# AOL Search Data Analysis

Rodger Byrd

May 5, 2019

## 1 Abstract

## 2 Introduction

In 2006 AOL released the search data for 20 Million queries for 650,000 unique users for an average of 30 searches per user. What kind of privacy impacts can result from the release of this specific data, and what are the impacts of search engines storing massive databases of queries from their users. The scale of the privacy implications of web search history is huge. Search engines are necessary due to the dynamic nature of the internet, and using them is how most users interact with the internet.

## 3 Design

First I had to download the data[1]. It consists of ten text files each with a portion of the search queries. Because the flat files are unwieldy, I planned to move the data to an SQLite database so I could query the data more easily. I ran into problems right away as this data is not well formatted for importing. It does not have a consistent delimiter between fields, it sometimes has spaces and sometimes has tabs, and some fields are just missing. I created a c++ program to parse the search data and insert it into the database, but it is not complete as it would have taken too much time to format and parse all the data. Because of the problems with the data, I decided to leave them as flat files and run text queries using grep and shell scripting. It ended up being a much simpler solution, and still resulted in fast searches. I thought it would be interesting to look for PII such as location, email addresses, or other interesting key words, and see what data could be correlated. The commands I ran are in `script.sh` file in the GitHub repository[3].

## 4 Evaluation

In evaluating the data I decided to look for common PII related data such as email, social security numbers, credit card numbers, and birthdays. I also did

a few searches to see if I could find anyone planning murder. There are some previously published notable examples that have turned up in the AOL data. One user, 44XXX49 was personally identified based due to the detailed nature of her search history[2]. There was a documentary created called I Love Alaska that chronicles the search history of user 71XX91[6].

### 4.a Email

First I searched for email addresses. This resulted in mostly false positives. The way the data is formatted each line contains the search and whatever link was followed from the resulting search. The email addresses showing up in the data were mostly mailto links in the format:
`http://mailto:email@address.com`.

### 4.b Social Security Numbers

There were a surprising number of social security numbers in the data. I only searched in the format NNN-NN-NNNN, Initially I had a regular expression formatted to accept dashes or no dashes, but there were too many nine-digit false positives in the results. Looking more closely at some of the search history for some of the results contain SSNs revealed even more data. User 51XXX34 searched for specific full names, addresses, and SSNs. If it wasnt their own username they inadvertently caused the release of the person they were searching for. User 40XXX51 looked up SSNs as well as medical conditions, medications and what appear to be local businesses.

### 4.c Credit Card Numbers

I didnt find any credit card numbers matching my regex. I found this somewhat surprising considering i found so many social security numbers.

### 4.d Birthday

It was difficult searching for birthdays in the data. There is a date field on every row, so simply searching for dates would return the whole data-set. I found a few interesting entries. User 11XXX27 has a St. Patricks day birthday and is located in the San Jose area.

### 4.e Murder

Searching for "how to kill" returned 700 results. Fortunately mostly were related to bugs and termites, but there were some troubling searches. It seems many people want to kill their pet dogs, cats, and hamsters. User 17XXXX39 is a somewhat famous example[5] with a very disturbing search history that was identified when the data was released. From his search history it looked like he was planning to kill his wife. I found more examples of people looking to kill their family members such as their brothers and sisters.

# 5 Discussion

The scope of this problem is huge. There are serious privacy implications related to searches and search history. Search engines store queries, clicks, IP-addresses, and other information about the users. As of December 2012, google had 1.17 billion users[8], receives 63,000 searches per second and has a market value of $739 billion.[7].

## 5.a How can one protect their privacy

When looking at what a user can do to protect their privacy there are a few options. A single proxy provides some protection but requires users to trust a third party. There are other solutions such as TrackMeNot and Tor. TrackMeNot works by obfuscation. It injects randomized search queries to make it more difficult to identify user profiles. Using Tor for searches makes it more difficult for the search engines to build a profile on the users. It was shown that user queries can be identified 48.88% of the time when using TrackMeNot and 25.95% of the time with Tor by performing testing on the AOL search logs[4]. The search engine `duckduckgo.com` doesn't maintain logs of users. They sell advertisements based on each search individually. They claim to not perform any tracking or advertisement targeting and don't sell users information because they don't keep any.

## 5.b What are engines doing with search history

Search logs enable better search performance.

## 5.c How should search history be released

It doesn't seem like there are any good ways to anonymize search data. It has been shown that anonyminzing methodologies such as k-Anonymity isn't enough protection [9]. I think this is due to the fact that the search data is personal in nature. I suspect that is the reason the only release of search data to date has been the 2006 AOL release.

# 6 Conclusions

It could be argued that search engines have stored the largest databases of personal information of all time. Search logs are a very personal view into peoples lives. Include implicit and explicit information about people are contained in search engine logs. Whether people are explicitly searching for social security numbers or medical conditions or information about them can be aggregated and compared with other publicly released data to implicitly identify them. Search histories contain data that should have privacy protections. The AOL release was very bad, but only averaged 30 searches per user. If the average user performs 3-4 searches per day, and only used Google

since it started in 1998, they would have performed approximately 30,660 searches! That is demonstrably going to reveal a great deal about any person.

# References


[1] AOL Search Data,
https://archive.org/download/AOL_search_data_leak_2006.

[2] M. Barbaro and T. Zeller,A Face is Exposed for AOL Searcher No. 4417749, New York Times,
http://www.nytimes.com/006/08/09/technology/09aol.html?ex=1312776000en=f6f61949c6da4d38ei=5090

[3] Project GitHub,
https://github.com/rodger79/CS5930-project

[4] Sai Teja Peddinti, and Nitesh Saxena. *Web search query privacy: Evaluating query obfuscation and anonymizing networks.* Journal of Computer Security 22, 2014.

[5] Frind, Markus: AOL Search Data Shows Users Planning to commit Murder,
https://web.archive.org/web/20080605041024/http://plentyoffish.wordpress.com/2006/08/07/aol-search-data-shows-users-planning-to-commit-murder/

[6] I love Alaska,
https://www.imdb.com/title/tt1455044/

[7] 63 Fascinating Google Search Statistics,
https://seotribunal.com/blog/google-stats-and-facts/

[8] 1.17 Billion People Use Google Search,
https://www.statista.com/chart/899/unique-users-of-search-engines-in-december-2012/

[9] Michaela Gtz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke, *Publishing Search LogsA Comparative Study of Privacy Guarantees.* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, MARCH 2012