Privacy Policies Through Time:
A Dataset for Privacy Policy History Analysis and Exploration

Jack Workman
School of Information
University of California, Berkeley
jackcworkman@berkeley.edu

1. Introduction

2018 was a monumental year for data privacy. In the European Union (EU), the new General Data Privacy Regulation (GDPR) brought sweeping changes to how companies can handle users' data, and, around the world, news outlets ran story after story covering the latest major data breaches. It is therefore not surprising that privacy and, specifically, data privacy is a growing concern inside the United States[1].

Unbeknownst to many, the field of data privacy law (often referred to as "information privacy law") is several centuries old. Concerns over privacy can be traced back to before the installation of the postal service to the invention of the telephone and now, most prominently, to the widespread adoption of the internet. With each new technological advancement, a new suite of privacy problems is introduced (Solove 2016). This paper focuses on a relatively recent addition to the field of data privacy: the privacy policy.

The privacy policy is a company's formal declaration of the personal information it collects, maintains, and uses from its customers. In the digital age, this information is often called "data" and can be gathered from anyone who visits a company's website or uses a company's service. Companies have been posting privacy policies since the late 1990s, and their policies have grown as their businesses expand and as new privacy legislation is passed. With this in mind, this paper presents a novel methodology and dataset for gaining insight into the evolution of privacy policies to better understand the events that promote policy change. This knowledge and data can be utilized in research to devise more comprehensive legislation and organizational privacy management. The final dataset is named Privacy Policies Through Time (PPTT) and contains 294 policies across 21 companies and 20 years.

This paper is organized as follows: Section 2 covers the background and related work of privacy policy analysis, Section 3 reviews the methodology used to build the dataset, Section 4 presents a brief dataset analysis and a novel visualization for exploring policy changes over time, Section 5 discusses possible use cases and significance of the dataset, Section 6 covers future work, and Section 7 is the conclusion.

[1] "SAS Survey: 67 percent of US consumers think government should do ...." https://www.sas.com/en_us/news/press-releases/2018/december/data-management-data-privacy-survey.html. Accessed 17 Apr. 2019.

2. Background & Related Work

Prior efforts to collect a dataset or corpus of privacy policies have largely focused on policy content and structure and not on historical change. Table 1 presents an overview and comparison of available privacy policy datasets.

Dataset	Policy Count	Specialty
OPP-115 Corpus (ACL 2016)	115	Annotated website policies
APP-350 Corpus (PETS 2019)	350	Annotated Android app policies
Opt-out Choice Dataset (EMNLP 2017)	102	Policies with opt-out choice labels
ACL/COLING 2014 Dataset	1010	Website policies spanning 12/2013 - 01/2014
The Privacy Policy Landscape After the GDPR	9100	Policies before/after GDPR
Privacy Policies Through Time (PPTT)	***294***	***History and connected revisions***

Table 1. Overview of available privacy policy datasets

Some privacy policy datasets focused on annotations to promote data analysis and research. The OPP-115 Corpus contains 115 privacy policies with the addition of 23K manual annotations of data practices referenced in the policy text (Wilson et al. 2016). The APP-350 Corpus is a dataset of 350 annotated Android app policies used to study the differences between acts described in the apps' policies and the behaviors of the app themselves (Zimmeck et al. 2019).

Natural language processing has also been applied to study and better understand the relationship between policies. The ACL/COLING dataset of 1010 privacy policies was collected and analyzed with Hidden Markov Models to attempt to align common sections to support empirical study of online privacy policies (Ramanath et al. 2014). Another work focused on the challenge of automatically identifying and extracting sections related to user choice from privacy policies and published a dataset of 102 privacy policies annotated with opt-out language labels named the Opt-out Choice Dataset (Sathyendra et al. 2017). These works are all either manually collected or leveraged crowdsourcing to collect and annotate their datasets.

The fifth dataset in Table 1, published by Linden et al, most resembles the PPTT dataset. In their work, Linden et al. analyzed the differences in privacy policies before and after GDPR. Linden describes a web scraping pipeline developed to automatically gather policies and conducts a number of analyses also used in this paper on the collected corpus to study the impact of the new rules and regulations introduced by GDPR (2018). This paper extends the scope of Linden's analysis from two revisions per company to all privacy policy revisions per company thus opening the door for broader research.

3. Methodology

This section discusses the system developed for identifying and retrieving historical privacy policies as well as the structure of the dataset.

3.1 Policy Retrieval

Policy retrieval is done on a per-company basis with the Internet Archive's Wayback Machine[2], a configuration file, and a python[3] web scraping script. The procedure for creating a configuration file can be found in Algorithm 1. For the algorithm used by the web scraping script, see Algorithm 2.

Note that privacy policies are commonly found in the footer of most websites under a heading similar to "Privacy" as mandated by the California Online Privacy Protection Act of 2003[4]. This methodology assumes that all sites comply with these instructions and does not provide for sites that do not.

1. Search Wayback Machine for company's website home page
2. Navigate to oldest snapshot of company's home page and click on "Privacy" link in footer
3. Save URL and start date in configuration file
4. Advance by Month forward through time until privacy policy URL changes
5. Use Wayback Machine's automatic redirect feature or return to company's website's home page at the date of policy URL change to find the new URL
6. Save end date in configuration file
7. Repeat steps 3-6 with new URL until desired end date is reached

Algorithm 1. Company privacy policy configuration file creation procedure

[2] "Internet Archive: Wayback Machine." https://archive.org/web/. Accessed 15 Apr. 2019.
[3] "Python.org." https://www.python.org/. Accessed 15 Apr. 2019.
[4] "Section 22577(a) - California Legislative Information." https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=8.&chapter=22.&lawCode=BPC. Accessed 15 Apr. 2019.

Without the Wayback Machine, collecting and creating this dataset would not have been possible. The Wayback Machine allows users to view web pages in their original form at any point in time when one of its web crawlers visited the site. For popular sites, this can amount to several times a day. This functionality is used to find all of the past URLs to a company's privacy policy.

The configuration file is a JSON[5] file that contains information on the company and a list of that company's privacy policy URLs with the date or dates that each URL was active. An alternative configuration format of only URLs is available for companies that reference historical privacy policies with direct URLs like Twitter[6] or Google[7].

```
1. For each URL in configuration file,
   a. For every month between URL start date and end date
      i.   Get HTML from Wayback Machine web API
      ii.  Extract policy text and publication date from
           HTML
      iii. If publication date is newer than previous
           publication date, save policy as new revision
2. Write paths to saved policy text files as new rows in company
   index file
3. Run build_master_index.py to combine all company index files
```

Algorithm 1. Policy retrieval algorithm psuedocode

The python web scraping script *wayback_search.py* uses the python requests[8] package to interface with the Wayback Machine's web API to retrieve the HTML of the privacy policy URLs stored within the configuration file. Once retrieved, a series of regular expressions and string formatting functions are used to process and extract the text and publish date of the policy. The script will then step forward month-by-month and query the Wayback Machine until the given URL's end date is reached.

3.2 Dataset Structure

Once a policy's text is retrieved and it is determined that it is a new revision, it is saved in a text file. The text file is stored in a directory of the same name as the company and named with the

[5] "JSON." https://www.json.org/. Accessed 15 Apr. 2019.
[6] "Previous Twitter Privacy Policies." https://twitter.com/en/privacy/previous. Accessed 16 Apr. 2019.
[7] "Updates: Privacy Policy – Privacy & Terms – Google." https://policies.google.com/privacy/archive?hl=en-US. Accessed 16 Apr. 2019.
[8] "Requests: HTTP for Humans™ — Requests ...." http://docs.python-requests.org/en/master/. Accessed 16 Apr. 2019.

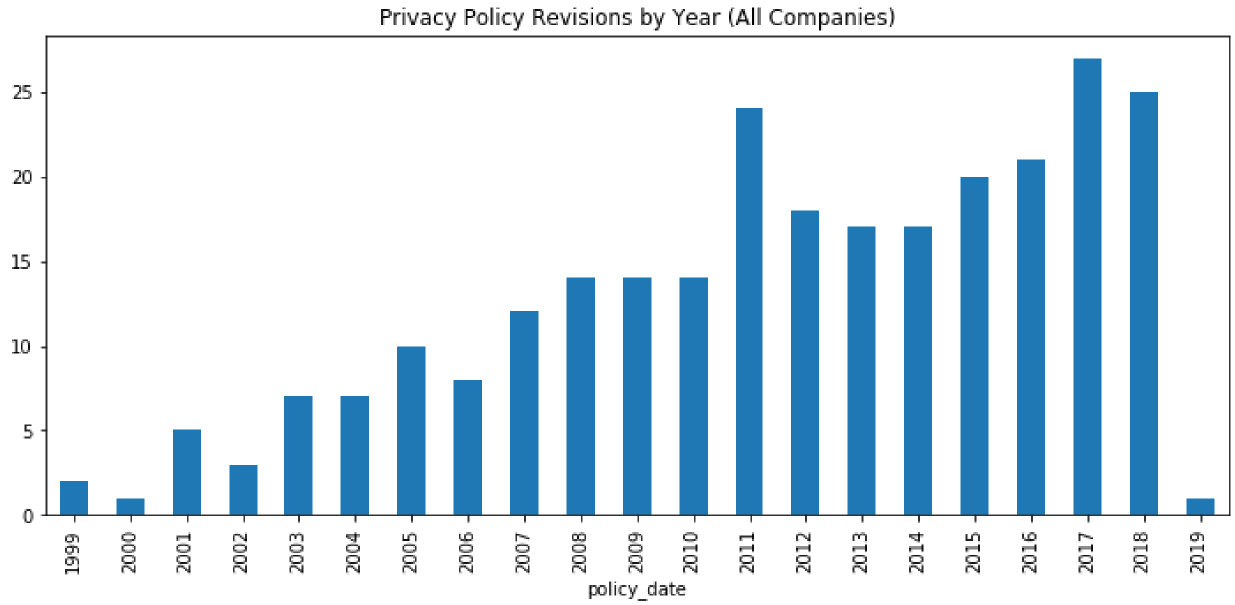
date of the policy's revision prefixed by the company's name (for example, facebook-2006-05-22.txt). The relative paths to these policy text files, their publication dates, and their URLs are saved in company-specific and then aggregate CSV[9] index files. The end result is a *master-privacy-policies-index.csv* file that can be used to programmatically peruse the available privacy policies and filter as desired by company or by date.

4. Analysis

In total, 294 privacy policies were collected from 21 companies over a span of 2 decades. This section demonstrates several avenues of analysis made available by this dataset as well as a novel visualization for viewing the change of privacy policies over time across companies.

4.1 Privacy Policy Revisions by Year

Perhaps the most simple and most direct application of this dataset is to track the number of privacy policy revisions (or in other words, the number of policies published) each year. One hypothesis might be that, as privacy becomes a bigger concern in the daily life of consumers, companies are updating and revising their policies more often. At first glance, this dataset might appear to confirm this hypothesis as demonstrated in Figure 1a. However, this view does not consider that more policy revisions might be the result of a higher number of active companies in recent years. This is affirmed by Figure 1b which has been normalized by count of active companies. Figure 1b shows that the rate of policy revisions is steady over the last two decades.



[9] "Comma-separated values - Wikipedia." https://en.wikipedia.org/wiki/Comma-separated_values. Accessed 16 Apr. 2019.

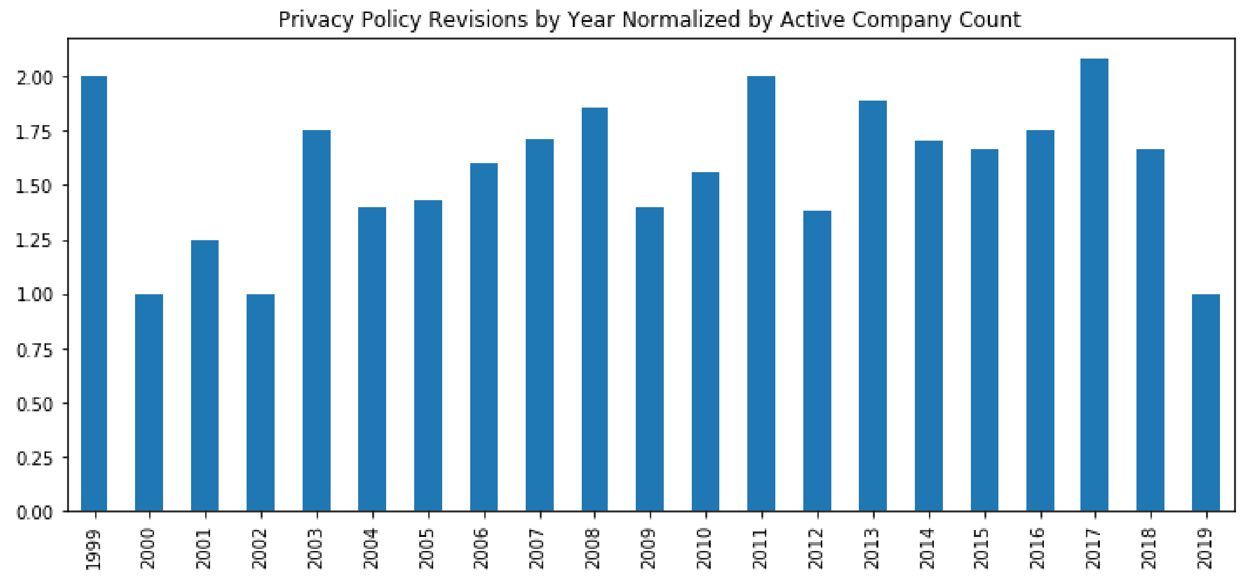


Figure 1a (top). Number of privacy policy revisions by year for all companies.
Figure 1b (bottom.) Number of privacy policy revisions by year normalized by count of active companies

4.2. Visualization of Key Privacy Events

The history of data privacy law has been punctuated by the passing and enforcing of several high-profile pieces of privacy legislation. This section focuses on the following:

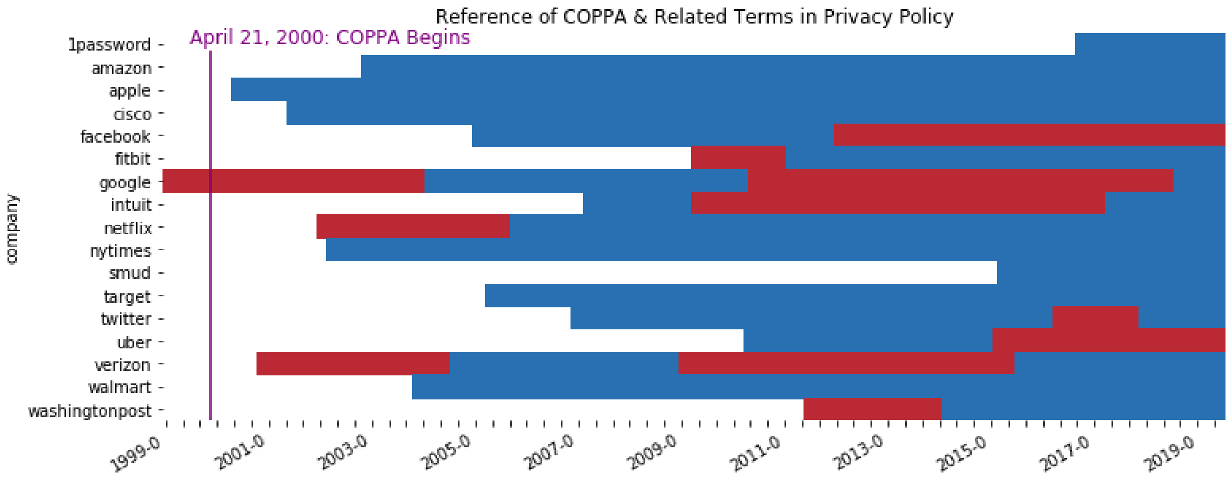
1. **Children's Online Privacy Protection Act (COPPA)** of 1998 effective 21 April 2000[10]
2. **International Safe Harbor Privacy Principles** developed between 1998 and 2000 and overturned 6 October 2015[11]
3. **EU-US Privacy Shield** agreements of 2 February 2016[12]
4. **General Data Protection Regulation (GDPR)** of 2016 effective 25 May 2018[13]

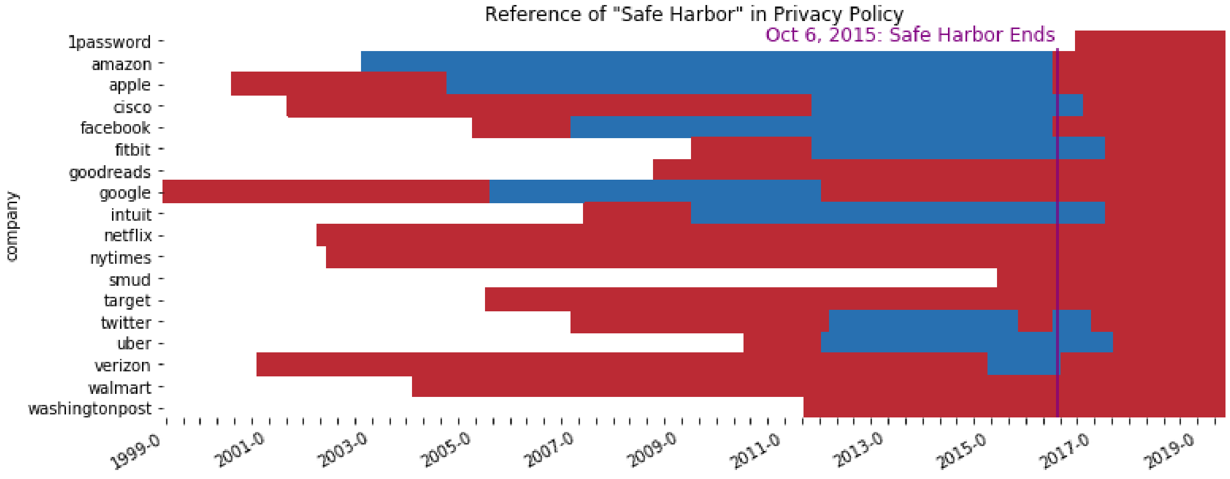
Each of the above were supposed to make a significant impact on how companies handle data and approach privacy. The PPTT dataset provides a means of validating each legislation's impact based on the assumption that companies would respond to the stipulations set forth by each with a revision to their policy. For example, after the EU-US Privacy Shield agreements in

[10] "Children's Online Privacy Protection Act - Wikipedia." https://en.wikipedia.org/wiki/Children%27s_Online_Privacy_Protection_Act. Accessed 17 Apr. 2019.
[11] "International Safe Harbor Privacy Principles - Wikipedia." https://en.wikipedia.org/wiki/International_Safe_Harbor_Privacy_Principles. Accessed 17 Apr. 2019.
[12] "EU–US Privacy Shield - Wikipedia." https://en.wikipedia.org/wiki/EU%E2%80%93US_Privacy_Shield. Accessed 17 Apr. 2019.
[13] "General Data Protection Regulation ...." https://en.wikipedia.org/wiki/General_Data_Protection_Regulation. Accessed 17 Apr. 2019.

early 2016, Amazon updated its privacy policy on 30 September 2016 to include the words, “Amazon.com, Inc. participates in the EU-US Privacy Shield framework.”[14]

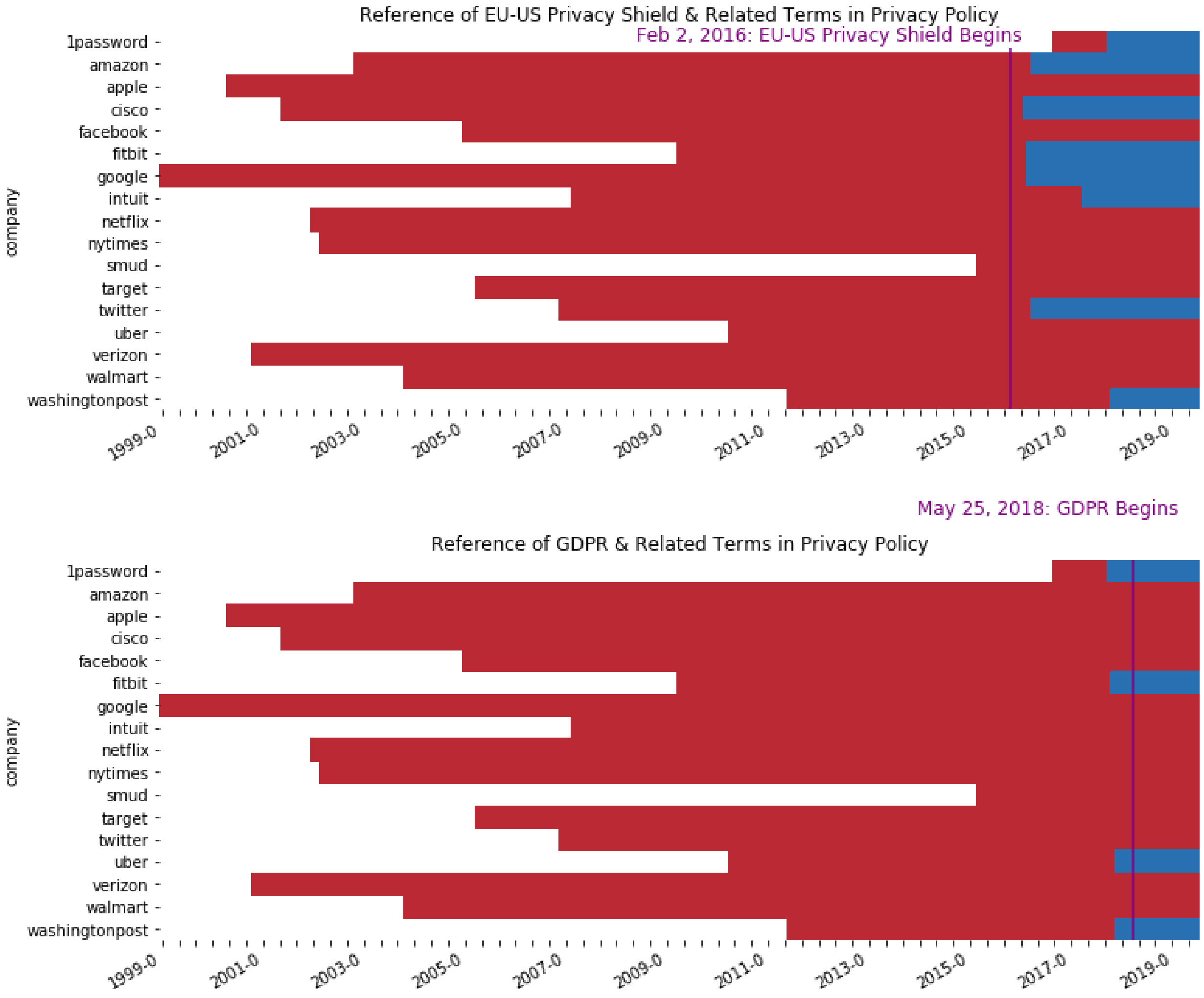
Using a heatmap, the impact of each of these key privacy events can be visualized across companies and time. This visualization is demonstrated in Figures 2a, 2b, 2c, and 2d. The heatmaps were constructed with python’s seaborn[15] package from data provided by the PPTT dataset. The blue bars signal that the active privacy policy for the company at that time contained a reference to the key event while the red signals that it did not. Table 2 lists the terms and permutations used for each key event when generating the heatmaps.


Reference of COPPA & Related Terms in Privacy Policy


Reference of "Safe Harbor" in Privacy Policy

[14] “Amazon Privacy Notice.” http://web.archive.org/web/20161001035323/https://www.amazon.com/gp/help/customer/display.html?ie=UTF8&nodeId=468496. Accessed 17 Apr. 2019.

[15] "Seaborn." https://seaborn.pydata.org/. Accessed 17 Apr. 2019.



From top to bottom:

Figure 2a. References found to COPPA in privacy policies across companies and time; White signals that a company has no privacy policy yet, red means that a reference was not found, blue translates to a reference was found.

Figure 2b. References found for the International Safe Harbor Privacy Principles across privacy policies.

Figure 2c. References found for the EU-US Privacy Shield agreements across privacy policies.

Figure 2d. References found for GDPR across privacy policies.

Of the four key privacy events analyzed, COPPA and Figure 2a yield the least decisive visualization followed by GDPR and Figure 2d. COPPA requires that companies who handle the data of minors disclose their practices in their privacy policies. One explanation for the lack of references observed in the red portions of Figure 2a is that, at that time, those companies did not offer their services to minors and thus did not track nor possess their data. This might explain why some companies like Netflix or Fitbit began as red and then transitioned to blue.

Another explanation might be that companies often do not put as much effort into their privacy policies when they are small, but, as they experience growth much as Netflix and Fitbit did in their early years, these companies revisit their policies and ensure that they follow all principles and guidelines to avoid litigation from their growing user base.

Key Event	**Policy Reference Filters**
Children's Online Privacy Protection Act	['children', 'child', 'minor', 'underage', 'teenager', 'Online Privacy Protection Act', 'coppa']
International Safe Harbor Privacy Principles	['safe\s*harbor']
EU-US Privacy Shield	['shield', 'EU-US\s*Privacy\s*Shield', 'EU-US']
General Data Protection Regulation	['gdpr', 'General Data Protection Regulation']

Table 2. Key privacy events and filters used to detect references in the text of privacy policies (all filters ignore case)

However, these explanations do not explain why Facebook's privacy policy moves from blue to red in 2012. Manually reviewing Facebook's privacy policy[16] (Facebook calls it a "Data Policy") confirms that no mention of COPPA or its requirements are present but reveals that Facebook maintains a second policy strictly for minors[17]. This second policy is linked in Facebook's navigation bar and thus not part of the text extracted from the website while building the dataset. Possible improvements to include these secondary policies are discussed in Section 6.

Comparing Figures 2b and 2c clearly shows when the Safe Harbor rules ended and the Privacy Shield rules began. Most interestingly are the companies such as Apple that published policies with references to Safe Harbor rules but did not update to include references to Privacy Shield. Manually reviewing Apple's current policy[18] confirms that no mention of the Privacy Shield agreements exists. This begs the question: why would Apple disregard the Privacy Shield but adopt Safe Harbor while it was active? This question and the related visualizations demonstrate the potential of the PPTT dataset to expose trends like these as areas of exploration for privacy experts.

[16] "Data Policy - Facebook." https://www.facebook.com/policy.php. Accessed 17 Apr. 2019.
[17] "Minors and safety - Facebook." https://www.facebook.com/about/privacy/minors. Accessed 17 Apr. 2019.
[18] "Legal - Privacy Policy - Apple." 4 Apr. 2019, https://www.apple.com/au/legal/privacy/en-ww/. Accessed 17 Apr. 2019.

The final heatmap, Figure 2d, showcases the shortcomings of the search-by-keyword approach to determining if a company adheres to the rules of regulations like GDPR. It is widely recognized that GDPR had a sweeping impact on how companies approached data privacy and spawned a wave of privacy policy revisions. The heatmap fails to depict this wave as it assumes that regulation adherence and references to "GDPR" are directly connected. Of course, common reasoning tells us that this is not so. A company can uphold GDPR values like readability and transparency without referencing GDPR directly. As an alternative to search-by-keyword, researchers can use readability metrics to judge how well policies adhere to these values as well as discover the influence on these values by regulation like GDPR. This technique was employed by Linden et al. in their paper with direct attention to observing GDPR's influence (2018). In the next section, Linden et al.'s readability metrics are applied to the PPTT dataset to assess readability of datasets over time.

4.3. Readability Over Time

Linden et al. reviewed the changes between privacy policy revisions made before and after GDPR. Part of their analysis included a set of readability metrics. Borrowing from their work, this section applies four of the same readability metrics to the much broader range of privacy policies available in the PPTT dataset. The readability metrics are:

1. **Lexicon Count**: number of words
2. **Syllables Count**: number of syllables
3. **Sentence Count**: number of sentences
4. **Flesch Kincaid Grade**: readability measure presented as U.S. grade level (Kincaid et al. 1981)

To limit noise and increase the value of this analysis, the companies analyzed were restricted to those commonly referred to as FAANG[19] (Facebook, Apple, Amazon, Netflix, Google) as these companies are large players in the world of data. In the PPTT dataset, 87 of the 294 privacy policies belong to these five companies.

4.3.1 Lexicon Count

While not a comprehensive measure of readability, the lexicon count tells the length of a document. Figure 3 shows the evolution of policy length over time as told by lexicon count of the FAANG companies. Examining the plot shows a gradual trend towards higher counts. This could be due to companies putting in extra effort to explain technical concepts, because their use and collection of data is growing and requires additional privacy policy sections to explain, or because recent legislation demands greater attention to the content and quality of a company's privacy policy.

[19] "Facebook, Apple, Amazon, Netflix and Google - Wikipedia." https://en.wikipedia.org/wiki/Facebook,_Apple,_Amazon,_Netflix_and_Google. Accessed 17 Apr. 2019.

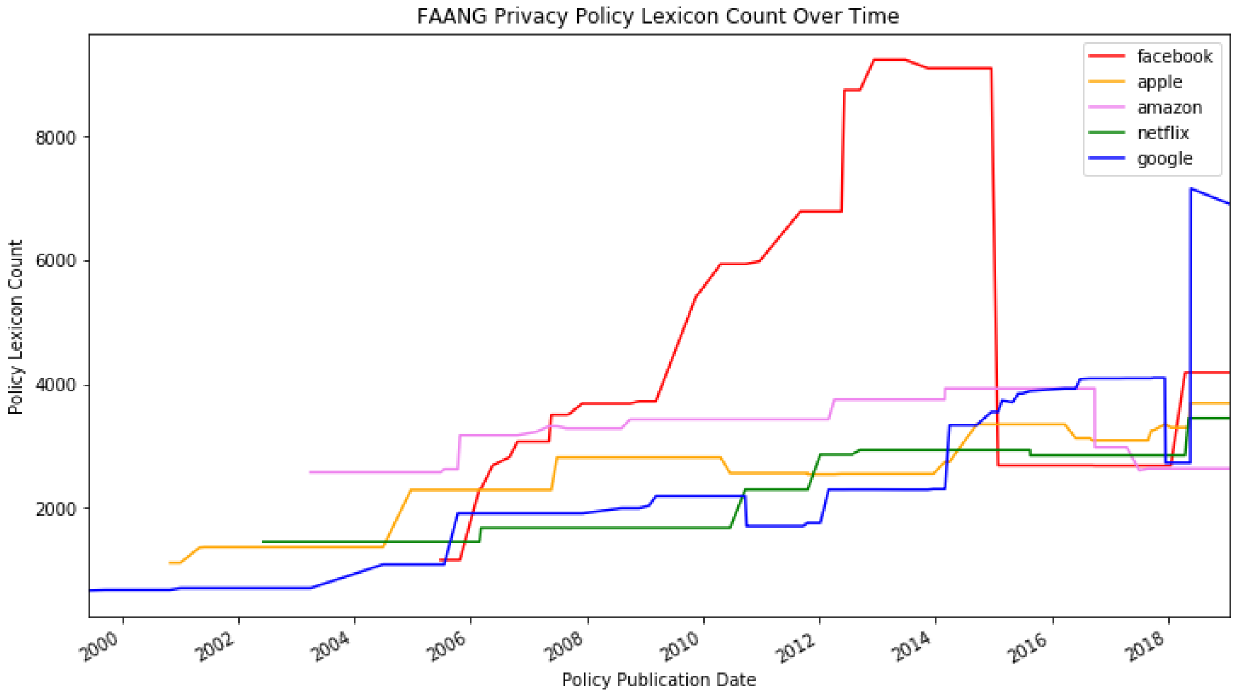


Figure 3. Lexicon Count of FAANG privacy policies over time

Of note in Figure 3 is the sudden drop of Facebook's lexicon count in 2015, and the sudden rise of Google's in 2018. A quick manual inspection of the latter's recent privacy policies reveal that Google made sweeping changes to its policy to prepare for GDPR and added extra sections and detailed descriptions all contributing to the higher word count. Facebook's word count drop is likely due to the company overhauling its approach to sharing data amongst apps and its services in 2015[20].

4.3.2 Syllables Count

Syllables count is another metric for assessing readability of a document. Figure 4 plots the syllables count of the captured privacy policy revisions over time of the FAANG companies. The plot by itself reveals no surprises not already revealed by lexicon count. Counting syllables might be of most use to researchers when comparing the syllable count to the lexicon count and exploring the differences. For example, the final policy revision by Google caused the lexicon count to go down and the syllable count to go up.

[20] "Facebook Is Shutting Down Its API For Giving Your ... - TechCrunch." 28 Apr. 2015, https://techcrunch.com/2015/04/28/facebook-api-shut-down/. Accessed 17 Apr. 2019.

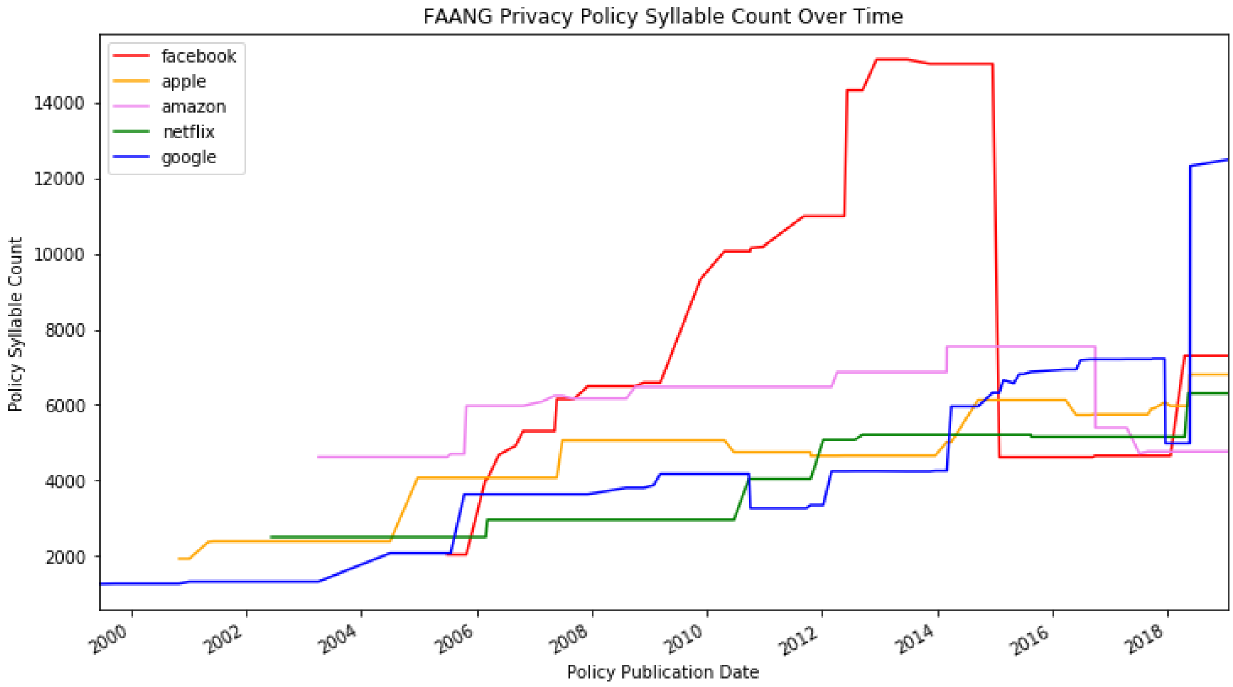


Figure 4. Syllable Count of FAANG privacy policies over time

4.3.3 Sentence Count

Sentence count is a close cousin of syllable count and word count as demonstrated by the plot in Figure 5. Just as with lexicon count and syllable count, sentence count follows the same general positive trend with the exception of Facebook's large drop in 2015. There seems to be more variation in sentence count compared to lexicon count and syllable count than between only lexicon count and syllable count. Also of note is that sentence and lexicon count drop with Google's last privacy policy update but syllable count rises.

4.3.4 Flesch Kincaid Grade

The Flesch Kincaid Grade Level measure is meant to judge a document's reading level and assign it a score analogous to the grades used in the US educational system. Recent legislation like GDPR advocates for privacy policies to be readable and transparent, so one would expect the Flesch Kincaid Grade Level scores to go down as a result. However, as shown in Figure 6, the opposite has occurred. In 2018, the year GDPR went into effect, all of FAANG except Amazon revised their policies and caused their Flesch Kincaid Grade Level score to skyrocket to all-time highs for each company. If this measure truly reflects the readability of grade levels, then Apple's and Google's privacy policies are at at 23rd grade reading level, followed by Facebook at 19 and Netflix at 17 (all college level and above).

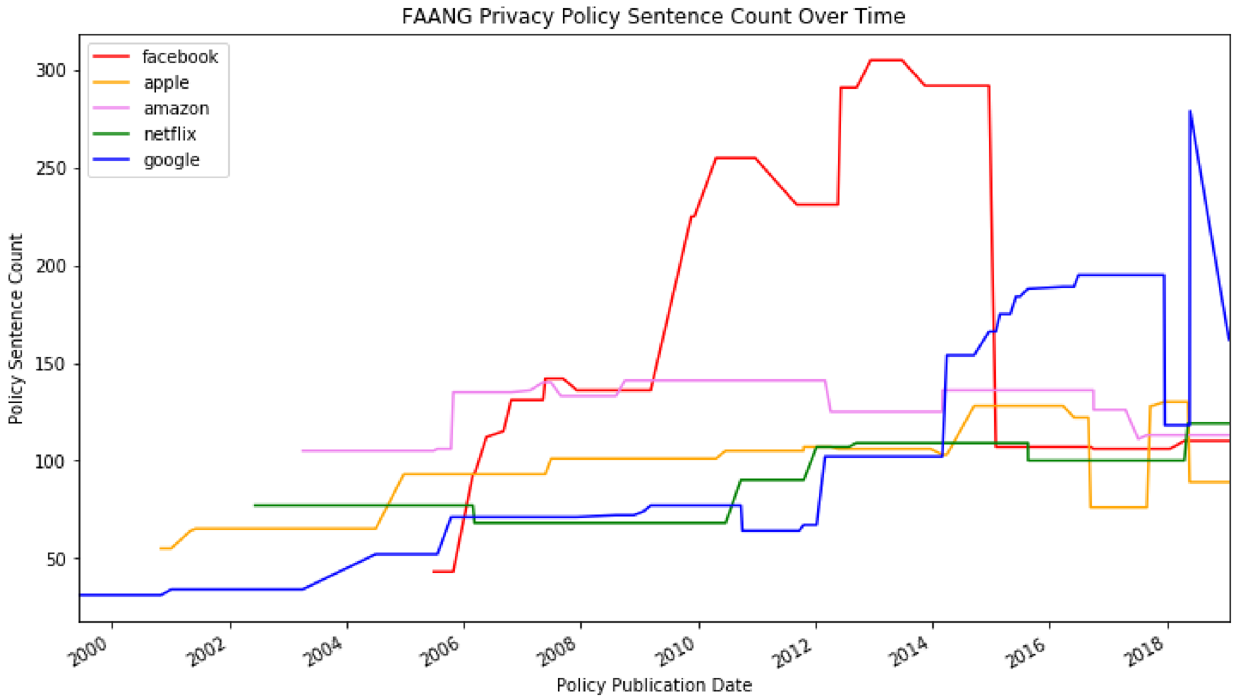


Figure 5. Sentence Count of FAANG privacy policies over time

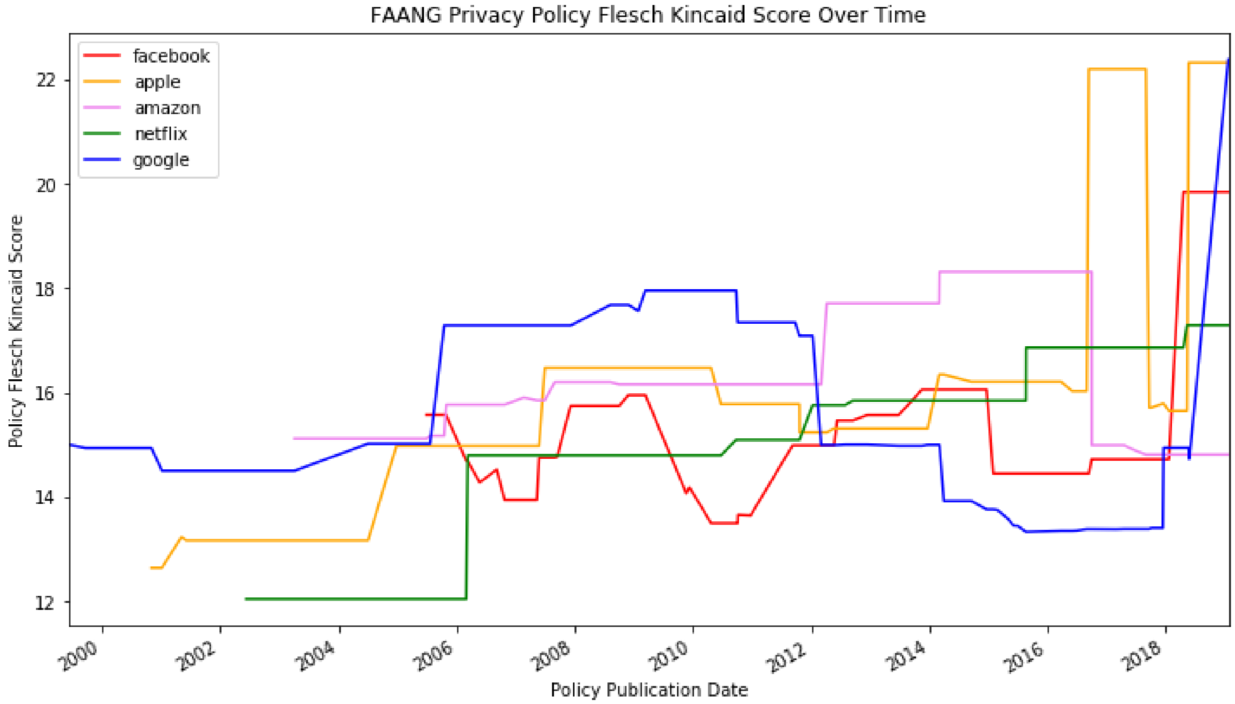


Figure 6. Flesch Kincaid Grade Level Score of FAANG privacy policies over time

5. Discussion

As reviewed in Section 2, the existing literature does not cover the change of privacy policies over time - a potentially valuable resource in understanding the impact of key privacy events at aggregate and per-company levels. The Privacy Policies Through Time (PPTT) dataset seeks to promote the use of this resource in research by making the past revisions of companies' privacy policies easily available.

Through analysis of past and current privacy policies as shown in Section 4, interesting insights such as the decline in influence from the transition of the International Safe Harbor Privacy Principles to the EU-US Privacy Shield and the gradual increase in both policy length and of the Flesch Kincaid Grade Level Score observed across companies can be uncovered. Connecting these trends over time with key events like GDPR can create a lens through which researchers can assess an event's impact or through which researchers can learn about how a company approaches privacy.

Furthermore, by open-sourcing the tools developed to create the PPTT dataset, other researchers in the field can contribute to and continue to grow the number of past privacy policies available. With luck, this dataset will become an archive to be utilized by those interested in researching how companies and privacy policies have changed over time and to explore, reveal, and apply the knowledge they uncover to the field of data privacy.

6. Future Work

There are several limitations with the dataset's methodology that could be addressed by future work. First, the lack of scientific procedure used when selecting companies from which to retrieve privacy policies should be remedied. At present, the dataset's companies represent the biases of the dataset's creator and is not comprehensive or an accurate representation of the privacy policy landscape as a whole.

Second, the process of identifying a company's historical privacy policy URLs for the configuration file could be automated. Instead of saving the URL to the privacy policy, the configuration could instead store the URL to the company's home page, and the *wayback_search.py* script could parse the home page for the privacy policy link to locate the policy. This would save time when creating the configuration files as a company's home page is much less likely to change over time than the link to its privacy policy.

Third, some companies have multiple privacy policies - one for each service or tool that they offer. This practice is less common today than in the 2000s as splitting your policy does not align well with modern regulation's definition of readable and accessible. This paper's approach does not account for these split privacy policies and thus might miss part of the picture of a company's privacy policy evolution. A possible improvement to this would be to detect when a company’s core website privacy policy links to other sub-policies and append the sub-policies’ text to the core policy’s text.

Fourth, the process of extracting policy text from privacy policies could be improved. With the current extraction scheme, some policies contain text from the website’s header, footer, or navigation bar. This is not ideal. Experiments were conducted with a popular article text extraction python package newspaper3k[21], but the package frequently returned text with portions of the beginning or end of the policy missing and some policies were returned entirely empty.

Finally, the number of policies and companies could be vastly expanded. With the script in place and the improvements mentioned above, crowdsourcing via a service such as Amazon’s Mechanical Turk[22] could be leveraged to quickly and efficiently gather the URLs and dates of company’s home pages for many more companies, thus creating a more comprehensive and ultimately useful dataset.

7. Conclusion

This paper has demonstrated a methodology for gathering and analyzing a dataset of companies’ historical privacy policies to promote further research in how privacy policies change over time. The dataset is called Privacy Policies Through Time (PPTT) and is available in an online and open-sourced repository[23] along with the scripts and configuration files used for the dataset’s generation.

As the use and collection of personal data continues to grow, so too will the ways that companies make use of our personal information. It is important that these companies remain transparent and honest about how our data is used and shared so that we as consumers can make informed decisions when registering for new services. As of now, privacy policies are the main conduit for companies to communicate their data-related behaviors to consumers, so the readability and accessibility of these documents is paramount. It is the earnest hope of this researcher that this dataset can be utilized to promote quality privacy policy practices and contribute to the education of others in understanding the change and transformation of privacy policies through time.

[21] "newspaper Documentation - Read the Docs." 24 Jan. 2018, https://readthedocs.org/projects/newspaper/downloads/pdf/latest/. Accessed 16 Apr. 2019.
[22] "Amazon Mechanical Turk." https://www.mturk.com/. Accessed 16 Apr. 2019.
[23] https://github.com/workmanjack/privacy-policies-through-time

References

Kincaid, J. Peter, et al. "Computer readability editing system." IEEE Transactions on Professional Communication 1 (1981): 38-42.

Linden, Thomas, Hamza Harkous, and Kassem Fawaz. "The Privacy Policy Landscape After the GDPR." arXiv preprint arXiv:1809.08396 (2018).

Ramanath, Rohan, et al. "Unsupervised alignment of privacy policies using hidden markov models." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. 2014.

Sathyendra, Kanthashree Mysore, et al. "Identifying the provision of choices in privacy policy text." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

Solove, Daniel J. “A Brief History of Information Privacy Law.” PROSKAUER ON PRIVACY, PLI (2006).

Wilson, Shomir, et al. "The creation and analysis of a website privacy policy corpus." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2016.

Zimmeck, Sebastian, et al. “MAPS: Scaling Privacy Compliance Analysis to a Million Apps”. Privacy Enhancing Technologies Symposium 2019.