Introduction

Background

Google. Less than 25 years ago, the term was entirely non-existent. Today, it is known as one of the largest technology conglomerates on Earth, primarily for it's unparalleled search engine capabilities. The search engine market is so deeply monopolized by it that even the Oxford English Dictionary (widely regarded as the most authoritative dictionary for the English language) defines it as a verb. Nevertheless, even the best softwares also have their own respective downsides. Regarding search engines specifically, this drawback manifests as the fact that nearly all of them harvest a plethora of data from their users. In today's society, information is power; the more information one has about someone, the more power one has over them. Thus, this brings into light a controversial question that has plagued search engines for decades: Should search engines be allowed to collect user data?

Data

In order to answer the aforementioned question, however, one must first understand the types of data that a search engine collects. As stated previously, Google has essentially monopolized the search engine market, making it a prime example for examining the types of data that are collected by many search engines; computer science Professor Douglas Schmidt at Vanderbilt University ran an experiment to intercept network traffic data sent from a personal device to Google servers, finding a large multitide of data associated with the user's Google account being sent frequently along with their other data. The researchers claimed that almost everything you do "online is collected and collated, from your morning routine (such as music tastes, route to work, and news preferences) to errands (including calendar appointments, webpages visited, and purchases made)" (Kelley, 2018).

When questioned about these excessively apparent amounts of data accumulation, Google responded by saying "it uses much of that data to improve its products. The information can lead to more relevant search reseults, for example" (Kelley, 2018). While this may be true, it is unknown to what degree the search results are made more relevant; in other words, using vastly greater amounts of additional user data may slightly benefit the accuracy of a search algorithm, but is that sliver of an advantage worth the price of that much privacy? By referencing past search algorithms, one can begin to understand how the accuracy/relevance and new user data requirements did increase, as well as how they should be increasing, with iterative advancements in such algorithms.

Arguments for User Data Collection

Search Algorithm Optimization

This is the explicit (and primary) purpose for which search engines claim to collect data, and can be further understood by exploring the evolution of the search algorithm during the childhood of the internet. According to Metaxas, "When the web was relatively small, Web directories were built and maintained that were using human experts to screen and categorize pages according to their characteristics." (Metaxas, 2010). This was the very first instance of the so-called 'search engine.' However, it was developed and maintained manually by humans, not unlike a file directory structure one might see on their local filesystem. As one can imagine, when the web boomed with new data, this algorithm quickly became obsolete: "By the mid 1990's, however, it was apparent that the human expert model of categorizing web pages would not scale" (Metaxas, 2010). People were then incentivized to move towards basic text searches and pattern matching, requiring a new text query to find results. Pattern matching was the first major leap in search engine algorithms, yielding a much higher releveance and number of results. A positive correlation between more data and higher relevance can thusly be established, such that the more data that a search engine has regarding a user's query, the more accurate results it can return. The unfortunate downside of this simplistic textual algorithm was its exact wording for effective searches, which was difficult to master for many search engine users.

Hence, this influenced the development of the modern search engine algorithm, known as 'PageRank.' PageRank is so successful as a search engine algorithm because it takes advantage of the properties of the modern web, which previous algorithms before it did not do. According to Arasu, the modern web can be represented as a graph data structure (a non-linear network of nodes, interconnected via the edges between the nodes): "We treat the HTML 'points-to' relation on Web pages as a directed graph $G = (V, E)$, where the set V of vertices consists of the n pages, and the set E of directed edges (i, j), which exist iff page i has a hyperlink to page j" (Arasu, Novak, Tomkins, & Tomlin, 2002). By collecting more data about a user's visits to individual pages, the powerful PageRank algorithm proves both its success (evidenced by the fact that it has existed for over 2 decades, and continues to be used today) and the correlation that more data yields better results. As a result, it is imperative that search engines are not hindered in their data acquisition process, and should be allowed to collect user data in order to maximize their accuracy.

Several variants of the PageRank algorithm currently exist, used differently by various search engines, such as Google, Bing, Yahoo, etc (Kumar, Duhan, & Sharma, 2011). These variants are what many people are concerned about, because they tend to collect large amounts of additional data than that of the original PageRank algorithm, Google most of all. Nevertheless, those very same algorithms show excellent results as well; Google, the search engine which collects the most data, is also the most accurate/relevant one. People want their tasks simplified over

time, shown by the evolution and usage of novel technologies such as personal assistants and voice recognition, made to make life easier. If search engines are disallowed from aggregating information, this trend of simplicity will be halted and reversed, preventing search engines from progressing into more advanced algorithms in the future.

In fact, there are already some new models and theories for futuristic search algorithms. The majority of these are creating entirely new meanings for the term 'search engine;' as their novel approaches involve collecting and storing so much data from the users that the algorithm becomes more akin to an intelligent personal assistant or chatbot, rather than an entry form for a search query. Several of these models include using advanced and emerging technologies such as Natural Language Processing (NLP) to enhance the interpretation of a phrase and to understand an individual user's writing or speaking style. Currently, the most realistic of these is the concept of a 'session-based search engine': "users often have to modify their queries several times before they can reach a page that meets their information need. During these interactions with the search engine, the user provides a lot of useful information to the search engine, which can be exploited to infer the user information need" (Sriram, Shen, & Zhai, 2004). This novel idea places all the user's searches within a context, helping the search engine to build a 'profile' of sorts that can be used to understand even highly ambiguous queries. It would essentially allow a personalized connection with the search algorithm, and require even greater amounts of user data to be stored and indexed, but it would also be exponentially more efficient at finding results. This system could be an entirely new phase for search engines, which could have the ability to intelligently understand questions and even hold a conversation. However, greater advancements such as this could no longer be even a possibility with the continued issues regarding search engine data collection; if we are to progress technologically, it is imperative for information sharing to become more relaxed.

Government Surveillance

When most people hear the term 'government surveillance,' it usually does not evoke a positive connotation. Today, people are very protective of their freedom, and tend to take great offense if they even think that their rights could be infringed upon (especially in the United States, where freedom is a driving principle for the country). Nonetheless, while many kinds of government surveillance may seem to be infringing on these 'rights,' their primary purposes are usually for a greater good.

In the case of the internet and data privacy, government surveillance can be highly advantageous in reducing all sorts of crimes, and could especially help with reducing terroristic attacks (potentially) worldwide. With access to users' search queries, a preliminary threat model could be constructed within a classified sector of the government to identify dangerous individuals, even before a crime would be committed so that the government could be alerted to

the potential threat. According to CNN, "the move toward 'home-grown' terror will necessarily require, by accident or purposefully, collections of U.S. citizens' conversations with potential overseas persons of interest..." even if "this truth is naturally uncomfortable for a country with a Constitution that prevents the federal government from conducting 'unreasonable searches and seizures' " (Sulmasy, 2013). Just as how a patient might share confidential medical information with a doctor for health reasons, the act of the government investigating its citizens' data could be thought of in a similar manner, to help reduce safety risks for a nation as a whole (i.e., the 'greater good').

Arguments Against User Data Collection

Data Sharing and Theft

Yet the key difference between a doctor and the government is that you have the choice to share knowledge with your doctor, whereas the government, not so much; any process of government surveillance would need to be universal and mandatory, not to mention completely classified, if it were to be successful, as such a "disclosure likely would perturb shareholders and subject the company [search engine] to public scrutiny and possibly legal action" (The Google-NSA Alliance: Developing Cybersecurity Policy at Internet Speed 2012). This excerpt, taken from a report regarding an alliance between Google and the National Security Administration to help Google secure its servers from foreign attacks, emphasizes the backlash that would ensue if any illicit activities originating from Google (or any other search engine, for that matter) were to ever occur. In any case, real attempts have already been made by the United States government for an initiative such as this: "In January 2006 it was revealed that the U.S. Justice Department asked a federal judge to compel the Web search engine Google to turn over records on millions of its users' search queries as part of the government's effort to uphold an online pornography law" (Zimmer, 2008). Once again, this had the correct intentions, but gaining access without users' permission is seen as such a great breach of privacy that the public outcry would be just as devastating, if not more. Therefore, it is in the best interest of users to withold as much of their data as possible from search engines.

The idea of data sharing in general, not just with the government, but with other third parties as well, can also be a very dangerous issue. Countless studies have shown that people, if given the chance to sell another user's data without their permission in order to benefit themselves, would take it a majority of the time, as long as they deem the data to be 'useless' or unimportant (Benndorf & Normann, 2018). Most people seem to have little regard for others data privacy on the internet, which is highly concerning, considering a great deal of a users' data relies in other people's hands.

Cybercrime and data theft are also important issues, as once the data has been obtained, no matter who it may be, there is no 100% guarantee that it will be safeguarded eternally.

Highly sensitive user data (such as credit card numbers, social security numbers, etc.) aren't stored with your search history, so many people feel as though they still do not have much to worry about when concerning theft of their search data. In reality, however, search data is very informative regarding a person's life and interests, and can help anyone with such data to launch a phishing attack where the truly critical data could be then stolen (Levinson, 2012).

The most critical issue with search engine data is not the way it is used by the search engine; rather, it pertains to the way it is stored. Recently, one of the largest incentives of mass data collection and storage for search engines has been the concept of 'big data' (Whittle, Eaglestone, Ford, Gillet, & Madden, 2007). Big data "refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods," and usually stored long-term for processing at a later date much in the future, when the computing resources become available (Thompson, n.d.). By storing so much sensitive user data for such a long time, search engines create an enormous attack vector surface for hackers to obtain that data. In the wrong hands, the dangers that these kinds of data present, help to answer the overarching question of this subject with a resounding no: search engines should absolutely be prevented from storing any of this data, for the sake of its users' protection.

Advertising

While phishing scams and other illict activites that use search engine data are extremely dangerous, the legal route of advertising can seem even more so at times. Targeted ads remain a very annoying problem in today's society, and are especially dangerous in the way that they can influence people in the same way that phishing scams might. When Cambridge Analytica harvested massive amounts of data from Facebook users, it then used the data to attempt to influence the 2016 U.S.Presidential Election (Graham-Harrison & Cadwalladr, 2018). The amount of influence that such ads can have tends to be terrifying to many internet users, even if they might not realize it immediately.

Furthermore, targeted ads also take away from the natural online market (relating to a user's search). For instance, if a search engine optimizes gardening ads to target a user who is searching about pottery, it detracts from the natural gardening market, which over a great deal of time and people, can be quite hurtful to the economy (Xing & Lin, 2006). Therefore, using user search data is immoral if it is for any other purpose than that which it was originally intended.

Concluding Remarks

Utilitarianism is an ethical outlook that essentially helps to investigate the net positive outcome of an action and determine whether that action is moral or immoral. We can use utilitarianism to summarize the advantages and disadvantages of search engines collecting data, and from there conclude whether the action is moral or immoral, in a very black-and-white manner.

The trend of simplicity and efficiency is very clear in modern advancements in technology; people want to make their own lives easier. Collecting more user data for this purpose is then justified, as the overall goal is to simplify tasks for users, and an optimized search engine algorithm does just that. Collecting more data also paves the way for novel search algorithms that could reshape the future of the search engine industry, such as the session-based search algorithm.

As stated previously, the power that information holds over people can be immensely terrifying. If even a users' search history were to land in the wrong hands, they could be the victim of merciless phishing attacks and other cybercrime. The way that even targeted ads, as evidenced by the Cambridge Analytica scandal, can influence users in a presidential election using data that could be considered 'unimportant' is unacceptable.

Considering all of these points through the ethical framework of utilitarianism, one can then say that the collection of user data by search engines is morally correct. The majority of users benefit from the search engine's more relevant and accurate results, as opposed to the slim margin of users who are targeted by scammers, cybercriminals, and malicious advertisements. Therefore, search engines should in fact be allowed to continue collecting user data.

References

Arasu, A., Novak, J., Tomkins, A., & Tomlin, J. (2009). PageRank Computation and the Structure of the Web: Experiments and Algorithms. Stanford University Computer Science Department.

Benndorf, V., & Normann, H. (2018). The Willingness to Sell Personal Data. The Scandinavian Journal of Economics, 120(4), 1260-1278. doi:10.1111/sjoe.12247

The Google-NSA Alliance: Developing Cybersecurity Policy at Internet Speed. (2012). OECD Digital Economy Papers. doi:10.1787/5k8zq92vdgtl-en

Graham-Harrison, E., & Cadwalladr, C. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. Retrieved from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

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

Kumar, G., Duhan, N., & Sharma, A. K. (2011). Page ranking based on number of visits of links of Web page. 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011). doi:10.1109/iccct.2011.6075206

Levinson, M. (2012, January 26). Are You at Risk? What Cybercriminals Do With Your Personal Data. Retrieved from https://www.cio.com/article/2400064/are-you-at-risk–what-cybercriminals-do-with-your-personal-data.html

Metaxas, P. T. (2010). Web Spam, Social Propaganda and the Evolution of Search Engine Rankings. Lecture Notes in Business Information Processing Web Information Systems and Technologies, 170-182. doi:10.1007/978-3-642-12436-5_13

Sriram, S., Shen, X., & Zhai, C. (2004). A session-based search engine. Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval - SIGIR '04. doi:10.1145/1008992.1009086

Sulmasy, G. (2013, June 11). Opinion: Why we need government surveillance. Retrieved from https://www.cnn.com/2013/06/10/opinion/sulmasy-nsa-snowden/index.html

Thompson, W. (n.d.). Big Data: What it is and why it matters. Retrieved from https://www.sas.com/en_us/data/what-is-big-data.html

Whittle, M., Eaglestone, B., Ford, N., Gillet, V. J., & Madden, A. (2007). Data mining of search engine logs. Journal of the American Society for Information Science and Technology, 58(14), 2382-2400. doi:10.1002/asi.20733

Xing, B., & Lin, Z. (2006). The impact of search engine optimization on online advertising market. Proceedings of the 8th International Conference on Electronic Commerce The New E-commerce: Innovations for Conquering Current Barriers, Obstacles and Limitations to Conducting Successful Business on the Internet - ICEC '06. doi:10.1145/1151454.1151531

Zimmer, M. (2008). The Gaze of the Perfect Search Engine: Google as an Infrastructure of Dataveillance. Web Search Information Science and Knowledge Management, 77-99. doi:10.1007/978-3-540-75829-7_6