Exploratory Analysis Of The Readability Of Information Privacy Statement Of The Primary Social Networks

Musa J. Jafar, West Texas A&M University, USA
Amjad Abdullat, West Texas A&M University, USA

ABSTRACT

The goal of a privacy policy statement of a web site is to inform users of the policies and procedures of a web-site as it relates to their collection, use, sharing, access, security and use of technology as it relates to collection of data (cookies and web beacons) and disclosure of personally identifiable information when a user visits the web site.

In this paper we perform exploratory data analysis of the historical evolution of the readability as well as the reading grade level of the privacy policy statements of Google, Yahoo, Myspace and Facebook. We used the Flesch-Kinkaid, Gunning Fog and SMOG reading grade analysis measures. We gathered summary statistics of the complexity of each privacy statement (count of 3^+ syllables words, count of 6^+ characters words, count of 20^+ word sentences).

We conclude that (1) Except for Yahoo.com, these privacy policy statements are currently written for web-users with a minimum of 2 years of college education. This is not the case for most of social networks users. (2) Using Yahoo.com as a benchmark, privacy policy statements can accomplish their goals and maintain a reading grade level of high school education or less. Accordingly, social networks can accomplish their goal of providing clear and concise privacy policy statements without having to complicate the policy statements with too many 3+ syllable words, 6+ characters words and 20+ word sentences. In summary, it is possible to write a legally binding privacy policy statement that is also clear and easy to read.

Keywords: Information Privacy Policy, Social Networks, Google, Myspace, Facebook, Yahoo, readability, reading grade level

INTRODUCTION

In this paper we perform exploratory data analysis of the historical evolution of the readability as well as the reading grade level of the privacy policy statements of Google, Yahoo, Myspace and Facebook. We used the Flesch-Kinkaid, Gunning Fog and SMOG reading grade analysis measures. We gathered summary statistics of the complexity of each privacy statement (count of 3^+ syllables words, count of 6^+ characters words, count of 20^+ word sentences).

In 1999 Scott McNealy the CEO of Sun Microsystems (now a division of Oracle Corporation) told a group of reporters that "You have zero privacy anyway, Get over it", neither the FTC nor the privacy watchdog groups were happy with his comments then (Sprenger, 1999). Since then, the usage of the internet has transformed to broadband, we are authoring, posting and sharing digital content. Digital mobile device allows us to communicate on the go, web-based digital social networking is part of every day's life.

As of December, 2008 (Figure 1) broadband subscription has increased to approximately 80 Million (OECD Portal, 2009). It is in line with the PEW Internet & American Life Project which claims that approximately

55% of adult Americans have broadband internet connection at home (Horrigan, 2008). 35% of American adult users have a profile on an online social network site, 65% of online American teens use social networks (Lenhart, Amanda, 2009). Figure 2 (quantcast.com, 2009) outlines the monthly visits to top social network sites. The FTC issued self-regulatory principles for online advertising practices of the "Network advertisers" of online Behavioral advertising companies that track and share surfers' online browsing history (FTC Staff Report, 2009). Gramm-Leach-Bliley ACT (GLB) on privacy was enacted in 1999, Health Insurance Portability and Accountability Act (HIPPA)was enacted in 1998 and Children's Online Privacy Protection Act (COPPA) was enacted in 1998.

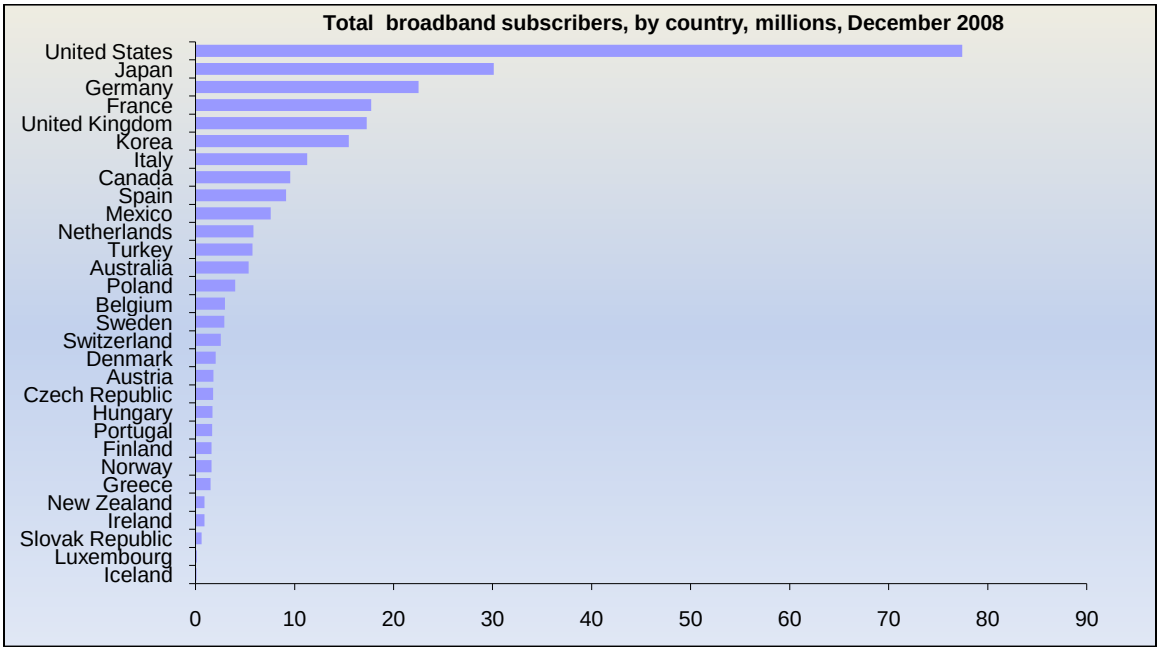

Figure 1: OECD Portal, 2009 Broadband Subscribers by Country

On top of that, 44.89% of the 18 years or over of the U.S. population attained a high school or less education (U.S. Census Bureau, 2009). With broadband penetration and its reach to all kind of wired and wireless devices, the ability of marketing to collect, share, analyze and cross-reference data on the fly, the regulation of the privacy of online information is at center stage. It is required to protect the privacy of individuals while allowing businesses to function properly. However, the main question remains whether the readability and the reading grade level of the privacy policy statements of social networks web sites that cater to the majority if not all of the high school or less education audience, is simple enough and clear enough for them to understand its content and comprehend it. In other words, Is it possible for a social network to provide a legally binding and comprehendible statement where the majority of the social networkers can understand?

INFORMATION PRIVACY POLICIES

Privacy is usually defined as the right to be left alone. This definition does not differentiate between the different clusters of privacy and the legal protections provided for each one of them. (Samuelson, 2008) Defines four arising clusters of privacy as spatial privacy, electronic communication privacy, individual information held electronically by a third party privacy and anonymity in public places privacy (surveillance cameras, Google street maps, etc). This paper is concerned with the third cluster (individual Information privacy). It is about the privacy of information electronically held by third parties such as social networks, ISP(s), search engines, etc. about us individuals that we do not think should be made public.

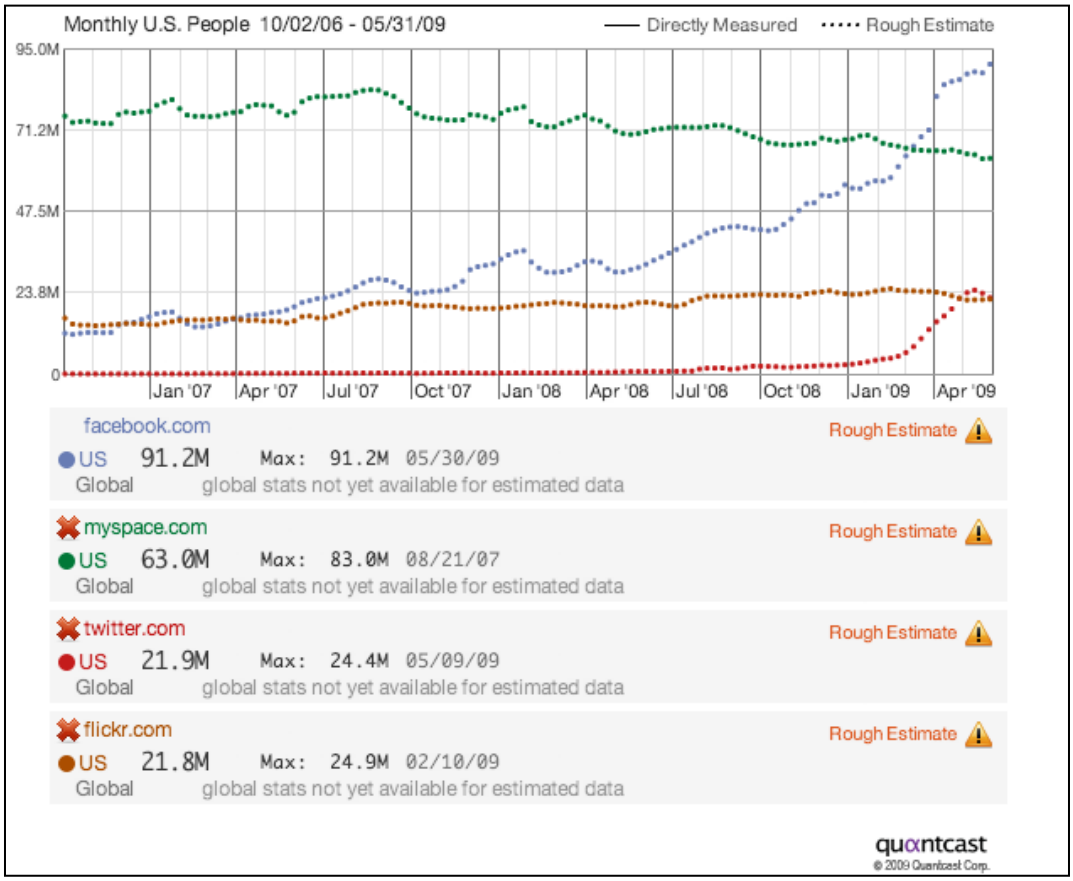

Figure 2: Monthly U.S. Visits Chart

The Fourth amendment protected spatial privacy. In the 1960s, the courts extended its protection to include wired communication. Wiretap laws protected doctor-patient privileges, attorney-client privileges. Probable cause court order law protected individual library records. Medical records are protected under HIPPA. Online privacy of children is protected under COPPA. However there are no comprehensive European Union style guidelines or laws that protect the overall information privacy of individual records held by third parties such as search engine records, social networks records, email records, browsing history records, etc. (soft individual information privacy records).

Historically, information privacy of individuals was not an issue that requires serious legal consideration or attention. Records were either not electronically stored or stored in sporadic and disconnected electronic databases that are owned by different companies. Until recent advancements in digital technologies, the internet, the World Wide Web, and its penetration into our daily life, it was either too costly or prohibitive for private entities to gather information about individuals. Until recently, the way information was gathered and stored, made it hard for it to be cross-referenced, searched, shared or widely utilized. However, current advancement in information technology and the usage of it, those barriers no longer exist. Technology provided opportunities for marketing to reach a larger audience, gather personally identifiable information, surfing session information and session history. Marketing companies are using sophisticated techniques, exploits and loopholes in the underlying communication protocols (HTTP, TCP/IP, POP3, IMAP, etc.) and the governing laws (State and Federal) to gather information about individuals that can be used to marketing products and to push advertisements and promotions.

A recognized Privacy and Personal Information Protection Act was introduced in 1998 in Australia, It defines information privacy as "The way in which governments and organizations handle our personal information such as our age, address, sexual preferences and son on". It gives individuals the right to exercise control over one's

personal information. Accordingly, the information privacy of an individual is violated when electronic personal information that was entrusted to third parties is electronically shared or cross-referenced with other parties without the consent of the individual. If a digital folder of an individual's personal data is created, shared and traded without the consent of that individual and without the ability of that individual to view and correct their information then their privacy has been violated.

According to the U.S. Department of Justice's Global Justice Information Sharing Initiative (Global), "A privacy policy is a written, published statement that articulates the policy position of an organization on how it handles the personally identifiable information that it gathers and uses in the normal course of business. The policy should include information relating to the process of information collection, analysis, maintenance, dissemination, access, expungement, and disposition" (Justice, 2008). So, the goal of a privacy policy statement of a web site is to inform users of the policies and procedures of a web-site as it relates to their collection, use, sharing, access, security, use of technology as it relates to collection of data (cookies and web beacons) and disclosure of personally identifiable information when a user visits the web site.

RELATED WORK

(Lewis, Colvard, & Adams, 2007) used Microsoft's word Flesh-Kincaid grade level testing tools to analyze the privacy statements of banks, credit card counseling and check cashing companies. They concluded that the reading grade level ranges between 10^{th} grade and 20^{th} grade indicating that the Gramm-Leach-Bliley Financial Services Modernization Act of 1999 had no impact on the wording of information privacy statements of financial institutions. (Proctor, Ali, & Vu, 2008) Examined users' comprehension of privacy policies for 100 different web sites in seven categories (Financials, Insurance, retail, technology, etc.) focusing on personally identifiable information, content, goal mining and to some extent readability. They concluded that college students have poor comprehension of privacy policy statements. They also analyzed the amount of personally identifiable information that those sites request from customers. They concluded that at least 71% of the web-sites in the study requested personally identifiable information such as email address, date of birth, first and last name, and postal address. They concluded that for the websites under test the reading grade level of the audience should have been 13-15 years of schooling, i.e. college level. Cadogan et al. (Cadogan, 2004) analyzed the friendliness and the reading grade level of a random selection of 3 companies (Amazaon.com, Dell.com and privacyalliance.com). Since then the reading grade level of Privacyalliance.com statement went up by .8 to 11.8, the Dell.com statement went up by 4.5 (4.5 more years of education) to 16.5 and the Amazon.com stayed at 18 which is post graduate level of education. Sheng et al. (Sheng & Cranor, 2006) went through an elaborative effort of gathering privacy policies and other information privacy statements of fifty companies in the US financial industry for the years between 1999 and 2005. They analyzed privacy policies from a compliance perspective. They concluded that legislation had modest impact on the privacy policy of financial institutions, the reading grade level using the Flesh-Kincaid score average to a freshman in college level.

DEMOGRAPHICS OF SOCIAL NETWORKS

As Figure 3, 4 and 5 indicate, millions of individuals visit and use web-based social networks on a daily basis, 47% of social networks visitors have no college degree, 20.2% are 12-17 years old, 40.8% are 18-34 years old, and another 21.8% are 35-49 years old. The data in the figures were compiled from statistics gathered from (quantcast.com, 2009).

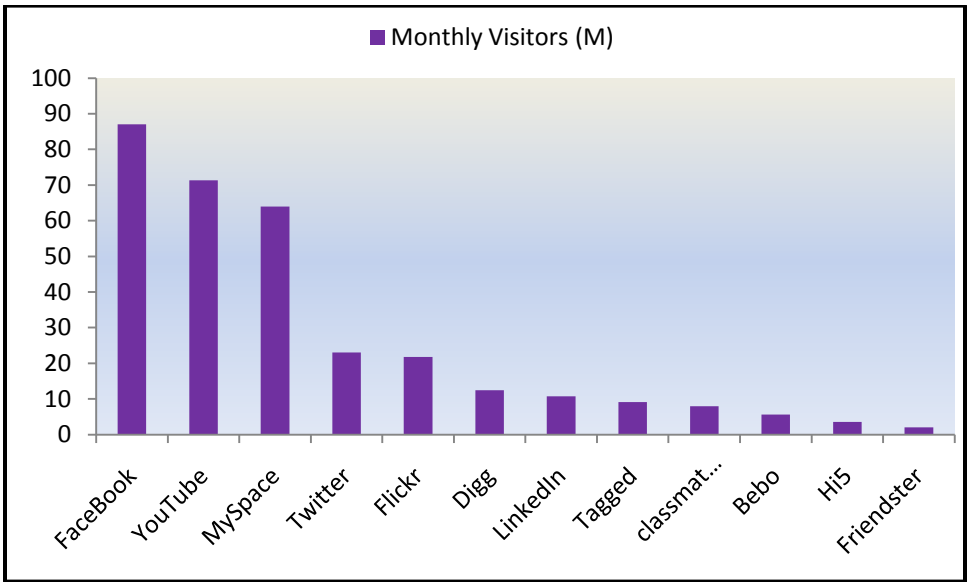

Figure 3: Visitors Count (in Millions) of Social Networks Sites

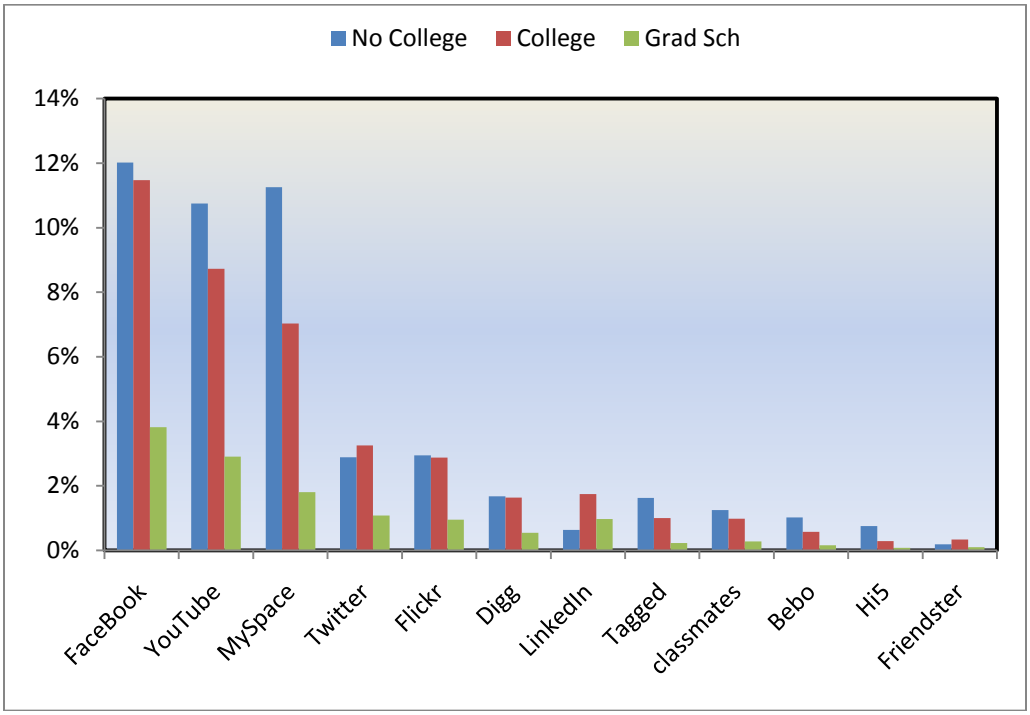

Figure 4: Weighted Education Demographics of Social Networks Sites

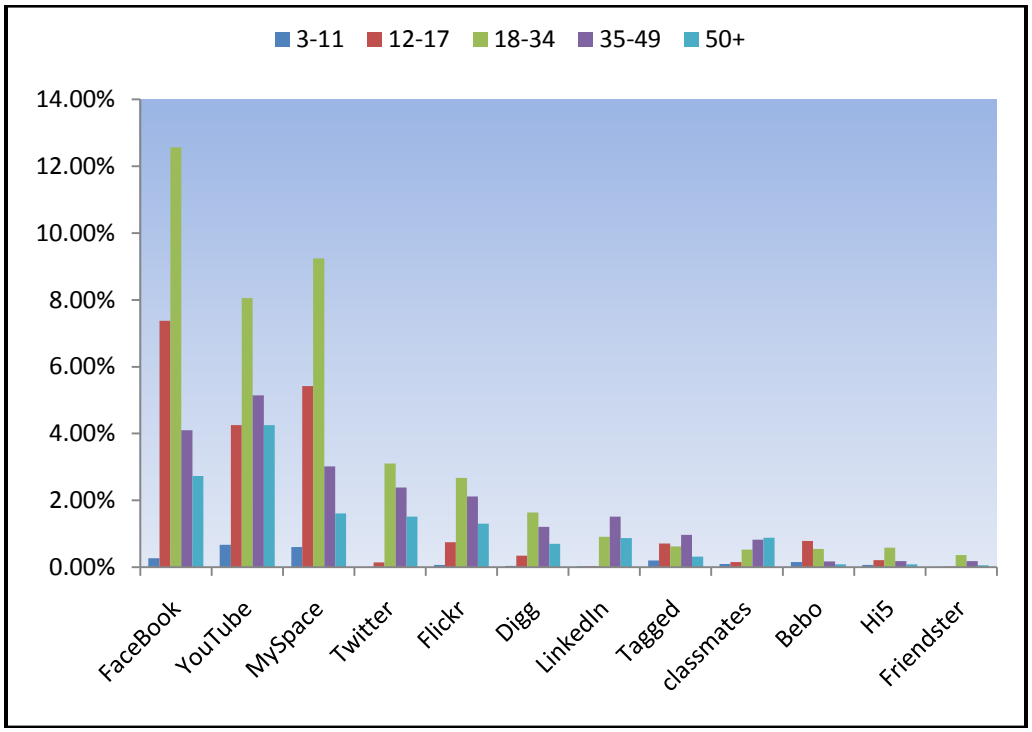

Figure 5: Weighted (% Within a Social Network) Age Demographics of Social Networks Sites

READING GRADE LEVEL MEASURES

A readability formula is a heuristic framework for calculating the minimum reading grade level of the targeted audience of a document for that audience to comprehend the content of the document. They basically use statistics of syllable counts per word, character counts per word and word count per sentence, sentence count, etc. to calculate a reading grade level of the underlying document. The validity, sensitivity and the correctness of these measures is beyond the scope of this paper. However it is worth mentioning that these measures are extensively used in the English language literature and software tools (public domain and for purchase) are built around these measures (Flesh-Kincaid, Gunning Fog, SMOG, Dale-Chall, Spache, Fry Graphs, etc.) to analyze the readability and the reading grade level of a document. The following is a summary of three measures that are used in this paper.

- The Flesh-Kincaid readability formula (Kincaid, Rogers, & Chissom, 1975) is calculated as:

$$\boldsymbol{G = (11.8 * B) + (0.39 * S) - 15.59}$$ where

  - G is the U.S. grade level
  - B is the Average number of syllables in a word in the document
  - S is the average length of a sentence in the document

- Gunning Fog readability formula (Gunning, 1952) is calculated as:

$$\boldsymbol{G = 0.4\left(\frac{number(words)}{number(sentences)} + 100\left(\frac{number(words\ with\ 3^{+}syllables)}{number(words)}\right)\right)}$$

- The SMOG readability algorithm is detailed in (Mc Laughlin, 1969) as:

$$G = 3 + \sqrt{count(3^{+}syllables\ word\ in\ three 10 sentences\ in\ a\ row\ of\ the\ document)}$$

One observation worth mentioning is that these measures are pre-intensive electronic computing processing (1975, 1969, 1952) era where it is hard to manually analyze counts (syllable, characters, words, sentences and phrases) in a document. Computing power was either not available, too expensive or even the know-how knowledge of programming it to perform these tasks was not available. Another observation, instead of coming up with a crisp measure of readability (G) as in the three formulas that we presented, a fuzzy or statistical approach based on the distribution of the variables in the formula is more expressive. For example, instead of concluding that the reading grade level of a document is **G**= 15.6, using the average of number of syllables and the average length of sentences in the Flesh-Kincaid case; the distribution of the 3^{+} syllables words combined with the distribution of 20^{+} word sentences could have been used instead. The G value would have been calculated as a distribution instead. The same could have been said about the SMOG and the Gunning Fog measures.

To track the evolution of a privacy policy, and to analyze differences between documents, we used KDiff3, an Open Source software package freely available from Sourceforge.org. To analyze the reading grade level, word counts, sentence counts, paragraph counts, long sentences count, syllable count per word, character-count per word etc. we used Readability Studio 2008 from Oleander.com.

ANALYSIS OF THE HISTORICAL EVOLUTION OF PRIVACY POLICY STATEMENTS

In the following sections, we analyze the evolution of the privacy policy statement of Google.com, Yahoo.com, facebook.com, and myspace.com. We analyze the reading grade level and the complexity of the historical statements within each of the four networks (3^{+} syllable counts of words, 6^{+} characters count of word and 20^{+} word counts of sentences). We also analyze the reading grade level and the complexity of the historical statements across the four networks.

According to (Sheng & Cranor, 2006) "In the US, there is no governmental entity responsible for collecting or maintaining an archive" of the privacy disclosures that they are required to make. We were able to gather the archived privacy statements from two sources: (1) Google.com provides a link to their archived privacy policy statements. (2) The WayBackMachine project (web.archive.org) provides archives of web sites way back to the late 1990(s). We searched the archives comparing the privacy policy statements of a given web-site using the KDiff3 Open Source software. We found that Google.com went through five, Yahoo went through five, Facebook went through eight, and Myspace went through 6 privacy policy statement updates. We analyzed each of the privacy statements using Readability Studio 2008 software package. It is an elaborate software package that provides in-depth analysis of text documents.

For Google.com we will perform an elaborate and detailed analysis of the historical privacy policy statements. In order for us not to repeat ourselves, for the rest of the networks (yahoo.com, facebook.com and myspace.com) we provide summary analysis.

HISTORICAL EVOLUTION OF THE PRIVACY POLICY OF GOOGLE.COM

Google's current privacy policy statement was last updated on March-11-2009. On the privacy statement web page, Google also provide links to their Oct-14-2005, Jul-1-2004 and Aug-14-2000. For our records, we downloaded all of the privacy statements. For each of the five statements, we used Readability Studio 2008 to gather summary statistics (frequency of syllable counts per word, frequency of character counts per word, frequency of sentence counts per paragraph, frequency of word counts per sentence). Using readability studio 2008 too, we calculated the reading grade level of each statement. Readability Studio allows us to choose up to nineteen reading grade level tests. We chose the 3 most common ones (Flesh-Kincaid, Gunning Fog and SMOG). According to Reading Grade level scoring measures, long sentences, 3 or more syllable words, 6 or more character words and 20 or more word sentences contribute to the complexity of the document. Each grade level test uses its own heuristic to calculate the measure and the corresponding age of the person assuming that they are continuing their education. In the following sections, we present figures and charts and analyze the gathered data Summary statistics and Reading grade levels.

Summary Statistics

From Tables 1, 2 and 3 and Figures 6, 7 and 8, it is clear that as the policy statement of Google evolved over time, the complexity of the statement grew until it stabilized around year 2005. As of 2005, approximately 63% of the sentences have twenty words or more and 43.8% of the words are six characters or more (Figure 6 and Table 1). These measures contributed heavily to the complexity of the statement.]

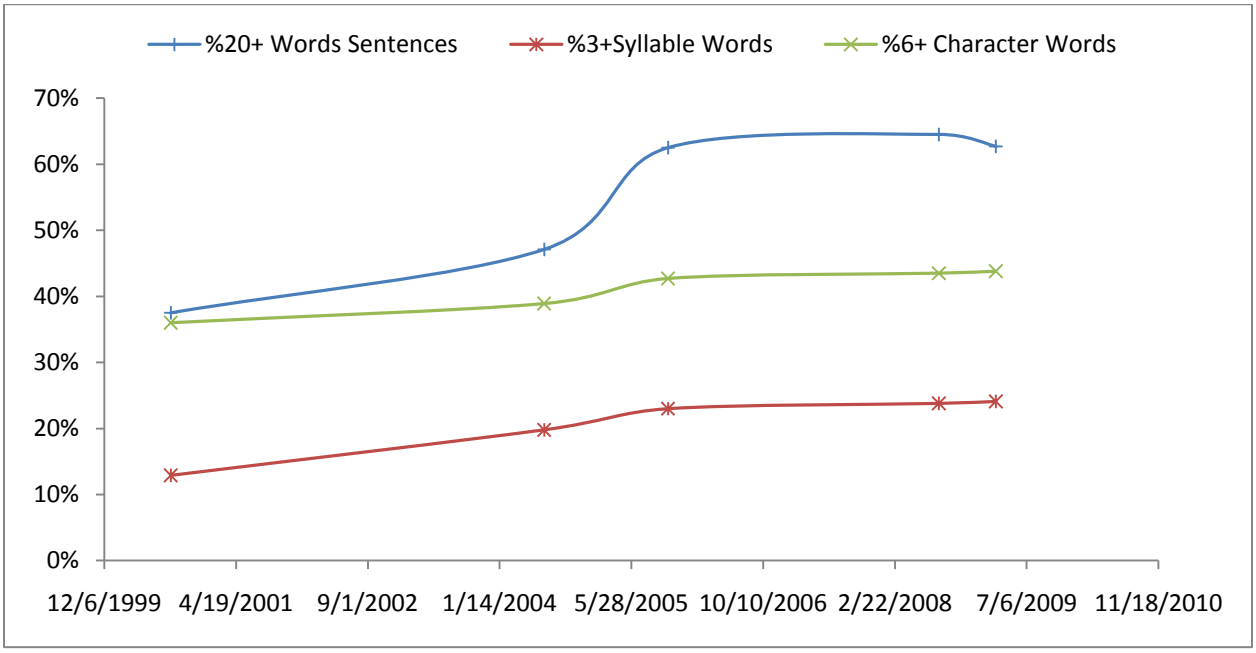

Figure 6: Evolution of the Complexity of Google Privacy Policy Statements

Table 1: Summary Statistics of Google Privacy Policy Statements

Statement Updates	Size (KB)	Word Count	Sentence Count	Paragraph Count	%20$^+$Word Sentences	%3$^+$Syllable Words	%6$^+$Character Words
8/14/2000	4	650	32	11	37.5	12.9	36
7/01/2004	7	1,036	51	23	47.1	19.8	38.9
10/14/2005	12	1,813	72	33	**62.5**	23.0	**42.7**
8/7/2008	12	1,897	76	35	**64.5**	23.8	**43.5**
3/11/2009	14	2,089	83	37	**62.7**	24.1	**43.8**

Table 2 is a count of the frequency of 20+ words for the various privacy policy statements of Google. For example, Google's privacy policy statement dated 8/14/2000 had three sentences with 27 words each. The tables and figures summarize the historical growth of the 20$^+$ word sentences. In 2000, long sentences ranged from 23 to 42 words in length for a total of 12 sentences or 37.5% of the total number of sentences in the document (Table 2). In 2009, long sentences ranged from 21 to 78 words per sentence for a total of 52 sentences or 62.7% of the total number of sentences in the statement. As the percentage of long sentences grew so did the complexity of the statements and the reading grade level of the privacy policy statement.

The Graph in Figure 7 is a plot of the data in Table 2. It provides a scatter plot of the frequency of the word counts for each privacy statement. For the 2000 policy, we connected the points with a smooth line for a better visualization of the range and distribution.

Table 2: Frequency of 20+ Word-Sentences of Google's Privacy Policy Statements

8/14/00	Count	7/1/04	count	10/14/05	Count	8/7/08	Count	3/11/09	Count
23	1	**21**	3	**21**	3	**21**	5	**21**	5
24	1	**22**	4	**22**	3	**22**	2	**22**	2
25	1	**23**	4	**23**	2	**23**	3	**23**	3
26	1	**25**	1	**24**	2	**24**	1	**25**	4
27	3	**26**	2	**25**	3	**25**	5	**26**	4
28	1	**27**	1	**26**	2	**26**	3	**27**	3
33	1	**28**	2	**27**	4	**27**	3	**28**	4
36	1	**29**	1	**28**	5	**28**	5	**29**	1
37	1	**30**	1	**30**	2	**29**	1	**30**	3
42	1	**34**	2	**32**	4	**30**	1	**32**	4
		36	1	**33**	2	**32**	4	**33**	1
		39	2	**35**	3	**33**	1	**34**	2
				36	3	**34**	2	**35**	5
				37	1	**35**	3	**36**	4
				39	2	**36**	4	**37**	1
				43	1	**37**	2	**45**	2
				58	1	**45**	1	**49**	1
				65	1	**58**	1	**58**	1
				79	1	**65**	1	**65**	1
						79	1	**78**	1

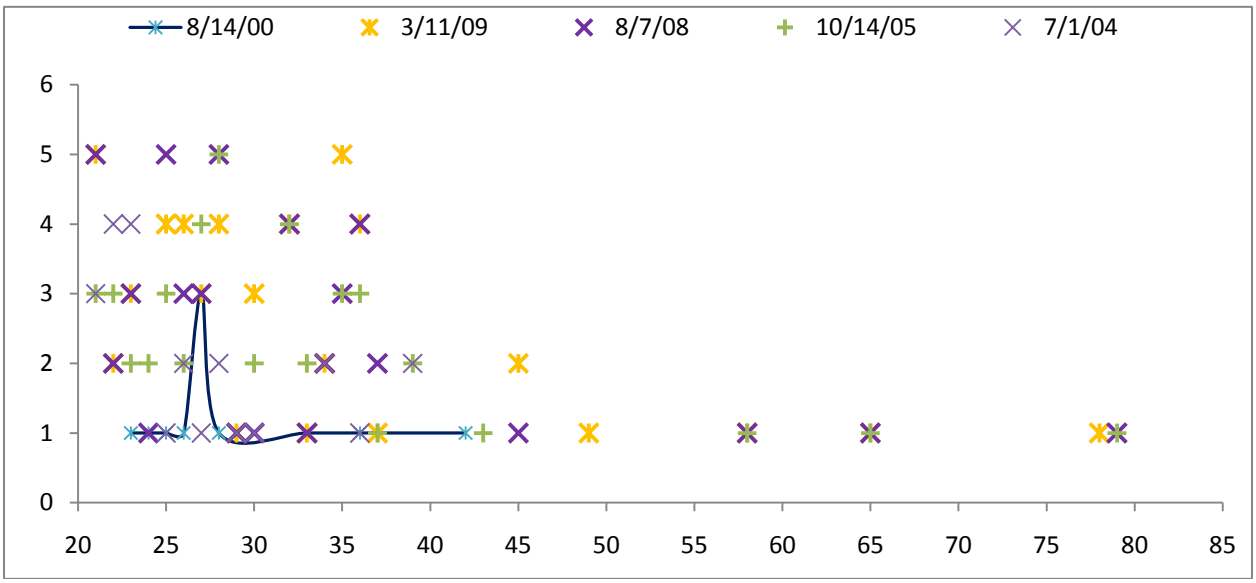

Figure 7: Plots of the Frequency of 20+ Word Sentences

To plot a histogram of the data, we grouped the word counts in Table 2 into intervals starting at 20 and ending at 80 with a range of 5 words for a total of 9 intervals. Table 3 is a grouping of the count data into intervals with the numbers in the cells corresponding to the count of the number of sentences in the interval range for the corresponding policy statement. Figure 8 is the histogram of the data in Table 3.

Table 3: Grouping of the Frequency of 20+ Word Sentences

Word Count Range	8/14/2000	7/1/2004	10/14/2005	8/7/2008	3/11/2009
20-24	2	11	10	11	10
25-29	6	7	14	17	16
30-34	1	3	8	8	10
35-39	2	3	9	9	10
40-44	1		1		
45-49				1	3
55-59			1	1	1
65-69			1	1	1
75-80			1	1	1
Long Sentences	**12**	**24**	**45**	**49**	**52**

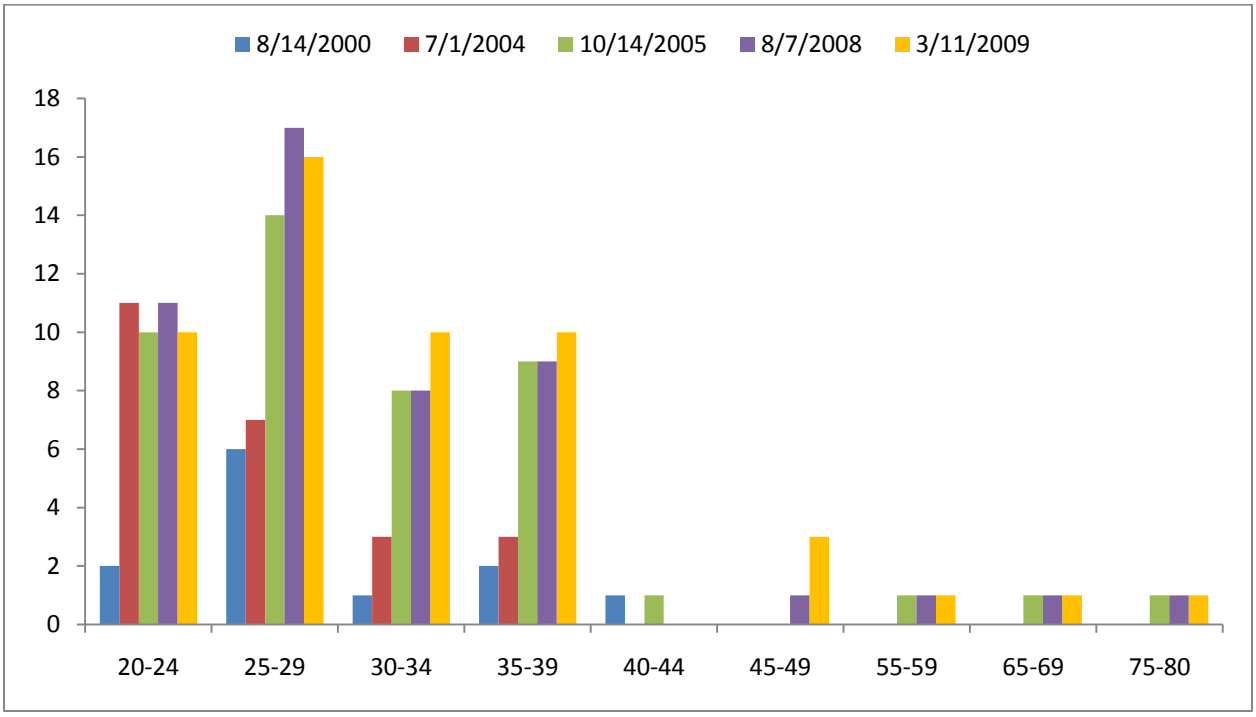

Figure 8: Histogram of the 20+ Word Sentences

Reading Grade Level Analysis

We analyzed the reading grade level of each privacy statement. We selected to use the Flesh-Kincaid, Gunning Fog and the SMOG readability measures. Table 4 provides the Flesch-Kincaid, Gunning Fog and SMOG numeric measures of each privacy policy statement (sorted by date). Figure 9 is a plot of the tabulated data. It is clear that as Google's privacy statement evolved through time, the reading grade level went up from 11.5, which is a high school level to 15.8, which is almost a college graduate, alternatively a postgraduate according to Gunning Fog and SMOG measures.

Table 4: Reading Grade Level of Google Privacy Statements

Google	Flesch-Kincaid	Gunning Fog	SMOG
8/14/2000	11.5	12.6	12
7/1/2004	12.9	15.5	14
10/14/2005	15.5	18.4	16
8/7/2008	15.6	18.1	16
3/11/2009	15.8	18.3	16

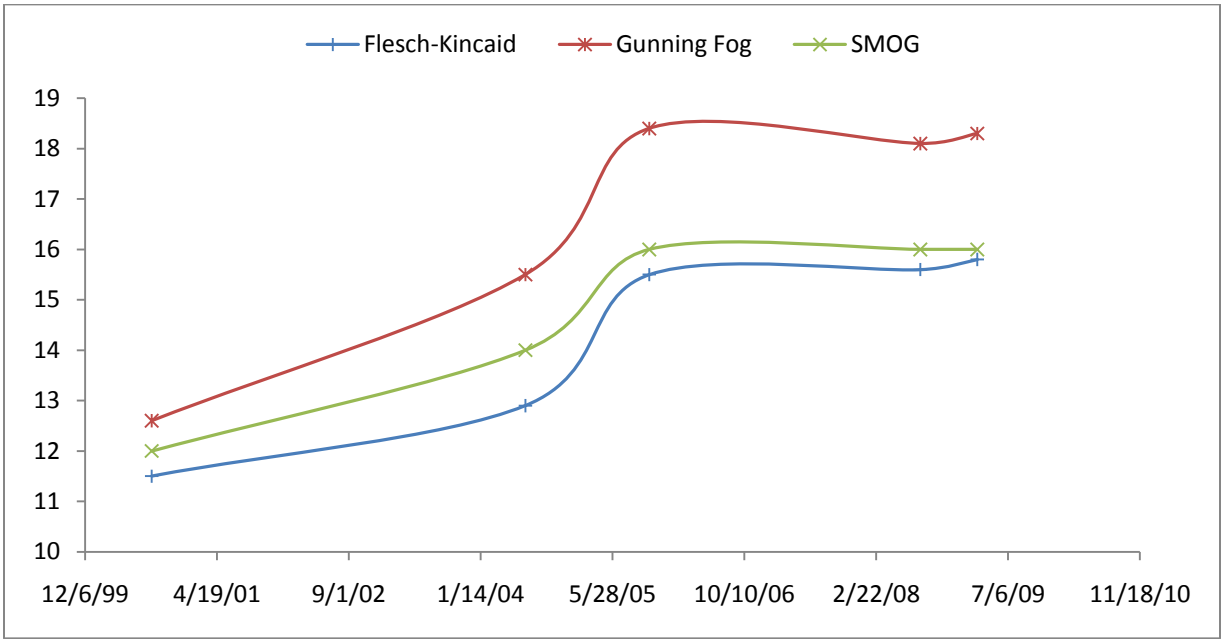

Figure 9: Reading Grade Level Measures of Google Privacy Statements

Summary

With time Google privacy policy statement got longer, which is understandable. However, there is no reason for the sentences to get longer. Reading grade level measures do not factor in the length of the document into their metric. However, they do factor in the complexity of the document. Accordingly as the privacy statement of Google evolved, it got more complex until it stabilized in 2005 to a reading grade level of 3rd year in college or more.

HISTORICAL EVOLUTION OF THE PRIVACY POLICY OF FACEBOOK

Facebook's current privacy policy statement was last been updated on November 26 2008. We used the WayBackMachine project (web.archive.org) website to gather Facebook's privacy statements. The archives collected Facebook websites starting on December-12-1998 until march-26-2008. The archives extensive coverage of Facebook was between 2005 and 2007. In 2005, 147 web pages were archived, **in 2006, 368 pages were archived**, in 2007 105 pages. Using KDiff3 software package to detect differences between the posted privacy statements, we searched the archives for the different privacy statements, tracking back from the current privacy statement. We found that the privacy policy statement of Facebook went through seven updates for a total of eight different privacy policy statements. Similar to Table 1, Table 5 provides a summary of each of Facebook's privacy statements. Figure 10 is a plot of the parameters that contribute to the complexity of the statements. In the following sections, we provide summary analysis similar to our analysis of Google.

Summary Statistics

The size of the privacy policy statement of Facebook tripled since 2005 (word count and size). Yet another proof that there is no correlation between the size of a document and its complexity. The Syllable count stayed the same, the long sentences count went up by 6%, and the long words count went up by 12%.

Table 5, Table 6 and Figure 10 show that the complexity of Facebook statement has been stable since its inception with some skewness to the left, all of the observations are within two standard deviations or less from the mean. It is worth noting that all Facebook's statements are post regulations. Contrary to Google's, None of Facebook's privacy statements are before 2005.

Table 5: Summary Statistics of Facebook Privacy Statements

Statement Updates	Size (KB)	Word Count	Sentence Count	Paragraph Count	%20+Word Sentences	%3+Syllable Words	%6+Character Words
6/28/2005	7	1,163	56	23	50.0%	19.5%	34.3%
2/27/2006	15	2,297	101	36	53.5%	20.9%	37.6%
5/22/2006	17	2,689	120	44	52.5%	20.6%	37.3%
9/5/2006	18	2,818	124	45	53.2%	20.7%	37.5%
10/23/2006	19	3,049	135	47	52.6%	20.1%	37.6%
5/24/2007	22	3,493	150	50	54.0%	20.5%	38.6%
9/12/2007	22	3,504	150	50	54.0%	20.5%	38.7%
11/26/2008	23	3,636	156	52	53.2%	19.7%	37.5%

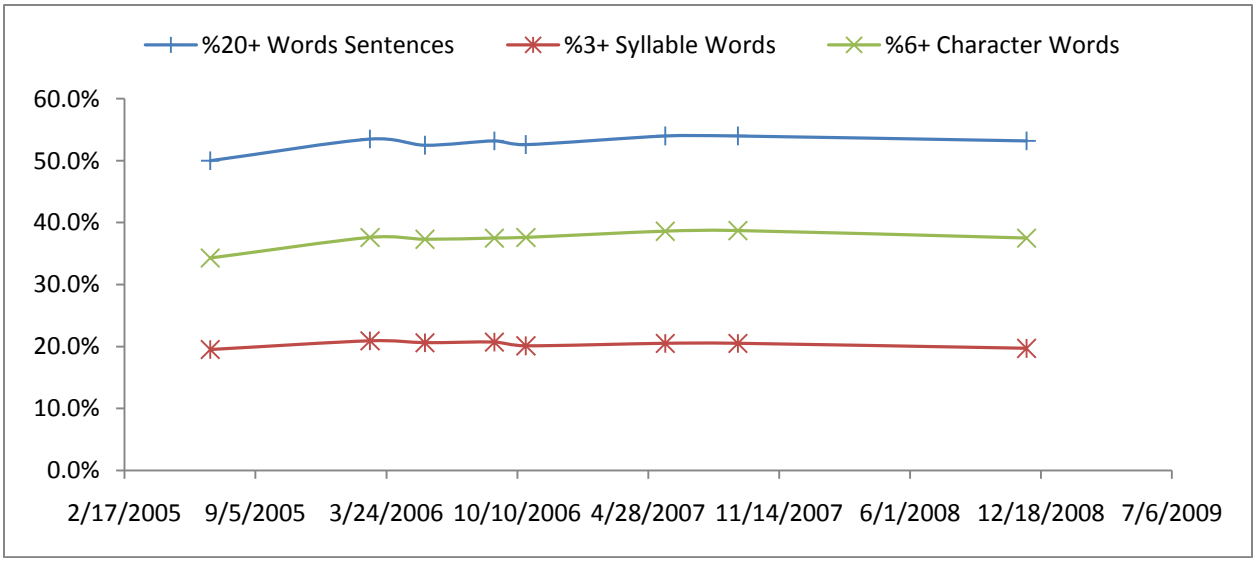

Figure 10: Complexities of Facebook's Privacy Policies

Table 6: Summary Statistics of the Complexities of Facebook's Privacy Statements

Summary Statistics	%20+Word Sentences	%3+Syllable Words	%6+Character Words
Mean	52.88%	20.31%	37.39%
Median	53.20%	20.50%	37.55%
Stdev	1.29%	0.50%	1.35%
Kurtosis	4.031	-0.798	4.869
Skewness	-1.84	-0.741	-1.942

Reading Grade Level Analysis

Table 7 and the graph in Figure 11 show that except for the first privacy statement, Facebook' reading grade levels of the historical privacy policy statements have been stable. On the Flesh-Kincaid scale, they average second year in college, on the Gunning Fog scale they averaged fourth year in college and on the SMOG scale they averaged third year in college.

Table 7: Reading Grade Level of Facebook's Privacy Statements

Reading Grade Level	Flesch-Kincaid	Gunning Fog	SMOG
6/28/2005	12.4	15.2	14
2/27/2006	13.7	15.9	15
5/22/2006	13.6	15.8	15
9/5/2006	13.7	15.9	15
10/23/2006	13.5	15.6	15
5/24/2007	14.1	15.9	15
9/12/2007	14.1	16	15
11/26/2008	13.7	15.8	15
Mean	**13.60**	**15.76**	**14.88**
Standard Deviation	**0.53**	**0.26**	**0.35**

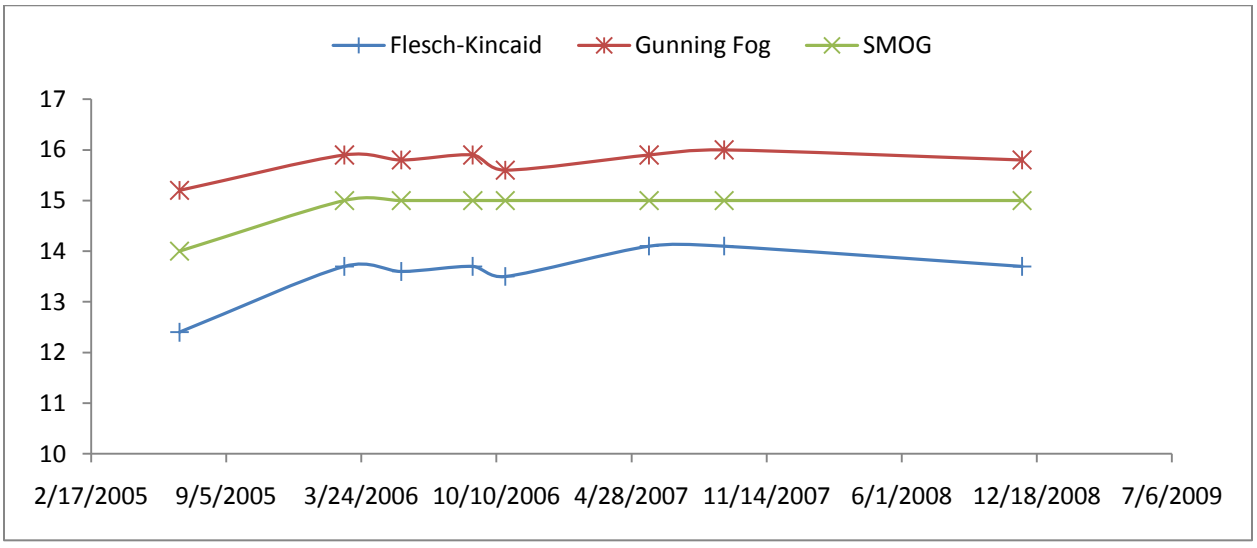

Figure 11: Reading Grade Level Measures of Facebook's Privacy Statements

HISTORICAL EVOLUTION OF THE PRIVACY POLICY OF MYSPACE

Similar to Facebook, Myspace current privacy policy statement was last been updated on February 28 2008. We used the WayBackMachine project (web.archive.org) website to gather Myspace's privacy statements. The archives collected Facebook websites starting on January-12-1997 until April-1-2008. The archives extensive coverage of Facebook was on 2001 and between 2004 and 2008. We searched the pages for the different privacy statements. We found that the privacy policy statement of Myspace went through five updates for a total of six different privacy policy statements. In the following sections, we provide summary analysis similar to our analysis of Google and Facebook.

Summary Statistics

Similar to Tables 1 and 5, Table 8 provides a summary of each of Myspace's privacy statements. Figure 12 is a plot of the parameters that contribute to the complexity of the statements. Judging from what learned from Google's and Facebook's analysis, The 2004 and beyond are the important changes that we should pay attention to in a privacy policy statement. Accordingly, Myspace's privacy policy statement has been stable. However its complexity is high. 63% of its sentences are long which is10 percentage points more than that of Facebook and inline with that of Google. 20% of its words are 3^+ syllables and 41% of its words are 6^+ characters. Accordingly it has high reading grade level too.

Table 8: Summary Statistics of Myspace's Privacy Statements

Statement Updates	Size (KB)	Word Count	Sentence Count	Paragraph Count	%20$^+$Word Sentences	%3$^+$Syllable Words	%6$^+$Character Words
9/21/2001	3	431	25	8	36.0%	16.7%	34.1%
10/20/2003	7	1,027	44	16	56.8%	18.8%	39.1%
3/5/2004	11	1,604	66	24	59.1%	19.8%	40.5%
2/25/2005	**11**	**1,659**	**69**	**24**	59.4%	20.2%	40.7%
8/26/2005	**11**	**1,524**	**50**	**19**	66.0%	19.2%	40.1%
2/28/2008	15	2,231	84	34	63.1%	19.9%	40.9%

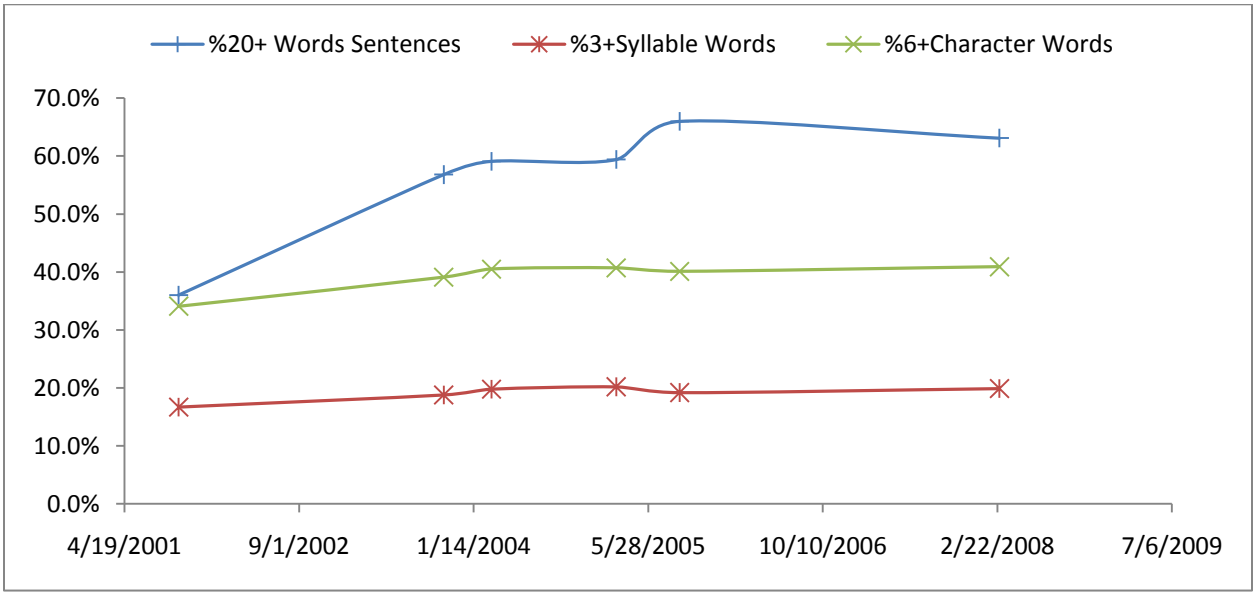

Figure 12: Complexity of Myspace's Privacy Policies

Reading Grade Level Analysis

Table 9 and the graph in Figure 13 show that the privacy statement of Myspace has been fluctuating. Although the size of the policy has increased from 2005 to 2008, the reading grade level went down by two grade levels according to the Flesh-Kincaid and the Gunning Fog measures.

Table 9: Reading Grade Level of Myspace's Privacy Statements

Myspace	Flesch-Kincaid	Gunning Fog	SMOG
9/21/2001	10.6	13.2	12
10/20/2003	14.3	15.5	14
3/5/2004	15	16.4	15
2/25/2005	15.1	16.4	15
8/26/2005	17.3	18.5	16
2/28/2008	15.4	16.6	16

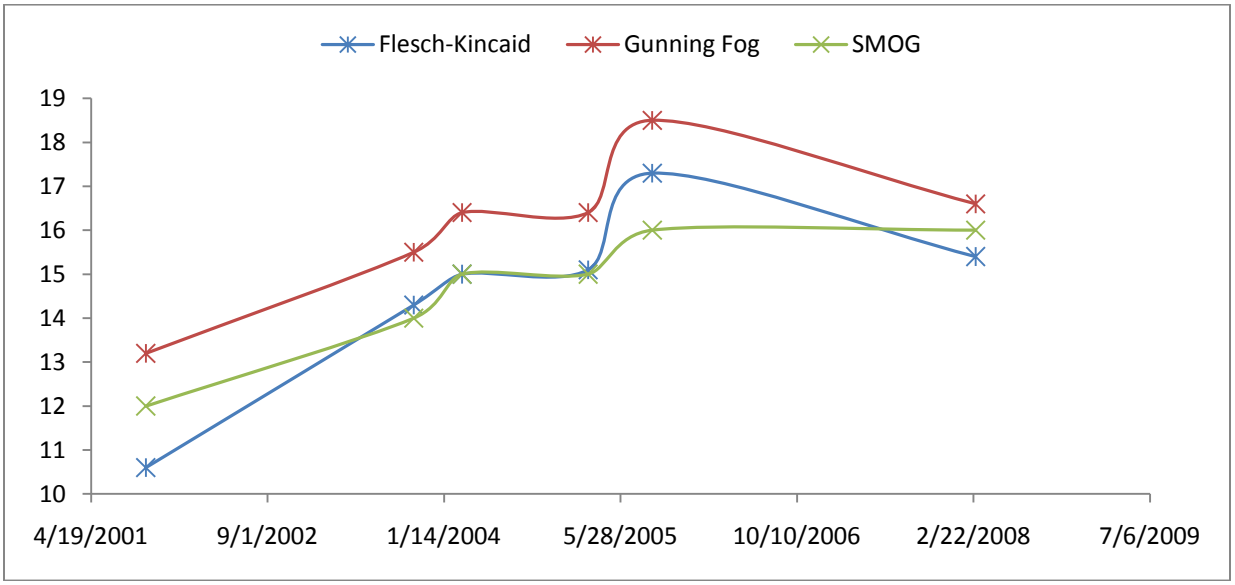

Figure 13: Reading Grade Level Measures of Myspace's Privacy Statements

YAHOO

Yahoo's current privacy policy statement was last updated on November 22 2006 (three years ago). We used the WayBackMachine project (web.archive.org) website to gather Yahoo's privacy statements. The archives started collecting Yahoo websites on October 17 1996 until April-1-2008. The archives extensive coverage of Yahoo was between 2000 and 2008. In the following sections, we provide summary analysis similar to our analysis of that of Google, Facebook and Myspace.

Summary Statistics

Similar to Tables 1, 5 and 8, Table 10 provides a summary of each of Yahoo's privacy statements. Figure 14 is a plot of the parameters that contribute to the complexity of the statements. Contrary to the other privacy policy statements, Yahoo maintained a privacy policy statement that is small in size with low complexity and low reading grade level. It is a proof that a large company with many related entities can provide and maintain a legally binding privacy policy statement that is neither complex, large in size nor high reading grade level.

Table 10: Summary Statistics of Yahoo's Privacy Statements

Statement Updates	Size (KB)	Word Count	Sentence Count	Paragraph Count	%20+Word Sentences	%3+Syllable Words	%6+Character Words
6/30/1998	10	1,522	137	47	18.2	19.0	34.3
3/28/2002	10	1,274	110	38	17.3	22.1	38.6
1/22/2003	9	1,202	108	38	15.7	21.8	38.6
1/1/2004	9	1,235	107	37	17.8	21.9	38.4
11/22/2006	9	1,346	114	38	17.5	21.8	38.8

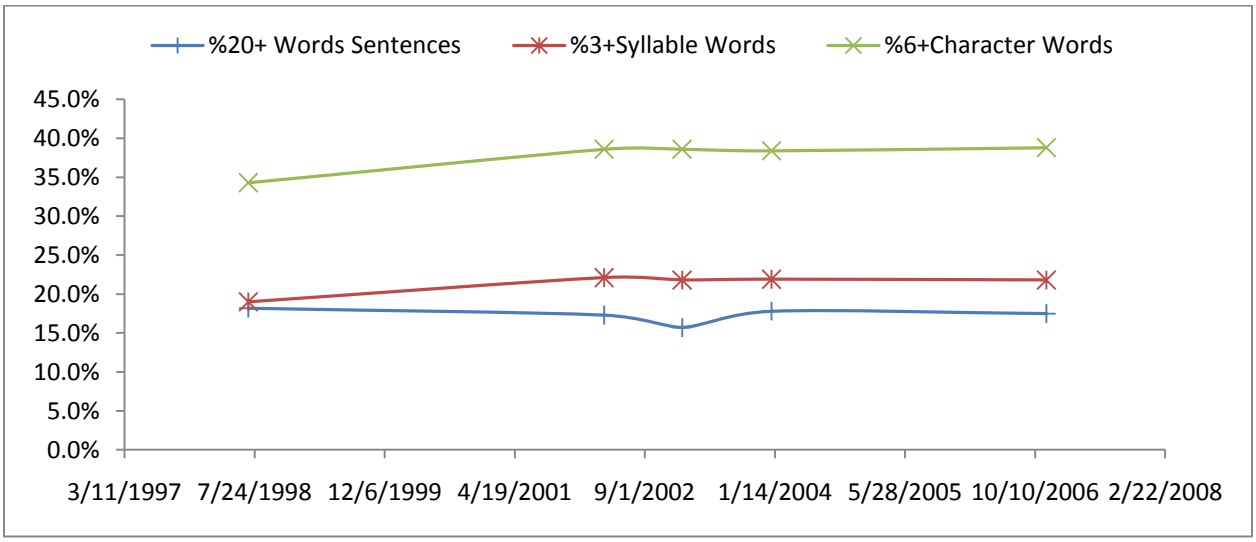

Figure 14: Complexities of Yahoo's Privacy Policies

Reading Grade Level Analysis

Table 11 and the graph in Figure 15 show that the privacy statement of Yahoo stabilized around 2003. It also shows that a social network can provide a privacy statement without have to make it complex.

Table 11: Reading Grade Level of Yahoo's Privacy Statements

Yahoo	Flesch-Kincaid	Gunning Fog	SMOG
6/30/1998	8.9	11.4	11
3/28/2002	10	13	12
1/22/2003	9.9	12.7	12
1/1/2004	10	12.8	12
11/22/2006	10.1	12.9	12

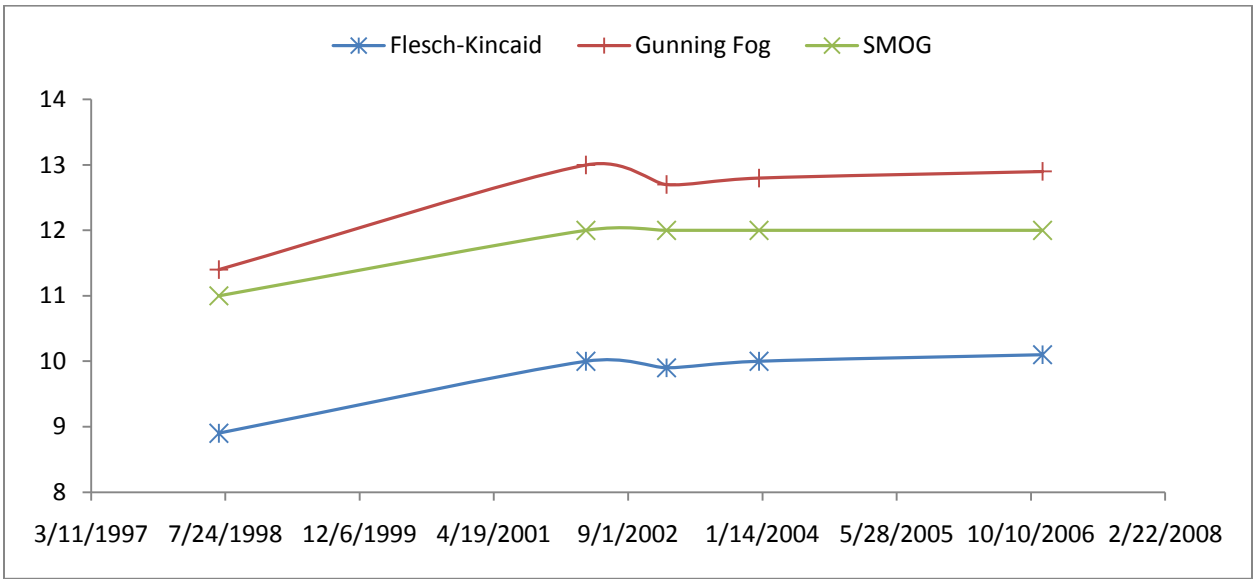

Figure 15: Reading Grade Level Measures of Yahoo's Privacy Statements

OVERALL COMPARISON OF GOOGLE, MYSPACE, FACEBOOK AND YAHOO STATEMENTS

From the individual analysis of the historic evolution of the privacy policy statements of Google, Myspace and Facebook, it is clear that as the policy evolved the complexity and the reading grade level became higher. This is contrary to the historical evolution path of Yahoo's privacy policy statement. Yahoo was able to maintain a privacy policy statement that is not complex and with a high school reading grade level. Table 12 presents a summary of the reading grade levels and the percentage of 20^{+} word sentence counts of Google, Facebook, Myspace and Yahoo privacy policy statements as they evolved over time.

Judging from the graphs in Figure 16 where we plotted the Flesch-Kincaid score (the Gunning Fog and the SMOG graphs are very similar), when evaluating the complexity and reading grade level of the privacy policy statement of a web site, we should start around 2004. The pre 2004 statements should be ignored. 2004, 2005 years are landmark years for privacy awareness, regulations and active privacy advocacy groups. In Figure 17, we plotted the Flesh-Kincaid readability score vs. $\%20^{+}$ word sentences of Table 12 for each of the social networks. Except for Yahoo privacy policy statements, Figure 17 shows that there is a direct positive correlation between the length of sentences and the reading grade level.

Table 12: Summary of the Reading Grade Level and Long Sentences Count of the Four Networks

	Flesch-Kincaid	Gunning Fog	SMOG	%20^{+}Word Sentences
Google				
8/14/2000	11.5	12.6	12	37.5%
7/1/2004	12.9	15.5	14	47.1%
10/14/2005	15.5	18.4	16	62.5%
8/7/2008	15.6	18.1	16	64.5%
3/11/2009	15.8	18.3	16	62.7%
	Flesch-Kincaid	**Gunning Fog**	**SMOG**	**%20^{+}Word Sentences**
Facebook				
6/28/2005	12.4	15.2	14	50.0%
2/27/2006	13.7	15.9	15	53.5%
5/22/2006	13.6	15.8	15	52.5%
9/5/2006	13.7	15.9	15	53.2%
10/23/2006	13.5	15.6	15	52.6%
5/24/2007	14.1	15.9	15	54.0%
9/12/2007	14.1	16	15	54.0%
11/26/2008	13.7	15.8	15	53.2%
	Flesch-Kincaid	**Gunning Fog**	**SMOG**	**%20^{+}Word Sentences**
Myspace				
9/21/2001	10.6	13.2	12	36.0%
10/20/2003	14.3	15.5	14	56.8%
3/5/2004	15	16.4	15	59.1%
2/25/2005	15.1	16.4	15	59.4%
8/26/2005	17.3	18.5	16	66.0%
2/28/2008	15.4	16.6	16	63.1%
	Flesch-Kincaid	**Gunning Fog**	**SMOG**	**%20^{+}Word Sentences**
	Yahoo			
6/30/1998	8.9	11.4	11	18.2%
3/28/2002	10	13	12	17.3%
1/22/2003	9.9	12.7	12	15.7%
1/1/2004	10	12.8	12	17.8%
11/22/2006	10.1	12.9	12	17.5%

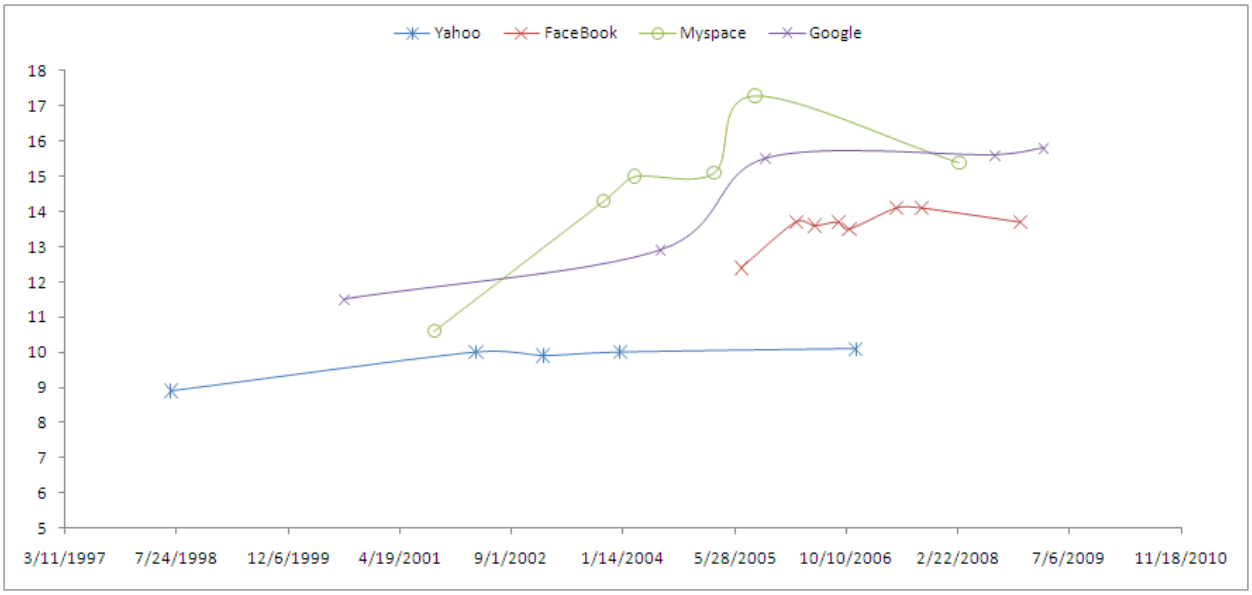

Figure 16: Overall Historical Flesch-Kincaid Readability Score

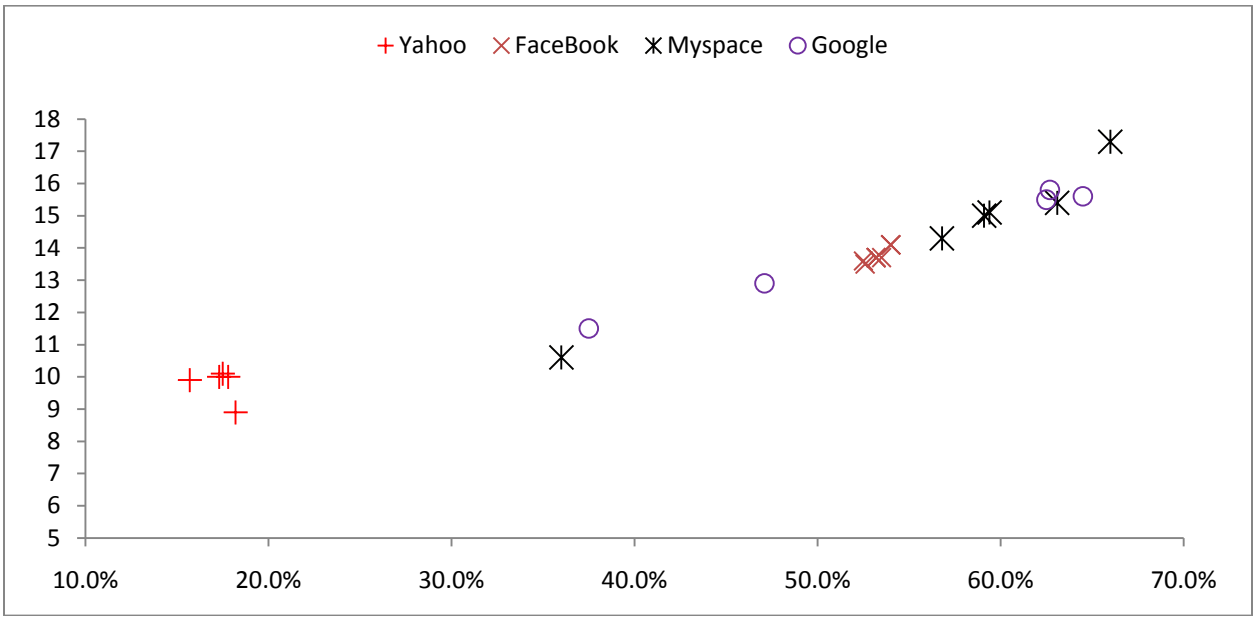

Figure 17: A Scatter Plots of Long Sentences vs. the Readability Score of A Policy by Network

SUMMARY AND CONCLUSION

Except for Yahoo, as the privacy statement of Google, Facebook and Myspace evolved, the more complex it got and the higher the reading grade level was required. The question has always been, if it is possible to write a legally binding privacy policy statement that is clear and easy to read? The answer is yes it is possible. Corporations should exert an effort to simplify their privacy policy statements. Yahoo's privacy policy statement is an example of a privacy policy statement that is not complex. It is not the character count, word count, sentence count or paragraph count that makes a document complex and makes it hard to read. It is the 3+ syllable counts per word, the 6+ characters per word count and the 20+ words per sentence that makes a document complex. In summary it is the choice of words and the length of sentences. Not the size of the document that makes it complex.

In summary, Except for Yahoo, the reading grade level of the privacy policy statement of the major social networks is beyond the reading grade level of the internet population in the United States. A legally binding policy that protects the right of the individuals and the social network can be achieved. Social networks need to exert more effort to articulate their policies. In the future, we will expand this research to include other social networks where we use these four networks as a bench mark.

AUTHOR INFORMATION

Musa J. Jafar is an Assistant professor of Computer Information Systems at West Texas A&M University. He worked as a senior software engineer for companies like IBM, US West Communications and Bell Communications Research

Amjad A. Abdullat is an Associate of Computer Information Systems and the Chair of Information and Decision Management Department at West Texas A&M University. Dr. Abdullat is one of the founding principles and member of the board of directors of Edmin.com. Edmin is a comprehensive technology enterprise providing learning organizations with the next generation of web-based products, applications and services.

REFERENCES

1. Cadogan, R. A. (2004). An Imbalance of Power: The Readability of Internet Privacy Policies. *Journal of Business & Economics Research* , 2 (3).
2. FTC Staff Report. (2009). *Self-Regulatory Principles For Online Behavioral Advertising.* U.S. Government, Federal Trade Commision.
3. Gunning, R. (1952). *The Technique of Clear Writing.* New York NY: McGraw-Hill International Book Co.
4. Horrigan, J. E. (2008). *Home Broadband Adopyion 2008.* PEW Internet & American Life Project. PEW/Internet.
5. Justice, U. D. (2008). *Privacy Policy Development Guide and Implementation Templates.* US Department of Justice.
6. Kincaid, P. J., Rogers, R. L., & Chissom, B. S. (1975). *erivation of Readability Formulas (Automated Readability Index, Fog Count and Flesh Reading Ease Formula.* U.S. Naval Air Station, Naval Technical Command.
7. Lenhart, Amanda. (2009). *PEW Internet Project Data Memo.* PEW/Internet, PEW Internet & American Life Project.
8. Lewis, S. D., Colvard, R. G., & Adams, N. C. (2007). A Comparison of the Readability of Privacy Statements of Banks, Credit counseling Companies, and Check Cashing Companies. *Academy of Organizational Culture, Communication and Conflict. 12(2),* pp. 23-38. Reno, NV: Allied Academies International Conference.
9. Mc Laughlin, H. G. (1969, May). SMOG Grading - A New Readability Formula. *Journal of "Developmental" Reading* , 639-646.
10. *OECD Portal*. (2009, May 20). Retrieved May 27, 2009, from Organization for Economic Co-Operation and Development: http://www.oecd.org/document/54/0,3343,en_2649_34225_38690102_1_1_1_1,00.html
11. Proctor, R. W., Ali, A. M., & Vu, K.-P. L. (2008). Examining Usability of Web Privacy Policies. *Journal of Human-Computer Interaction , 24* (3), 307-328.
12. *quantcast.com*. (2009). Retrieved June 14, 2009, from quantcast.com: http://www.quantcast.com/
13. Samuelson, P. (2008, Spring). *Info 205 Information Law and Policy Video Lectures*. Retrieved from Webcast.berkeley: http://webcast.berkeley.edu/course_details.php?seriesid=1906978514
14. Sheng, S., & Cranor, L. C. (2006). AN Evaluation of the Effect of US Financial Privacy Legislation Through the Analysis of Privacy Policies. *I/S: A Journal of Law and Ploicy for the Information Society , 2* (3).
15. Sprenger, P. (1999, 1 26). Sun on Privacy: 'Get Over It'. Retrieved 5 27, 2009, from Wired Magazine: http://www.wired.com/politics/law/news/1999/01/17538
16. U.S. Census Bureau. (2009). Educational Attainment in the United States: 2008. U.S. Department of Commerce, U.S. Census Bureau. Washington D.C.: U.S. Census Bureau.