Pwned: The Risk of Exposure From Data Breaches

Gaurav Sood[*]
gsood07@gmail.com

Ken Cor[†]
mcor@ualberta.ca
The University of Alberta
Edmonton, Canada

ABSTRACT

News about massive data breaches is increasingly common. But we do not know what proportion of Americans are exposed in these breaches. We combine data from a large, representative sample of American adults (n = 5,000), recruited by YouGov, with data from *Have I Been Pwned* to estimate the lower bound of the number of times an average American's private information has been exposed. We find that at least 82.84% of Americans have had their private information, including account credentials, Social Security Number, etc., exposed. On average, Americans' private information has been exposed in at least three breaches. The better educated, the middle-aged, women, and Whites are more likely to have had their accounts breached than the complementary groups.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

Security risk, Privacy risk, Data breaches, Digital divide

ACM Reference Format:
Gaurav Sood and Ken Cor. 2019. Pwned: The Risk of Exposure From Data Breaches. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3292522.3326046

1 INTRODUCTION

On the Internet, nobody knows you're a dog. So the adage goes. But increasingly, others know that you like dog food and hate cats. Many of us have made our peace with this new reality. A slew of massive account breaches in recent years, however, threatens to pull the rug from under all illusions of anonymity [9].[1]

But there is little existing research on how frequently Americans' private information is part of such breaches. Much of the research on data breaches has focused on the downstream impact on corporations, e.g., [5, 14, 16], and people, e.g., [2, 3, 10]. Such research is vital—it informs data breach notification policy and laws, e.g., [6, 8, 12]. But absent from the literature is data that is important for developing effective public policy and laws on corporate liability for data breaches—data on the average American's risk of their private information being exposed in a data breach. In this note, we shed light on this mostly unexplored question.

Using a unique dataset, we estimate the lower bound of the average number of breached online accounts per person. We merge data from a large representative sample from YouGov (n = 5,000) with data from Have I Been Pwned (HIBP). We check whether the email associated with the YouGov account is part of the 293 public breaches cataloged by HIBP. We also study how exposure to breaches varies by socio-economic factors including ethnicity, sex, age, and education.

2 DATA AND METHODS

In July 2018, YouGov drew a nationally representative sample of 5,000 adult Americans. YouGov draws the sample as follows: it starts with a random sample of a high-quality sample of American adults, e.g., Current Population Survey, and then finds people on its panel that match the drawn sample most closely [13]. Some research suggests that the quality of samples drawn by YouGov is comparable to those drawn using probability sampling [1]. The sample that YouGov drew here, however, is better than its traditional survey samples. Non-response bias in our sample is zero because YouGov did not have to send out surveys; it used the emails associated with the accounts to collect the data. (YouGov never shared the emails with us.) Table 1 presents the marginals on key socio-demographic variables (see here for the codebook). (Table SI 1.1 presents the comparison between the Current Population Survey (CPS) and YouGov on key demographic variables. The upshot is that on key marginals, the difference between YouGov and CPS is less than 5%.)

After drawing the sample, YouGov used the emails associated with the accounts to query the HIBP API. (YouGov did the lookups so that it didn't have to share the email IDs.) HIBP is a non-profit clearinghouse of information about online account breaches. HIBP's stated aim is to provide a way for people to check if they are at risk from online breaches. It currently carries data from 293 breaches covering 278 unique domains and 5,235,843,322 accounts, including data from prominent breaches like the two Yahoo! breaches covering nearly 1.5 billion accounts. The HIPB data are, however, not comprehensive. Security researchers believe that there are many breaches that the companies are unaware of and at least a few cases where a company doesn't share information about a breach it knows about. HIBP also refuses to provide data on sensitive breaches—breached accounts where a person's inclusion may adversely affect

[1] On September 22, 2016, for instance, Yahoo! revealed that 500M accounts had been compromised in a breach [4]. Less than three months later, on December 14, 2016, Yahoo! announced that data had been stolen from nearly 1B user accounts in a different breach [11]. In all, Wikipedia lists 272 separate breaches between 2004 and 2018 (see https://en.wikipedia.org/wiki/List_of_data_breaches)

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326046

Table 1: YouGov Sample Characteristics

	proportion
race	
white	.67
hispanic/latino	.13
black	.12
asian	.03
middle eastern	.02
mixed race	.01
native american	.01
other	.00
sex	
female	.54
age	
(18, 25]	.09
(25, 35]	.19
(35, 50]	.26
(50, 65]	.28
(65, 100]	.18
education	
no hs	.06
hs grad.	.32
some college	.20
2-year college degree	.11
4-year college degree	.19
postgrad degree	.11

them—from their public API[2]. So data from HIBP only gives us a lower bound.

HIPB provides an easy way to get all the breached accounts associated with a particular email ID—you just need to make a simple API call passing the email that you want to get data on. This method gives us data on all the breaches logged by HIPB for all the 5,000 profiles. There is one caveat. Our YouGov sample provides data associated with only one email ID, the email people used to register with YouGov. People often have multiple email IDs. And that is another reason why all we get from this data is a lower bound. The actual number of breached accounts per person is likely much higher.

With each request, HIBP returns some metadata on the kind of breaches. (See the codebook for details about all the data that it returns.) Two pieces of information are material here. HIBP classifies each breach as verified or unverified. And it defines unverified breaches as breaches whose "legitimacy" it cannot "establish beyond [a] reasonable doubt." HIBP includes these unverified breaches because "they still contain personal information about individuals who want to understand their exposure on the web." The other material column that HIBP returns relates to whether a breach is part of a "spam list." HIBP defines *SpamList* as cases where "large volumes of personal data are found being utilized for the purposes of sending targeted spam." HIBP adds, "This often includes many of the same attributes frequently found in data breaches such as names, addresses, phones numbers and dates of birth. The lists are often aggregated from multiple sources, frequently by eliciting personal information from people with the promise of a monetary reward." And the reason HIBP includes these data is: "whilst the data may not have been sourced from a breached system, the personal nature of the information and the fact that it's redistributed in this fashion unbeknownst to the owners warrants inclusion here."

3 RESULTS

In all, 14,979 breaches are associated with the 5,000 emails on file. Or on average, there are three breaches per person. The median is also three. And at least 82.84% of Americans' accounts have been breached at least once.

The relationship between the number of breaches and socio-economic is counter to what focusing on traditional concerns around the digital divide would lead us to believe. If anything, the data suggest that people who use online services more are somewhat more likely to have their accounts breached. (See SI 1.1 and SI 1.2 for corresponding regressions and figures illustrating group-wise means along with the 95% confidence intervals.)

The number of breaches increases roughly monotonically with education (see Table SI 1.4 and Figure SI 1.3). The average number of breaches among people with no high school degree is 2.35. Compare this to postgraduates, who are part of 3.20 breaches on average (or 1.3 times the average of people with no high school degree).

In contrast to the relationship between education and the number of breaches, the relationship between the number of breaches and age is curvilinear (see Table SI 1.5), with young people's and seniors' accounts least likely to be breached, and middle-aged adults' accounts most likely to be breached. But, as the loess illustrates (see Figure SI 1.4), the relationship is modest.

When we compare the average number of breaches among men and women, we find that women's accounts are 1.12 times more likely to be breached than men's (see Table 2 and SI 1.3; $p < .05$). Analyzing breaches by ethnicity, Blacks' and Whites' accounts are most frequently breached. The mean number of breaches associated with the emails for Blacks and Whites is 3.12 and 3.16 respectively. For Hispanics/Latinos, the corresponding number is 2.5 (see Table SI 1.2; $p < .05$). And for Asians, the mean is 2.82.

To assess the source of the exposure, we checked the source of the breaches. The 14,979 breaches stemmed from 156 different sites, but there was a sharp skew with 21 sites with more than 100 breaches accounting for 11,783 of the breaches. Table 3 lists the 21 sites. Prominent websites like linkedin.com, adobe.com, dropbox.com, lastfm.com, among others feature on the list.

In the analysis presented until now, we don't distinguish between different kinds of breaches. But not all breaches are equally grave. So next, we shed light on the type of breaches. Of the 15,837 breaches, 14,979 or 94.58% were part of verified breaches. And about a third of the 15,837 breaches are categorized as *SpamList*. In all, we have 10,188 breaches that are verified and not categorized as *SpamList*. We focus our attention on these plausibly graver breaches, checking

[2] HIBP website notes that it does not share whether or not an account has been part of the breach at Adult Friend Finder, Ashley Madison, Beautiful People, Bestialitysextaboo, Brazzers, CrimeAgency vBulletin Hacks, Fling, Florida Virtual School, Freedom Hosting II, Fridae, Fur Affinity, HongFire, Mate1.com, Muslim Match, Naughty America, Non Nude Girls, Rosebutt Board, The Candid Board, The Fappening, xHamster and 1 more.

Table 2: Frequency of Account Breaches By Socio-economic Factors

	mean	se
Age		
(18,25]	1.96	0.10
(25,35]	3.12	0.09
(35,50]	3.34	0.08
(50,65]	3.29	0.07
(65,100]	2.95	0.07
Missing	1.19	0.16
Education		
No HS	2.35	0.12
HS Grad.	2.89	0.06
Some College	3.04	0.09
2-year College Degree	3.07	0.10
4-year College Degree	3.22	0.09
Postgrad Degree	3.20	0.11
Sex		
Female	3.17	0.05
Male	2.82	0.05
Race		
White	3.12	0.05
Black	3.16	0.11
Hispanic/Latino	2.50	0.08
Asian	2.82	0.21
Native American	2.96	0.26
Middle Eastern	2.66	0.24
Mixed Race	2.45	0.22
Other	2.92	1.32

whether the relationship with socio-economic variables we see above hold in this smaller subset.

When we look at education, the pattern holds up. Once again, the number of breached accounts per person for people with a college degree or more is higher than for people who only got as far as high school (see Table 4). Moving to sex, the pattern is more attenuated with women just nudging ahead of men—the mean for women and men is 2.15 and 2.05 respectively. The general pattern for age remains roughly similar to what we saw above, with the middle-aged more likely to have their accounts breached compared to people younger than 25 and older than 65. Breaking down by race, we see some interesting changes. Asians join Whites near the top of the pile, with means of about 2.2. Accounts of Hispanics or Latinos are less likely to be part of verified non-spam-list breaches (mean = 1.73, $p < .05$). The big relative change is for Blacks; African-Americans are likelier to be part of unverified, *SpamList* breaches.

4 CONCLUSION

Nearly 83% of Americans' have had their accounts breached at least once. In total, the 5,000 email accounts on file are associated with 14,979 breaches. Or, on average, people's accounts have been breached thrice. This number, though, is the lower bound for three reasons. First, not all breaches are made public. Second, HIBP doesn't allow access to data on sensitive breaches—breached online accounts on services that may have reputational consequences for people—via its public API. Third, many Americans have multiple email accounts. We only had one email ID per person.

Table 3: Most Frequently Implicated Domains

domain name	n
rivercitymediaonline.com	2,913
linkedin.com	1,089
modbsolutions.com	1,067
myspace.com	1,059
data4marketers.com	996
cashcrate.com	856
adobe.com	609
disqus.com	570
ticketfly.com	393
tumblr.com	340
dropbox.com	288
dailymotion.com	255
last.fm	248
evony.com	171
clixsense.com	150
cafemom.com	145
imesh.com	144
kickstarter.com	140
edmodo.com	130
zomato.com	112
neopets.com	108

We also find that the kinds of people who are most likely to use online services—the better educated, Whites, etc.—are generally the most exposed. This finding is consistent with Laohaprapanon and Sood, who find that the better educated, people with higher incomes, and racial majorities spend a smaller proportion of time online on problematic sites, but because they are online more often, they end up visiting more such sites [7]. This is contrary to the traditional narrative about the digital divide [15].

Table 4: Frequency of Verified, Non-SpamList Account Breaches By Socioeconomic Factors.

	mean	se
Age		
(18,25]	1.63	0.10
(25,35]	2.44	0.08
(35,50]	2.37	0.07
(50,65]	2.16	0.06
(65,100]	1.78	0.05
Missing	0.91	0.13
Education		
No HS	1.53	0.09
HS Grad.	1.91	0.05
Some College	2.22	0.08
2-year College Degree	2.10	0.08
4-year College Degree	2.37	0.08
Postgrad Degree	2.30	0.08
Sex		
Female	2.15	0.04
Male	2.05	0.05
Race		
White	2.21	0.04
Black	2.03	0.08
Hispanic/Latino	1.73	0.07
Asian	2.16	0.18
Native American	1.85	0.18
Middle Eastern	2.05	0.21
Mixed Race	1.70	0.19
Other	2.69	1.19

REFERENCES

[1] Stephen Ansolabehere and Brian F Schaffner. 2014. Does survey mode still matter? Findings from a 2010 multi-mode comparison. *Political Analysis* 22, 3 (2014), 285–303.

[2] Cassandra Cross, Megan Parker, and Daniel Sansom. 2019. Media discourses surrounding "non-ideal" victims: The case of the Ashley Madison data breach. *International Review of Victimology* 25, 1 (2019), 53–69.

[3] Shelby R Curtis, Jessica Rose Carre, and Daniel Nelson Jones. 2018. Consumer security behaviors and trust following a data breach. *Managerial Auditing Journal* 33, 4 (2018), 425–435.

[4] Seth Fiegerman. 2016. Yahoo says 500 million accounts stolen. https://money.cnn.com/2016/09/22/technology/yahoo-data-breach/

[5] 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* 82, 2 (2018), 85–105.

[6] McKenzie L Kuhn. 2018. 147 Million Social Security Numbers for Sale: Developing Data Protection Legislation After Mass Cybersecurity Breaches. *Iowa L. Rev.* 104 (2018), 417.

[7] Suriyan Laohaprapanon and Gaurav Sood. 2018. Domain Knowledge: Predicting the Kind of Content Hosted by a Domain. http://www.gsood.com/research/papers/domain_knowledge.pdf

[8] Daniel J Marcus. 2018. The Data Breach Dilemma: Proactive Solutions for Protecting Consumers' Personal Information. *Duke LJ* 68 (2018), 555.

[9] David McCandless. 2017. World's Biggest Data Breaches & Hacks. http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

[10] Vyacheslav Mikhed and Michael Vogan. 2018. How data breaches affect consumer credit. *Journal of Banking & Finance* 88 (2018), 192–207.

[11] Lily Hay Newman. 2016. Hack Brief: Hackers Breach A Billion Yahoo Accounts. A Billion. https://www.wired.com/2016/12/yahoo-hack-billion-users/

[12] Bernold Nieuwesteeg and Michael Faure. 2018. An analysis of the effectiveness of the EU data breach notification obligation. *Computer Law & Security Review* 34, 6 (2018), 1232–1246.

[13] Douglas Rivers. 2010. Decennial Census Surname Files (2010, 2000). Sample Matching: Representative Sampling from Internet Panels, https://github.com/themains/pwned/lit/rivers.pdf.

[14] Pierangelo Rosati, Peter Deeney, Mark Cummins, Lisa Van der Werff, and Theo Lynn. 2019. Social media and stock price reaction to data breach announcements: Evidence from US listed companies. *Research in International Business and Finance* 47 (2019), 458–469.

[15] Alexander Van Deursen and Jan Van Dijk. 2011. Internet skills and the digital divide. *New media & society* 13, 6 (2011), 893–911.

[16] Kimberly A Whitler and Paul W Farris. 2017. The impact of cyber attacks on brand image: Why proactive marketing expertise is needed for managing data breaches. *Journal of Advertising Research* 57, 1 (2017), 3–9.

SI 1 SUPPORTING INFORMATION

SI 1.1 Tables

The missing entries reflect cases where we do not have commensurate categories in our data. The largest differences we see are on race and ethnicity, a variable where our coding differs from CPS in meaningful ways.

Table SI 1.1: Comparison Between YouGov and CPS 2018

	cps	yg	diff
Age			
18 to 25	0.14	0.13	0.01
26 to 35	0.18	0.18	0.00
36 to 50	0.25	0.23	0.02
51 to 65	0.25	0.26	-0.01
66 to 80+	0.19	0.18	0.01
Sex			
Male	0.48	0.51	-0.03
Female	0.52	0.49	0.03
Race			
White alone	0.78	0.64	0.14
Black or African American alone	0.13	0.12	0.01
American Indian and Alaska Native alone	0.01	0.01	0.00
Asian alone	0.06		
Native Hawaiian and Other Pacific Islander alone	0.00		
Two or more races	0.02		
Educational Attainment			
No high school diploma	0.11	0.07	0.04
High school or equivalent	0.29	0.33	-0.04
Some college, less than 4-yr degree	0.28	0.31	-0.03
Bachelor's degree or higher	0.32	0.29	0.03

Table SI 1.2: Number of Breaches by Race/Ethnicity

	Dependent variable:
	Number of Breaches
Black	.04
	(.12)
Hispanic/Latino	$-.62^{***}$
	(.10)
Asian	$-.30$
	(.23)
Native American	$-.16$
	(.36)
Middle Eastern	$-.46^{**}$
	(.24)
Mixed Race	$-.67^{**}$
	(.29)
Other	$-.20$
	(.73)
Constant	3.12^{***}
	(.05)
Observations	5,000
R^2	.01
Adjusted R^2	.01
Note:	*p<0.1; **p<0.05; ***p<0.01

Table SI 1.3: Number of Breaches by Sex

	Dependent variable:
	Number of Breaches
Male	$-.35^{***}$
	(.07)
Constant	3.17^{***}
	(.05)
Observations	5,000
R^2	.004
Adjusted R^2	.004
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table SI 1.4: Number of Breaches by Education

	Dependent variable:
	Number of Breaches
HS Grad.	.54***
	(.16)
Some College	.69***
	(.16)
2-year College Degree	.72***
	(.18)
4-year College Degree	.87***
	(.17)
Postgrad Degree	.85***
	(.18)
Constant	2.35***
	(.14)
Observations	5,000
R^2	.01
Adjusted R^2	.01
Note:	*p<0.1; **p<0.05; ***p<0.01

Table SI 1.5: Number of Breaches by Age

	Dependent variable:
	Number of Breaches
ns(age, 2)1	2.23^{***}
	(.20)
ns(age, 2)2	$-.90^{***}$
	(.21)
Constant	2.01^{***}
	(.09)
Observations	5,000
R^2	.03
Adjusted R^2	.03
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

SI 1.2 Figures

Figure SI 1.1: Relationship Between Race and Number of Breaches

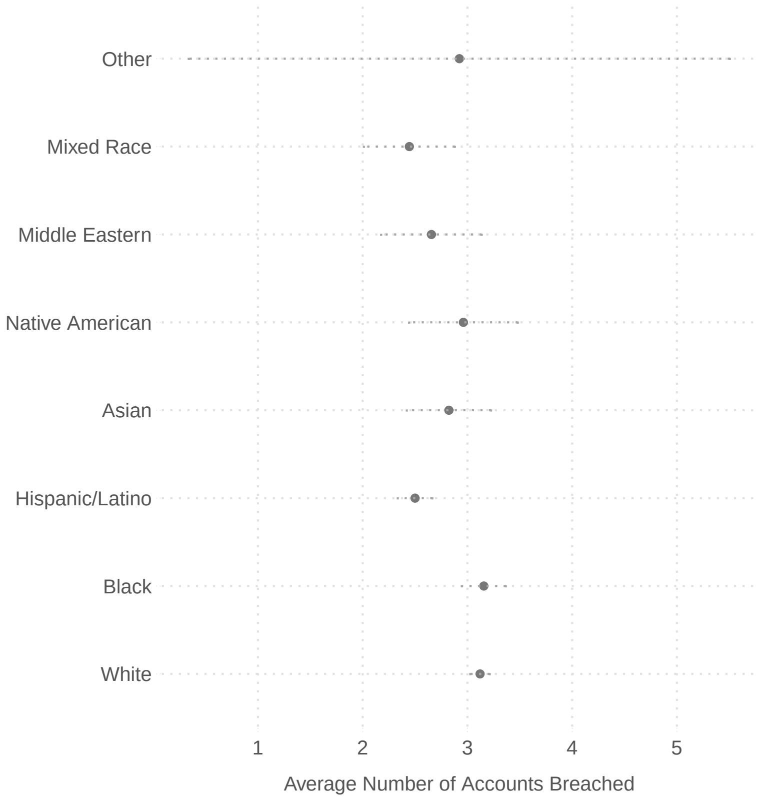


Figure SI 1.2: Relationship Between Sex and Number of Breaches

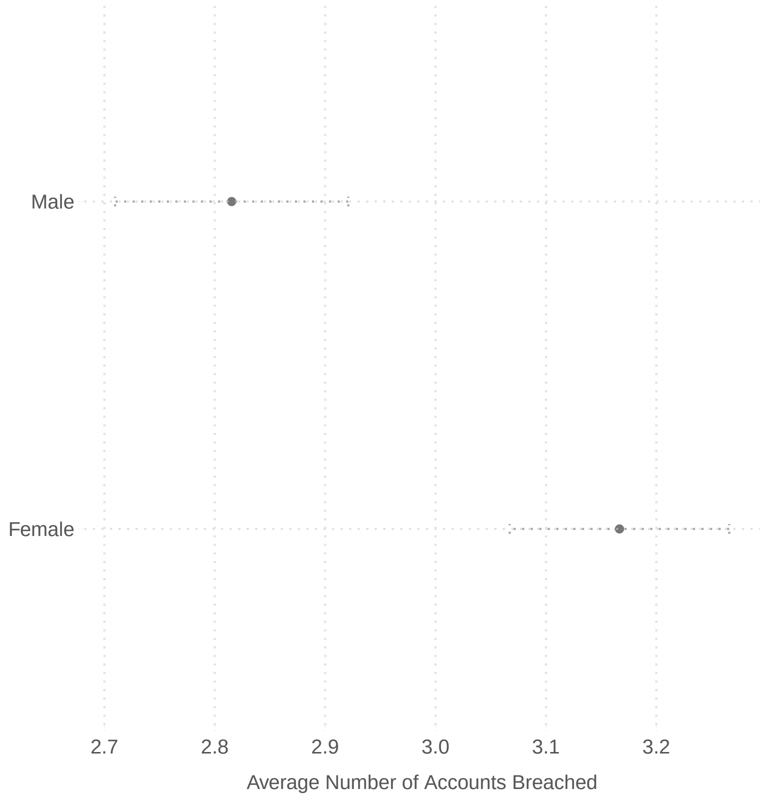


Figure SI 1.3: Relationship Between Education and Number of Breaches

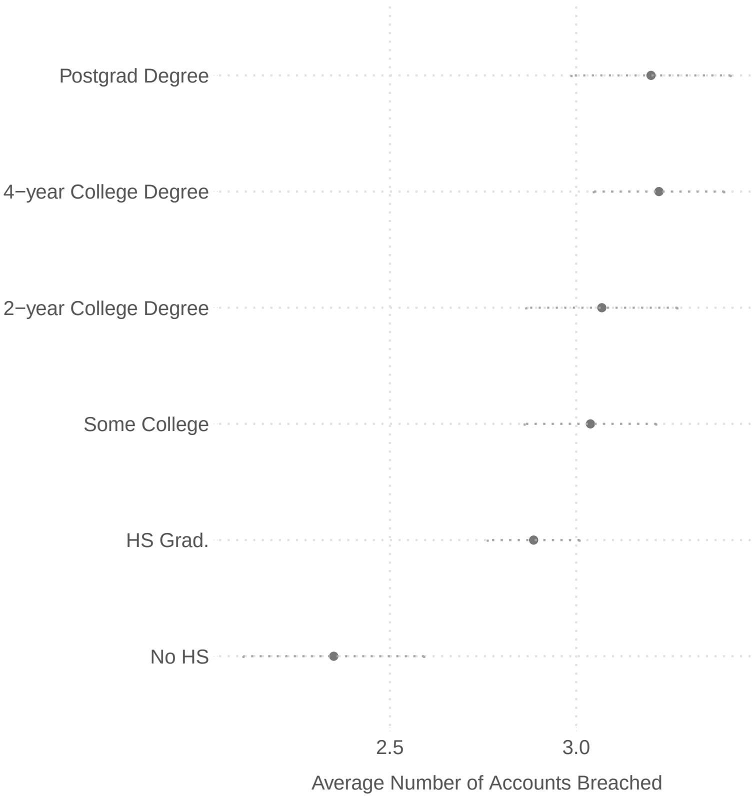
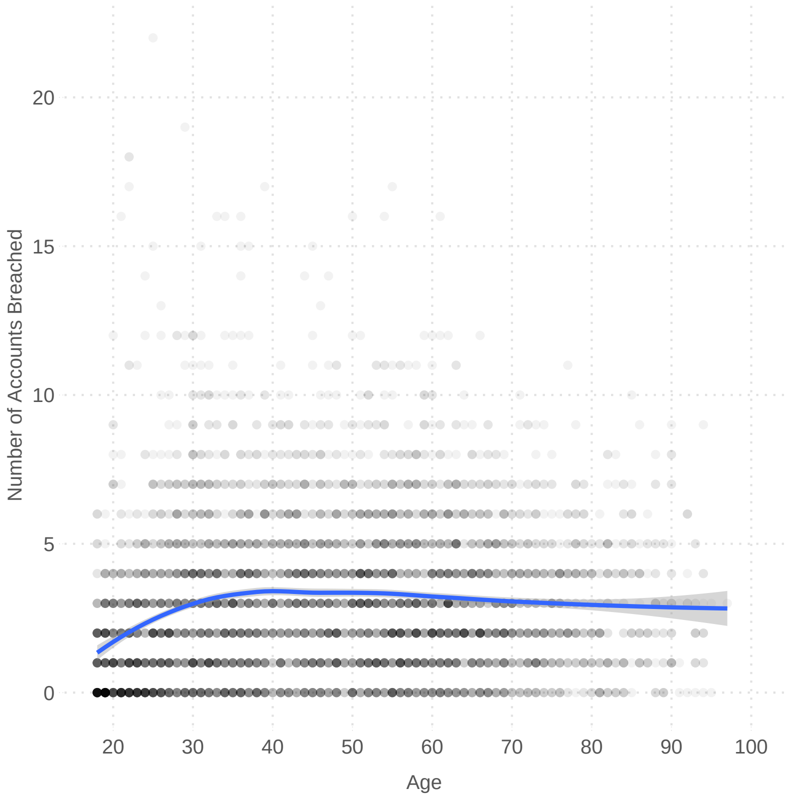


Figure SI 1.4: Relationship Between Age and Number of Breaches


Number of Accounts Breached
Age