish?: A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *CHI*. http://doi.acm.org/10.1145/1753326.1753383

32. L.D. Stanley. 2003. Beyond Access: Psychosocial Barriers to Computer Literacy Special Issue: ICTs and Community Networking. *The Information Society* (2003). DOI:http://dx.doi.org/10.1080/715720560

33. J. van Dijk and K. Hacker. 2003. The Digital Divide as a Complex and Dynamic Phenomenon. *The Information Society* (2003). DOI: http://dx.doi.org/10.1080/01972240309487

34. R. Wash and E. Rader. 2015. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *SOUPS*. https://www.usenix.org/conference/SOUPS2015/proceedings/presentation/wash