JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Volume 13, Issue. 2
July 2020
ISSN: 1946-1836

In this issue:

4. **Digital Transformation and Geospatial Analytics**
Ileana Abot, Xavier University
Thilini Ariyachandra, Xavier University

13. **Examining Organizational Security Governance (OSG) Objectives: How strategic planning for Security is undertaken at ABC Corporation**
Sushma Mishra, Robert Morris University

25. **An Exploratory Study of the Perceptions of Library Faculty and Patrons on Library Resources**
Kyle Herman, Georgia Southern University
Hayden Wimmer, Georgia Southern University
Carl Redman Jr., University of San Diego

33. **Developing a GIS Dashboard Tool to Inform Non-Profit Hospitals of Community Health Needs at the Neighborhood Level**
Evan Copello, University of North Florida
Jason Smith, University of North Florida
Anna Mease, Mayo Clinic
Karthikeyan Umapathy, University of North Florida
Dan Richard, University of North Florida
Ann-Marie Knight, Mayo Clinic
Monica Albertie, Mayo Clinic
Hana Hamadi, University of North Florida
Aaron Spaulding, Mayo Clinic
Emma Apatu, McMaster University

54. **A Longitudinal Study of Google Privacy Policies**
Alan Peslak, Penn State University
Lisa Kovalchick, California University of Pennsylvania
Mauir Conforti, Geisinger Health System

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three issues a year. The first date of publication was December 1, 2008.

JISAR is published online (http://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (http://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of EDSIG who perform the editorial and review processes for JISAR.

2020 Education Special Interest Group (EDSIG) Board of Directors

Jeffry Babb West Texas A&M President	Eric Breimer Siena College Vice President	Leslie J Waguespack Jr. Bentley University Past President
Jeffrey Cummings Univ of NC Wilmington Director	Melinda Korzaan Middle Tennessee State Univ Director	Lisa Kovalchick California Univ of PA Director
Niki Kunene Eastern Connecticut St Univ Treasurer	Li-Jen Lester Sam Houston State University Director	Michelle Louch Carlow University Director
Rachida Parks Quinnipiac University Membership	Michael Smith Georgia Institute of Technology Secretary	Lee Freeman Univ. of Michigan - Dearborn JISE Editor

Copyright © 2020 by Information Systems and Computing Academic Professionals (ISCAP). Permission to make digital or hard copies of all or part of this journal for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial use. All copies must bear this notice and full citation. Permission from the Editor is required to post to servers, redistribute to lists, or utilize in a for-profit or commercial use. Permission requests should be sent to Scott Hunsinger, Editor, editor@jisar.org.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Editors

Scott Hunsinger
Senior Editor
Appalachian State University

Thomas Janicki
Publisher
University of North Carolina Wilmington

2020 JISAR Editorial Board

Wendy Ceccucci
Quinnipiac University

Ulku Clark
University of North Carolina Wilmington

Christopher Davis
Univ of South Florida, St. Petersburg

Gerald DeHondt
Ball State University

Ed Hassler
Appalachian State University

Muhammed Miah
Tennessee State University

James Pomykalski
Susquehanna University

Christopher Taylor
Appalachian State University

Karthikeyan Umapathy
University of North Florida

Peter Wu
Robert Morris University

Jason Xiong
Appalachian State University

A Longitudinal Study of Google Privacy Policies

Alan Peslak
arp14@psu.edu
Information Sciences and Technology
Penn State University
Dunmore, PA 18512 USA

Lisa Kovalchick
kovalchick@calu.edu
Mathematics, Computer Science and Information Systems
California University of Pennsylvania
California, PA 15419 USA

Mauri Conforti
maconforti1@geisinger.edu
Geisinger Health System
Scranton, PA 18508 USA

Abstract

Google has come a long way since its founding by Larry Page and Sergey Brin in 1998. From a Stanford University PhD research project, to the world leader in Internet search engines, Google has changed and grown exponentially. "Google it" has become a household phrase. Its core search engine product has approximately 246 million unique US visitors per month. Its market share in the US, in search engines, is estimated at 63%. In addition, it has expanded product offerings to include the ubiquitous YouTube Internet video platform, Google Home smart speakers, Chrome web browser, Android operating system, Pixel phones, and Gmail, among other products. As a result, Google collects a tremendous amount of data from its users. With its growing popularity and the growing user privacy concerns, following the recent data breaches, Google is constantly updating its privacy policy. Our manuscript examines Google privacy policies from 2000 to present day. The policies were collected via the Internet Archive and represent a selected day from each available year from 2000 through 2018. A comprehensive qualitative, linguistics and sentiment analysis on these policies was performed. Our review finds significant similarities and differences in Google privacy policies over the years. Overall, Google privacy policies have become increasingly wordy and legalistic, but also more positive in sentiment and more personal in approach. Though the grade level of the documents has declined slightly they remain at a 12th grade level. Implications and opportunities for further research are presented.

Keywords: Sentiment analysis, Google, Linguistic analysis, LIWC

1. INTRODUCTION

Through the use of the web, people are now able to easily search for any type of information by typing a few keywords into an Internet search engine and the information they are seeking is at their fingertips. Google is the most popular search engine in the US. According to Statista (2019), in the month of December 2018, there were 246 million unique visitors to the search engine home page. This is startling considering there were 328

million people in the US in December 2018 (Census, 2019).

Google also has an estimated 63% market share in the United States for Internet searches. In the global market, their market share is even higher, estimated at 90%. (Statista, 2019). "Seventy-four percent of U.S. adults currently say they use Google in a typical week" (Jones, 2018).

The original Google product was the online search engine, but its product lines have expanded greatly over the years and they all share their data across platforms. Their products now include YouTube, which has been accessed by 68% of the population. They have developed online advertising, and sponsored searches. (Statista, 2019). As reported by Kelly (2018), Google "is the world's largest digital advertising company". They control the Android operating system, which has an 86% market share, globally, for operating systems. They own and control Google Maps, Gmail, the Chrome web browser and Chrome OS. They have in home devices such as Chromecast, Google Home, and Google Now and Assistant. Recently, they expanded into mobile phones with Nexus and Pixel brands (Statista, 2019).

Kelly (2018) performed a detailed journalistic study of Google and found many interesting and chilling statistics; she reports that "Google collects far more data than Facebook".
Have you ever been sitting next to someone talking about making a purchase, maybe for something like a new pressure washer, then, the next time you take a look at the web, there it is, an advertisement for a pressure washer? As Kelly (2018) reports, this is not coincidence; she quotes a study by Digital Context Next which reports "Google can collect data even if you aren't using your phone. The study says that a dormant Android phone with Chrome running in the background sent location data to Google servers 340 times in one 24-hour period. ….But Google collected two-thirds of its data without any input at all from users in the researcher's experiment."

If you are concerned about your privacy and want to limit the information that Google collects, you may think this would be a simple task. All you need to do is to adjust the privacy settings in the myriad of Google products that you use. But hold on, it is not that simple. "A recent investigation by the Associated Press found that the company continued to record location data, even after a user disables the Location History option. Google said the data is used to improve services but has updated the wording of the setting to make it clear that location information is still collected." (Kelly, 2018).

Kennedy chronicled some incredible statistics in 2008, statistics which have only grown over the past decade. In 2008, Google processed 40,000 search queries each second or 1.2 trillion searches yearly, worldwide. In 2008, Google processed over 20 petabytes of data per day through an average of 100,000 MapReduce jobs spread across its massive computing clusters. A petabyte is a million gigabytes. This is a massive amount of data being collected and stored about its myriad of users.

As a result of this huge data collection, there is increasing concern over what use Google makes of this data. States attorneys general and the federal government have been investigating what and how Google stores and shares its data, and specifically, how Google shares information across all its programs and devices further reducing privacy. The 2018 Google Privacy Policy specifically states "We may combine the information we collect among our services and across your devices for the purposes described above."

With the emergence and rapid increase in Internet usage, along with the recent increase in privacy concerns due to data breaches that exposed the personal information of many Americans, some states have created recommendations with regard to online privacy policies. In 2004, California was the first state in the U.S. to enact online privacy policy legislation, the California Online Privacy Protection Act (CalOPPA), which requires commercial websites and online services to post a privacy policy. Since CalOPPA went into effect, the California Attorney General has set forth several recommendations regarding the construction of such privacy policies. One such recommendation addresses the readability of such policies; it recommends the use of "plain, straightforward language. Avoiding technical or legal jargon" (Harris, 2014). Turow, J., Hennessy, M., & Draper, N. (2018) begin to explore the problems with privacy policies and examine Americans' misunderstanding of the function of privacy policies.

In order to explore the current status of Google privacy, we need to examine the privacy policy that they post and to which they are legally required to adhere. Our study reviews the evolution of this policy from 2000 to the present day. By examining the policy and its evolution, we hope to gain insight into how the policy has

changed and what its implications are for the myriad of Google products users.

2. LITERATURE REVIEW

Over the years, a number of studies have examined the readability, content, and complexity of various website privacy policies. Graber, D'Alessandro, and Johnson-West (2002) examined 80 Internet health websites and found that these privacy policies are "not easily understandable by most individuals in the United States and do not serve to inform users of their rights." Jensen and Potts (2004) analyzed 64 website privacy policies and found that "only 6% of policies are readable by the most vulnerable of the population, and that 13% of policies were only readable by people with a post-graduate education". Proctor, Ali, and Vu (2008) examined the privacy policies of 100 websites and found that "although the readability analysis showed that a person with 13 years of education should be able to comprehend the policies, college students were able to answer correctly only about 50% of questions they were asked about specific policies".

Several studies have examined snapshots of website privacy policies for various reasons. Opsahl (2010) examined snapshots of Facebook's privacy policy over the years to demonstrate what he perceived as the disappearance of user's privacy. Warzel and Ngu (2019) examine snapshots of Google privacy policies over the past 20 years, in order to demonstrate how the Internet has changed and to attempt to understand underlying reasons for major changes to the Google privacy policy over the years.

Sentiment evaluation and linguistic analysis are commonplace techniques of studies in communication analysis. The use of word frequency and word analysis, though not perfect, is well established in the literature as a tool for corpus analysis including Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). The utilization of linguistic analysis and especially the use of a LIWC (Linguistic and Word Count) software program for research functions has been substantial. Back, Kufner, and Egloff (2011), Cordova, Cunningham, Carlson, and Andrkowski (2001), and Robinson, Navea, and Ickes (2013) all used LIWC analysis.

Holtman et al. (2018) used LIWC to study linguistic patterns of narcissism and corelated these with sports, second-person pronouns and swear words. Hawkins et al. (2017) used LIWC to study dream content.

LIWC software (Pennebaker, Booth, Boyd, and Francis, 2015) is described as such: "The way that the **L**inguistic **I**nquiry and **W**ord **C**ount program works is fairly simple. Basically, it reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech" (Pennebaker Conglomerates, 2015).

Overall, sentiment analysis has been an increasingly important qualitative analysis tool. As Liu (2012) defines: "Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes,". Sentiment analysis is the review of written or other forms of communication or qualitative data to determine a quantifiable and comparable measure of some form of feeling in the communication or data. Pak and Paroubek (2010) studied Twitter feeds for sentiment analysis. Pang and Lee (2008) analyzed whether textual information had a positive or negative sentiment.

Google has also been a valid and frequent subject of study in the literature. For example, Wu and Brynjolfsson (2015) studied Google Trends to predict changes in housing prices and sales. Many studies have been performed on privacy policies of Internet sites. The authors previously studied the Privacy Policies in several manuscripts (Peslak, 2016, Peslak, 2017, Peslak, 2018). After a comprehensive Google Scholar search, we could find no instances of sentiment analysis or qualitative mining of Google privacy policies in the literature.

3. METHODOLOGY

To study the Google privacy policies over time, it was first necessary to obtain past and current Google privacy policies. The current Google privacy policy was obtained from their website, and past policies were retrieved from the Internet Archive (https://archive.org/web/) also known as the Wayback Machine. As noted, we selected one dated page from each available year from 2000 to 2018 (2001-2003 were unavailable) and retrieved the archived privacy policy from Google from that date. Though there may have been years where the policy may have changed at other times within the year, we believed a once a year selection provided a reasonable representation of the volatility of Google privacy. Once we extracted the policies, we next needed to determine how to analyze these policies. We chose three areas of analysis based on past work and other qualitative relevant literature.

The three general areas studied include: Overall Content, Specific Word and Key Word Content, and Sentiment and Linguistic Analysis. To analyze Overall Content, we utilized several tools; Microsoft Word was used to determine reading grade level, complexity, and word count. For sentiment and linguistic analysis, two tools were used. LIWC was used to determine key variables over time including clout, analytic, tone, and authenticity. IBM Watson Sentiment Analysis was used for sentiment (positive/negative) evaluation to determine degree of positive and negative content. For keyword and other specific content, Voyant Tools was utilized, as well as, specific author reviews of each policy. Another tool used in the study was Microsoft Excel for charting and other analyses.

In addition, we imported the privacy policies into LIWC (Linguistic and Word Count). LIWC software produces unique measures for linguistic analyses. For the most part these are expressed by a percentage of total words mapping to the dictionary category of each measure. The exceptions are several relating to word counts, as well as, calculated emotional measures. Appendix 1 includes all the measures, including LIWC variables used. Analytic reflects logical thinking versus narrative, authentic reflects honest versus guarded, and tone reflects upbeat versus sad.

4. RESULTS

The first finding of our study is that the Google privacy policy has become more difficult to read. One measure of complexity is the reading grade level of the policies. As Figure 1 depicts, Google has always had a high reading grade level for its privacy policies. (Note that all raw data in charts is shown in Appendix 1). Although reading grade level decreases slightly over the years, the expectation is still at a 12th grade level or above. According to the Clear Language Group (n.d.) "For the general public, text should be written at the 8th grade level or lower". An example of online documents with much better readability index come from a study by Leroy et al. (2008), who found online consumer health sites have a readability index of 10.5. Google's own guide to "Finding your way around YouTube" instructions only requires a Flesch-Kincaid reading grade level of 9.3 (Google, 2019). The current Google privacy policy now includes video explanations and additional imagery; thus it seems that Google may be attempting to make the content of its privacy policy more accessible to the general public.

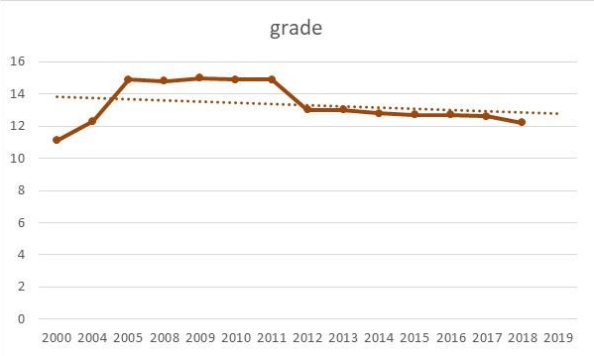

Figure 1. Reading Grade Level

Our readability findings are aligned with previous findings. Protcor et. al. (2008) studied the privacy policies of 100 different websites and found that the policies were at a 13th grade reading level. Jensen and Potts (2004) examined 64 website privacy policies and found that the average reading grade level was 14.15. Graber et. al. (2002) studied the privacy policies of 80 Internet health websites and found that, of the sites that had a privacy policy, the average readability level was that of grade 14.

The increasing complexity of the Google privacy policy over the years is shown by an ever-increasing word count (see Figure 2). In 2000, the word count of the privacy policy was 657, with just over one page of information. In 2009, the word count increased to 2,140, with four pages. Multiple new sections were added since 2000, such as Choices for Personal Information, Information Sharing, Information Security, Data Integrity, and Accessing and Updating Personal Information, and Enforcement. Also, in 2000, sections included What Information Do We Collect and the "Google and Cookies" sections had three paragraphs that expanded to fourteen paragraphs in 2009, under the heading Information We Collect and How We Use It.

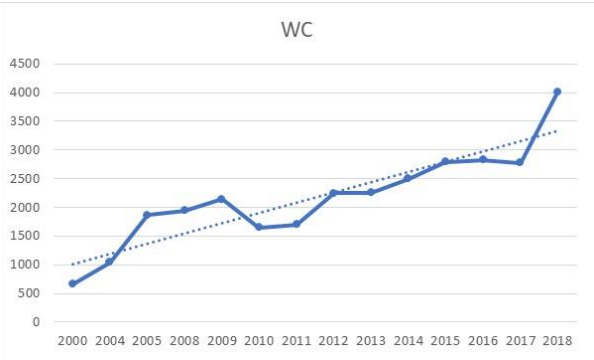

Figure 2. Word Count

In addition, the 2009 policy added uses for personal information including: Customized Content and Advertising, Auditing, Research and Analysis, Ensuring Technical Function, Protecting the Rights or Property of Google or Our Users, and Developing New Services. There is also a notice that Personal Information can be processed in the United States or on servers in other countries. Choices for Personal Information in 2009, include: Your Consent and Changes to the Privacy Policy.

Google acquired the Internet advertising company DoubleClick in 2008 and shortly afterwards, the Information Sharing section of their privacy policy increased from one paragraph to four paragraphs previously titled, With Whom Does Google Share Information? Additional new information includes:

- Require consent to share personal information
- Personal information can be used by trusted businesses on Google's behalf.
- Legal requests, terms of service enforcement, security reasons.
- Can be transferred if a merger or sold after notice is given.
- Third party sharing will not identify who you are, just your interest.

Several other sections were added in 2009, including the following. The Information Security section describes security measures in place, to restrict unauthorized access to personal data. The Data Integrity section describes how personal information is used according to the privacy policy and reviewed for accuracy; users must update their own information, when needed. The Accessing and Updating Personal Information section describes your access to personal information and the procedures used to correct or delete it. The Enforcement section, previously titled Who Can I Ask if I Have Additional Questions, informs users how to contact Google with questions or concerns and how they reply.

In 2010, the word count decreased to 1,652. Sections condensed include: Introduction, Choices for Personal Information's title was changed to Choices, and Changes to this Privacy Policy. The omitted items were Gadgets, Links, and Data Integrity. A Unique Application Number section was added. Some external services identify you with this ID, but this number is not linked to your Google personal information.

In 2018, the word count reached 4,009, the peak word count of all years. This increase came as Google re-wrote its privacy policy, in response to the Europe Union's (EU) General Data Protection Regulation (GDPR). "Simply put, the GDPR mandates a baseline set of standards for companies that handle EU citizen's data to better safeguard the processing and movement of citizens' personal data" (De Groot, 2019).

New sections that were added include the following (*in italics*). *Why Google Collects Data*, which states this is done to provide, customize, and deliver better services.

Your Privacy Controls, which defines controls the user can access to manage, review, and adjust privacy settings, such as links to Privacy Check Up and Product Privacy Guide.

Compliance & Cooperation with Regulators, which is reviewed frequently, to ensure compliance. Even though servers may be outside the country with different protection laws, Google provides the same protection, no matter where the server is located.

About This Policy, which explains that all the services offered by Google are covered by the privacy policy, such as YouTube and third-party sites. However, the privacy policy does not apply to the practices of other companies or vendors.

Related Privacy Practices, which added 18 links for additional information about Google privacy notices.

Sharing Your Information, which is on the 2010 policy with less content, than 2018. One of the new sections pertains to Domain Administrators. This policy is different, since it refers to students and employees under an organization that uses Google services, instead of individual users.

Next, we examine the analytic score of the various Google privacy policies. Pennebaker et. al. (2015) state that a high analytic score indicates formal, logical and hierarchical thinking; whereas, a low analytic score indicates more informal and personal thinking. As shown in Figure 3, a major decrease in the analytic score was found from 2011 to 2012, where a more personalized experience emerged. Google began using the words "you", "your", and "our users" to give a more individual-centered experience.

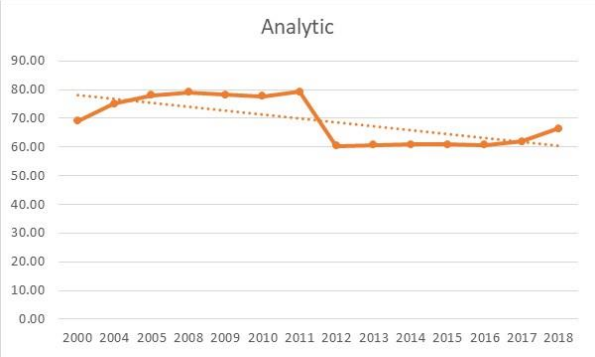

Figure 3. Analytic Scores

A re-worded 2012 introduction personalizes their services, products, and websites.

- "We strive to develop innovative services to better serve our users"
- "We recognize privacy is an important issue, so we design and operate our services with the protection of your privacy in mind"

Several sections were added in the How We Use Information We Collect section giving a more individualized experience.

- "We use the information we collect ……to offer you tailored content – like giving you more relevant search results and ads."
- Information from cookies and technologies like pixel tags is used "to improve your user experience and the overall quality of our services".
- Google combines personal information from its services "to make it easier to share things with people you know".

In 2018, the analytic rating finished slightly below where it began on the chart in 2000. The personalized tone shown in 2012 gradually changed to a more legalistic approach with less personal touch by 2018.

According to Pennebaker et. al. (2015), a high authenticity score indicates a more honest and personal text; whereas, a low authenticity score indicates a more guarded and impersonal text. As shown in Figure 4, authenticity scores decreased significantly over the 18-year span except for one year, with a more guarded, legalistic, less personal approach. In 2000, the score was 48.0, in 2004, it increased slightly to 51.1, in 2018, the most recent privacy policy, the score fell to 19.4. In 2004, they added some verbiage using "our" and "your" in the introduction and the addition of "we design and operate our services with the protection of your privacy in mind." This most likely accounts for the slight increase in the numbers. In 2012, authenticity dropped slightly, with the addition of the following sections: Transparency and Choices, Information Security, and Application of the Privacy Policy.

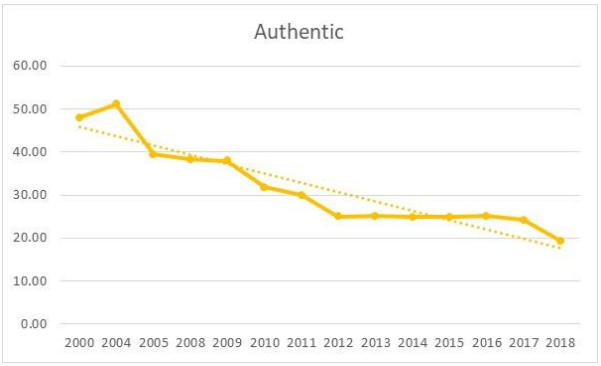

Figure 4. Authentic Scores

In 2018, as a result of the re-write to address the EU's GDPR, several new sections were added addressing security concerns, which most likely affected the authenticity score. Your Privacy Controls describes the controls for privacy management on Google. Links in the introduction are Privacy Checkup, Product Privacy Guide, and Google Account to Review and Update Information. Additional links included are: Activity Controls, Ad Settings, Information You Share, My Activity, Google Dashboard, Your Personal Information About You, Shared Endorsements and Export Your Data. In this section alone, in the content, 2018's policy has 26 links to assist in controlling your privacy compared to 0, in 2000 and 11, in 2010. Exporting & Deleting Your Information describes controls allowing you to export or delete some or all of your data. There are measures in place to protect the information you are deleting from being maliciously deleted. Related Privacy Practices has 18 additional links for access to more specific resources on how Google practices and their privacy policies such as Chrome, Payments, Privacy Checkup, Google's Safety Center, Technology and Principles, and How Google Uses Data When You Use Our Partners Sites or Apps, among others.

Over the years, the sentiment of Google privacy policies has shown a positive, friendlier, more enjoyable tone. Figure 5 shows the increase in sentiment based on IBM Watson. The exception is the 2018 privacy policy sentiment score, which decreased after 8 years of consistent increases. The highest increases are from 2000 to 2004. In 2004, specific positive passages were added:

- "when we require personally identifying information, we will inform you about the

types of information we collect and how we use it."

- "we hope this will help you make an informed decision about sharing your personal information with us."
- "we may share the information submitted … in order to provide you with a seamless experience"

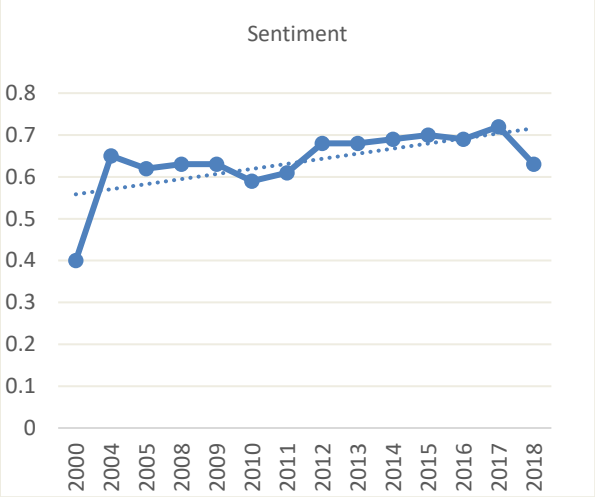

Figure 5. Sentiment Scores

From 2017 to 2018, there is a slight decrease in sentiment taking the sentiment to the side of a more negative experience.

The following 2017 passages that give a positive, friendly tone were removed in the 2018 policy.

- Using Google services with your information "we can make those services even better"
- "we tried to keep it as simple as possible, but if you're not familiar with terms like cookies, IP addresses, pixel tags and browsers, then read about them first. Your privacy matters to Google so whether you are new to Google or a long-time user, please do take time to know our practices – and if you have any questions, contact us."

A common tool for visualizing large text material is a word cloud. A word cloud shows the relative frequency of words in a document by the size of the word, itself. The word clouds for the collected Google privacy policies of 2000 and 2018 are displayed in Figure 6 and 7, respectively.

Figure 6. 2000 Google Privacy Word Cloud

Figure 7. 2018 Google Privacy Word Cloud

A review of the two word clouds shows large differences between the original and current privacy policy. Initially, cookies frequently appeared in the policy; however, in 2018, the term cookies, has much less prominence. Information and services dominate in 2018; whereas, in 2000, Google and search were very prominent. Use and account also become much more prominent in 2018, reflecting that detailed description of use of data and the inclusion of an account, in the recent policy.

Another fertile area for data visualization is network graphs that show collocation of words. These are readily available via Voyant Tools. A "Collocated Graph represents keywords and terms that occur in close proximity, as a force directed network graph." Keywords are shown in blue and collocated words (words in proximity), are shown in orange.

Figures 8 and 9 depict the collocated graphs for the collected Google privacy policies for the years of 2000 and 2018, respectively.

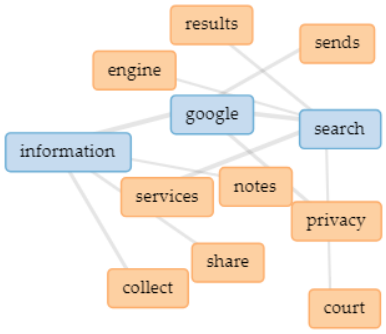

Figure 8. 2000 Collocation Graph

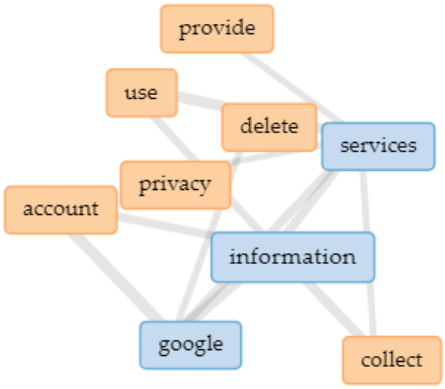

Figure 9. 2018 Collocation Graph

In 2000, the keywords were Google, information, and search. Google Collocated words were Google sends, Google privacy, and Google privacy court. With the keyword Information, collated words were Information collect, Information share, and Information notes. Search engine, Search results, Search privacy, Search privacy court, and Search services are the collocated Search words.

In 2018, we have an entirely different graph. Search is no longer a keyword. This, perhaps, suggests the de-emphasis on just the Google search engine, as their product offerings have expanded. Instead of search, we now have Services as a keyword. We also have dropped Share as a word in the graph; it seems to have been entirely replaced by use. Some new Google collocates are: Google account, Google delete, Services provide, Services collect, Information privacy, Information account, and Services collect, use, delete.

The final chart, displayed in Figure 10, comes from Voyant tools and is a charting of the five fair information practices and their inclusion in the privacy policies over the years. Though terms do not show the entire story, they do provide some measure of how organizations' privacy policies are easily mappable to Federal Trade Commission Fair Information Practices. The Y axis is shown as relative frequency, to adjust for the increasing word count over the years. In 2000, the FTC (2000) published a document entitled "Fair information practices" and suggested a voluntary standard for privacy policies, recommending what they should contain. The 5 areas to be included were: access, security, notice, choice, and enforcement. Over the years, the inclusion of these terms has varied widely for Google privacy policies. Notice, as a specific term was important in early years, but has dwindled significantly over the years. Security was initially not mentioned but has been moderately included all other years. Access was not noted until 2004 but rose strongly and has remained the most frequent term, in most years. Choice was first included in 2010 but has low mentions, in most years. Finally, enforcement, as a term, started in 2005 and was included until 2014, when it was dropped altogether.

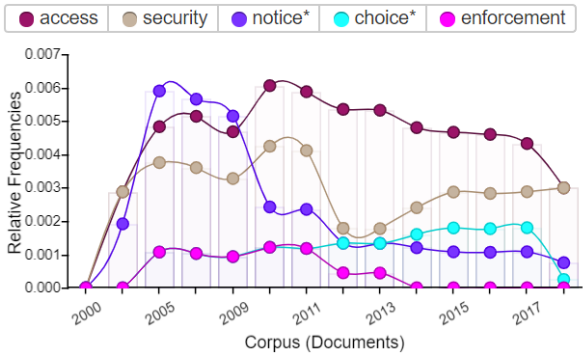

Figure 10. Fair Information Practices

5. SUMMARY AND CONCLUSIONS

In this paper, the authors have examined Google privacy policies from 2000 to 2018. We have performed a detailed review through a variety of qualitative and data visualization tools. The results can be used by students, faculty, practitioners, and researchers, to understand the evolution of Google privacy over time. The study can also be used as a basis for comparing other privacy policies. The study can serve as a model for comparing any written documents for similarities and differences.

Some of the key findings from this study include:

- The complexity of the Google policy has increased over time, due to a more than

sixfold increase in word count, as well as, the continuation of a 12th grade reading level.
- The Google privacy policy has evolved to a more personalized document, suggesting an attempt to make the document appear friendlier.
- The policy sentiment also improved over the years using more positive and enjoyment laden words; however, this is offset by a 50% decrease in authenticity score, which actually depicts a more legalistic and guarded content.
- There were no signficiant differences in tone over the period.
- Google has further obfuscated their privacy policy, by adding links to other pages. For example, the 2018 policy has 26 links to assist in controlling your privacy compared to 0, in 2000, and 11, in 2010.
- The word clouds have similarities between 2000 and 2018, but also have many key differences. Services supplanted search and use became a very frequenty used word. Cookies virtually were elimiated as a keyword, by 2018.
- Collocation network graphs show differences in keywords and words in proximity for 2000 and 2018. Some of these differences include importance of services in 2018 and the rise of use, delete, and account in 2018; along with the elimination of share as a significant collocated word from 2000 to 2018.
- Of the five fair information principles, access is the most frequently mentioned, followed by security, choice, and notice. Enforcement has been dropped from the Google privacy policy.

Overall there has been great change in Google privacy policies over the past two decades. The greatest takeaway for the authors is that, even though the policy gives the appearance and words associated with friendlines and positive sentiment, the policy is actually highly legalistic and allows for cross platform sharing and using of data for primarily any purpose. As noted by Kelly, the information we provide to Google allows the organization to track our day from beginning to end.

Although websites such as Google offer their services to users free-of-charge, these services actually do come at a cost to the user. This cost is realized in the collection of user data, which is utilized by advertisers and marketers, and other groups. Google's free services come at a high price for privacy. Additional research is proposed to further explore these polcies and their full impact on consumers, users, and society, at large. The authors plan to further pursue this fertile area with future studies.

6. REFERENCES

Back, M. D., Küfner, A. C., & Egloff, B. (2011). Automatic or the people? Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, *22*(6), 837-838.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, *28*(2), 15-21.

Clear Language Group (n.d.). Readability. Retrieved from http://www.clearlanguagegroup.com/readability/

Cordova, M. J., Cunningham, L. L., Carlson, C. R., & Andrykowski, M. A. (2001). Social constraints, cognitive processing, and adjustment to breast cancer. *Journal of consulting and clinical psychology*, *69*(4), 706.

De Groot, J. (2019, July 15). What is the General Data Protection Regulation? Understanding & Complying with GDPR Requirements in 2019 [Blog post]. Retrieved from https://digitalguardian.com/blog/what-gdpr-general-data-protection-regulation-understanding-and-complying-gdpr-data-protection

Federal Trade Commission. (2000). *Privacy Online: Fair Information Practices in the Electronic Marketplace, A Report to Congress*. Available at http://www.Federal Trade Commission.gov/reports/privacy2000/privacy2000.pdf

Google (2000 -2018) Google Privacy Policy.

Google (2019) Find your way around YouTube https://support.google.com/youtube/answer/2398242?hl=en&ref_topic=9257409

Graber, M.A., D'Alessandro, D.M., & Johnson-West, J. (2002). Reading level of privacy policies on Internet health Web sites. Journal of Family Practice, 51, 642-645.

Harris, K.D. (2014, May). Making your privacy practices public. Retrieved from https://oag.ca.gov/sites/all/files/agweb/pdfs/cybersecurity/making_your_privacy_practices_public.pdf

Hawkins, I. I., Raymond, C., & Boyd, R. L. (2017). Such stuff as dreams are made on: Dream

language, LIWC norms, and personality correlates. *Dreaming*, *27*(2), 102.

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C., große Deters, F., ... & Mehl, M. R. (2018). Linguistic Markers of Narcissism: An Exploratory LIWC Analysis of 15 Samples. *PsyArXiv. August 8*.

Hutchinson, Baird, and Garg (2016) Federal Trade Commission. (2000). *Privacy Online: Fair Information Practices in the Electronic Marketplace, A Report to Congress*. Available at http://www.Federal Trade Commission.gov/reports/privacy2000/privacy2000.pdf

Jensen, C., & Potts, C. (2004). Privacy policies as decision-making tools: An Evaluation of online privacy notices. CHI 2004, 6, 471-478.

Jones, J. (2018, April 11) Facebook Users' Privacy Concerns Up Since 2011. Retrieved from https://news.gallup.com/poll/232319/facebook-users-privacy-concerns-2011.aspx

IBM (2019) IBM Watson Natural Language Understanding https://www.ibm.com/watson/services/natural-language-understanding/

Internet Archive (2019) https://archive.org/web/

Kelly H. (2018, August 21). Google's data collection is hard to escape, study claims https://money.cnn.com/2018/08/21/technology/google-data-collection/index.html

Kennedy, N. (2008) Google processes over 20 petabytes of data per day https://www.niallkennedy.com/blog/2008/01/google-mapreduce-stats.html

Leroy, G., Helmreich, S., Cowie, J. R., Miller, T., & Zheng, W. (2008). Evaluating online health information: Beyond readability formulas. In *AMIA Annual Symposium Proceedings*(Vol. 2008, p. 394). American Medical Informatics Association.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Microsoft Word Readability Statistics (2019) Microsoft Corp.

Opsahl, K. (2010, April 28). Facebook's Eroding Privacy Policy: A Timeline. *Electronic Frontier Foundation*. Retrieved from https://www.eff.org/deeplinks/2010/04/facebook-timeline

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

Pennebaker Conglomerates (2015) *LIWC How It Works*. http://liwc.wpengine.com/how-it-works/

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015*). Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX : Pennebaker Conglomerates (www.LIWC. net).

Peslak, A. (2018). Facebook Fanatics: A Linguistic and Sentiment Analysis of the Most "Fanned" Facebook Pages. Journal of Information Systems Applied Research, 11(1) 23-33.

Peslak (2017). Sentiment Analysis and Opinion Mining: Current State of the Art and Review of Google and Yahoo Search Engines' Privacy Policies V10 N3 Pages 38-47 December 2017 Journal of Information Systems Applied Research

Peslak, A. R. (2016). Do Social Networking Sites Privacy Policies Differ? A Linguistic Analysis Of The Ten Most Popular Social Networking Sites *Issues in Information Systems* (17), 7.

Proctor, R.W., Ali, M.A., & Vu, K.L. (2008). Examining usability of web privacy policies. *International Journal of Human-Computer Interaction*, 24(3), 307-328.

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology*, 0261927X13476869.

Sinclair, Stéfan and Geoffrey Rockwell, 2016. *Voyant Tools*. Web. http://voyant-tools.org/

Statista (2019) Google - Statistics & Facts https://www.statista.com/topics/1001/google/

Turow, J., Hennessy, M., & Draper, N. (2018). Persistent misperceptions: Americans' misplaced confidence in privacy policies, 2003–2015. *Journal of Broadcasting & Electronic Media*, *62*(3), 461-478.

US Census (2018) National and State Population Estimates

https://www.census.gov/newsroom/press-kits/2018/pop-estimates-national-state.html

Voyant Tools (2019). http://www.voyant-tools.org

Warzel, C., & Ngu, A. (2019, July 10). Google's 4,000-word privacy policy is a secret history of the Internet. The *New York Times*. Retrieved from https://www.nytimes.com/interactive/2019/07/10/opinion/google-privacy-policy.html

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89-118). University of Chicago Press.

Editor's Note:

This paper was selected for inclusion in the journal as an CONISAR 2019 Distinguished Paper. The acceptance rate is typically 7% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2019.

Appendices

Appendix 1. Word Count, LIWC variables, Grade Level

Year	WC	Analytic	Authentic	Sentiment	Grade Level
2000	657	69.02	48.00	0.4	11.1
2004	1044	75.11	51.12	0.65	12.3
2005	1866	77.98	39.47	0.62	14.9
2008	1947	79.04	38.25	0.63	14.8
2009	2140	78.24	37.93	0.63	15
2010	1652	77.75	31.81	0.59	14.9
2011	1703	79.23	29.97	0.61	14.9
2012	2246	60.35	25.04	0.68	13
2013	2256	60.80	25.11	0.68	13
2014	2499	60.86	24.85	0.69	12.8
2015	2789	60.88	24.83	0.7	12.7
2016	2829	60.75	25.06	0.69	12.7
2017	2779	62.00	24.21	0.72	12.6
2018	4009	66.42	19.38	0.63	12.2