HUME CENTER FOR NATIONAL SECURITY AND TECHNOLOGY
VIRGINIA TECH

ELECTRONIC SYSTEMS LAB

WHITEPAPER

Use & Abuse of Personal Information

Executive Summary

On a daily basis, we are inundated by a wide array of digital content, some of which we request, and many times material that we do not. Ever wondered how that excess content gets to you?

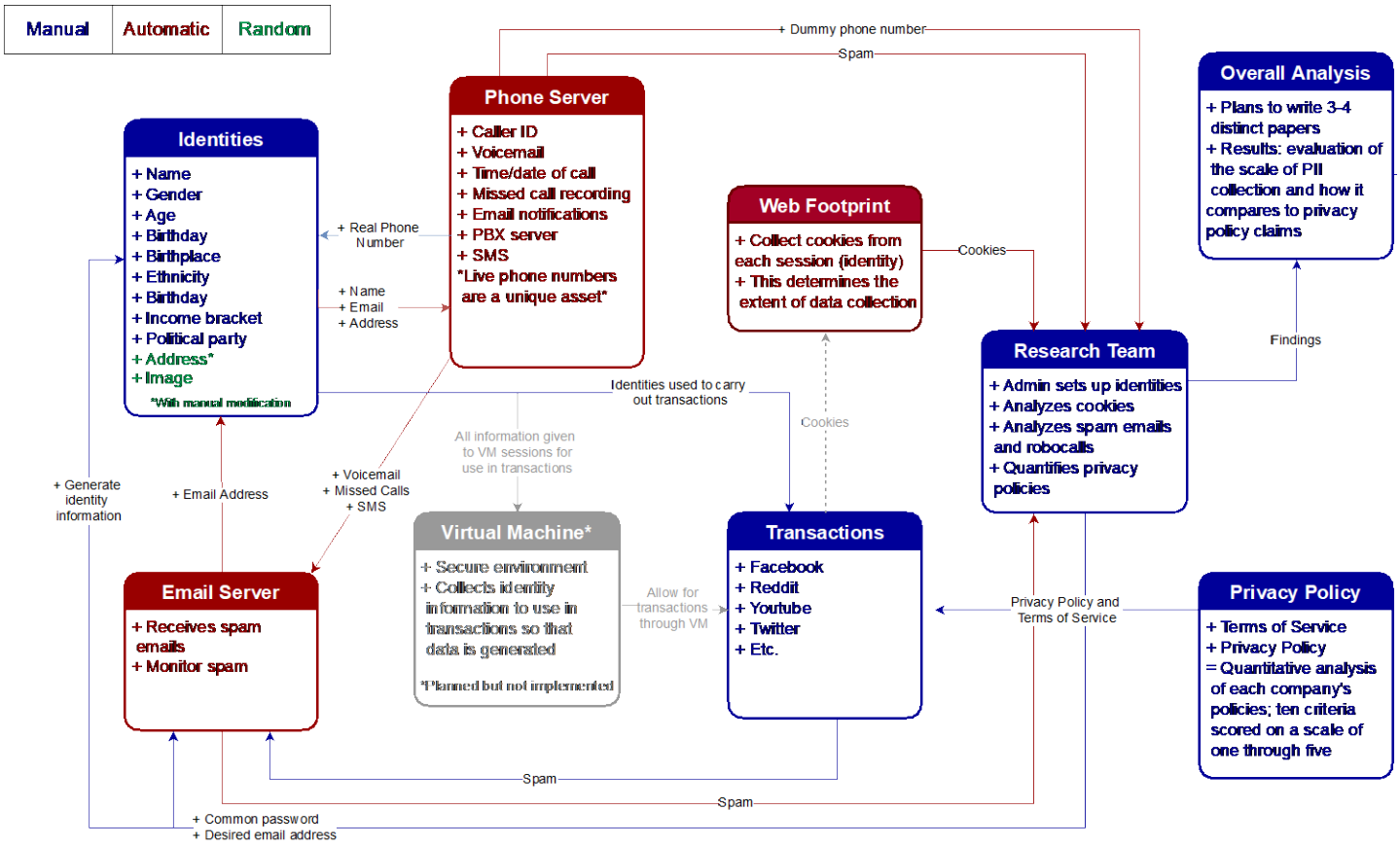
Over the last 16 months, our multi-disciplinary *Use & Abuse* research team, which consisted of 15 students from 10 different majors, has sought to answer that question. To do this, we created a rich source of falsified personal information representing 300 unique identities that we used to perform one-time interactions with a batch of 188 distinct organizations across the Internet. The one-time nature of the interaction, in which we offered any and all personal information that the second party would accept, offers an element of traceability since that single transaction may be traced as the source of all subsequent traffic arising from our online presence. Key elements of this personal information included name, email, a live phone number that supported phone calls and texts, and demographics representative of national averages – it is worthwhile to recognize that this was a relatively small experiment, with assignment of identities to organizations performed with controlled randomness. Organizations were selected to understand behaviors across different industries, different thematic goals, different countries, etc. However, it is an important conclusion and recognition that our selection of companies was probably too conservative – in fact, 290/300 identities showed no evidence of sharing email information, although phone data appears to be more highly shared, yet comes with less paper trail. Most telling was the sheer quantity of traffic generated from the one-time interaction, with one news agency leading the pack with 2436 emails in a 9-month span (and 44 the day before the election)! In most cases, a single identity was selected per organization, and that becomes a unique data point for the company's sharing behaviors, although in many other cases, a falsified web browsing history was leveraged to induce a political identity that can be harvested by cookie inspection, leading to three identities per organization (one left, one right, one neutral). In still other cases, discrete financial transactions were made to ensure that our information landed on customer rolls.

Additional research questions explored using this dataset include those associated with political and social events, which interactions lead to receipt of malicious content, whether evidence exists of foreign interest in the recent 2020 election cycle, whether there are trends across industries or geography, and whether or not a company's privacy policy / terms & conditions give an indication as to how extensively they will use or share our information? We made every attempt in the experimental setup to prevent our own potential biases from skewing the results, so enforced balance across party lines and/or issues; potential variations on these selections will require a larger sample set. On the latter question, our team performed a structured review of 171 published privacy policies against a quantitative privacy policy rubric that evaluates 10 different dimensions of privacy protections from a consumer perspective. The short answer is that the lawyers publishing the legalese and the coders handing the information do not appear to be correlated, though some industry-based trends exist, with political organizations offering the least in terms of promises to protect our data.

Ultimately, the lessons learned from this experiment will help us design a larger experiment that is less conservative in terms of which organizations we analyze, that expands the personal information sources to ~100K identities, and continues the analytical tool development intended to automate how we track incoming content. In addition, we are opening our entire dataset (~20K distinct artifacts) and in-process analysis code on GitHub to invite independent testing as we move towards the large-scale experiment.

Experimental Setup

To set the context for our experiment, our collection of 300 fake identities were composed of randomized personal information derived from a statistical atlas and unique email addresses; in approximately half the cases, identities were also assigned live phone numbers that were also capable of receiving SMS texts, a unique capability as compared to earlier studies on personal information propagation. We used online name, age, and [as necessary] face generation services allowing users to randomly assign age and provides relevant names depending on the user's stated origin. These generator established most of the names, ages, birthdates, and email address names used for our collection process. Emails were established using a custom Rainloop-based mail server, with all received email organized into a single account, yet sorted into distinct folders based upon recipient email; in this way, we also tested for variations on email names and/or generic emails to the server (none were received). Phone and text services were established via virtual phone numbers rented from Zadarma; the online-managed PBX server converted call records / texts to emails that were subsequently forwarded to our email server. Voicemail service was particularly limited as a result of the transcription process cutting off 10-12 seconds at the beginning of voicemails. Additional tools were established to manufacture false web browsing histories and to pre-process the received content based upon known qualifiers of that identity or transaction. Finally, we attempted to develop predictive and quantitative tools for analyzing privacy policies and account terms to identify if those correlate with the observed personal information use and sharing behaviors. A summary of this setup is shown below, with color-coded segmentation between manual, automatic, and random selections.



In an attempt to elicit the use and sharing behaviors of our online transactions, we made semi-random assignments of the 300 identities to a broad selection of 188 companies with representation from multiple industries, multiple sizes, etc. The selected companies were intentionally selected to target sharing behaviors of *mostly reputable* organizations,[1] qualified by a conservative interpretation of content and transaction adhering to all applicable university IT policies and public laws (e.g., no use of dark web resources, avoidance of pornographic sites or other sites likely to send objectionable content, any use of falsified SSNs, and/or Federal Election Commission rules prohibiting anonymized political donations). Finally, a set of 33 foreign organizations were selected, primarily focusing on social media and/or news sites, to evaluation foreign interest in the 2020 election. A complete list of the organizations and their categorizations is included below.

Companies have been color coded according to their economic sectors:					
Consumer Staples/Defensive	Communication Services	Industrials	Political Organizations	Real Estate	Other
Online Retail/Cyclical	Hospitality	News/Media	Software/Technology	Restaurants	

Domestic Organizations						
7-11	CNET	Flikr	Kohls	Papa John's	Target	Yelp
ACLU	CNN	Food Lion	Kroger	Pepsi	The Guardian	Yidio
ACM	Coca Cola	Food Network	LA Times	PETA	Tiktok	YouTube/Google
Adidas	Collegiate Times	Fox	LinkedIn	Pinterest	Tim Kaine (VA Senator)	Zillow
Alesis	Communist Party	Free Movies	Lowes	Planned Parenthood	Toyota	Zoom
Amazon********	Consumer Report*	g2a	Lyft	Player Auctions*	Trip Advisor	
American Airlines	Costco	Glassdoor	Macy's	Poshmark	Trulia	
Apple	Csgv.org	Go fund me	Marriot	Pro-Life Action League	Tumblr	
Atlanta JCC	CVS	Godaddy	Match	Putlocker	Twitch	
Autotrader	Dccc.org	Gop.gov	McDonalds	Quizlet	Twitter	
Autozone	Delta	Green Peace	Medium	Quora	Uber	
BBC	Denver Post	Groupon	Miami Herald	Realtor	UNICEF	
Bed Bath and Beyond	Discord	Healthline	Michaels	Reddit*	US News	
Best Buy*	Dollar tree	Hi5	Microsoft	Retail Me Not	USA Today	
Bleacher Report	Dominos	Hilton	Mitch McConnell	Roanoke Times	VA Citizens Defense League	
Breitbart	DonaldJTrump.com	Home Depot*	Motor Mile	Rotten Tomatoes	Walgreens	
bstock	Dunkin Donuts	Huffington Post	Moviesjoy	Safeway	Walmart	
Business Insider	Ebay	IKEA	MSN	Sheetz	Wayfair	
Carmax	eHarmony	IMDB	NAACP	Slack	WeatherBug	
Cars.com	ESPN	Indeed	Netflix	Spotify	WebMD	
Carvana	Etsy	Instagram	New York Times	Squarespace	Wendys*	
CD Keys	Expedia	Jimmy Johns	Newegg	Starbucks	Whatsapp	
Chicago Tribune	Exxon	Joebiden.com	Nike	Steam	Wish*	
Chick-Fil-A	Facebook	Kickass Torrents	NPR	StubHub	Wix	
Chipotle	Family Research Council	Kinguin	Panera	Taco Bell*	Yahoo News	

Foreign Organizations	
20 Minutes	PesaPal
Alibaba	Rakuten
Asos	RuTube
Badoo	Shaadi
Cookpad	Sputnik
Discovery Store	Stuff
Douban	Taringa
Goalzz (KOOORA)	Tokopedia
Hatena	Toutiao
Hudson Bay	Twoo
JD Sports	VZ
KrisShop	XING
Leboncoin	Yandex
Lefigaro	Yandex Disk
Millat Facebook	Zhanqi
Ouest France	

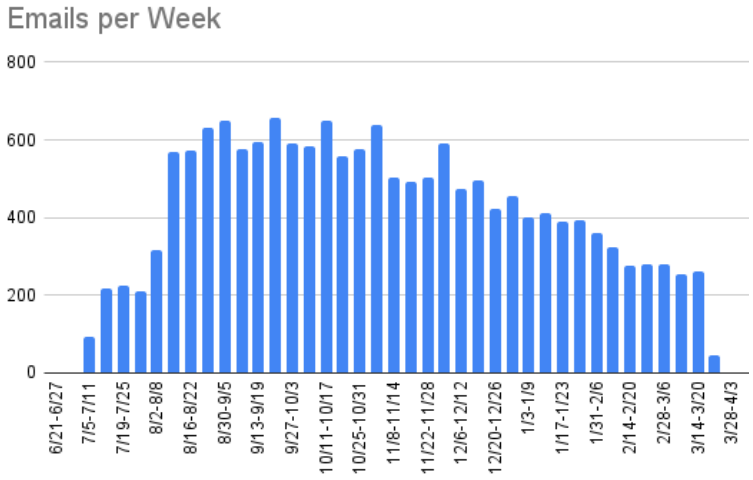
* indicates a financial transaction
Multiple asterisks indicate multiple transactions, each by a different fake identity

The remainder of the whitepaper focuses on the methodology, data analysis, results, and attempts to extract relevant conclusions from the received dataset. This includes data analysis of email traffic, voicemails, and SMS texts for raw quantities, content, and trends across data. In total, there were over 20,000 raw data records. Additionally, the focus on privacy policies and their correlation to observed personal information use and sharing behaviors is then analyzed, with an attempt to provide a quantitative basis (as yet, uncorrelated).

Finally, there is wide recognition that our analysis represents that of a relatively small dataset – most online transactions included only a single identity per second-party organizations. Thus, sweeping conclusions for specific companies are limited to the observed traffic of one individual and there is limited ability to determine if those behaviors change based on demographic information or other account characteristics. The lessons learned from this project are intended instead to seed the design of a large-scale experiment (e.g., 100K identities) that will explore those questions in more detail. Finally, the raw and processed datasets, data analysis, and preliminary tools developed in this project will be open sourced on our GitHub page for any readers interested in the dataset or in processing to develop other conclusions.

[1] Interpret this with the same level of skepticism as *mostly harmless*.

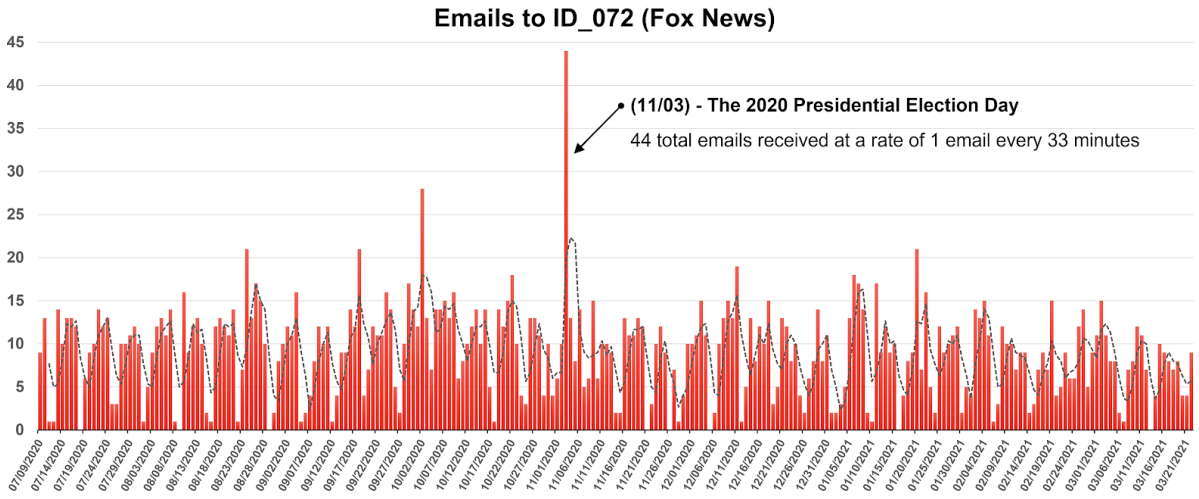
Data Collection/Analysis



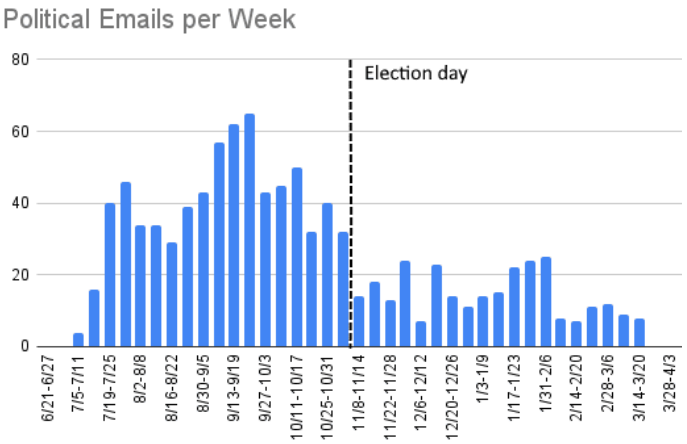
Emails

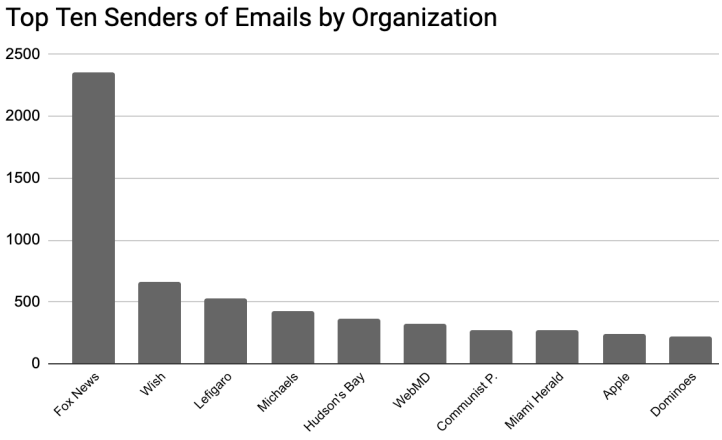
Using the emails collected over our 9-month data collection period, we analyzed the number of emails received per week. A total of 16,540 emails were sent over nine months, with an average of 2 emails per week for each account. Since we used a single online transaction (July 2020) to spur this activity, and proceeded in a receive-only fashion, it is also noteworthy that the email traffic decayed over time as online entities sensed the inactivity on our side.

The winner for excessive use of this email contact information was Fox News, who sent 2,356 emails throughout the project, averaging 9 emails per day. On the day prior to the election, we averaged one email every 33 minutes. For comparison, the next largest sender of emails was online retailer Wish, with a total of 658. The average number of emails received per account was 55.



Given the highly contested election, we also isolated content from political organizations, which were primarily focused on encouraging recipients to donate money to campaigns and vote, so the vast majority of emails from these organizations were received before election day. Other emails sent by these organizations were responses to significant events that had happened, like the passing of Justice Ruth Bader Ginsburg, the subsequent nomination of Amy Coney Barret, and Donald Trump's second impeachment trial.



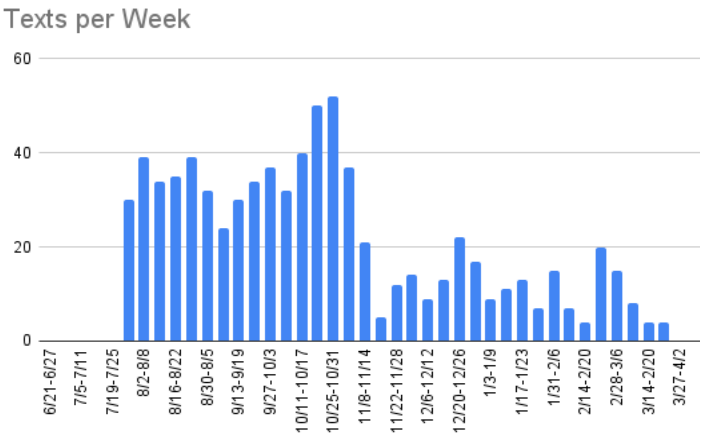


In addition to trends over time, other dynamics were observed in the email data. No specific correlations within industry exists for the excessive users of email information (top 10 shown to the right). Additionally, two foreign companies -- Lefigaro and Hudson's Bay -- were among the top ten senders, which is roughly proportional to the number of foreign entities studied overall. Next, we sorted received emails to determine if the website each respective fake identity had subscribed to matched the sender's email address for the emails each identity received. If they did not match, it was established that that organization had shared user PII with a third party. From our review of the emails, we found that most companies (290/300) did not noticeably sell our personal information, as the senders' email addresses virtually always matched the fake identities' assigned companies, allowing for emails from different servers at the organization. When determining how many emails were sent by third parties, we sorted by identity and then by company, labeling in-group emails as 0 and out of group emails as 1. We erred on the side of the in-group in terms of labeling. For example, we received emails from NPR with a sender of membership@vt.edu. We presume that NPR recognized that we were from Virginia Tech due to our email ending in ".vt.edu," and put us on that mailing list in lieu of sending individual emails. Thus, membership@vt.edu was counted as an in-group, despite it differing from NPR account holder's actual email address.

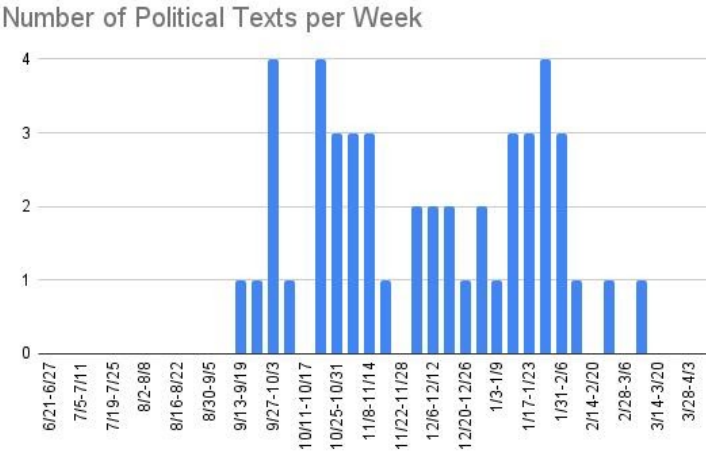
Potential scams include three emails to the account signed up with G2A from "hulu@hulumail.com," as well as four emails to the account signed up to Michaels asking us to verify our GMail email addresses. We received messages advertising Etsy from our email account that we signed up for B-Stock, leading to a presumption that B-Stock sold our personal information. We also received emails to the account signed up to Cookpad from Badoo. Additionally, we found that political organizations shared user information between candidates and national organizations. We also found that the identity assigned to the Communist Party USA received an email from the domain of member@surveymonkey.com (as well as other affiliated websites), but it contained a survey the Communist Party wanted its members to take. Thus, it should be noted that SurveyMonkey likely has CPUSA members' contact information due to that survey.

We detected some amount of PII sharing from Twitter, Facebook, Reddit, and TikTok. Unfortunately, due to incorrect setup (direct sharing of email to political organizations rather than just generating a web history), we have excised the results of ID_179 and ID_156, which were assigned to Facebook and Reddit, respectively. Thus, the results for Facebook and Reddit's information sharing were considered invalid since we did not see similar behaviors across other accounts with the same organizations. However, with a reasonable degree of certainty, we found that our fake identity's PII leaked from Twitter to the Republican party, and likewise, PII from TikTok to the Democratic party; from the configuration of those accounts and the seeding of political identities, we posit that sharing occurred through cookie tracking and falsified browser histories.

SMS Messages



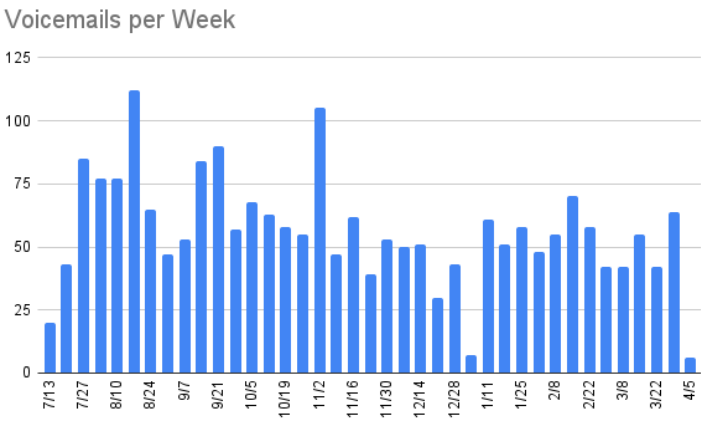
The fake identities received fewer SMS text messages than emails and phone calls; however, a few trends can be noted. We received messages saying "call me back" or "please text me," and one providing a Facebook confirmation code we had not requested. These were assumed to have been sent to the wrong number and intended for someone else. The fake identities also received a proportionally large number of malicious content and scam messages, likely because phones have less protection than computers and are thus easier to target. Also contributing to the malicious content was borrowed phone numbers that may have already become associated with scam messages before we began using them. Throughout the 9-month data collection period as a whole, we received 774 text messages distributed unevenly amongst 150 phone lines.



For the political subset, we found that political SMS messages extended well past the election, unlike the observed behaviors for email. In terms of political alignment, we found that the accounts subscribed to Republican organizations received far more SMS texts than those subscribed to Democratic organizations. Looking at a timeline of political events, and specifically the *RealClearPolitics* betting odds on the election outcome, we found reasonable correlation with the current underdog sending the most traffic; in fact, Biden's traffic nearly ceased prior to the election. While the five Democratic websites we studied -- JoeBiden.com, the NAACP, and others -- received three text messages over the entire period of study, the five Republican-leaning websites studied received 47 messages total; 42 of the 47 Republican messages received were sent by the Family Research Council, and all three Democratic messages were sent by the NAACP. However, it should be noted that some other websites sent political SMS content, despite not being listed as political websites in our data collection process. These include Reddit and Facebook, whose accounts have been removed from our analysis due to a mistake in the account creation process, and TikTok and Twitter.

Voicemails

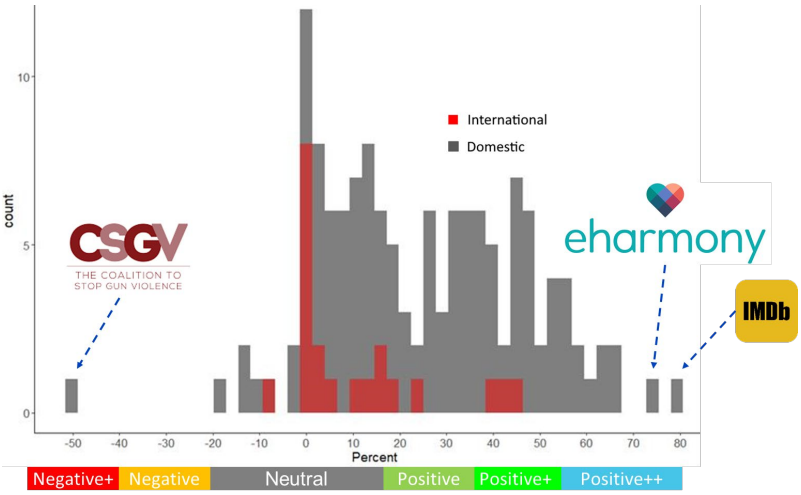
The fake identities received a total of 887 voicemails over the nine-month data collection period, 40% of which were *silence*. We were limited in our ability to draw firm conclusions on this data since the phone numbers that we rented from Zadarma had been used before, and we have no way of knowing what had been done previously with the numbers. Trends do indicate a positive correlation through between email information users and phone information users.



Likewise, the classification of the voicemails was limited by Zadarma chopping off 12 seconds of each voicemail. This means that we received a large number of *silent* messages as the caller was ending their call; and calls for a custom/improved PBX server in future experiments. Additionally, repeated robocalls comprised a more significant proportion than presumably expected, given the rest of the data. However, from what we gathered, we saw that 10% [2] of all voicemails were the same "extended vehicle warranty" scam, and 17% of all voicemails were malicious.

Surprisingly, we did not receive any voicemails that could be classified as either “Political Donation” or “Political Call-to-Action.” However, fake identities registered with political organizations did receive voicemails that were a variety of classifications. Since none of these voicemails were inherently political in nature, we cannot draw any conclusions about how those organizations protect or share their information due to the ease with which hackers can spoof phone numbers.

Additional analyses on the data included a Natural Language Processing (NLP)-based evaluation of emails using the VADER Sentiment Analysis model, which quantifies the emotional polarity of phrasing in text passages. The two overlapping colors show the distributions between the international companies and their policies and the domestic companies. The most extreme outliers for this analysis include the Coalition to Stop Gun Violence (selling fear) at the bottom and upper extremes of Eharmony (selling future love) and IMDB (they apparently really love their movies!), all of which match a common sense prediction of their content.



Our biggest takeaway from the raw data is that, while there are some notable extremes, the content received offers much more in the way of lessons learned and design of a future experiment than in specific soundbites of attributable sharing behaviors. The conclusions are fundamentally limited given the sparse coverage (300 fake IDs split across 188 organizations) and bounded inclusion of organizations fitting the previously described *mostly harmless* content limitations. Nevertheless, this dataset offers a tangible basis to develop future automated tools for ingesting and processing information sharing behaviors.

[2] The true value may have been more, but our manual data classification techniques led to these reported numbers.

Privacy Policy Analysis

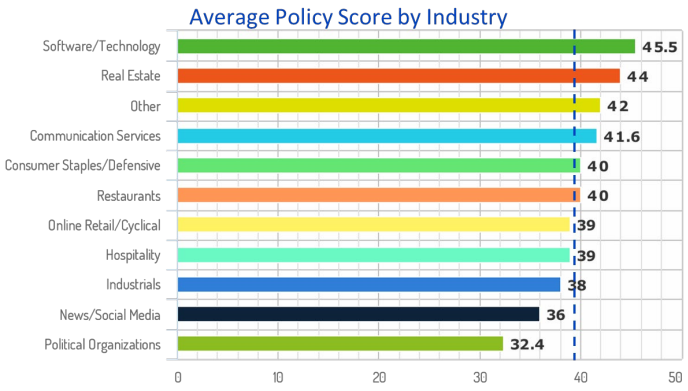
To analyze the privacy policies of all collected services, we created a standardized system for qualitatively grading the linguistic patterns present in various companies' policies and how they relate to privacy concerns. This allows a clear comparison to be drawn between different services in terms of how much consideration a given company claims to provide for its end users' privacy.

This system takes the form of a categorized ranking table, which was used to assign numeric values (1-5) for ten categories of linguistic patterns found in a given company's policy documents. These comparison phrase-value mappings were determined based on recurring themes amongst the many privacy policies being analyzed, focusing on the end user's privacy and how certain policies limit access to their PII or increase PII exposure to multiple sources. Further, established studies in this area highlight the importance of how companies use this data, how long it is stored, and how it is being protected (Barker et al., 2009).

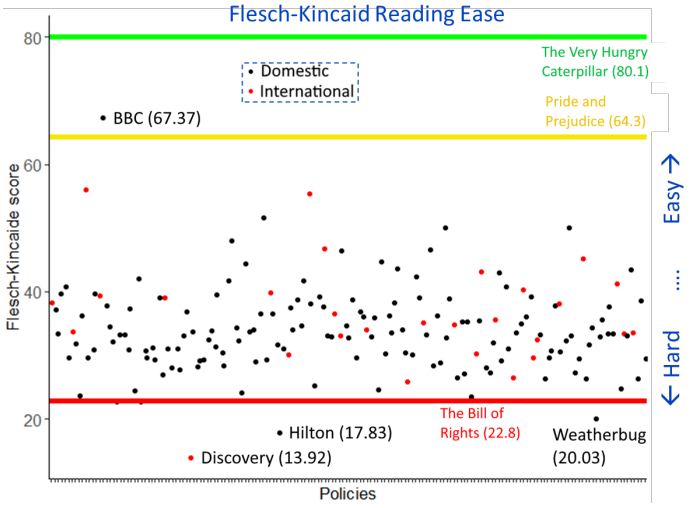
The wide pool of companies yielded diversity in privacy policy content; however, the basic structure for many policies consists of the same sections. For example, every policy category in Reddit's privacy policy is similarly shared in the policy of Apple and Bed Bath and Beyond. This is primarily due to legal requirements to inform the consumer of the stated legal rights of every individual, especially concerning what information they collect, how they distribute that information to third parties, and how they intend to retain this information (Zimmeck et al., 2019). Thus, to accurately score the privacy policies, we conducted a literature review of policy models and found the most important and relevant criteria that policies could be scored by; this allowed us to construct a privacy policy quantification table that accurately evaluated privacy policies. Specifically, TOS:DR (Terms of Service: Didn't Read) was especially useful in determining which categories should be used for our analysis. Similarly, studies done by Carnegie Mellon have established comparable evaluation metrics (Wilson et al., 2016). Using criteria as determined through our literature search, we created categories within each section, attempting to segregate policies equally within a five-tier rubric, with one being the lowest score and five being the highest.

Once the policy evaluation metric was established, 188 international and domestic companies were selected for evaluation based upon general user sentiment and general traffic. To analyze all 188 companies' records as well as compensate for variance in scoring between reviewers, pairwise combinations of the four student reviewers scored the policies to catch any errors. Any score pairs that differed between reviewers were resolved through one-on-one discussion or arbitration discussions with the advisor. This ensured that no individual's interpretation of the scoring system differed greatly from the others throughout the process.

To determine a correlation between policy score and industry, we sorted companies into their respective industries as determined by the stock market index category. Companies were placed into 11 categories, including cyclical, defensive, industrials, and technology companies. Political organizations were separated into their own group. Out of all the industry scores, the highest industry average on the policy quantification table was Software/Tech



with a score of 45.5; the lowest industry average was the political organizations with a score of 32.4, which was lower than the general policy average of 39.8.



For our quantitative results of the privacy policies, we found that, based on the Flesch-Kincaid readability score, the BBC Privacy Policy was by far the easiest to read with a score of 67.4 - slightly higher than *Pride and Prejudice*'s score of 64.3; Such a score corresponds to a reading grade level of 7.2. Three policies rate as more difficult to read than the Bill of Rights. The policy that had the longest word count was Indeed. Assuming an average reading speed of 250 words per minute, it would take 280 minutes to read Indeed's entire policy. However, it should be noted that Indeed's Terms of Service contained the terms for subsidiary programs, employers, and users on the same page, most of which do not apply to the average consumer. Overall, the average privacy policy requires 100.6 minutes to read as compared to the average 58 minutes worth of content received over 9 months from an account.

A summary of our privacy policy dimensions are listed below, with complete definition of the quantitative rubric on the next page.

- **Changing Terms** - If the underlying website/service states they have the right to change their policy without notifying the user, this makes it hard for the consumer to determine suitable follow-up action.
- **Holding Service Harmless** - If the privacy policy of a particular website/service states that the user must defend the service in spite of a related lawsuit, this is cause for concern.
- **Ignores Do Not Track (DNT) devices** - If the privacy policy of a website/service claims to ignore the DNT setting put in place by many modern browsers, we see this is potentially concerning.
- **Personal Identifiable Information (PII) used for Ads** - If the website/service uses information entered by the user to target ads, it can be a concern. While there are cases where information is sold, this criteria focuses more on internal use of the information to target ads/services back onto the user.
- **Release of information to third parties** - If a website/service releases or sells information to a third party of a user without explicit notification to the user, it undermines a consumer's personal privacy.
- **Signing away moral rights** - Not to be confused with moral values, when one signs away moral rights, they lose the ability to have total control over their work or, in this case, "presence" on a particular website. By signing these rights away, the overlaying company can use your work/status however they please.
- **Retention of Personal Data** - When a user leaves a service/company, we see this as a privacy concern if the service mentioned above retains the user data for profit.
- **Information being sold due to Bankruptcy** - As a result of bankruptcy, some companies sell users' information as assets, which can be concerning to personal privacy as it signals increased retention.
- **Puts sole risk on users for breach of PII** - No data is completely safe. Companies constantly share or sell this information for profit which leaves it open for malicious hacking. Despite these concerns, if a company/service places the responsibility on the user for data security, it can be harmful to a consumer's personal privacy.

Score	Changing Terms	Holding Service Harmless	Ignores Do Not Track (DNT) Devices	Personal Identifiable Information (PII) used for Ads	Release of information to third parties
1	Change privacy policy with changes applicable retroactively	User must defend the service against any claims/costs/liabilities if any lawsuit arises	Does not acknowledge or mention DNT signals	The service internally collects any available information of the user to sell and/or create targeted ads.	The service consistently sells/distributes PII to associated third parties for any purpose, including undefined "business purposes."
2	Change privacy policy without notification, but changes are forward-looking.	User is responsible for defending the service in cases where the user violated the company's privacy policy.	Complete recognition of these signals and denies the user the right to the website and/or continues to track the user without notification.	Collects a significant amount of PII (i.e., address, contacts, and site browsing activity). Does not collect all available PII but more than as specified in Category 3	The service collects and sends PII to third parties for them to sell advertisements or for defined "business purposes."
3	Claims to give notice but provides vague distribution details	User is responsible for defending the service in cases where the user violated others rights/broke the law, not from policy violation	Acknowledges DNT signals and continues to track only due to lack of infrastructure to support these settings/lack of standard	Collects a 'normal' amount of PII, including name, email address, log data, general location data ascertained from IP address, etc.	The service only releases information to third parties if the user requests a service/ more information from the initial website
4	Clear notification of changes in the privacy policy	User is responsible for defending the service in cases where the user violated others; however, the service can remain accountable if they played any role in the digression	Acknowledges DNT signals and complies; however, the service does not allow full access to all of the present features	Service provides a menu to disable all but necessary cookies and collects a normal/less than normal amount of PII as defined above	The service releases PII only with previous consent from the user to show the user more relevant content
5	User permitted to opt-out of privacy policy changes/ allows for extensive copies of previous policies to ensure changes.	Service assumes the risk and takes liability away from the user if a lawsuit arises.	Service complies with DNT signals and allows the user access to the full features of the service.	Minimal to no PII is collected or used for internal targeted products or services. The user still has access to the full features of the product.	The service releases little to no information to third parties regardless of user consent and maintains internal consistency with the user's PII.

Score	Signing away moral rights	Retention of Personal Data	Deletion of PII upon request	Information being sold due to Bankruptcy	Puts sole risk on users for liabilities
1	A complete dismissal of these rights and liability of suit when the user agrees to a particular privacy policy	Full retention of all data indefinitely after a user deactivates their account	The service does not offer such a feature or continues to retain information despite a request from the user	The company/service will sell and contribute all stored customer data as the result of being bought out or merging with another company	Puts total risk on the user for any liabilities, and the service as mentioned above is not held accountable
2	The user obtains some say over their content; however, the particular service maintains most of the control	Service holds information for as long as they deem necessary/after a predefined extended period of time longer than a year	User is unable to request or delete any information; however, the service will allow less information to be collected	User is notified of acquisition; however, no action can be taken by the user to limit data being transferred	User maintains soles risk on every aspect of the site; however, service can be held liable to distribute cash compensation up to twenty dollars or in extreme cases
3	Rights are waived; however, the privacy policy places some liability on the company, and users maintain almost equal control	Service temporarily holds a reduced quantity of information or retains PII in case of potential reactivation	A user is able to request their information; however, they are unable to delete any information or request to delete is not honored	In merger or asset sales, data is sent to receiving company under the pretense of equivalent or improved privacy standards	The user and the service are mutually responsible. The service uses good faith to ensure data security and information accuracy. Will not claim responsibility for negligence
4	Waiving moral rights is optional; however, the service still has the final say over user content on the service	Information is stored after deletion of account only to comply with applicable regulations. A scheduled deletion is still in place with no intention of prolonged storage	A user is able to request to delete all their information; however, they may not be able to delete most of their information only some	User is notified that their data is forfeit due to bankruptcy/merger; however, they may only be able to delete certain aspects of their PII. Some will be transferred over to the acquisition company	Data breaches caused by the user are not protected; however, if the service experiences a breach in their databases or any other circumstance, the user is not held liable. User is protected on service negligence
5	The privacy policy states you are not required to sign away your moral rights.	Either a user can delete all PII upon deactivation or request, or companies collect no user PII (in which case retention is impossible)	A user is able to request all their information and delete it upon request with assurance from the service that the information will be rightly processed.	Either all information is forfeit and not part sold to the takeover company, or the user is notified and has an opportunity to delete their data before it is sold.	The service is completely responsible for breaches on their end and/or not all risk is placed on the user.

Conclusions from *Use & Abuse* Analysis

After collecting phone and email information for nine months, and analyzing the privacy policies, here is a consolidated list of takeaways:

- Very few companies transferred our personal information in ways that we can confirm; *abuse*, however, is evident in terms of the over-use of personal information in many cases. Email lists appear to be generally retained and protected by the second party, while phone numbers appear to yield much more suspect results, though conclusive tracing is limited by the virtual phone service.
- The average amount of emails per company was 55 emails over 9 months. Likewise, the average time spent per company consuming their proffered content was around an hour.
- While we did not find correlations between the Flesch-Kincaid score, our grading score, or word count, we did find that the average time to read a privacy policy and relevant documents was 46 minutes without hyperlinks and 100 minutes with one minute added per hyperlink.
- We found that use of email lists does not appear to have strong trends based on industry, geography, or other immediately identifiable factors. Outliers (e.g., Fox News) can be rather extreme though.
- While we did not have a large enough sample size to detect any non-extreme differences, there does not appear to be a significant difference between foreign and domestic companies in terms of number or frequency of emails sent, stated interest in election outcomes, or privacy policies.
- We found that ~10% of voicemails we received were the same vehicle warranty scam; however, Zadarma cut off 12 seconds of each voicemail, removing half of our voicemail data as around half of the voicemails received were 12 seconds or less.
- On the political side, we found a nearly 2x ratio of content sent by Republicans over that from Democrats, with approximate timing of this content aligned with the current underdog of the election based on time-varying betting odds. Political content dropped significantly in emails after the election, yet stayed constant substantially longer via SMS activity.
- There was a general decline in emails over time, indicating second parties pruning our falsified email addresses for lack of activity.
- We can deduce that the following companies have released our information: FreeMovies, G2A, Michaels, B-Stock, Cookpad, CPUSA, Twitter, and TikTok as fake identities associated with these organizations received emails from third parties.
- Finally, as a small-scale experiment that evaluated sharing behaviors of a conservatively selected set of companies, we will need to significantly open the aperture in a future scaled experiment to incorporate less restrictions on the companies, their content, and our collection infrastructure to find the true sources of spam. Our core assumptions as to information sharing are also being re-considered via more extended literature surveys. This process is ongoing and is planned for deployment in 2022.

Lessons Learned and Future Research

This 12-month-long project has been a learning experience for all involved, as we discovered which aspects of our experimental setup have worked well, those that could have been done better, and certain unexpected complications that substantially affected our findings. It is particularly worthwhile to note that small variations in the process can have significant impacts to the observed results. Moving forward, with project scalability and repeatability in mind, we have compiled an overview of technical challenges overcome, lessons learned and suggestions for future work extending from our research.[3]

- **Rainloop server**: in establishing the email server, there was a default limit of 50 different filters. This made it challenging to ensure all the emails are appropriately routed to their corresponding identity folders. The inbox served as the central meeting point for all emails. However, secondary searches in these extra profiles indicated hidden emails sent by studied transaction-based websites.
- **Collection infrastructure**: due to a limitation in the Zadarma PBX server, we only received content for about 900 of the voicemail files, which were manually categorized onto a spreadsheet, and the recorded audio files we did receive had the first ~12 seconds cut off. Additionally, it appears you set off no-fly list style warning bells when buying clusters of virtual phone numbers.
- **Account creation**: With the creation of 300 Identities, we wanted to create a diverse portfolio of websites and marketplaces that we believed would be attracted to our information. Amazon was quick to catch on to our account creation. After the initial transaction, we completed three purchases on separate accounts, and the last profile was flagged regardless of preemptive routed connections. Facebook was similar; however, it caught on before the initial accounts could be created. After submitting the assigned Zadarma number for account creation, the account required an uploaded photo with a manual representative review. Finally, we believe Google had the best filtering for web-based phone numbers as it would not allow account creation with the assigned Zadarma number. Google remained consistent across multiple accounts with denying the verification numbers
- **International accounts**: the international profiles posed multiple barriers. Some websites, which have been previously researched, have changed names/owners, making it harder to find specific websites. In some cases, the websites would allow the initial transaction to be completed. However, we would not receive any confirmation emails/phone calls, initiating a state of radio silence with the transaction site. Finally, most of the Asian transaction sites required a local cell phone to interact with the app. Websites like this included QQ, Qzone, Tencent Video, WeChat, Toutaio, etc. Using the pre-ordered U.S.-exchange online numbers made using these sites impossible, so additional research was required to find transaction-ready sites compatible with our collection.
- **Account establishment**: the initial process of creating the Fake IDs' PII, website accounts, and political history was extremely time-consuming and potentially prone to human error given the divide & conquer approach with many students; any up-front errors propagate significantly through the collection period.
- **Inactivity**: over time, we noticed that many companies reduced interest in our fake IDs, based on how many emails they were sending us. We attributed this to the senders using "read" receipts and/or remote tracking content to detect that we weren't opening any of them, resulting in decreased message frequency over time. Future experiments must also stimulate false activity.
- **Privacy policies**: the manual policy analysis phase was also very time consuming and prone to human error; we mitigated this as much as possible, yet a scaled experiment will require an automated method of "reading" and scoring the policies. We ended up re-grading privacy policies at the end of the nine months, which may not be accurate if they had updated their privacy policy. Ideally, we would have started by creating the policy document files to lock in the version, but the way we did it had some websites update their policies mid-way through analysis.

[3] For more in-depth documentation, please refer to "appendix" documents in our public GitHub repository

Future Work

As described so far, the basic results and lessons learned from the *Use & Abuse* experiment will be the foundation for a future scaled experiment. We intend to spend the next year automating analysis tools and refining the design of the larger experiment. In many cases (email sorting when filters failed, voicemail analysis, reading privacy policies), straightforward manual approaches were used, which do not scale to the larger experiment, but the results will provide us a concrete dataset and a baseline for independent verification of the developed tools. This tool creation is anticipated to include improved up-front design of email filters, implementation of our own PBX server for voice and text content, speech-to-text converters (voicemail analysis), and enhanced NLP-based analysis of email content. We are not confident that existing NLP-based tools will be able to replicate the assessment of privacy policies, yet improved key word searches may be able to accelerate the process.

Other future work will be aimed at (1) the creation and hosting of confirmed sharing networks, whereby causal traces of email sharing will be identified for anyone to see, (2) expanded coverage of second parties to include a far-less conservative selection, (3) individual coverage of every major political race (U.S. House, Senate, and presidential elections) to illuminate any sharing abuses, (4) non-invasive methods of stimulating activity in our fake accounts that pass the smell test of inactivity sensors, and (5) evaluation of differential sharing behaviors based on user demographics.

Data Available for Open-Source Archive

To facilitate other understanding the data and performing their own experimentation, we will host our raw dataset, collection method description, and preliminary tools on GitHub at https://humeesl.github.io/Use-and-Abuse-PII/.

- From the experimental outline, we will be providing access to the .xlsx spreadsheet containing the full sets of demographic and PII generated for our 300 Fake Identities, as well as summary statistics on data collected for each profile. Documentation for our design process and experimental methods will also be available.
- From the Data Collection Phase, we have compiled a large database of .xlsx spreadsheets with .mbox files exported from our email/phone server, which includes the fully parsed message and metadata for all 20,330 / 21,641 emails, voicemails, phone call logs, and SMS text messages collected over the 9-month study period. We are also making available all of the 948 voicemail .mp3 audio files received through Zadarma, accompanied by the spreadsheet used for the brute-force classification effort manually done by researchers. All spreadsheet data entries are conveniently indexed by date and include the Fake ID numbers associated with the PII used. Additionally, we have a lighter-weight Database Summary spreadsheet with logs of time-stamped data entries for every email, voicemail, call, and text message received - designed for time-based frequency analysis, to be easily navigated or parsed through by users or data processing software.
- From our Privacy Policy Analysis, we have both plaintext and Word-formatted offline file versions of the Terms of Service and Privacy Policy documents we used for each company studied, since the versions of these documents available online are routinely updated. We are providing a spreadsheet of the qualitative scores assigned to each company and quantitative readability metrics derived from their policy documents.
- From our Political Timeline Analysis, we have a spreadsheet for our central Timeline of Politically-Charged Events with frequency data of emails/vmails/calls/sms received per day, as well as several specialty spreadsheets breaking down the data by political party/organization for certain IDs over time.
- Lastly, we will also provide all Python source code files used for automating certain steps of our study, such as exporting data from our email server with IMAP and creating database spreadsheets from it, as well as conducting statistical, readability, and sentiment analysis on our wealth of accumulated data. Please note that any and all hyperlinks embedded in emails and SMS messages are unfiltered and since potentially malicious links have not been thoroughly scanned for viruses, **DO NOT CLICK ON OR OPEN** any links found in our dataset unless you have taken appropriate precautions (our dataset is provided as-is, so we assume no responsibility for potential damages).

Works Cited

Scaling Privacy Compliance Analysis to a Million Apps. Ken Barker, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam Majedi, Sampson Pun, and Adepele Williams. A data privacy taxonomy, 2009.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichan-der, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and NormanSadeh. Maps: Scaling privacy compliance analysis to a million apps.Proceedings on Privacy Enhancing Technologies, 2019(3):66–86, 20

(Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014).

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Rus-sell, et al. The creation and analysis of a website privacy policy corpus. InProceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1330–1340, 2016

Link(s) to Appendix Content

The *Use & Abuse* public repository is located at: https://humeesl.github.io/Use-and-Abuse-PII/

humeESL / Use-and-Abuse-PII

in a read-only format. While we are attempting to open this data to other researchers, we intentionally removed some content (e.g., passwords, personal information of the research team) and disabled all of the accounts and collection servers to mitigate nefarious use.

Acknowledgements:

The authors wish to thank the Hume Center's Electronic Systems Lab, the Virginia Tech Institute for Critical Technology and Science (ICTAS), the Commonwealth Cyber Initiative (CCI) Southwest Virginia node, and the MITRE Corporation for supporting elements of the team during this research and presentation. The positions, opinions, and viewpoints are those of the authors alone and do not reflect those of the sponsoring groups.

Contributing Authors:

Lauren Anderson, Harrison Bui, Cara Dunnavant, Kiernan George, Piper Hancock, Joe Harrison, Joshua Lyons, Maya Jackson, Clare Mathewes, Lauren Maunder, Paul O'Donnell, Sarah Ramboyong, Allie Schliefer, Brian Timana-Gomez, Brandon Vanek, and Alan J. Michaels

Corresponding Author:

To contact the team, please reach out to Dr. Alan J. Michaels, who serves as the team advisor and Director of the Electronic Systems Lab, at ajm@vt.edu.