How Often is Politicians' Data Breached? Evidence from HIBP*

Lucas Shen† Gaurav Sood‡

March 2025

Abstract

Data breaches involving politicians are concerning because of the threat of impersonation and blackmail, among other nefarious things. To shed light on the concern, we estimate how frequently politicians' data is compromised. Using a dataset of 12,384 emails of politicians from 59 countries spanning three decades, we check whether these emails are part of breaches by using *Have I Been Pwned*, a widely used online service for searching public breach data. A third of the politicians have had their data breached at least once. More alarmingly, over one in five have had their sensitive data, such as bank account numbers, biometric data, browsing history, chat logs, credit card CVV, etc., breached. These numbers are still too optimistic for several reasons, including the fact that we do not have all the email addresses used by politicians. Accounting for some of the biases suggests that more than half the politicians have suffered a serious breach.

Keywords: Cybersecurity, Have I Been Pwned (HIBP), Elites, Data breaches, Security and privacy

*The replication materials are posted on https://github.com/themains/pwned_pols.

†Institute for Human Development and Potential, Agency for Science, Technology, and Research, lucas@lucasshen.com.

‡Independent, gsood07@gmail.com.

1 Introduction

Data breaches can reveal login credentials, personal details, and service usage patterns, providing avenues for impersonation, extortion, and blackmail. When data breaches involve politicians, the stakes are higher (Harding, 2016; BBC News, 2020; Witman and Mackelprang, 2022). Politicians make policies and have access to sensitive data and influential people.

Data breaches are also exceedingly common. A study of a large random sample of Americans found that more than 80% of people have had their data breached (Sood and Cor, 2019).[1] While some of the breaches are innocuous, involving already public information, many of the breaches involve private data. For instance, nearly 72% of the public breaches include passwords. Even when only login credentials are compromised, the threat is sizable. Because many people reuse their passwords or only make minor tweaks, attackers using credential tweaking attacks can compromise 83% of the accounts (Li et al., 2019) (see also Chintalapati and Sood (2022)). The threat is nearly as grave even when leaks involve hashed passwords. Attackers can use resources like `hashes.org`, which contains clear text values for passwords to crack passwords. In 2021, a study found that nearly 99% of the hashed passwords in the *Have I Been Pwned* (HIBP) database can be recovered (Kanta et al., 2021).

This paper studies how often politicians' data is exposed in breaches. We assemble a large dataset of politician emails spanning 59 countries over three decades. We then check if these emails are associated with data breaches using HIBP, a large repository of public data breaches. The numbers are sobering. More than one in three politicians have had their data breached at least once. Concerningly, nearly one in five have had a "serious" breach, in which private data like credit card numbers, etc., has been leaked. Concerning as these numbers are, they are gross underestimates, partly because we do not have all the email

[1] A follow-up study using the Florida voter registration data found a similar percentage of people had had their data breached (Sood, 2023).

addresses used by politicians. Accounting for some of the biases suggests that the true rate of serious breaches is north of 50%.

2 Research Design, Data, and Measures

Our estimand of interest is the frequency with which politicians' data, especially sensitive data, is breached. Our population of interest is all politicians who have held a legislative or executive role at a national, state, or local level in the last three decades. Our definition of population excludes politicians who have never served the government.

To study the question, we assemble a large, diverse convenience sample. Our final data set has thousands of politicians from 59 countries spanning three decades.

To measure how often politicians' data is leaked, we check public breach data using politicians' emails. We do this because email addresses allow us to confidently connect a breach to a person. The downside is that our method only enables us to estimate a conservative lower bound of the frequency with which data are leaked for three reasons. First, the breached data must have email addresses. Some public breach data may not; for instance, it may only have usernames and passwords (which can, based on other breaches, be connected back to the politician). Second, we need all the email addresses used by the politician to know all the public data breaches in which a politician's data has been exposed. We only have official email addresses for many of the politicians. There is also reason to expect that exposure estimated using official email addresses is lower than estimated using personal emails. We expect politicians to be less likely to use official emails when opening online accounts, as the official email accounts for politicians (in many Democracies) are likely temporary. We also expect politicians to be reluctant to associate their official emails with private activities because of disclosure laws and official policies restricting the use of official emails. Lastly, not all breaches become public.

Our data on data breaches comes from HIBP, which collates data from more than 850 public breaches and classifies breached data into nearly 150 categories. Some data breaches are benign, revealing mainly public information; others are more serious, revealing sensitive data. To better understand the actual threat posed to politicians, we classify breaches as serious based on the kind of data breached. Our measure leaves out the most serious breaches. HIBP doesn't provide data on sensitive breaches—breaches that can be embarrassing, e.g., adult websites like Adult FriendFinder (2015), Ashley Madison, etc.—via its public API. It only allows people who can verify the emails to see those breaches (Have I Been Pwned, 2018).

In all, we start by providing a conservative lower bound of the threat faced by politicians. Next, to account for the attenuating bias stemming from private vs. official emails, we craft a measure distinguishing the two and model the probability of being connected to a data breach. Using the model that controls for confounds, we arrive at a more realistic assessment of the threat politicians face. We also describe the variation in threat across politicians based on country and gender. Lastly, we analyze how politicians' data comes to be leaked. We examine which breaches are responsible for the leaks.

2.1 Data on Politicians

Our data on politicians comes from two sources. The first is https://Everypolitician.org/ (mySociety, 2018). The data has been used extensively by researchers (see, for e.g., Eshima and Smith, 2022; Stockemer and Sundström, 2018; Martini and Walter, 2024; Kosinski et al., 2024) and as a source for building other databases (Kurz and Ettensperger, 2023; Stockemer and Sundström, 2022). The data covers legislatures from across the world between 1997 to 2019.

From the data, we exclude any legislature with fewer than 30 politicians with a listed email address. In all, we have 8,538 unique email addresses from 61 legislatures across 55

countries.

We supplement this data with data gathered by scraping official parliamentary sites of the Argentinian, Brazilian, Danish, Greek, Indian, Nigerian, Norwegian, and Singaporean parliaments (or Senates) and the Indian state legislatures of Bihar, Delhi, Himachal Pradesh, Tamil Nadu, and Uttar Pradesh. We scraped the data in January 2025.

We collate the two datasets and check if the email addresses conform to expected email formats (e.g., RFC 5322-compliant addresses like *name@domain.com*). To check, we first preprocess email strings by stripping whitespace and normalizing character representation by converting them to Unicode (Tauberer, 2024). Next, we check if the domain exists via DNS resolution and check deliverability by looking up DNS MX records. Emails that fail these checks are removed from the data because querying HIBP with an invalid email (e.g., *an@invalid.email*) will result in a "No breach found" status, biasing our estimates downwards.

Our final dataset has 12,384 unique politician email accounts from 67 legislatures across 59 countries from 1997 to 2025, representing close to 50% of the global population living under electoral regimes (Lührmann et al., 2018; Herre, 2022) and 36% of the global population overall. Table 1 summarizes our data.

2.2 Official Vs. Personal Emails

We classify email addresses as official or personal based on domain patterns. We classify an email as official if it belongs to a government entity as identified through standard government domains, e.g., *.gov*, *.gouv*, *.gob*, *admin.ch*, *nic.in*. We also classify an email as official if the domain has words like *parliament*, *senate*, *congress*, *assembly*, and their foreign-language equivalents, e.g., *parlament*, *senado*, *congreso*, *assemblee*, etc. We classify emails as personal if they are not official and if they are associated with prominent providers, e.g., *gmail*, *yahoo*, or are known commercial domains, e.g., *.com*, *.net*, *.biz*, etc.

Table 1. Summary of politician email data by country

			Summary of coverage				
		Country	Emails	Years	Chamber(s)	Legislative body	Pop.
1	ALB	Albania	140	2009, 2013, 2017	Unicameral	Kuvendi	2.7
2	AND	Andorra	31	2015	Unicameral	Consell General	0.1
3	ARG	Argentina	71	2025	Upper	Parliament	46.9
4	ARM	Armenia	119	2019	Unicameral	National Assembly	2.8
5	AUS	Australia	177	2004–2016	Lower, Upper	House of Representatives, Senate	26.9
6	BEL	Belgium	149	2014	Lower	Chamber of Representatives	11.9
7	BGR	Bulgaria	205	2013, 2014, 2017	Unicameral	National Assembly	6.4
8	BIH	Bosnia	42	2014	Lower	House of Representatives	3.2
9	BLR	Belarus	59	2016	Unicameral	House of Representatives	9.1
10	BMU	Bermuda	33	2017	Lower	Parliament	—
11	BRA	Brazil	81	2025	Upper	Parliament	217.6
12	BTN	Bhutan	46	2013	Lower	National Assembly	0.8
13	CAN	Canada	432	2011, 2015	Lower, Upper	House of Commons, Senate	40.4
14	CMR	Cameroon	104	2013	Lower	Assemblée Nationale	29.4
15	COL	Colombia	169	2014, 2018	Lower	Cámara de Representantes	52.3
16	CYP	Cyprus	56	2016	Unicameral	House of Representatives	1.3
17	DNK	Denmark	332	2001–2025	Unicameral	Folketing	6.0
18	EST	Estonia	101	2011, 2015, 2019	Unicameral	Riigikogu	1.4
19	FIN	Finland	121	2003, 2007, 2011	Unicameral	Eduskunta	5.6
20	GBR	UK	832	1997–2017	Lower, Unicameral	Commons, Senedd, Scottish Parliament	68.6
21	GEO	Georgia	145	2012, 2016	Unicameral	Parliament of Georgia	3.8
22	GGY	Guernsey	39	2016	Unicameral	States	—
23	GRC	Greece	539	2004–2025	Unicameral	Hellenic Parliament, Parliament	10.3
24	GRL	Greenland	38	2014	Unicameral	Inatsisartut	—
25	GTM	Guatemala	152	2012, 2016	Unicameral	Congress	17.9
26	HKG	Hong Kong	55	2012, 2016	Unicameral	Legislative Council	7.5
27	HUN	Hungary	184	2014, 2018	Unicameral	Országgyűlés	9.4
28	IND	India	3334	2014, 2025	Lower, State	Lok Sabha, State Legislature	1441.7
29	IRN	Iran	104	2012, 2016	Unicameral	Majles	89.8
30	ITA	Italy	299	2013, 2018	Upper	Senate	58.6
31	JEY	Jersey	60	2014	Unicameral	States	—
32	KEN	Kenya	253	2013	Lower	National Assembly	56.2
33	KOR	South Korea	251	2012, 2016	Unicameral	National Assembly	51.6
34	LKA	Sri Lanka	221	2015	Unicameral	Parliament	22.1
35	LUX	Luxembourg	60	2013, 2018	Unicameral	Chamber of Deputies	0.7
36	MDA	Moldova	44	2014	Unicameral	Parlament	2.4
37	MKD	Macedonia	101	2014, 2016	Unicameral	Sobranie	1.8
38	MLT	Malta	45	2013	Unicameral	Parliament	0.6
39	NAM	Namibia	69	2015	Lower, Upper	National Assembly, National Council	2.6
40	NGA	Nigeria	200	2015, 2025	Lower, Upper	House of Representatives, Senate	229.2
41	NIC	Nicaragua	84	2012	Unicameral	National Assembly	7.1
42	NLD	Netherlands	144	2012, 2017	Lower	Tweede Kamer	17.9
43	NOR	Norway	174	2025	Unicameral	Storting	5.6
44	NPL	Nepal	264	2014	Unicameral	Constituent Assembly	31.2
45	NZL	New Zealand	117	2008–2017	Unicameral	New Zealand Parliament	5.3
46	PNG	Papua New Guinea	52	2012	Unicameral	National Parliament	10.5
47	PYF	French Polynesia	57	2013	Unicameral	Assembly	0.3
48	ROU	Romania	147	2016	Lower	Chamber of Deputies	18.8
49	RWA	Rwanda	76	2013	Lower	Chamber of Deputies	14.4
50	SGP	Singapore	370	2001–2025	Unicameral	Parliament	6.0
51	SUR	Suriname	57	2015	Unicameral	National Assembly	0.6
52	SVK	Slovakia	164	2006–2016	Unicameral	National Council	5.3
53	SYC	Seychelles	32	2011	Unicameral	National Assembly	0.1
54	TZA	Tanzania	403	2005, 2010, 2015	Unicameral	National Assembly	69.4
55	UGA	Uganda	158	2011	Unicameral	Parliament	49.9
56	URY	Uruguay	119	2015	Lower	Chamber of Deputies	3.4
57	ZAF	South Africa	380	2014	Lower	National Assembly	61.0
58	ZMB	Zambia	66	2011, 2016	Unicameral	National Assembly	21.1
59	ZWE	Zimbabwe	34	2013	Lower	House of Assembly	17.0

Note: The last column reports World Bank 2024 country population estimates (in millions) (World Bank, 2025). The UK sample includes Scotland and Wales (unicameral) and House of Commons (lower). The India sample includes Lok Sabha and five state legislatures (Bihar, Delhi, Himachal Pradesh, Tamil Nadu, and Uttar Pradesh).

2.3 Measuring Exposure to Data Breaches

To evaluate whether politicians' data has been breached, we check emails against the *Have I Been Pwned* (HIBP) database. Researchers frequently use HIBP to study breaches (see, for, e.g., Hien et al., 2025; Sood and Cor, 2019; Sood, 2023). For each breach, HIBP reports the type of data exposed, for example, email addresses, hashed passwords, sexual orientation, email messages, etc., and includes metadata such as the year of the breach, its scope (number of records affected), and the industry of the breached online service. HIBP covers breaches dating back to 2007, though publicly available records begin in 2013 (Table SI1).

2.4 Serious Breaches

HIBP classifies breached data into 146 categories. The most frequently compromised data are email addresses, passwords, usernames, names, and IP addresses (see Table SI2). Of these, we identified 38 data types as particularly serious because of the security and privacy risks associated with them (see Table SI3 for the complete list of these data classes). These include authentication credentials (e.g., passwords, PINs, mother's maiden names), financial details (e.g., credit card information, bank account numbers), personal identifiers (e.g., social security numbers, biometrics, MAC addresses), data about health, and private communications. We classify a breach as serious if it exposes any of these data types.

3 Results

3.1 Prevalence of Politician Email Breaches

Nearly one in three politicians' emails is linked to at least one public breach (Panel A, Table 2). However, there is a sharp skew in the distribution. While the mean number of breaches per account is 1 ($\hat{\sigma} = 3.2$), the median is 0, with 19.5% of accounts breached

multiple times, some as many as hundreds of times.

Serious data breaches were associated with 21.6% of politicians' emails (Panel B, Table 2). Of the breached accounts, 65.5% were implicated in a serious data breach (Section 2.4). Again, we see a substantial skew. Most accounts are never seriously breached, but a smaller proportion are breached multiple times.

Table 2. Summary of the number of data breaches

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				Percentiles					Percentage	
	n	Mean	SD	Min	25p	Med	75p	Max	≥ 1	≥ 2
Panel A: Data breaches										
All emails	12,384	1.0	3.2	0	0	0	1	256	33.0%	19.5%
Official emails	9,371	0.0	1.0	0	0	0	1	18	28.4%	16.1%
Personal emails	3,013	1.0	5.0	0	0	0	2	256	47.3%	30.3%
Panel B: Serious data breaches										
All emails	12,384	0.5	2.1	0	0	0	0	189	21.6%	10.6
Official emails	9,371	0.0	0.0	0	0	0	0	9	17.2%	7.6
Personal emails	3,013	1.0	3.0	0	0	0	1	189	35.2%	20.0

Note: This table reports the number of data breaches by official versus personal emails (Section 2). The last two columns report the percentage of emails with at least one and at least two breaches. Serious data breaches are those that include at least one serious data type exposed in the breach (Table SI3).

The most common data exposed in breaches involving politicians are email addresses, names, phone numbers, job titles, and physical addresses (Table 3). While these are mostly innocuous information, passwords are the most common serious data class exposed. Nearly 63% of the accounts had their password exposed. In nearly 1.6% of the account breaches, credit card information was revealed. Bank account numbers were compromised in 1.0% of the breaches, and government-issued IDs in 1.3%. Some breaches also exposed private communications, such as emails (2.3%) and other private messages (0.7%).

3.2 Official Vs. Personal Emails

Table 2's numbers paint too rosy a picture. As we note above, for many politicians, we only have their official government email addresses. To shed light on the issue, we analyze how

Table 3. Breakdown of Compromised Data in Politician Email Breaches

	Data type	#	(%)	Srs.		Data type	#	(%)	Srs.		Data type	#	(%)	Srs.
1	Email addresses	4090	100.0%		21	Marital statuses	99	2.4%		41	Payment histories	38	0.9%	
2	Names	3479	85.0%		22	Religions	99	2.4%		42	Survey results	34	0.8%	
3	Phone numbers	3200	78.2%		23	Email messages	96	2.3%	✓	43	User website URLs	31	0.8%	
4	Job titles	2841	69.4%		24	Password hints	81	2.0%	✓	44	Telecommunications carrier	30	0.7%	
5	Physical addresses	2701	66.0%		25	Ethnicities	77	1.9%		45	Nationalities	27	0.7%	✓
6	Social media profiles	2644	64.6%		26	Home ownership statuses	77	1.9%		46	Private messages	27	0.7%	✓
7	Passwords	2596	63.5%	✓	27	Auth tokens	75	1.8%	✓	47	Relationship statuses	21	0.5%	
8	Geographic locations	2090	51.1%		28	Occupations	75	1.8%		48	Company names	20	0.5%	
9	Employers	1868	45.7%		29	PINs	69	1.7%	✓	49	Deceased statuses	18	0.4%	
10	Genders	1714	41.9%		30	Partial credit card data	67	1.6%	✓	50	Website activity	16	0.4%	
11	Dates of birth	1593	38.9%		31	Credit status information	64	1.6%	✓	51	Professional skills	15	0.4%	
12	IP addresses	1573	38.5%		32	Family structure	61	1.5%		52	Credit cards	14	0.3%	✓
13	Usernames	1004	24.5%		33	Financial investments	61	1.5%		53	Passport numbers	9	0.2%	✓
14	Salutations	419	10.2%		34	Net worths	61	1.5%		54	Profile photos	8	0.2%	
15	Education levels	414	10.1%		35	Personal interests	61	1.5%		55	Support tickets	8	0.2%	
16	Purchases	246	6.0%		36	Government issued IDs	55	1.3%	✓	56	Cryptocurrency wallet addresses	5	0.1%	
17	Spoken languages	220	5.4%		37	Device information	50	1.2%		57	Social security numbers	5	0.1%	✓
18	Political donations	154	3.8%		38	Browser user agent details	43	1.1%		58	Account balances	4	0.1%	
19	Income levels	134	3.3%		39	Bios	41	1.0%		59	Health insurance information	4	0.1%	✓
20	Job applications	113	2.8%		40	Bank account numbers	40	1.0%	✓	60	Loyalty program details	4	0.1%	

Note: This table breaks down the 60 most common types of compromised data involving accounts across 562 breaches, showing the number (#) and percentage (%) of the 4,091 emails with known data breaches. See Table SI2 for a similar breakdown of data types covering all HIBP breaches.

the number of breaches varies by official vs. personal emails.

Table 2 tabulates the number of breaches by personal and official email accounts. Of the approximately 12,000 politicians' emails, we have personal emails for nearly 3,000. Politicians are almost twice as likely to have at least one breach associated with a personal email than with an official email account (Table 2). The breach rate for official emails is 28.4% vs. 47.3% for personal emails (Panel A, Table 2). The relative difference for serious data breaches is more pronounced, with the respective numbers being 17.2% and 35.2% (Panel B, Table 2). The exposure rate of personal emails provides a more realistic picture of politicians' exposure.

3.3 Who Gets Breached?

The personal vs. official email results may be driven by selection bias. Politicians who use personal email providers may be more lax about their security or anticipate shorter political stints. To better address the confounds and to study the variation in exposure more generally,

we model breaches as a function of politician characteristics. Given the skew in the data (Table 2), we model the likelihood of an email breach rather than the number of breaches. In all, we estimate the following equation:

$$\begin{aligned}\text{Breach}_{i,t} \;=\; \beta_1\,\text{PersonalEmail}_i \;+\; \beta_2\,\text{Female}_i \;+\; \beta_3\,\text{SocialMedia}_i \\ +\; \gamma_{c(i)} \;+\; \tau_t \;+\; \theta_i \;+\; \lambda_{c(i),i} \;+\; \varepsilon_{i,t}, \quad (1)\end{aligned}$$

where i indexes emails, t indexes year, and c indexes country. Our base specification models the risk of a breach (or serious breach) as a function of email type (personal vs. official), country, and decade-fixed effects (based on the legislative start year). For the EveryPolitician data, we have more variables available. We take advantage of the richer data by adding controls for politician gender (β_2), whether or not the politician has a known Twitter or Facebook account (β_3), legislature type (lower house, upper house, unicameral legislature, $\lambda_{c(i),i}$), and political party (θ_i). Since we expect observations within the same country to share unobserved characteristics that may induce correlation in the error terms, we cluster the standard errors at the country level. We estimate the models using *fixest* in R (Bergé, 2018).

Before turning to the results, a necessary caveat: because of potential confounds, cross-group differences cannot be interpreted as stemming from some essential aspect of the group. For instance, one of the many potential confounds that could vitiate comparisons is that the risk of exposure is correlated with political tenure. If, for instance, male legislators may have longer tenures than female legislators, the coefficients on gender will partly reflect differences in tenure (which is not modeled).

Across all specifications, personal emails are strongly linked to a higher probability of breach (see Table 4). Personal emails have a 24–30 percentage points greater chance of having a breach (serious breach) associated with them ($p < .001$, Table 4 Columns (1)–(4)).

Table 4. Probability of data breach

	Dependent variable			
	Breach		Serious breach	
	(1)	(2)	(3)	(4)
Personal email	0.286***	0.241***	0.264***	0.295***
	(0.016)	(0.031)	(0.012)	(0.042)
Female		0.009		0.023+
		(0.012)		(0.012)
Social media		0.086***		0.078***
		(0.021)		(0.015)
# Country	59	54	59	54
# Decade	4	3	4	3
# Legislature type	–	3	–	3
# Political party	––	465	––	465
Dependent variable mean	0.330	0.386	0.216	0.245
Observations	12,384	7,188	12,384	7,188
R^2	0.303	0.518	0.276	0.466
Country fixed effects	✓	✓	✓	✓
Decade fixed effects	✓	✓	✓	✓
Legislature type fixed effects		✓		✓
Political party fixed effects		✓		✓
Sample: Pooled	✓		✓	
Sample: EveryPolitician		✓		✓

Note: The dependent variable in Columns (1) and (2) is whether the email has been in a breach. The dependent variable in Columns (3) and (4) is whether the email has been in a *serious* breach. Columns (2) and (4) include politician and legislature attributes available only in the EveryPolitician sample. All models are linear probability models. The omitted categories are official email for email type, male for gender, and no recorded social media account (Facebook or Twitter) for social media presence. Standard errors clustered at the country level are reported in parentheses. Significance levels: *** 0.001 ** 0.01 * 0.05 + 0.1.

In all, given that personal emails better reflect online behavior (Section 2), we believe a more realistic estimate of the percentage of politicians with a serious breach is over 50%.

Next, we describe the variation in the likelihood of being involved in a breach by other characteristics of the politician. Female politicians are more likely to have a serious breach than male politicians, with a 2.3 percentage point higher probability ($p < .1$, Table 4 Column (4)). The kinds of politicians for whom we have public social media profiles are likelier to have had their data breached. There is a 7.7 percentage points increase in the likelihood of a serious breach ($p < .001$, Table 4 Column (4)).

Figure 1 reports the mean-centered fixed effects for countries with the highest and

lowest breach rates conditional on the email type and time fixed effects. The top five countries with the highest probability of having accounts breached are Australia, the UK, Jersey, Hong Kong, and Denmark. The bottom five are Iran, Korea, Papua New Guinea, Albania, and Nepal. Figure 1 (right panel) also confirms that older accounts have higher probabilities of breach.

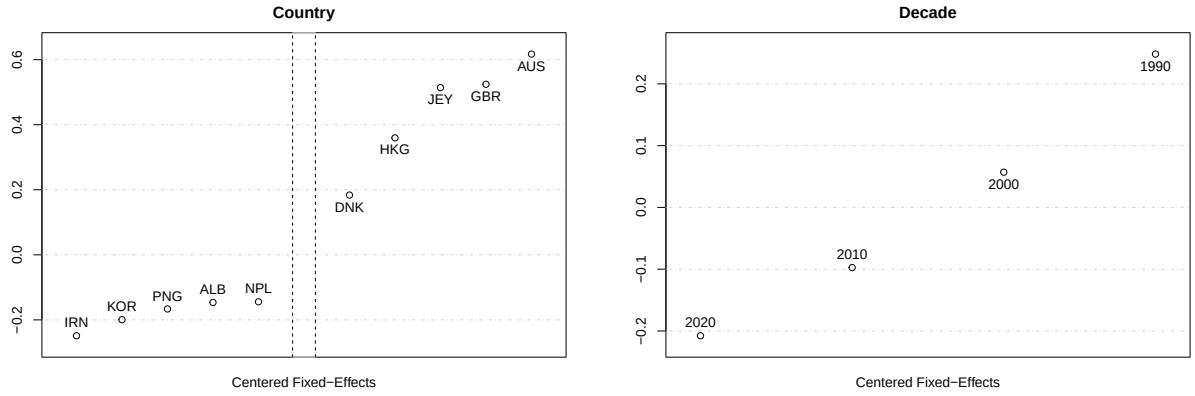

Figure 1. Country and decade mean-centered fixed effects of the probability of a serious breach. This figure reports the top five and bottom five estimated fixed effects for the country (left panel) and the fixed effects for a decade (right panel) from Model (3) in Table 4). All plotted effects are centered such that the mean is zero. See Table SI4 for all countries' estimated effects.

3.4 Which Breaches Are Most Responsible?

Table 5 reports the most common data breaches involving politicians. The two most common breaches involving politicians are the *db8151dd* (2020) and the *OnlinerSpambot* (2017) breaches. The *db8151dd* breach, originating from Covve (a Customer Relationship Management app), leaked fairly benign information, such as email addresses, job titles, names, phone numbers, physical addresses, and social media profiles. (The same six data types are also the six most common kinds of data that are compromised in breaches more generally Table 3.)

OnlinerSpambot was a more serious breach. The *OnlinerSpambot* incident was not just another breach but a live cybercrime operation, with the stolen email access weaponized to

Table 5. Top 20 data breaches involving politicians

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Breach name	Emails	%	Domain	Breach date	Public date	Yrs till public	Total leaks	#data classes	Srs.
1	db8151dd	1440	35.2%	covve.com	2020-02-20	2020-05-15	0.2 years	22.8M	6	
2	OnlinerSpambot	1230	30.1%	—	2017-08-28	2017-08-29	0.0 years	711.5M	2	✓
3	PDL	1128	27.6%	—	2019-10-16	2019-11-22	0.1 years	622.2M	7	
4	VerificationsIO	1037	25.3%	verifications.io	2019-02-25	2019-03-09	0.0 years	763.1M	10	
5	LinkedIn	489	12.0%	linkedin.com	2012-05-05	2016-05-21	4.0 years	164.6M	2	✓
6	LinkedInScrape	385	9.4%	linkedin.com	2021-04-08	2021-10-02	0.5 years	125.7M	7	
7	Apollo	340	8.3%	apollo.io	2018-07-23	2018-10-05	0.2 years	125.9M	8	
8	Intelimost	294	7.2%	intelimost.com	2019-03-10	2019-04-02	0.1 years	3.1M	2	✓
9	Cit0day	271	6.6%	cit0day.in	2020-11-04	2020-11-19	0.0 years	226.9M	2	✓
10	Twitter200M	269	6.6%	twitter.com	2021-01-01	2023-01-05	2.0 years	211.5M	4	
11	Collection1	256	6.3%	—	2019-01-07	2019-01-16	0.0 years	772.9M	2	✓
12	ExploitIn	233	5.7%	—	2016-10-13	2017-05-06	0.6 years	593.4M	2	✓
13	AntiPublic	226	5.5%	—	2016-12-16	2017-05-04	0.4 years	458.0M	2	✓
14	DemandScience	205	5.0%	demandscience.com	2024-02-28	2024-11-13	0.7 years	121.8M	7	
15	Gravatar	196	4.8%	gravatar.com	2020-10-03	2021-12-05	1.2 years	114.0M	3	
16	TelegramCombolists	164	4.0%	—	2024-05-28	2024-06-03	0.0 years	361.5M	3	✓
17	Nitro	160	3.9%	gonitro.com	2020-09-28	2021-01-19	0.3 years	77.2M	3	✓
18	FairVoteCanada	154	3.8%	fairvote.ca	2024-03-02	2024-10-21	0.6 years	0.1M	5	
19	YouveBeenScraped	147	3.6%	—	2018-10-05	2018-12-06	0.2 years	66.1M	6	
20	NotSOCRadar	129	3.2%	—	2024-08-03	2024-08-09	0.0 years	282.5M	1	

Note: Column (1) is the unique identifier of a data breach incident. Columns (2) and (3) report the number and percentage of politician emails compromised. Column (4) is the associated domain. Columns (5) and (6) are the breach date and date when added to HIBP. Column (7) is the lapse in years between the date of the breach and the date added to the HIBP repository. Column (8) is the total number (in millions) of the general population accounts that were compromised. Column (9) is the number of data classes leaked (see Table SI2 for all data classes). Column (10) indicates whether the breach is serious (Section 2.4).

spread malware that steals bank account credentials and credit card details (Hayashi, 2017; Kelion, 2017; Trend Micro, 2017; Have I Been Pwned, 2018). Both email addresses and passwords were also released.

The *OnlinerSpambot* breach was detected within a day. Other serious breaches like the *LinkedIn* breach, which is the fifth-most common and leaked email addresses and passwords, went undetected for four full years (Column (7), Table 5) until the leaked data went on sale on a dark market site (Have I Been Pwned, 2018). Some breaches in HIBP took as long as 12 years before the public found out (Table SI1). Notable, many breaches such as *OnlinerSpambot*, unlike *db8151dd* and *LinkedIn*, are not tied to any single online service or domain, and it is still unknown where bad actors sourced the credentials from (Kelion, 2017; Trend Micro, 2017; Have I Been Pwned, 2018).

3.5 Leaks to Floods? Breaches Over Time

Finally, we investigate how the exposure from data breaches grows over time. We start with the 9,200 email accounts where legislative start dates are available. Figure 2 plots the percentage of emails with a breach associated with them over time. The middle and bottom panel of Figure 2 plots the total number of politician emails and breach incidents, respectively.

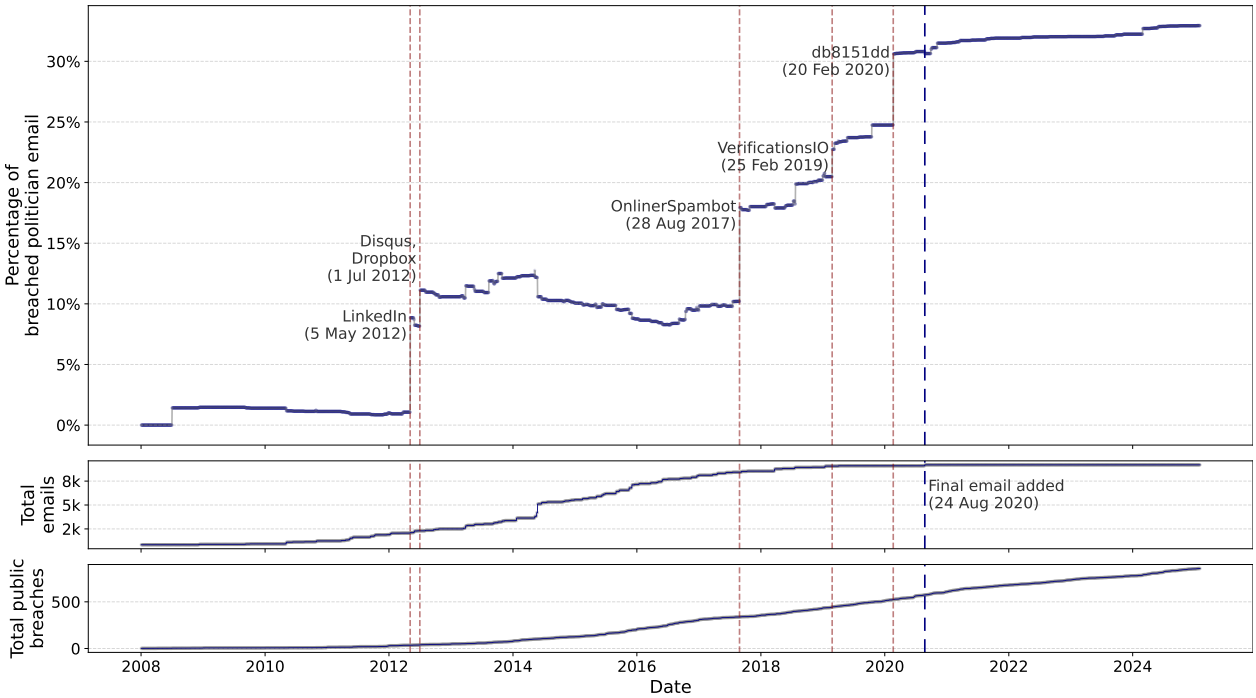

Figure 2. Breach rate of politician emails over time. This figure plots the percentage of breached politician email accounts ($n = 9,200$) based on the cumulative number of extant email accounts (based on the observed earliest legislative start date) and publicly-known data breaches. Each point represents the percentage of total extant emails found in breaches as of that date. The top panel annotates the five largest percentage jumps and the corresponding breach incidents (dashed vertical red lines). *Disquis* and *Dropbox* happened the same day. The middle panel tracks the cumulative number of emails. The bottom panel tracks the cumulative number of data breaches.

The breach rate starts at zero, but by January 2025, it reaches 33%, the number reported in Table 2 (Panel A) using the full sample. While the bottom panel of Figure 2 shows a gradual increase in publicly-known breaches over time, breach rates do not follow that pattern. Instead, increases in exposure are concentrated around a few incidents, as indi-

cated by many sharp stepwise increases in breach rate among extant politician emails. Five incidents with an outsized impact on politician breaches are the *LinkedIn* (2012), *Dropbox* (2012), *OnlinerSpambot* (2017), *VerificationsIO* (2019), and *db8151dd* (2020) breaches (see also Table 5). Overall, Figure 2 suggests that a small number of large-scale breaches drive exposure.[2]

4 Discussion

Data breaches expose people to theft, extortion, impersonation, and blackmail. When the target is a politician, the social cost of a breach is likely considerably higher. Politicians handle sensitive government information, influence policy, and engage with influential figures. Their public statements shape discourse. The corresponding risks from impersonation, blackmail, etc., are hence greater (for example, Harding, 2016; Almasy, 2017; BBC News, 2019, 2020; Witman and Mackelprang, 2022). The threat is amplified when we account for systematic efforts to exploit information or election interference, misinformation campaigns, and undermining of democracy (Harding, 2016; CrowdStrike, 2016; Almasy, 2017; BBC News, 2017, 2019; Bizga, 2020; Stahie, 2020; BBC News, 2020).

In this study, we assemble a large database of politicians' emails to assess the threat faced by them. The results are alarming. Conservatively, more than one in five politicians have had their sensitive data breached at least once. A more realistic rate of sensitive data being compromised is more than 50%, though even that is a gross underestimate.

There is a sharp skew in exposure, with a small proportion of politicians repeatedly

[2]The small dips in the percentage of accounts with a breach reflect those cases where a new cohort of emails joins the pool. Figure SI3 follows a small and fixed cohort of emails that existed before 2007 and shows a monotonic pattern that is similar to the pattern we see in Figure 2, with a few incidents driving most of the breaches, and with breach rate reaching close to 70% by Jan 2025. This aligns with Figure 1 (and Figure SI2) confirming that earlier email accounts have higher breach probability.

compromised. This pattern could arise from several reasons, including credential reuse, weak passwords, and attacks targeting certain politicians (Harding, 2016; Almasy, 2017; BBC News, 2019).

Correspondingly, there is also a sharp skew in the number of politicians entangled in each breach. A few large breaches account for most of the exposure. This pattern mirrors that of the general population (Sood and Cor, 2019). Part of the pattern is founded in leaks from a few widely used online services, such as LinkedIn and Dropbox.

Policymakers and political organizations may want to prioritize cybersecurity training and implement protocols to safeguard sensitive data. To prevent credential reuse, one avenue may be systems that present politicians with a data breach notification. As Albayram and Walker (2024) finds, it may lead people to change their password. Other solutions may include using password meters as a visual aid (Ur et al., 2012) and enforcing stricter two-factor authentication for government email accounts, as the UK Parliament did in the aftermath of their 2017 breach (BBC News, 2017; UK Parliament, 2017).

In summary, frequent data breaches of politicians' data underscore an urgent need for stronger digital security practices. By addressing these vulnerabilities, public institutions can better protect their officials and, by extension, the integrity of political discourse.

References

Albayram, Yusuf, and Jaden Walker. 2024. "Investigating Effectiveness of Informing Users About Breach Status of Their Email Addresses During Website Registration." *International Journal of Human–Computer Interaction* 0 (0): 1–20. 10.1080/10447318.2024.2404721, DOI: 10.1080/10447318.2024.2404721.

Almasy, Steve. 2017. "Emmanuel Macron's French presidential campaign hacked." https://edition.cnn.com/2017/05/05/europe/france-election-macron-hack-allegation/index.html, https://edition.cnn.com/2017/05/05/europe/france-election-macron-hack-allegation/index.html.

BBC News. 2017. "Iran blamed for Parliament cyber-attack." https://www.bbc.com/news/uk-41622903, www.bbc.com/news/uk-41622903.

BBC News. 2019. "German politicians targeted in mass data attack." https://www.bbc.com/news/world-europe-46757009, www.bbc.com/news/world-europe-46757009.

BBC News. 2020. "Major US Twitter accounts hacked in Bitcoin scam." https://www.bbc.com/news/technology-53425822, www.bbc.com/news/technology-53425822.

Bergé, Laurent. 2018. "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm." https://ideas.repec.org/p/luc/wpaper/18-13.html.

Bizga, Alina. 2020. "Cybercriminals Target Norwegian Parliament; Email Accounts of Elected Members and Employees Compromised." https://www.bitdefender.com/en-us/blog/hotforsecurity/cybercriminals-target-norwegian-parliament-email-accounts-of-elected-memb www.bitdefender.com/en-us/blog/hotforsecurity/cybercriminals-target-norwegian-parliament-email-accounts-of-elected-members-and-employees-compromised.

Chintalapati, Rajashekar, and Gaurav Sood. 2022. "Pass-Fail: Using a Password Generator to Improve Password Strength." July, https://github.com/themains/password, github.com/themains/password.

CrowdStrike. 2016. "CrowdStrike's work with the Democratic National Committee: Setting the record straight." https://www.crowdstrike.com/en-us/blog/bears-midst-intrusion-democratic-national-committee/, www.crowdstrike.com/en-us/blog/bears-midst-intrusion-democratic-national-committee/.

Eshima, Shusei, and Daniel M. Smith. 2022. "Just a Number? Voter Evaluations of Age in Candidate-Choice Experiments." *The Journal of Politics* 84 (3): 1856–1861. 10.1086/719005, DOI: 10.1086/719005.

Harding, Luke. 2016. "Top Democrat's emails hacked by Russia after aide made typo, investigation finds." https://www.theguardian.com/us-news/2016/dec/14/dnc-hillary-clinton-emails-hacked-russia-aide-typo-investigation-finds, www.theguardian.com/us-news/2016/dec/14/dnc-hillary-clinton-emails-hacked-russia-aide-typo-investigation-finds.

Have I Been Pwned. 2018. "API v2." https://haveibeenpwned.com/API/v2.

Hayashi, Kaoru. 2017. "Banking Trojans: Ursnif Global Distribution Networks Identified." https://unit42.paloaltonetworks.com/unit42-banking-trojans-ursnif-global-distribution-networks-identified/, unit42.paloaltonetworks.com/unit42-banking-trojans-ursnif-global-distribution-networks-identified/.

Herre, Bastian. 2022. "The world has recently become less democratic." https://ourworldindata.org/less-democratic.

Hien, Ton Nguyen Trong, Adisak Sangsongfa, and Noppadol Amm-Dee. 2025. "Discovering Personal Data Security Issues: Insights from "Have I Been Pwned"." In *Advances in Computing and Data Sciences*, edited by Singh, Mayank, Vipin Tyagi, P. K. Gupta, Jan Flusser, Tuncer Ören, Amar Ramdane Cherif, and Ravi Tomar 259–269, Cham: Springer Nature Switzerland, , DOI: 10.1007/978-3-031-70906-7_22.

Kanta, Aikaterini, Sein Coray, Iwen Coisel, and Mark Scanlon. 2021. "How viable is password cracking in digital forensic investigation? Analyzing the guessability of over 3.9 billion real-world accounts." *Forensic Science International: Digital Investigation* 37 301186. https://doi.org/10.1016/j.fsidi.2021.301186, DOI: 10.1016/j.fsidi.2021.301186.

Kelion, Leo. 2017. "Huge spam list with 711m addresses discovered." https://www.bbc.com/news/technology-41095606, https://www.bbc.com/news/technology-41095606.

Kosinski, M., P. Khambatta, and Y. Wang. 2024. "Facial recognition technology and human raters can predict political orientation from images of expressionless faces even when controlling for demographics and self-presentation." *American Psychologist* 79 (7): 942–955. 10.1037/amp0001295, DOI: 10.1037/amp0001295.

Kurz, Kira Renée, and Felix Ettensperger. 2023. "Introducing a New Dataset: Age Representation in Parliaments on the Party-Level." *Statistics, Politics and Policy* 14 (3): 357–374. doi:10.1515/spp-2023-0014, DOI: 10.1515/spp-2023-0014.

Li, Lucy, Bijeeta Pal, Junade Ali, Nick Sullivan, Rahul Chatterjee, and Thomas Ristenpart. 2019. "Protocols for Checking Compromised Credentials." In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19 1387–1403, New York, NY, USA: Association for Computing Machinery, . 10.1145/3319535.3354229, DOI: 10.1145/3319535.3354229.

Lührmann, Anna, Marcus Tannenberg, and Staffan Lindberg. 2018. "Regimes of the World (RoW): Opening New Avenues for the Comparative Study of Political Regimes." *Politics and Governance* 6 (1): 60–77. 10.17645/pag.v6i1.1214, DOI: 10.17645/pag.v6i1.1214.

Martini, Marco, and Stefanie Walter. 2024. "Learning from precedent: how the British Brexit experience shapes nationalist rhetoric outside the UK." *Journal of European Public Policy* 31 (5): 1231–1258. 10.1080/13501763.2023.2176530, DOI: 10.1080/13501763.2023.2176530.

mySociety. 2018. "EveryPolitician." https://everypolitician.org/, https://everypolitician.org.

Sood, Gaurav. 2023. "Have I Been Pwned? Yes. Evidence from Florida Voter Registration Data." August, https://github.com/themains/reg_breach, github.com/themains/reg_breach.

Sood, Gaurav, and Ken Cor. 2019. "Pwned: The Risk of Exposure From Data Breaches." In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19 289–292, New York, NY, USA: Association for Computing Machinery, . 10.1145/3292522.3326046, DOI: 10.1145/3292522.3326046.

Stahie, Silviu. 2020. "Finnish Parliament Was Targeted in Cyberattack in 2020." https://www.bitdefender.com/en-us/blog/hotforsecurity/finnish-parliament-was-targeted-in-cyberattack-in-2020, www.bitdefender.com/en-us/blog/hotforsecurity/finnish-parliament-was-targeted-in-cyberattack-in-2020.

Stockemer, Daniel, and Aksel Sundström. 2018. "Age representation in parliaments: Can institutions pave the way for the young?" *European Political Science Review* 10 (3): 467–490. 10.1017/S1755773918000048, DOI: 10.1017/S1755773918000048.

Stockemer, Daniel, and Aksel Sundström. 2022. "Introducing the Worldwide Age Representation in Parliaments (WARP) data set." *Social Science Quarterly* 103 (7): 1765–1774. https://doi.org/10.1111/ssqu.13221, DOI: 10.1111/ssqu.13221.

Tauberer, Joshua. 2024. "python-email-validator." github.com/JoshData/python-email-validator.

Trend Micro. 2017. "Onliner Spambot Leverages 711M Email Accounts for Massive Campaigns." https://www.trendmicro.com/vinfo/sg/security/news/cybercrime-and-digital-threats/onliner-spambot-leverages-711m-email-accounts-for-massive-campaigns, www.trendmicro.com/vinfo/sg/security/news/cybercrime-and-digital-threats/onliner-spambot-leverages-711m-email-accounts-for-massive-campaigns.

UK Parliament. 2017. "Update following cyber-attack on Parliament." https://www.parliament.uk/mps-lords-and-offices/offices/commons/media-relations-group/news/update-following-cyber-attack/, www.parliament.uk/mps-lords-and-offices/offices/commons/media-relations-group/news/update-following-cyber-attack/.

Ur, Blase, Patrick Gage Kelley, Saranga Komanduri et al. 2012. "How Does Your Password Measure Up? The Effect of Strength Meters on Password Creation." In *21st USENIX Security Symposium (USENIX Security 12)*, 65–80, Bellevue, WA: USENIX Association, , August, https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/ur, www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/ur.

Witman, Paul D., and Scott Mackelprang. 2022. "The 2020 Twitter Hack – So Many Lessons to Be Learned." *Journal of Cybersecurity Education, Research and Practice* 2021. 10.62915/2472-2707.1089, DOI: 10.62915/2472-2707.1089.

World Bank. 2025. "World Development Indicators." https://databank.worldbank.org/source/world-development-indicators.

Supplementary Information

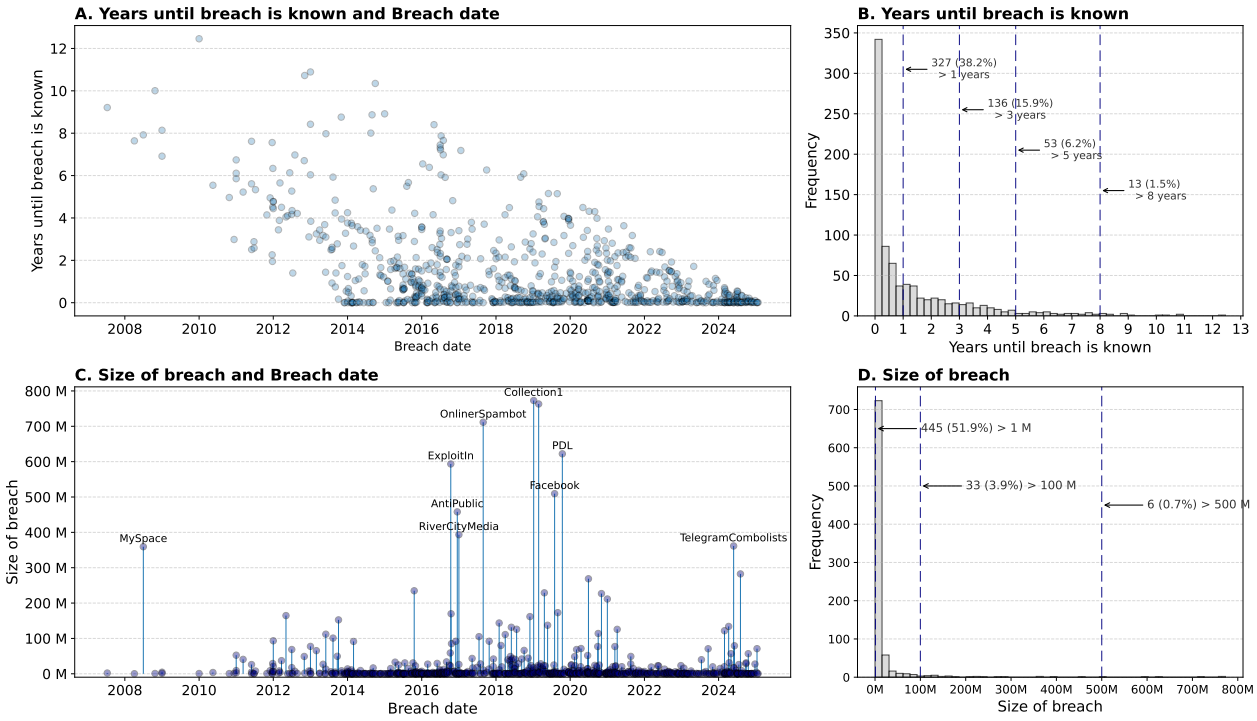

Figure SI1. Summary of data breaches recorded in HIBP. The subfigures summarize the time lag between the breach occurrence and its public disclosure on HIBP (*Years until the breach is known*) and the number of accounts compromised (*Size of breach*).

Table SI1. Summary of HIBP breaches ($n = 857$)

	Mean (1)	Std dev (2)	Min (3)	25p (4)	Median (5)	75p (6)	Max (7)
Breached accounts (in 100,000s)	169.59	677.67	0.01	2.7	11.29	58.15	7729.05
Number of compromised data types	5.29	2.76	1.0	4.0	5.0	7.0	25.0
Breach date (yyyy-mm-dd)	2018-10-11	—	2007-07-12	2016-02-19	2018-12-26	2021-04-23	2025-01-24
Date added to HIBP (yyyy-mm-dd)	2020-02-21	—	2013-11-30	2017-05-16	2020-02-20	2023-01-02	2025-02-02
Years till breach is public on HIBP	1.36	1.95	0.0	0.06	0.51	1.94	12.46

Note: Years till the breach is public on HIBP is the lapse between the date of the breach and the date added to the HIBP repository.

Table SI2. Types of data exposed in breaches ($n = 857$)

1–50				51–100				101–146			
	Data type	#	(%)		Data type	#	(%)		Data type	#	(%)
1	Email addresses	851	99.3%	51	Vehicle details	6	0.7%	101	Cellular network names	1	0.1%
2	Passwords	618	72.1%	52	Credit status information	5	0.6%	102	Charitable donations	1	0.1%
3	Usernames	431	50.3%	53	Family members' names	5	0.6%	103	Citizenship statuses	1	0.1%
4	Names	428	49.9%	54	Historical passwords	5	0.6%	104	Clothing sizes	1	0.1%
5	IP addresses	353	41.2%	55	Home ownership statuses	5	0.6%	105	Comments	1	0.1%
6	Phone numbers	276	32.2%	56	Partial dates of birth	5	0.6%	106	Company names	1	0.1%
7	Dates of birth	234	27.3%	57	Personal health data	5	0.6%	107	Customer feedback	1	0.1%
8	Physical addresses	218	25.4%	58	Social connections	5	0.6%	108	Customer interactions	1	0.1%
9	Genders	164	19.1%	59	User website URLs	5	0.6%	109	Deceased date	1	0.1%
10	Geographic locations	121	14.1%	60	Age groups	4	0.5%	110	Deceased statuses	1	0.1%
11	Website activity	77	9.0%	61	Chat logs	4	0.5%	111	Delivery instructions	1	0.1%
12	Purchases	66	7.7%	62	Ages	3	0.4%	112	Device serial numbers	1	0.1%
13	Social media profiles	49	5.7%	63	Nicknames	3	0.4%	113	Device usage tracking data	1	0.1%
14	Private messages	37	4.3%	64	Personal descriptions	3	0.4%	114	Eating habits	1	0.1%
15	Job titles	30	3.5%	65	Photos	3	0.4%	115	Employment statuses	1	0.1%
16	Partial credit card data	29	3.4%	66	SMS messages	3	0.4%	116	Encrypted keys	1	0.1%
17	Employers	27	3.2%	67	Sexual fetishes	3	0.4%	117	Fitness levels	1	0.1%
18	Browser user agent details	22	2.6%	68	Browsing histories	2	0.2%	118	Flights taken	1	0.1%
19	Device information	22	2.6%	69	Credit card CVV	2	0.2%	119	HIV statuses	1	0.1%
20	Salutations	20	2.3%	70	Cryptocurrency wallet addresses	2	0.2%	120	Licence plates	1	0.1%
21	Government issued IDs	18	2.1%	71	Driver's licenses	2	0.2%	121	Living costs	1	0.1%
22	Spoken languages	18	2.1%	72	Drug habits	2	0.2%	122	Login histories	1	0.1%
23	Marital statuses	15	1.8%	73	Financial investments	2	0.2%	123	Loyalty program details	1	0.1%
24	Security questions and answers	13	1.5%	74	Financial transactions	2	0.2%	124	MAC addresses	1	0.1%
25	Bios	12	1.4%	75	Health insurance information	2	0.2%	125	Mnemonic phrases	1	0.1%
26	Education levels	12	1.4%	76	Homepage URLs	2	0.2%	126	Mothers maiden names	1	0.1%
27	Income levels	12	1.4%	77	IMEI numbers	2	0.2%	127	Parenting plans	1	0.1%
28	Profile photos	12	1.4%	78	IMSI numbers	2	0.2%	128	Password strengths	1	0.1%
29	Physical attributes	10	1.2%	79	Loan information	2	0.2%	129	Payment methods	1	0.1%
30	Account balances	9	1.1%	80	Net worths	2	0.2%	130	Places of birth	1	0.1%
31	Auth tokens	9	1.1%	81	PINs	2	0.2%	131	Purchasing habits	1	0.1%
32	Ethnicities	9	1.1%	82	Partial phone numbers	2	0.2%	132	Races	1	0.1%
33	Nationalities	9	1.1%	83	Password hints	2	0.2%	133	Recovery email addresses	1	0.1%
34	Payment histories	9	1.1%	84	Personal interests	2	0.2%	134	Reward program balances	1	0.1%
35	Religions	9	1.1%	85	Political donations	2	0.2%	135	School grades (class levels)	1	0.1%
36	Sexual orientations	9	1.1%	86	Political views	2	0.2%	136	Spouses names	1	0.1%
37	Avatars	8	0.9%	87	Professional skills	2	0.2%	137	Tattoo status	1	0.1%
38	Email messages	8	0.9%	88	Support tickets	2	0.2%	138	Taxation records	1	0.1%
39	Family structure	8	0.9%	89	Survey results	2	0.2%	139	Travel habits	1	0.1%
40	Instant messenger identities	8	0.9%	90	Telecommunications carrier	2	0.2%	140	Travel plans	1	0.1%
41	Bank account numbers	7	0.8%	91	Address book contacts	1	0.1%	141	User statuses	1	0.1%
42	Credit cards	7	0.8%	92	Appointments	1	0.1%	142	Utility bills	1	0.1%
43	Drinking habits	7	0.8%	93	Apps installed on devices	1	0.1%	143	Vehicle identification numbers (VINs)	1	0.1%
44	Occupations	7	0.8%	94	Astrological signs	1	0.1%	144	Warranty claims	1	0.1%
45	Passport numbers	7	0.8%	95	Audio recordings	1	0.1%	145	Work habits	1	0.1%
46	Smoking habits	7	0.8%	96	Beauty ratings	1	0.1%	146	Years of professional experience	1	0.1%
47	Social security numbers	7	0.8%	97	Biometric data	1	0.1%				
48	Time zones	7	0.8%	98	Buying preferences	1	0.1%				
49	Job applications	6	0.7%	99	Car ownership statuses	1	0.1%				
50	Relationship statuses	6	0.7%	100	Career levels	1	0.1%				

Note: This table breaks down the 146 data types compromised across the 857 breaches, listing the number (#) and percentage (%) of breaches involving each data type. A breach can expose multiple data types.

Table SI3. Classification of the 38 Serious Data Classes in Breaches

#	**Category**	#	**Category**
1	Audio recordings	20	MAC addresses
2	Auth tokens	21	Mothers' maiden names
3	Bank account numbers	22	Nationalities
4	Biometric data	23	Partial credit card data
5	Browsing histories	24	Partial dates of birth
6	Chat logs	25	Passport numbers
7	Credit card CVV	26	Password hints
8	Credit cards	27	Passwords
9	Credit status information	28	Personal health data
10	Drinking habits	29	Photos
11	Driver's licenses	30	PINs
12	Drug habits	31	Places of birth
13	Email messages	32	Private messages
14	Encrypted keys	33	Security questions and answers
15	Government issued IDs	34	Sexual fetishes
16	Health insurance information	35	Sexual orientations
17	Historical passwords	36	SMS messages
18	HIV statuses	37	Social security numbers
19	Login histories	38	Taxation records

Note: List of data classes that, if exposed or leaked in data breaches, will pose serious security risks to the individual (e.g., used to gain unauthorized access to other accounts, commit identity theft, or personally sensitive information).

Table SI4. Estimated country fixed effects (ranked by serious breaches)

		Country	Serious breaches	Breaches
1	AUS	Australia	0.891	0.941
2	GBR	UK	0.798	0.830
3	JEY	Jersey	0.788	0.879
4	HKG	Hong-Kong	0.633	0.659
5	DNK	Denmark	0.457	0.687
6	FIN	Finland	0.427	0.695
7	NOR	Norway	0.420	0.722
8	NLD	Netherlands	0.406	0.737
9	NZL	New-Zealand	0.391	0.599
10	ZAF	South-Africa	0.385	0.519
11	IND	India	0.356	0.652
12	URY	Uruguay	0.339	0.355
13	SGP	Singapore	0.334	0.499
14	LUX	Luxembourg	0.333	0.407
15	NIC	Nicaragua	0.330	0.313
16	PYF	French-Polynesia	0.330	0.321
17	KEN	Kenya	0.302	0.581
18	SYC	Seychelles	0.294	0.284
19	UGA	Uganda	0.294	0.316
20	CAN	Canada	0.293	0.865
21	MLT	Malta	0.291	0.612
22	BRA	Brazil	0.289	0.716
23	ARG	Argentina	0.279	0.483
24	BEL	Belgium	0.278	0.430
25	NGA	Nigeria	0.264	0.377
26	BTN	Bhutan	0.263	0.254
27	GGY	Guernsey	0.261	0.274
28	GRC	Greece	0.260	0.491
29	ZWE	Zimbabwe	0.244	0.212
30	EST	Estonia	0.244	0.324
31	HUN	Hungary	0.241	0.319
32	ITA	Italy	0.232	0.354
33	CYP	Cyprus	0.208	0.503
34	BIH	Bosnia-and-Herzegovina	0.202	0.217
35	MDA	Moldova	0.200	0.214
36	SUR	Suriname	0.198	0.188
37	COL	Colombia	0.196	0.193
38	AND	Andorra	0.193	0.182
39	BMU	Bermuda	0.185	0.206
40	NAM	Namibia	0.183	0.174
41	GRL	Greenland	0.181	0.172
42	GTM	Guatemala	0.169	0.166
43	GEO	Georgia	0.168	0.256
44	BGR	Bulgaria	0.163	0.207
45	ARM	Armenia	0.163	0.297
46	ZMB	Zambia	0.154	0.161
47	RWA	Rwanda	0.154	0.145
48	MKD	Macedonia	0.154	0.145
49	CMR	Cameroon	0.154	0.145
50	LKA	Sri-Lanka	0.154	0.145
51	ROU	Romania	0.149	0.159
52	BLR	Belarus	0.146	0.136
53	TZA	Tanzania	0.134	0.126
54	SVK	Slovakia	0.130	0.190
55	NPL	Nepal	0.129	0.132
56	ALB	Albania	0.127	0.120
57	PNG	Papua-New-Guinea	0.107	0.115
58	KOR	South-Korea	0.074	0.084
59	IRN	Iran	0.025	0.045

Note: This table reports the estimated country fixed effects (the $\hat{\gamma}_{c(i)}$'s in Equation (1)), partialing out Personal/Official emails and the decade fixed effects, ranked by the Serious Breach country fixed effects. The Serious Breach column reports the fixed effects from Model (3) of Table 4. The Breach column reports the fixed effects from Model (1) of Table 4.

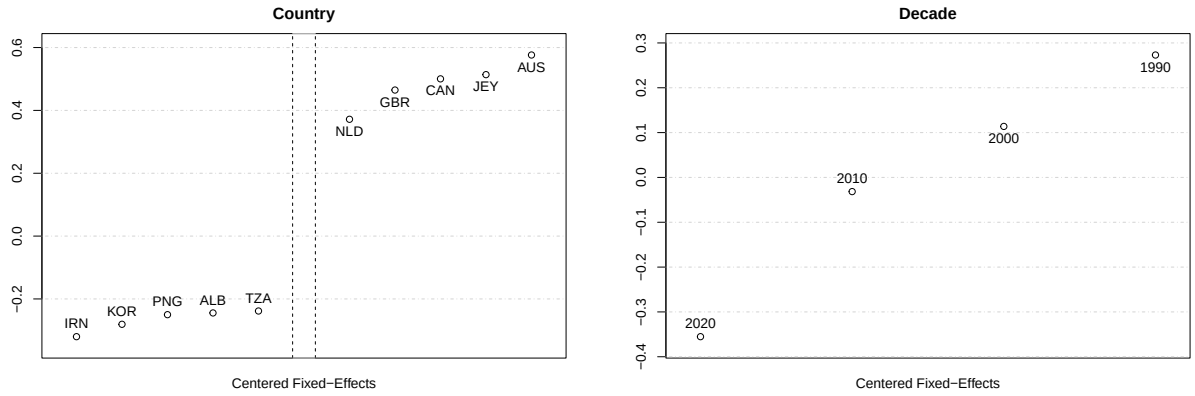

Figure SI2. Country and decade mean-centered fixed effects of the probability of a breach. This figure reports the top five and bottom five estimated fixed effects for the country (left panel) and the fixed effects for a decade (right panel) from Model (1) in Table 4. All plotted effects are centered such that the mean is zero. See Figure 1 for the same plot for serious breaches. See Table SI4 for all countries' estimated effects.

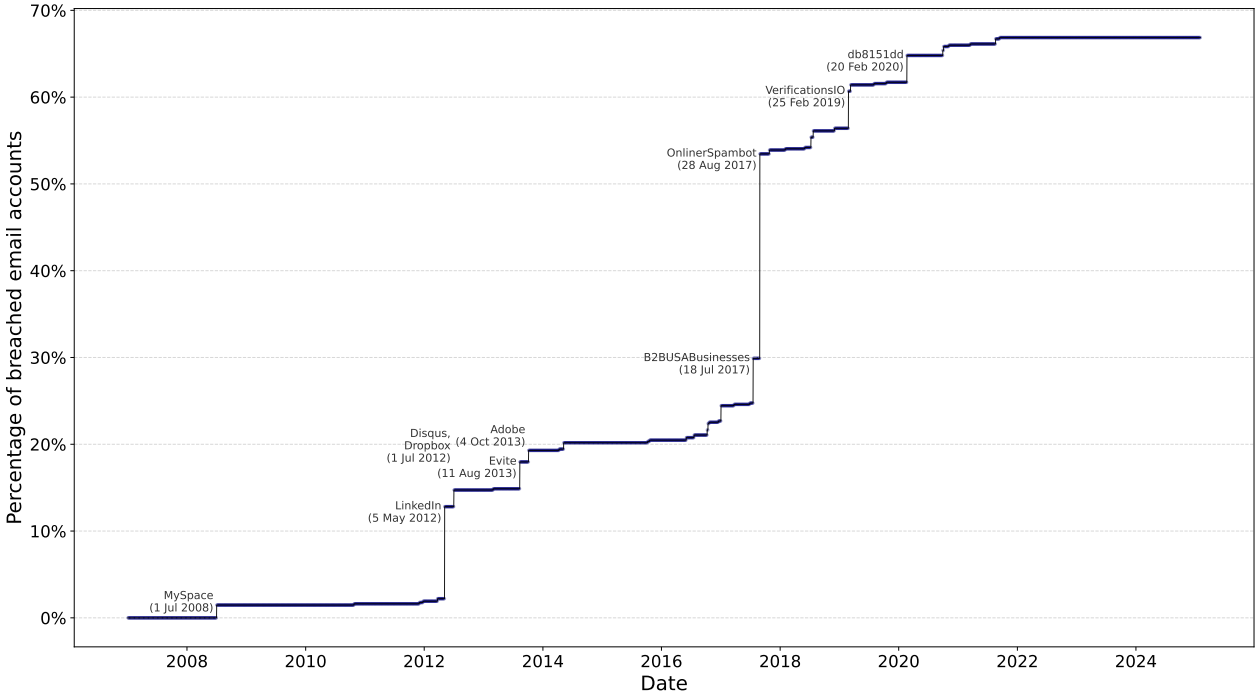

Figure SI3. Breach rate of a fixed cohort of politician emails over time. Similar to Figure 2, except that this plot follows a fixed cohort of n = 679 politician emails where the earliest known date is before 2007 (the first recorded HIBP breach is on 12 July 2007, Table SI1). The top panel annotates a few of the largest jumps in percentage and the corresponding breach incidents.