Pwned: Lower Bound of Number of Breached Online Accounts Per Person*

Ken Cor† Gaurav Sood‡

July 28, 2018

Abstract

News about large online breaches is increasingly common. But there has been little good data on how exposed people are because of these breaches. We combine data from a large, representative sample of adult Americans (n = 5,000) with data from *Have I Been Pwned* to estimate the lower bound of the average number of breached online accounts per person. We find that at least 82.84% of Americans have had their accounts breached. And that on average Americans' accounts have been breached at least three times. Better educated, the middle-aged, women, and Whites are more likely to have had their accounts breached than the complementary groups.

*Data and scripts behind the analysis presented here can be downloaded from https://github.com/themains/pwned.

†Ken can be reached at: mcor@ualberta.ca

‡Gaurav can be reached at: gsood07@gmail.com

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, threaten to pull the rug under all illusions of anonymity (**?**). In this note, we shed light on this threat. Using a unique dataset, we estimate the *lower bound* of the average number of breached online accounts per person.

To shed light on the question, 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 find that nearly 83% of Americans' accounts have been breached at least once. In total, the 5,000 email accounts on file are associated with 14,979 breaches. Or, on average, people are part of three breaches. This number, though, is the lower bound. People generally have more than one email, and we only use one here. And HIBP only catalogs data on a tiny chunk of the total breaches, and only provides data on breaches that do not harm the reputation of the person via its API. We also study how exposure to breaches varies by socio-economic factors and find that the kinds of people who are most likely to use online services—the better educated, Whites, etc.—are generally the most at risk.

Data

In July 2018, YouGov drew a nationally representative sample of 5,000 adult Americans. The general process that YouGov uses when drawing a sample is as follows: it starts by taking 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 (for more details, see Rivers 2010). Some research suggests that the quality of samples drawn by YouGov is comparable to those drawn using probability sampling (Ansolabehere and Schaffner 2014).

However, the sample that YouGov drew here is different from its traditional survey sample in one key aspect. Non-response bias in our sample is 0 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 http://github.com/themains/pwned/data/ for the codebook).

Table 1: YouGov Sample Marginals.

	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

After drawing the sample, YouGov used the emails associated with the accounts to query the `haveibeenpwned.com` (HIBP) API. 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 not comprehensive. Security researchers believe that there are many more breaches that the companies are unaware of and at least a few cases where a company doesn't share information about the breach even when it knows. HIBP also refuses to provide data on sensitive breaches—breaches from accounts where a person's inclusion may adversely affect them—from their public API.[1] So data from HIBP only gives us a lower bound.

HIPB provides an easy way to query all the accounts with a particular email ID have been breached. For instance, to check which accounts associated with the email gsood07@gmail.com (one of the authors' email) have been breached, you can query https://haveibeenpwned.com/api/v2/breachedaccount/gsood07@gmail.com. This method gives us all the breaches that HIPB data on for all the email IDs associated with each of the 5,000 profiles. But our YouGov sample provides only one email ID. People often have multiple. So that is another reason why all we get from this data is an absolute lower bound. The actual number is likely much higher than the number we obtain here.

With each request, HIBP returns some metadata on the kind of breaches. (See the https://github.com/themains/pwned/data/hipb_codebook.xlsx for details about all the data that it returns.) Two pieces of information are material here. HIBP catalogs each breach as verified or unverified. And it defines unverified breaches as breaches whose "legitimacy" it cannot "establish beyond reasonable doubt." HIBP includes these unverified breaches because "they still contain personal information about individuals who want to understand their expo-

[1]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."

sure 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 utilised 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."

Results

At least 82.84% Americans' accounts have been breached at least once. In total, the 5,000 email accounts on file are associated with 14,979 breaches. Or on average, there are three breaches per person. The median is also 3.

The relationship between how frequently emails are part of breaches and socio-economic factors suggests that on the whole who are more likely to use online services are more likely to have had their accounts breached. Though there are a few exceptions. Women's accounts are 1.2 times more likely to be breached than men's (see Table 2). Analyzing breaches by race, African Americans' and Whites' accounts are most frequently breached. The mean number of breaches their email is part of is 3.12 and 3.16 for African Americans and Whites respectively. For Hispanics/Latinos, the corresponding number is 2.5. And for Asians, the mean is 2.82, about 0.90 times that of the average for Whites.

Looking at education, the relationship is roughly monotonic, with the mean number of breaches increasing with education. The average number of breaches people with no HS are part of is just 2.35. Compare this to postgraduates, with a mean of 3.20 or 1.3 times as likely.

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

	mean
race	
white	3.12
hispanic/latino	2.50
black	3.16
asian	2.82
middle eastern	2.66
mixed race	2.45
native american	2.96
other	2.92
sex	
female	3.17
male	2.82
age	
(18, 25]	3.07
(25, 35]	3.49
(35, 50]	3.29
(50, 65]	2.29
(65, 100]	3.41
education	
no hs	2.35
hs grad.	2.89
some college	3.04
2-year college degree	3.07
4-year college degree	3.22
postgrad degree	3.20

Lastly, for age, the relationship is curvilinear, with young people's and seniors' accounts least likely to be breached, and mid-aged adults' accounts most likely to be breached.

To assess the source of the exposure, we checked the source of the 14,979 breaches. The 14,979 breaches stemmed from 156 different sites, but there was a sharp skew with 21 sites with more than 100 breaches alone 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.

Table 3: Most Frequently Implicated Domains.

domain name	n
rivercitymediaonline.com	2913
linkedin.com	1089
modbsolutions.com	1067
myspace.com	1059
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

The analysis until now hasn't distinguished between different kinds of breaches. 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 breaches, checking 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 gender, the pattern is more attenuated with women just nudging ahead of men (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 there in verified non-spam-list breaches (mean = 1.73). The big relative change is for Blacks and that suggests that they are likelier to be part of unverified, spam-list breaches.

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

	mean
race	
white	2.21
hispanic/latino	1.73
black	2.03
asian	2.16
middle eastern	2.05
mixed race	1.70
native american	1.85
other	2.69
sex	
female	2.15
male	2.05
age	
(18, 25]	1.99
(25, 35]	2.63
(35, 50]	2.53
(50, 65]	2.30
(65, 100]	1.93
education	
no hs	1.53
hs grad.	1.91
some college	2.22
2-year college degree	2.10
4-year college degree	2.37
postgrad degree	2.30

Conclusion

At least 83% of Americans' accounts have been breached at least once. And on average, people's accounts have been breached thrice. This is a 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.

And generally speaking, the people most at risk are those who are the likeliest to use online services—the better educated, Whites, etc. This sets it apart from the traditional narrative about the digital divide. Sometimes, the people holding the shorter end of the stick are those who use online services more.

References

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

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