Investigating COPPA Notification Compliance

Are App Developers Ensuring Compliance for the Sake of Children’s Privacy?

Mark Brom, Lydia Esbaum, Evan Lemker, Peter Mertka, Sebastian Rivera

ABSTRACT

In 1998, the United States Congress passed the Children’s Online Privacy Protection Act (COPPA), the first bill of its kind that addressed the need to protect the privacy of children. Due to their age, it is crucial that regulation provides some sort of rules and guidance as to what can and cannot be collected from children who are often more susceptible to marketing targeted ads resulting from data collection. COPPA itself requires many different practices be followed by those who collect children’s information, with our study focusing on the notification of collection via an entity’s privacy policy.

In order to comply with what is stated in COPPA’s regulations, three things must be displayed somewhere on a website regarding the data collection of children. First, there must be some sort of contact information, including name, address, phone number, and email address of either all parties collecting data, or of one designated data monitoring officer. Second, there must be a clear description of what kinds of data is being collected from users who are thirteen years old or younger. Lastly, it is necessary that either a direct link to or a process describing how a parent can review or request deletion of their child’s data must be available.

In this study, we begin to investigate how app developers on the Google Play Store with apps directed specifically at this young age group are complying with the necessary notification requirements of COPPA. By utilizing a one-time web crawl, we collect the privacy policies of over five hundred apps that are found in the “Kids” section of the Google Play Store. Using these policies, we can analyze the content to see which portions of notification are being included and to what extent. Using our results, we can better understand just how much of this bill is being followed twenty-five years after its initial passing for a medium that did not exist during the bill’s initial passing.

KEYWORDS

COPPA, Privacy Policies, Google Play Store, Compliance

1 Introduction

On October 21st, 1998, the Children’s Online Privacy Protection act was signed into law, bringing necessary regulation and protection to reign in an ever expanding and evolving online ecosystem. The act went into effect on April 21st, 2000, allowing for commercial websites and online services to ensure compliance with all the aspects of the bill [1]. Twenty-three years later, privacy is still a hotly debated issue and has become increasingly concerning as the number of avenues to collect data expands.

One avenue that hadn’t even existed at the time of passage of the law is that of mobile applications. With the rise of smartphones, tablets, and internet of things device use over the past ten years, it has become much easier for children below the age of thirteen to not only access the internet, but to use it without supervision. With this has come a significant number of mobile applications that have been solely developed and geared towards children. With these apps, developers have the ability to collect wide swaths of information that can be leveraged for profit against the user via targeting marketing and user profiling. This practice, however, is the exact type of behavior that COPPA is designed to protect against, and thus it is necessary to evaluate and ask just how many of these developers are in compliance with the exact rulings of the law?

This question is pertinent to our world today, as new technologies continue to evolve and grow at a rapid rate that is bound to outpace our current legal constraints, so it is key that one of the most prominent and popular mediums of internet access for children is following through with the necessary compliances, as by understanding where companies and developers are failing, it will be possible to provide adequate recommendations as to how these policies can be better enforced, for both current and future technologies.

For this study, we focused on one specific mandate of COPPA, specifically mandate 312.4d. The entirety of COPPA is a large and complex document that outlines numerous requirements and regulations that applicable companies are to follow, but we felt that it would be simplest to identify some of the most explicit requirements of the act itself. By focusing on 312.4d, we can better identify compliance or the lack there of, while also taking into consideration the constraints imposed by this style of audit. A wider scope would lead to a less organized study considering the time requirements of this study. To better understand what we are investigating, it is vital to know what 312.4d says. The exact text of mandate is as follows:

**312.4d (Official Statement): *Notice on the Web site or online service.*** In addition to the direct notice to the parent, an operator must post a prominent and clearly labeled link to an online notice of its information practices with regard to children on the home or landing page or screen of its Web site or online service, *and,* at each area of the Web site or online service where personal information is collected from children. The link must be in close proximity to the requests for information in each such area. An operator of a general audience Web site or online service that has a separate children's area must post a link to a notice of its information practices with regard to children on the home or landing page or screen of the children's area. To be complete, the online notice of the Web site or online service's information practices must state the following:

(1) The name, address, telephone number, and email address of all operators collecting or maintaining personal information from children through the Web site or online service. *Provided that:* The operators of a Web site or online service may list the name, address, phone number, and email address of one operator who will respond to all inquiries from parents concerning the operators' privacy policies and use of children's information, as long as the names of all the operators collecting or maintaining personal information from children through the Web site or online service are also listed in the notice;

(2) A description of what information the operator collects from children, including whether the Web site or online service enables a child to make personal information publicly available; how the operator uses such information; and the operator's disclosure practices for such information; and

(3) That the parent can review or have deleted the child's personal information and refuse to permit further collection or use of the child's information and state the procedures for doing so [2].

In section 312.4d, websites or online services must include the presence of clear identifiable links tied to information notices regarding children on the site’s home or other landing pages. Each site needs three components: contact information of those collecting the data (names, address, phone number), a section stating what is being collected, and a section devoted to the review and deletion of child’s data. Our web scrape of the Google Play Store should look for these policies and grab them from the web for further analysis. From there, our text segmentation scripts and natural language processing model will identify these three components and determine whether or not an app is in compliance with COPPA. If an app service is missing one of the three components, we can prove that the app is not fully COPPA compliant. Knowing this, we aim to answer the following research questions:

-**RQ1:** How many companies are posting the necessary data collector contact information so that parents can get into contact with the proper data authorities?

**-RQ2:** Which developers are clearly defining what is being collected of children by their applications?

**-RQ3:** What portion of companies in our dataset are providing clear and proper instruction for parents to review and delete their children’s data?

It is necessary that we analyze this compliance because there have been companies in the past that have not followed through on these regulations, thus making it incredibly difficult for parents to assert necessary control and judgement on behalf of their child [3]. The ability to review and understand what sort of information is being collected from your children is one that is vital to ensuring that companies remain accountable to the regulations COPPA provides.

In order to answer these questions, we utilized a three-step process that is able to crawl various sections of the Google Play Store’s “Kids” section. Firstly, we acquire the content of various children’s applications’ webpages from the Google Play Store website. From there, each page’s results is scraped to isolate the company’s directly posted privacy policy thanks to Google Play’s built in privacy policy tab listed underneath each developer’s contact information. From there, each policy is split into manageable one to three sentence chunks so that it can be properly classified by our natural language model. Lastly, each chunk is parsed by a combination of pattern matching expressions and our model in order to directly state to what degree are companies complying with each aspect of 312.4d.

Overall, our results have not been fully developed yet. We are in the process of finalizing the code that will return the final results.

2 Background

We begin our process by first reviewing prior works that have either audited similar laws or investigated different aspects of COPPA itself.

**2.1 Prior Work**

The Children’s Online Privacy Protection Act (COPPA) is a set of requirements that online services must abide by that deal with how data from children under 13 years of age is managed. The purpose of COPPA is to protect children on the Internet by regulating what personal information online services can gather, use, and share about children. COPPA also gives parents control over how their children’s personal information is gathered and shared by attaining verifiable parental consent. COPPA only applies to services that are either directed toward children under the age of 13, services that are directed to the general public but know they have collected information from children under 13, or services that know they have collected information directly from users of another online service directed to children [4].

There have been several prior studies that have analyzed COPPA and apps’ compliance with COPPA, and our group has chosen to loosely model our own project based on one of these studies, “‘Won’t Somebody Think of the Children’ Examining COPPA Compliance at Scale”. Researchers of this study created an automated evaluation framework for the privacy practices of Android apps. Specifically, the top 5,855 apps geared toward children that are governed by COPPA from Google’s Play Store in the U.S. were used in the analysis. Unlike many approaches that aim to identify potential COPPA violations but fail to do so because they do not observe actual violations or do not scale, the framework used in this study allowed researchers to supervise apps’ behaviors in real time and at scale [4].

Essentially, the methodology of the study included retrieving apps from a corpus of free, children-directed apps on the Google Play Store, running each app, and analyzing the information that was collected about each app’s access to personal information and communication to third parties. During analysis, parsing and extracting certain pieces of information, like whether an app accessed Android-guarded resources, was an automated process while obtaining other pieces of information, like checking for personal information in network transmissions, was a manual process. Similar to this approach, our group will be automating parts of our analysis as well as manually analyzing the data. We will also be using apps that are children-directed from the Google Play Store in our project, only we will be examining a mere 500 different applications. Our project complements this study by focusing on whether apps comply with a specific section of COPPA, 312.4(d) Notice on the website or online service, rather than analyzing if any section of COPPA is violated.

Another study that closely resembles the research we hope to do is "Analyzing privacy policies through syntax-driven semantic analysis of information types." This research paper focuses on creating a program that can automatically analyze hard-to-understand privacy policies and generate easy-to-understand summaries of what is being collected and shared in the policy. Our goal is similar to this as we want to create a program that can automatically analyze privacy policies to find the presence or lack of required components under COPPA regulation. While the formerly mentioned research project helps identify key concepts and categories relevant to users' privacy concerns, we aim to identify shortcomings in regulatory requirements relevant to companies covered by COPPA [5].

The researchers in this study designed a program using natural language processing techniques to perform syntax-driven semantic analysis of each part of the privacy policies. These techniques included heavily focusing on information types present in the privacy policy, such as names, email addresses, phone numbers, locations, etc., to identify what pieces of information were being used and how they were being used. Similarly, we will have critical information we will be searching for within privacy policies, such as web links, email addresses, and phone numbers. However, rather than using natural language processing only to understand what the company does with this information, we will be using it to understand whether or not these links and email addresses can be used to access the online notices and data collection contact information that is required under COPPA for companies' privacy policies to contain.

Lastly, another research group at the University of Iowa: The Security, Privacy, and Anonymity Research Team, or SPARTA Lab is investigating a different law but via a similar method. This lab is headed by Dr. Rishab Nithyanand and is currently working on several projects related to online privacy and regulation. Within these projects, there is one aiming to analyze privacy policies just as we are for compliance with regulatory frameworks. Specifically, they use the same natural language processing guided approach we used in our research. However, instead of using it to determine compliance with COPPA, their methods are concerned with determining compliance with the California Consumer Privacy Act (CCPA). This is significant because we will be able to collaborate with members of the research team, such as Maaz Bin Musa, a Ph.D. candidate at the University of Iowa studying under the supervision of Dr. Nithyanand, in order to gain insights into the design and use of different natural language processing techniques to find the key results we are searching for.

**2.2 Applicability**

The key first question to ask before we could even begin collecting privacy policies was to understand who COPPA applied to. Thankfully, unlike some other privacy laws such as the CCPA, the text of COPPA clearly outlines who needs to be in compliance with the specifications. Per the FTC’s rules, COPPA encapsulates all websites and online services (such as mobile apps) that are directly targeted at children who are age 13 or younger. Furthermore, anyone with knowledge of collecting, using, and disclosing personal information of children is also included, even if the data is being collected from a different site [1].

Another key distinction that is necessary when it comes to identifying who COPPA applies to is that the country of origin of the website or app’s controller does not exempt them from the outlines of the law. As long as the website or service is targeted at and used by U.S. children who are 13 years old or younger, they are required to follow all necessary requirements [1]. Thankfully, this meant that our web scrape and app selection did not need to take into account whether or not the app itself was expected to comply to the law. By virtue of the app being available in the “Kids” section of the U.S. version of the Google Play Store, that application is being targeted towards U.S. children. This is because the “Kids” section is explicitly for applications that are marketed at children who are 12 years old or younger. Thus, there is no need to exclude any companies that are collected during our web scrape.

**3 Experimental Design**

To gather the data necessary for the project, we will be scraping the top 1,000 apps in the Kids category of the Google Play Store. To do this we first need to set up a web scraping tool, something like Octoparse or Scrapy, that we can use to automate this process. We will be looking for relevant data regarding the application name, developer, and privacy policy information and then placing this data into an Excel sheet for later review by our group.

When getting the privacy policy information, we will need our scraper to look through the Google Play Store application page, find the “Data Safety” title and click the button to go into more detail. This will bring up a page that, if they have one, includes a link to the developer’s privacy policy. If it is listed, the scraper will take this link and add it to the excel file. We also intend to use the web scraper to look for common words or phrases that would automatically verify privacy policy compliance after we manually comb through enough to be confident on what words or phrases to be looking for.

Once we have scraped the top 1,000 apps in the Kids category of Google Play, we will begin the more fine-tuned reviewal process for each app developer. As it is likely that a developer will have more than one top 1,000 app, we will only review the privacy policies of the developers present in our scrape. Furthermore, any developers found to have not posted their privacy policy on the Google Play app store itself will have to be reviewed manually to identify their privacy policy to provide the most accurate results possible in the analysis. This is to ensure that companies that are in theory compliant with COPPA but have not posted their policy officially can be properly recorded.

Using the text that is returned from the results of the web scrape, we can begin auditing each company for the three different components of compliance that we are testing for. The first requirement, that the contact information of either all parties involved or some designated data operator be listed will be the simplest to test for as well as the most complex to evaluate full compliance. We can identify the presence of the necessary contact information: name, addresses, phone numbers, and email addresses, using regular expressions to test for what information is present in the policy itself. However, this will not guarantee compliance, as initial testing showed that most companies do not include all four. Thus, we quantify how much of the first regulation is being complied with, as simply qualifying compliance as an all or nothing approach will result in data that claims that almost no developers are currently compliant with this specific section.

Testing for the presence of the second and third parts of COPPA compliance is more of a challenge. There does not exist a stipulated language that is to be used by companies when identifying what information is being collected from children and how a parent is to proceed if they want to review or remove their child’s data. As a result of this, our identification is based on identifying selections of text that match a certain similarity threshold to predefined terms that our group selected as being associated with each of these two sections. To identify what sort of phrases are more closely tied to sections two and three of the COPPA regulation, our team manually reviews sample privacy policies to select phrases that appear to be complaint with the rule.

These selected key phrases will then be parsed and compared to the different headings and subsequent text that appears in each policy in order to determine if there is a section that can be considered sufficient to what COPPA stipulates in rule 312.4d. Each requirement will be tested separately, to ensure that no developer is penalized for having one section compliant but not the other. Before deciding that the selected phrases are sufficient for deducting compliance, manual error validation occurs to ensure that there is not a concerning level of false positives or negatives that our evaluation missed. Our results thus demonstrate to what extent developers are complying with our chosen rule. Using the data created from this analysis, it is then possible to investigate any potential trends that relate between app position and COPPA compliance.

REFERENCES

[1] FEDERAL TRADE COMMISSION, 2020. Complying with COPPA: Frequently Asked Questions. https://www.ftc.gov/business-guidance/resources/complying-coppa-frequently-asked-questions

[2] CODE OF FEDERAL REGULATIONS, 2013. Part 312 – Children’s Online Privacy Protection Rule. https://www.ecfr.gov/current/title-16/chapter-I/subchapter-C/part-312

[3] O’MELVENY, 2023. FTC Obtains Record Penalties from Video Game Company Amidst Growing Privacy and Consumer Protection Enforcement Trends. http://www.omm.com/resources/alerts-and-publications/alerts/ftc-obtains-record-penalties-from-video-game-company/#:~:text=The%20US%24275%20million%20penalty,affirmative%20consent%20from%20their%20parents.

[4] REYES, I, ET AL. 2018. “Won’t Somebody Think of the Children?” Examining Coppa Compliance at Scale*. Proceedings on Privacy Enhancing Technologies vol. 2018 no. 3* 63-83.

[5] BREAUX, T.D., ET AL. 2021. Analyzing privacy policies through syntax-driven semantic analysis of information types. *Information Software and Technology vol 138*

*Conference Name:ACM Woodstock conference*

*Conference Short Name:WOODSTOCK’18*

*Conference Location:El Paso, Texas USA*

*ISBN:978-1-4503-0000-0/18/06*

*Year:2018*

*Date:June*

*Copyright Year:2018*

*Copyright Statement:rightsretained*

*DOI:10.1145/1234567890*

*RRH: F. Surname et al.*

*Price:$15.00*