Privacy Policies with a Higher Count of 'Positive Data Actions' also have a Higher Count of Clauses Related to Data-regulated Users

William J. Trefiak

26/04/2021

Abstract

The past five years have marked a significant shift in the work of privacy policy researchers. Enabled by the Usable Privacy Project, myriad datasets in the form of annotated privacy policy corpora have been developed and further research is only accellerating. This project briefly tracks the efforts made by The Usable Privacy Project before turning to an experiment of it's own, which aims to measure whether emerging privacy legislation such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have an impact on the quality of privacy policies. Using a multiplenegative binomial regression, this project aims to speak on potential causality between legal language geared towards a specific regulatory audience and the quality of the privacy policy. Findings are somehwhat inconclusive given the nature of the data as well as large discrepancies in the count of variables observed as well as the data's overall prominence. All data and relevant documents for this project can be found via. github at: https://github.com/will-trefiak/usable_privacy

1 Introduction

Over the past several years, research in the communities of digital privacy and data protection have taken unprecedented strides. While this is strongly reflected in recent actions taken by governments to revamp their digital privacy legislation, the vastness of platforms with their own privacy policies provides a wellspring of data for researchers to dive into. The most significant research program in this field is undoubtedly the Usable Privacy Policy Project (Sadeh et al. 2013), which marks an international and multidisciplinary effort to collect the data found within privacy policies as corpora, and then provide these corpora to researchers for further data analysis. Work on this project began in 2013, and to date seven data sets and 62 publications have been produced that cover relevant areas such as scaling the collection of privacy policy corpora (Zimmeck et al. 2019), the reliability of machine learning models in classifying specific "data actions" within corpora (Kumar et al. 2019), and empirically measuring the privacy risks of a representative sample of privacy policies (Bhatia and Breaux 2018).

The following project and experimental design draws heavily on the methods and techniques developed by the Usable Privacy Project in addition to utilizing the Usable Privacy datasets for further statistical analysis. Namely, the data used in this experiment is the OPP-115 corpus, which is collection of 115 annotated privacy policies that were pre-selected through sector-based sub-sampling by the principal researchers of this project (Wilson et al. 2016). Given the relative complexity of this dataset from gathering through to analysis and modeling, this project also provides a brief survey of literature relevant to the Usable Privacy Project as to better acquaint the reader with methods considered state-of-the-art in digital privacy research.

While our internet, and the platforms that collect data from it, operate on a global scale, the means by which the internet is regulated still mostly falls into the hands of national actors working within specified jurisdictions. While researchers in the field of political science and international relations have an extremely firm understanding of how this can lead to policy inconsistency in other spaces that require collective action, relatively little is written on this relationship as it speaks to privacy and data protection. Although this project is not necessarily concerned with addressing issues of collective action, it is certainly concerned with investigating the relationship between regulators who have taken action and regulators who have not. More

specifically, this project is guided by the following research question: *Do privacy policies with a higher count of 'positive' data actions have a correspondingly high count of clauses written for international and specific audiences?*

This research question uses somewhat indirect variables to accomplish it's objective, but can be generally thought of as a way to measure how effective digital privacy legislation is at creating higher quality privacy policies, which are considered the *de jure* standard for notifying Internet users of applicable privacy practices" (Zimmeck et al. 2019). Although there are some limitations to this approach, the experiment will be conducted utilizing metrics from (Wilson et al. 2016) et al., that provide a count of particular "data actions" (Wilson et al. 2016) that appear in a typical privacy policy. While some of these data actions can most certainly be considered consumer adverse (collection actions), there are an equal number of data actions that can be considered consumer friendly (access, edit and deletion) or relatively benign (policy changes). In addition, a final data action identified by the principal researchers captures language addressed to a particular audience covered by regulations - most often for audiences in California or the EU. Overall, this experiment investigates whether a high number of instances where an international (read: regulated) audience is mentioned is related to an increased amount of data practices that are consumer friendly in a given privacy policy. Before explaining this modeling in detail, however, it is first important to provide some context on the Usable Privacy Project as well as the data used in this experiment.

2 Data

2.1 The Usable Privacy Policy Project

In many respects, this particular experiment draws from the body of research associated with The Usable Privacy project initially launched in 2013. While the OPP-115 corpus (Wilson et al. 2016) is the original dataset to come out of this project, six further datasets have been developed that both have enriched the original annotation scheme as well as scaled the capabilities of the OPP-115 corpus into the hundreds of thousands of privacy policies. The most recent significant contribution to the Usable Privacy Project comes from (Zimmeck et al. 2019), which scaled the initial annotation scheme of the OPP-115 to 441,626 privacy policies using web crawling and automated annotation to classify data actions with unprecedented accuracy - a process the authors title the MAPS method. Datasets geared specifically towards investigating opt-out choices in privacy policies as well as GDPR-specific annotation schemes are also products of The Usable Privacy Project (**Kumar?**).

It is also worth noting that the vast majority of data work done within The Usable Privacy Project was programmed with Python. Given that we have decided to conduct this experiment with R (**citeR?**), there were some inherent limitations in terms of how we could realistically engage with the data given its formatting in some sections. In addition, because of the iterative nature of The Usable Privacy project, much subsequent research including the MAPS method and larger datasets require the use of web crawlers and tools either built for python or not made publicly available. In a practical sense, this means state-of-the-art developments in The Usable Privacy Project could not be leveraged for our own research and therefore utilizing the least modified data (the OPP-115 Corpus) from this project seemed a prudent decision. To our knowledge, this is among the first experiments conducted in R using data from The Usable Privacy Project and therefore a secondary contribution of this project is the development a tidyverse-formatted dataset containing policy-by-policy counts of distinct OPP-115 corpus data actions. This dataset is made publicly available via Github here:

2.2 The R-Adapted OPP-115 Corpus - Data Sheets Approach

This experiment uses a version of the OPP-115 corpus adapted for R. As a means of making this adaptation process as transparent as possible, a breakdown of the transformed OPP-115 corpus is provided in accordance with the "Data Sheets" method developed by Kate Crawford et al. (Gebru et al. 2018). This part-method-part-tool provides people along the entire data workflow with a systematic approach for annotating datasets with a sheet listing and briefly explaining the specific conditions of a dataset and what these conditions imply. In essence, data sheets are a data-appendage gaining relatively fast adoption because they provide another

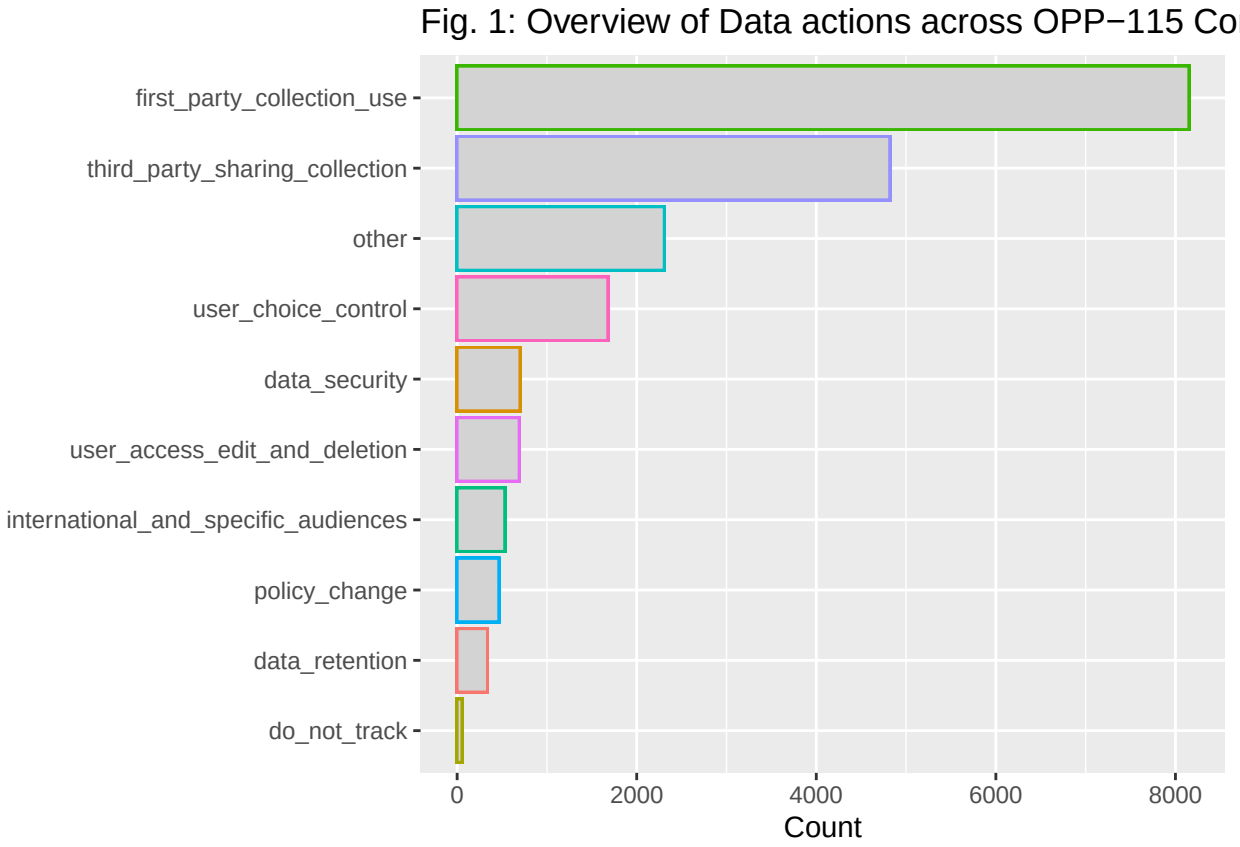
dimension upon which researchers and data scientists can inform their judgments. The following subsections are a breakdown that employs the data sheets method.

2.2.1 Motivation

The R-adapted OPP-11R corpus (herafter referred to as `opp115`) was created with the purpose of developing potential functionality of the original OPP-115 for R. Given the adapted nature of this dataset, factors such as the privacy policies included and data actions recorded could not be significantly modified without compromising the entirety of the data. Regardless, `opp115` was created for the specific task of investigating the relationship between counts of performed data actions.

2.2.2 Composition

Each instance in `opp115` is a count of a specified data action organized by the 115 URLs that contain a given company's privacy policy. These counts are integers that span between 0-220 and are noted for 10 unique data actions; in total, there are 1150 of these data action counts in `opp115`. Because of this, composition-wise `opp115` can be thought of as a basic frequency table whereby variables are readily sectioned into a number of counts according to company. These data actions are the result of annotations by paid privacy experts and subsequently condensed to reduce redundancy across annotations Fig. 1 provides a bar chart showing the count of all data actions across the `opp115`.



2.2.3 Collection Process

The `opp115` dataset created for this project uses data collected from (Wilson et al. 2016) et als. OPP-115 Corpus. As a result, much of the original collection process, and by extension sampling strategy, was dictated by the selection criteria established from the original OPP-115. Based on the manual.txt found in the original OPP-115 repository in addition to the principal research of (Wilson et al. 2016), the collection process follows

a sampling strategy that leverages Google analytics to target and sample the five most frequently visited websites in the U.S. across 15 sectors. From here, web crawlers and scraping tools were utilized to pull the samples from the web for annotation and analysis.

2.2.4 Preprocessing/Cleaning/Labeling

To leverage the data from the Usable Privacy Project in our own work, substantial transformations were performed on the data using the `tidyverse` package and associated methods in R. Before any work on the OPP-115 corpus could be even started, however, the experimenter was required find a memory-friendly solution for reading 115 separate privacy policy `.csv` files into their environment. Once a function was developed to make this task easier for both now and the future, data transformation was undertaken. This process includes additional support from packages such as `gtools` as well as `reshape2` for the proper merging and formatting of variables. All of the data found in `opp115` was created by extracting the information of the only two R-readable columns in the combined dataset, and counts of each particular data action were derived after rearranging the data into a format that makes every data action into a distinct variable for analysis.

2.2.5 Distribution

The original usable privacy corpus is publicly available online at usableprivacy.org/data for the intended purpose of enriching the value privacy experts can bring to their domains. In a similar spirit, the `opp115` dataset created here is also publicly available via Github repository with the hopes that other researchers and privacy enthusiasts find it as interesting as we do.

2.2.6 Maintenance

To our knowledge, the OPP-115 dataset is not maintained actively as no prominence metrics are provided, but given the scale of the Usable Privacy Project which hosts their data via domain, some infrequent/indirect maintenance likely occurs. In addition, the number of unique contributions found in the Usable Privacy Project create data maintenance in an iterative sense. When a particular Usable Privacy dataset (such as the OPP-115) is adapted to fulfill the objectives of a given research program, it is in some way being maintained or updated in accordance with the judgments of different research teams. Overall, the data in this community may not be actively maintained, but research begets a form of maintenance in itself.

3 Ethical Considerations

Like many other data science projects, the analysis conducted here is the product of a series of judgments and deliberations made by a single researcher who built this dataset and modeled the proceeding experiments. As a result, there are a number of ethical considerations in this project worth mentioning that may either introduce ethical biases or force a specific decision to be made about inclusion and omission of information. This semi-reflective practice of auditing the ethics of one's own work is heavily inspired by a cluster of data/communication research from scholars and industry practitioners such as danah boyd, Kate Crawford, and Solon Barocas, whose work emphasizes the importance of weaving ethics into the practice of data science rather than handling ethical considerations as top-down prescriptive measures. Although semi-ironically speaking in terms of "rules" in this article, (Zook et al. 2017) et al. present 10 action-based principles that advocate for a responsible data work *habitus*, namely:

**1. Acknowledge that data are people and can do harm* **2. Recognize that privacy is more than a binary value* **3. Guard against the reidentification of your data* **4. Practice ethical data sharing* **5. Consider the strengths and limitations of your data; big does not automatically mean better* **6. Debate the tough, ethical choices* **7. Develop a code of conduct for your organization, research community, or industry* **8. Design your data and systems for auditability* **9. Engage with the broader consequences of data and analysis practices* **10. Know when to break these rules* (Zook et al. 2017)

While many of these principles address a scope of data work larger than found in this particular project, we felt it prudent to include the entire list as to indicate some of the considerations data professionals can

think about writ large across enterprises. For our purposes, however, Principles *1.*, *2.*, *4.*, *5.*, *6.*, and *9.* are the key tenets that bear relevance worth discussing in terms of the ethical judgments made during this project. More specifically, we can discuss these principles in terms of three key considerations; the regional bias of the OPP-115 dataset, data transformation deliberations, and variable selection deliberations. These considerations are discussed in turn:

3.1 Regional bias in OPP-115

In terms of introducing possible substantial bias into this analysis, it must be noted the original OPP-115 corpus, and as a result the R-adapted `opp115`, includes only english-language privacy policies of companies based in the United States. As a result, the data used in this project cannot provide us with inferences about non-english speaking audiences and only extends to the quality of privacy policies written by American companies. While an investigation into privacy policy accessibility for non-english speakers is a timely and crucial metric for measuring the overall quality of a privacy policy, given the lack of non-english data available in general from the Usable Privacy Project these are metrics that cannot necessarily be evaluated in the present experiment.

Although analyzing non-english data is an opportunity currently unavailable, some measures within the data transformation process did help to reduce regional bias in `opp-115`. Frankly speaking, determining a way to measure the impact of privacy legislation on privacy policies within a relatively homogeneous legal jurisdiction was a central problem this researcher grappled with frequently. As mentioned, a slight workaround to this issue was developed in the data transformation process, whereby the variables of a specified column within the python-formatted OPP-115 dataset contained the 10 data actions found throughout each privacy policy. The entries from this column were uniquified and subsequently cast as column names in their own right, with the uniquified company URLs comprising the first column of data proceeded by a count of how many data actions are noted in each privacy policy, denoted by 115 rows. Luckily, one of the variables extrapolated from this python-formatted OPP-115 list is `international_and_specific_audiences`, which measures the count of clauses related specifically to audiences where a certain body of legislation applies such as the GDPR and CCPA. At the top of this paper, I spoke of 'somewhat indirect variables' being used to measure the quality of a privacy policy, and this specific variable was one of the four being alluded to. While it does capture the action of particular company *addressing* these privacy-regulated audiences, it does so in a way that is admittedly imperfect given any inferences about a particular relationship may not hold for privacy policy data for non-American organizations.

3.2 Data transformation deliberations

The next key area of ethical consideration is in the deliberations made when transforming the data into an r-formatted dataset. While obviously necessary to even begin this project, wrangling remains a step along the data workflow notorious for allowing bias to creep in. The "implicit ethical deliberations" (Zook et al. 2017) made during this stage are critical to be mindful of specifically because, functionally speaking, data wrangling is the process of making data machine readable, creating new variables for analysis, as well as the omission of data deemed 'irrelevant.' Of course, all three of these steps are best left up to the judgment of researchers, but only on the condition that researchers have a strong awareness of the impact these judgments may have on their subsequent work as well as the people hidden underneath the columns of most datasets.

In terms of the work carried out in this project, ethical deliberations during data cleaning were consciously noted by the researcher at specific intervals where an inclusion, omission, or change was made to the data. Specifically, given there were no named columns for any variables in the 115 orignal privacy policy `.csv`'s, deliberations were made on what specific variables represented in accordance with information provided in the documentation `manual.txt`. In addition, a number of tables providing the counts of each data action were already interspersed throughout the original `.csv`s, but all use complex coding and labeling schemes that could not be found in the documentation's `manual.txt`. As a result, when creating the r-adapted `opp115`, this researcher was forced to make deliberations that omitted much of the original data found in the original OPP-115 `.csv`s. This includes columns with inconsistent labels, unclear counting schemes, as well as a number arrays that contain descriptions of each data action - sadly nested in python-formatted tuples. These

variables were removed for the simple fact that little information of value could be determined from them given the cross-language limitations of this dataset as well as technical constraints faced by the researcher in "unnesting" particular columns with potentially insightful data. Overall, however, these deliberations were made consciously and with a degree of understanding about the nature of each variable being used in the subsequent analysis and modeling.

3.3 Variable selection deliberations

This consideration speaks to variable selection not in a transformation sense, but in the sense of why the particular variables of interest were derived and selected for this experiment. Already, I have noted this project heavily weighs how regulation 'magnitude' could be measured. Again, this researcher opted to use an indirect variable `international_and_specific_audiences` to capture this value given the logic found in the *regional bias* section above. Intuitively, identifying `international_and_specific_audiences` as the explanatory variable was relatively straightforward compared to identifying variables geared towards measuring the quality of a privacy policy.

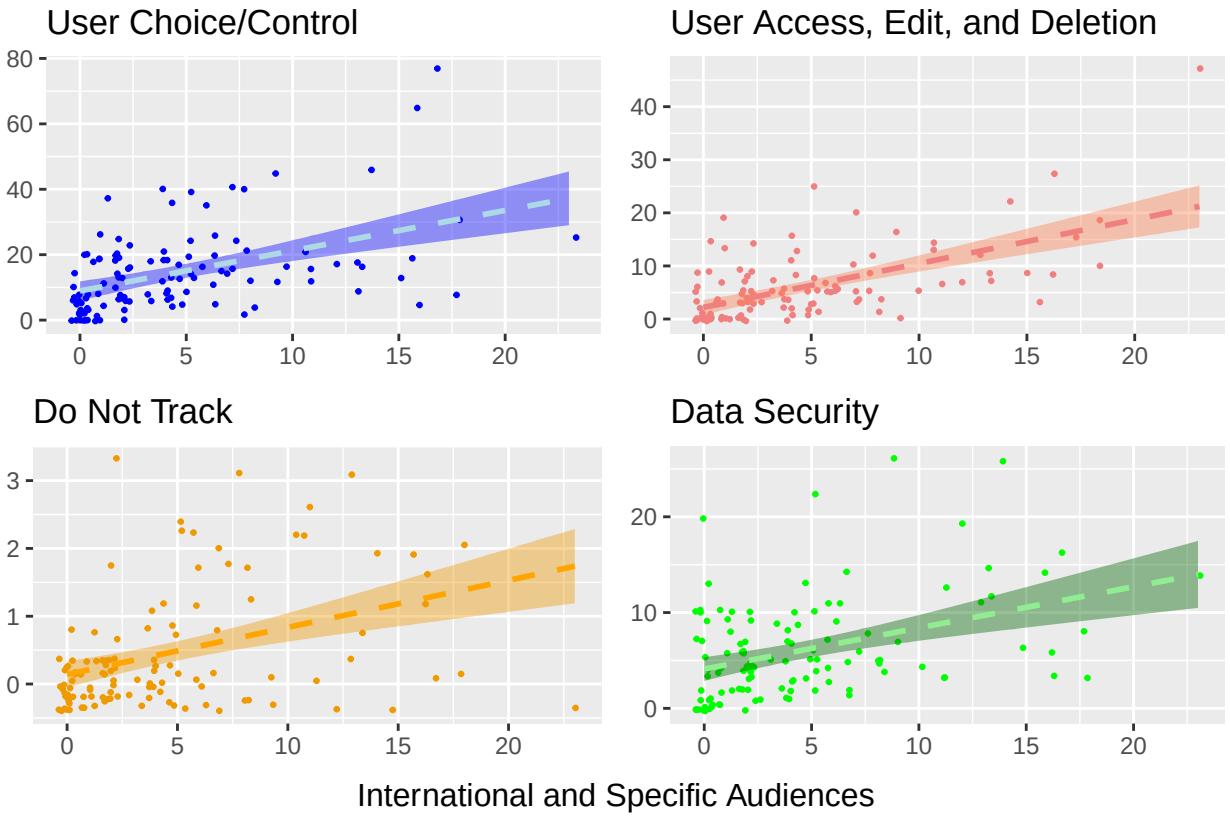
Given that the large majority of data actions found across the `opp115` corpus are **collection/use** related, four other variables specifically stood out to the researcher as examples of data actions that are indicative of a higher quality privacy policy. These were derived from a criteria of three questions; *1. is this data action associated with preserving people's rights? 2. does the data action provide people with agency over their information?* and *3. is this data action associated with information security?*. Based on this criteria, `data_security`, `do_not_track`, `user_access_edit_and_deletion`, and `user_choice_control` were all selected to establish a means of better measuring the overall 'quality' of a privacy policy. Again, however, the author would like to note the indirect, and therefore imperfect, nature of using these variables and the subjective deliberations that resulted in their selection.

The 'quality' of a privacy policy is somewhat subjective, and while the expertly-annotated OPP-115 provides the data actions used in for this experiment, it is entirely possible to assert there are more optimal methods for evaluating privacy policies such as document language complexity or reading level. In fact, The Usable Privacy Project itself has a number of such interpretations on what constitutes a 'high quality' privacy policy. I make this point not to diminish the merit of the work being done here, but rather to forward this researcher's own interpretation of privacy policy 'quality' (that being, a modeled score of the four variables identified) as another interpretation available for research and experimentation.

4 Experimental design and Model

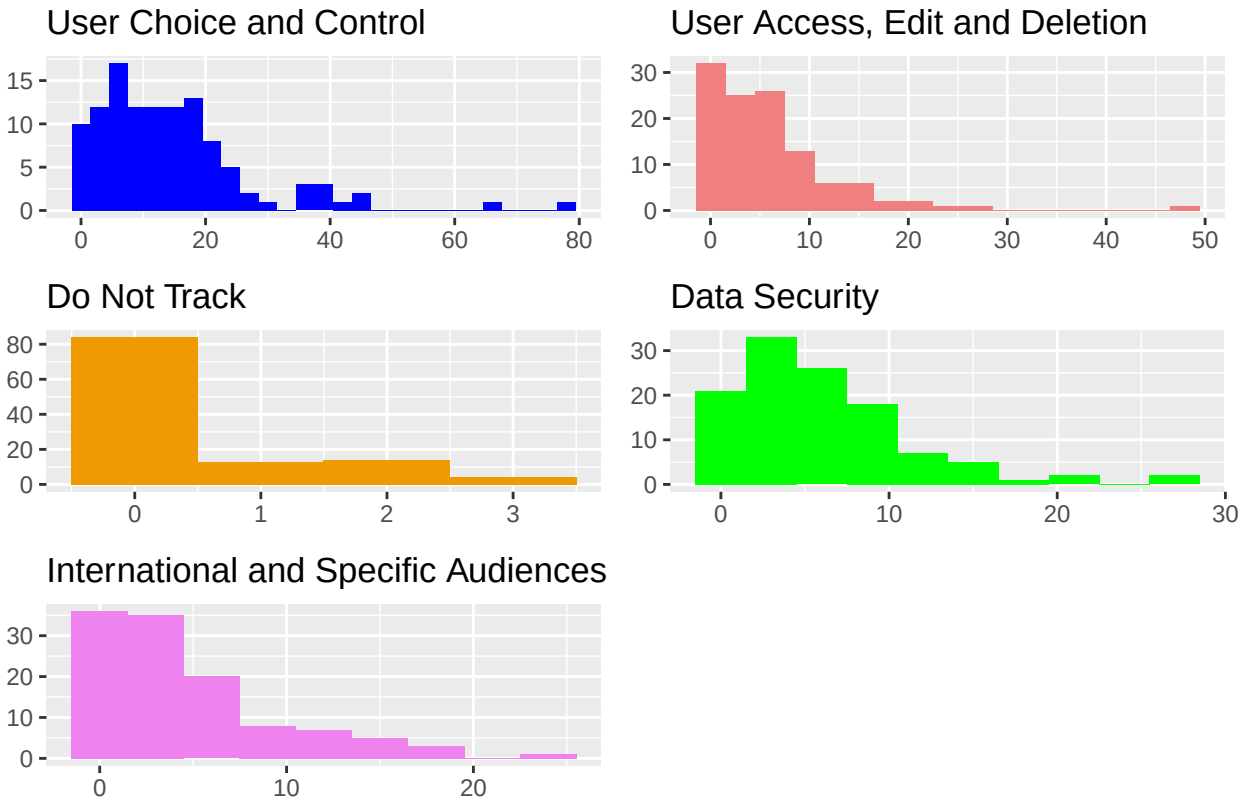
While it has been touched on in most sections prior, the experiment carried out in this project aims to answer the question of whether a higher count of `international_and_specific_audiences` in a privacy policy can be predicted by metrics of a 'higher quality' privacy policy, measured by counts of `data_security`, `do_not_track`, `user_access_edit_and_deletion`, and `user_choice_control`. Given that all the variables used in this experiment are count data, there are a number of statistical modeling approaches to consider based on information from the data itself. While certainly imperfect for robust analysis, Fig. 2.1 includes four exploratory plots; each one plotting the individual variables of privacy policy 'quality' against `international_and_specific_audiences` and fit with a basic `glm()` call. Coefficient values are omitted as these relationships will be measured statistically with an inversion of variables. Each point on the jitter plots in **Fig 2.1** represents one of 115 privacy policies that are measured as counts of the stated data actions.

Fig. 2.1: Exploratory plots of variables identified:



At a glance, it appears a possible relationship can be noted from plots above, but this cannot be determined conclusively without employing more rigorous statistical modeling. Keep in mind the difference of scale for each data action, with some having far fewer noted instances than others overall. This is especially the case with the `do_not_track` variable, which is largely absent from most privacy policies in the OPP-115 and only appears a maximum of thrice in any privacy policy overall. Additionally, *Fig. 2.2* displays a series of histograms, which are dedicated to highlighting the distributions of counts this researcher is concerned with.

Fig 2.2: Histograms of all variables relevant to this experiment



There are a couple of points I would like to discuss about the shape of this data. Primarily, a positive skew can be noted; this has implications for the experimental data, as specific forms of count-specific regressions perform more optimally based on particular statistical metrics. In general however, the vast majority of count data is analyzed *via.* some iteration of a Poisson regression, which requires an inversion of measurement than the one found in the exploratory plots above. In a general sense, Poisson regressions are primarily concerned with predicting the probability of variable μ_i (here derived from `international_and_specific_audinces`) given the counts of k predictors (x_k) (here `do_not_track`, `data_security`, `user_choice_control`, and `user_access_edit_and_deletion`). *Table 1* provides a list of summary statistics for each variable in the dataset. Across the dataset, we can confirm earlier suspicions regarding the positive-skewness of the data given the reported means of each variable are larger than their respective medians. In terms of power, it is also worth nothing the p-values of all variables of interest compared to `international_and_specific_audiences` are significantly lower than the 0.05 threshold, meaning the null hypothesis H_0 : There is no relationship between coefficient μ_i and variables $\beta_k x_k$. The p-values from a chi-squared test are listed in *Table 2*.

	mean	var	sd	sum	median
data_retention	2.9304348	19.9775744	4.4696280	337	1
data_security	6.0782609	28.4411899	5.3330282	699	5
do_not_track	0.4608696	0.7067887	0.8407073	53	0
first_party_collection_use	70.8608696	2095.6471396	45.7782387	8149	66
international_and_specific_audiences	4.6000000	24.6807018	4.9679676	529	3
other	20.0521739	237.6639207	15.4163524	2306	17
policy_change	4.0260870	11.4817696	3.3884760	463	3
third_party_sharing_collection	41.8869565	878.5923722	29.6410589	4817	36
user_access_edit_and_deletion	6.0000000	47.7192982	6.9079156	690	5
user_choice_control	14.5826087	169.1575896	13.0060597	1677	13

user_choice_control	user_access_edit_and_deletion	do_not_track	data_security
7.451109e-09	1.392948e-17	0.01867845	7.387578e-20

Furthermore, the reported variances being significantly larger than the reported means implies the data is over-dispersed. Because of this, a form of Poisson regression known as negative binomial regression makes up the foundation of this experiment, as it integrates an extra parameter into the model to account for this over-dispersion. While the standard Poisson regression assumes that reported means must be equal to their variances, the negative binomial variation loosens this assumption through an additional variable that adds gamma noise to the regression, a process which prepares the model for expecting high dispersion. Functionally, this is controlled by using a negative binomial regression from the `MASS` R package. Modeling was drawn upon from (Nabel 1997) The probability distribution of the negative binomial regression model can be expressed as:

$$\frac{\Gamma(y_i, \alpha^{-}1)}{\Gamma(\alpha^{-}1)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}^{\alpha-1}\right)\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

Where Γ is a coefficient included signify the relationship between y_i, an instance of our dependent variable, and α, which is equal to 1 divided by the sample variance v. Additionally, the negative binomial regression model:

$$\mu_i = exp(ln(y_i)) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Here, μ_i is the mean rate of incidence for a particular unit of exposure, which is denoted by $exp(ln(y_i))$, or more specifically our `international_and_specific_audiences` variable converted into a log-linear form. Further, $\beta_1 x_1$, $\beta_2 x_2$, $\beta_3 x_3$, and $\beta_4 x_4$ are the predictor variables selected for this experiment - `do_not_track`, `data_security`, `user_choice_control`, and `user_access_edit_and_deletion`. Given the relatively small n of this sample, a standard Poisson regression was also modeled and then scored with likelihood ratios to fully determine which model provides a more optimal understanding of the given variables. Regression outcomes are plotted with assistance from the `stargazer`.

5 Results

For referential purposes, please take a look at *Table 3* and *Table 4*. In addition, *Fig 3.1* can be found slightly further into this section. These three inclusions highlight coefficients of interest, confidence intervals, as well as visually plot the regression outputs from the negative binomial regression.

Before discussing on the possible inferences of this model, It is critical to note the non-parametric testing that happened 'behind the scenes' when running the above regressions. Focusing specifically on the Log Likelihood metric, which serve as a non-parametric estimator of model 'fitness' by running a standard chi-squared test measuring the difference between the Poisson and negative binomial regression. In essence, when comparing both log likelihood scores, we can see the negative binomial regression has a higher log likelihood meaning it is likely a 'tighter' model fit and was probably the best choice available to the researcher given the data. As a result, this section will be primarily focused on discussing implications of the values found in this negative binomial model as well as the possible limitations wrought by this decision. Given the metrics of *Table 3.*, we can also assert with 95% confidence that 95% of the confidence intervals would include the true coefficients:

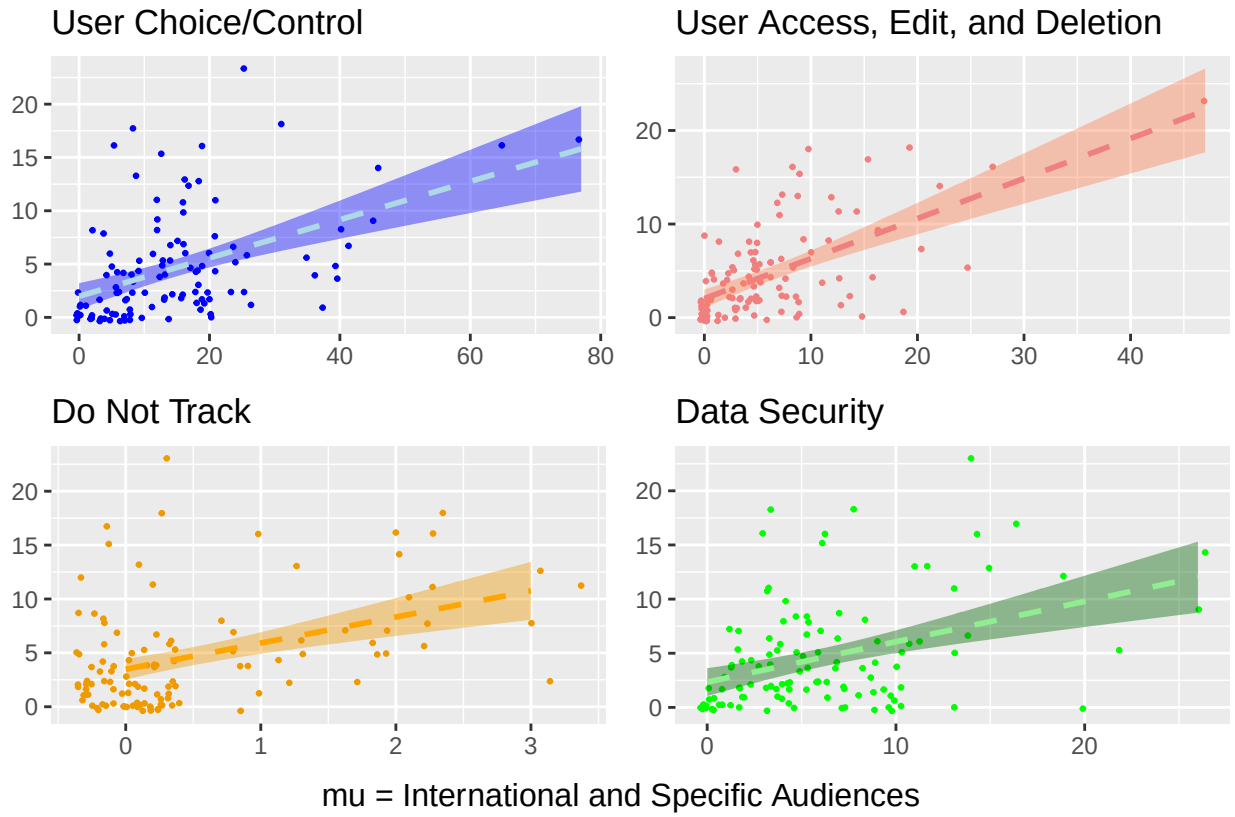
Table 3:

	Dependent variable:			
	international_and_specific_audiences			
	negative binomial		*Poisson*	
	(1)	(2)	(3)	(4)
data_security	0.027	0.057	0.012	0.057
	(0.021)	(0.056)	(0.009)	(0.037)
do_not_track	0.461***	0.239	0.362***	0.209*
	(0.122)	(0.179)	(0.048)	(0.117)
user_access_edit_and_deletion	0.042***	0.046	0.040***	0.050**
	(0.016)	(0.038)	(0.005)	(0.025)
user_choice_control	0.019**	−0.005	0.014***	−0.008
	(0.009)	(0.017)	(0.004)	(0.011)
Constant	0.392**	0.750**	0.667***	0.818***
	(0.174)	(0.339)	(0.090)	(0.223)
Observations	93	22	93	22
Log Likelihood	−222.610	−54.257	−267.389	−57.531
θ	1.581*** (0.387)	3.615* (2.073)		
Akaike Inf. Crit.	455.221	118.513	544.778	125.062
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4:

	Estimate	2.5 %	97.5 %
(Intercept)	0.392	0.015	0.763
data_security	0.027	-0.014	0.069
do_not_track	0.461	0.219	0.723
user_access_edit_and_deletion	0.042	0.013	0.077
user_choice_control	0.019	0.002	0.039

Fig. 3.1: Plotted Negative Binomial Regressions:



5.1 Discussion

Primarily focusing on column (1) of *Table 3.* as this column represents non-binomial regression run on a larger sample, we can see there is a relationship of statistical significance between some of our independent variables and our variable of interest `international_and_specific_audiences`. More specifically however, the results from our modeling indicate that for every one unit increase in the predictor variables ($\beta_k x_k$), we should expect to see a corresponding change in the difference in the logs of expected counts of our response variable. In the model, this log difference is denoted by $exp(ln(y_i))$, and the coefficient is denoted by μ_i; this means our μ_i represents the magnitude of unit change for `international_and_specific_audiences` given the corresponding unit increases of our ($\beta_k x_k$) variables. However, speaking in terms of the variables themselves, and assuming all variables are kept constant:

- For every unit increase in `do_not_track`, the difference in the logs of expected counts for `international_and_specific_audiences` is expected to increase by the number of corresponding units in *Table 3* In addition, this relationship is strongly significant.
- For every unit increase in `user_access_edit_and_deletion`, the difference in the logs of expected counts for `international_and_specific_audiences` is expected to increase by the number of corresponding units in *Table 3.* In addition, this relationship is strongly significant.
- The level of inconsistency in modeling for `data_security` and `user_choice_control` gave inconsistent measures of significance whenever the model was run, wavering between slight significant and no significance at all.
- In addition, the `Constant` variable gives us an understanding of our response variable when all other model variables are evaluated as zero. In other words, when none of the data action variables are identified in a given privacy policy, the log of expected counts for `international_and_specific_audiences` is expected to increase by the number of given units in *Table 3.*

So, what do these values mean for privacy policy quality in a practical sense? How does this relate to the content of privacy policies in a concrete manner? Well, for starters, we can certainly make the inference that an increase of two 'positive' data actions do in fact have a positive affect on the counts of clauses written for international and specific audiences. Interestingly enough, the two variables with the highest significance for our variable of interest were `do_not_track` and `user_access_edit_and_deletion`, both of which are largely covered by sections two (Information and access to personal data) and three (Right to erasure) of the GDPR (Council of European Union 2014). As a result, it is in some ways unsurprising that these variables have the strongest relationship to `international_and_specific_audiences` as the GDPR itself is not necessarily concerned with the technical aspects of data protection as measured in `data_security` and it is likely that clauses of `user_access_edit_and_deletion` have a high level of similarity to the language being used in `user_access_edit_and_deletion`, likely making these data actions especially difficult for the expert annotators to parse.

While I am not necessarily comfortable making causal claims based on some metrics in the dataset, it does seem there is a slight relationship to be observed between do not track clauses, edit and deletion clauses, and international and specific clauses. One possible causal factor to consider here is simply the size of each privacy policy, which vary widely in terms scope, complexity, and therefore sheer volume of data actions. In addition, the correlation noted between `do_not_track` and `international_and_specific_audiences` must also be called out, as these data actions were largely absent from the dataset except for entries where another 'positive' data action was mentioned, which means the reported statistical significance is also likely due to low numbers.

5.2 Limitations

I deigned to make too many causal claims above simply because I need to save space to discuss limitations as to why this experiment cannot and should not speak grandly about causality. For starters, the original OPP-115 dataset was created in 2016 with a level of prominence that is relatively unknown. In addition, the number of observations being 115 is somewhat out of the "big" scope we have come to associate with "big data" approaches to analysis. This is further compounded by the very real issue of both the GDPR and CCPA being passed as regulation a respective two and three years after the OPP-115 was created. Given the python-formatted nature of the original datasets, this researcher was frankly not keen on taking a large risk from a technical standpoint, and therefore limited themselves to one of the simpler datasets in the OPP-115 that was somewhat more intuitive to format for r. Of course, the consequence of this is the analysis provided in this project is only limited to a scope of data that is largely outdated for the purpose it was intended to be used for.

5.3 Future Research

Stemming from limitations, further research in this domain *specifically as it applies to R* would be to first create a common, updated, and dedicated research program with the aim of porting functionality from the python formatted usable privacy datasets into an r-friendly paradigm. While this project established the most minuscule of precedents for this effort, scholars in the fields of machine learning, applied statistics, and computer science are likely far better equipped to handle this monumental challenge than this researcher could.

In terms of next steps or possible avenues from research that could stem from the particular variables in question, a model with an updated version of the OPP-115 would likely suffice, but would also be further enriched through a secondary model which also accounts for excess zeroes in the `do_not_track` variable by adding a zero inflated parameter. Overall, future research for the usable privacy project is likely best done using datasets were developed and released with legislative actions accounted for.

Overall, this project was an attempt to highlight the potentiality of relationships between specific dimensions of a privacy policy. While this was the primary objective, a secondary contribution comes in the form of the `opp115` dataset. It is my hope that this analysis, and the dataset created as consequence, can assist policymakers and other researchers who are interested in engaging with the usable privacy project.

References

Bhatia, Jaspreet, and Travis D. Breaux. 2018. "Empirical Measurement of Perceived Privacy Risk." *ACM Transactions on Computer-Human Interaction* 25 (6): 1–47.

Council of European Union. 2014. "General Data Protection Regulation."

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. "Datasheets for Datasets." *CoRR* abs/1803.09010. http://arxiv.org/abs/1803.09010.

Kumar, Vinayshekhar Bannihatti, Abhilasha Ravichander, Peter Story, and Norman Sadeh. 2019. "Quantifying the Effect of in-Domain Distributed Word Representations: A Study of Privacy Policies." *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*.

Nabel, Michael. 1997. "NCSS." *The American Statistician* 51 (1): 97.

Sadeh, Norman, Ro Acquisti, Travis D. Breaux, Lorrie Faith Cranor, Aleecia M. Mcdonalda, Joel R. Reidenbergb, Noah A. Smith, et al. 2013. "The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to SemiAutomatically Answer Those Privacy Questions Users Care About." The Usable Privacy Project.

Wilson, Shomir, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, et al. 2016. "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)*. The Usable Privacy Project.

Zimmeck, Sebastian, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. 2019. "MAPS: Scaling Privacy Compliance Analysis to a Million Apps." *Proceedings on Privacy Enhancing Technologies* 2019: 66–86.

Zook, Matthew, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, et al. 2017. "Ten Simple Rules for Responsible Big Data Research." Public Library of Science San Francisco, CA USA.