Manipulation of Search Engine Results during the 2016 US Congressional Elections

Panagiotis Takis Metaxas and Yada Pruksachatkun

Department of Computer Science
Wellesley College
Wellesley, MA 02481
Email: pmetaxas@wellesley.edu

***Abstract*—Web spammers are individuals who attempt to manipulate the structure of the Web in such a way that a search engine (SE) will give them higher ranking location (and thus, greater visibility) in search results than what they would get without manipulation. Typically, Web spammers aim to promote their own financial, political or religious agendas exploiting the trust that users associate with SE query results. Over the last ten years, search engines have taken steps to defend against spammers with some success. Arguably, Web spamming is crucial during election times, when voters are likely to use search engines to get information about electoral candidates. At times of elections, spammers could succeed in spreading propaganda manipulating SE query results of candidates' names. In a symmetric but, arguably, less likely scenario, SEs might influence elections by manipulating their own results to favor one candidate over another. In fact, some have suggested that SEs (Google in particular) should be proactively regulated to avoid such a possibility. In this paper, we investigate to what degree the SE query results related to searches of electoral candidates names were altered by anyone (Web spammers or SEs) during the 2016 US congressional election, an election that saw the rise of "fake news" sites. Our results indicate that different SEs had different degree of success defending against spammers: Google gave preference to reliable sources in the first 6 of the top-10 search results when queried with the name of any electoral candidate. Also, Google did not allow much variation in the ranking of the top-10 results and did not allow "fake news" sites to appear at its organic results. Bing and Yahoo, on the other hand, did not have as good a record. This is even more apparent in the autocomplete box "suggest" options presented to the user while forming the query.**

Keywords–Search Engines; US Elections; Web Spam; Fake News; Google; Yahoo; Bing.

I. INTRODUCTION

Web spammers are individuals who are trying to manipulate the structure of the Web in such a way to control search engines ranking algorithms to give them higher ranking in search results than what they would get without the alteration [1]. This way, Web spammed pages will get higher visibility in the eyes of unsuspected users searching for the targeted terms. They do that by manipulating the SE ranking methods aiming to influence the user's opinion about their site's quality. In this respect, they behave very similarly to social propagandists who are trying to alter a citizen's mental trust network in ways beneficial to the propagandist [1].

Web spam has a long history of manipulating search results of SEs that starts with the creation of the first search engine, back in 1995. Usually, their intentions are:

- financial: turning the attention of users to particular products they are promoting, or gaining from online advertisement;
- political: helping elect the candidates they support;
- religious: helping promote the religion they support.

Even though their activities are not known to most Web users, spammers have had a significant role in the evolution of SEs because they have forced SEs to keep changing their ranking methods [2]. Ranking methods, such as the well known PageRank [3] used to be a well-understood, studied and evaluated set of mathematical functions of information retrieval, while today they are a secret, fluid, complicated and intentionally difficult to predict set of factors [4]. Since SE ranking methods is one of the most important factor that any Web site marketer, advertiser, or propagandist needs to understand, there is a whole $65 billion industry that is studying them [5].

A. Background and Prior Work

Researchers have followed the election-related Web spamming attempts in the past twelve years or so [6]. The first recorded attempt was in 2006 when spammers openly called for the promotion of negative information related to some candidates for the senate and recorded the results online on the myDD.com Web site. However, their initial success draw the attention of Google that reportedly tried to defend against their efforts, since their actions were compromising its reputation as a reliable search engine.

In particular, [7] and [8] studied to what degree Googles electoral search results were manipulated during the six months prior to the 2008 and the 2010 congressional elections. Their findings suggest that starting in 2008, Google tried to protect its search results by reducing the weight that the PageRank algorithm had on searches with queries the names of US electoral candidates. In 2012, Google started employing a vertical split-screen interface in which the left side of the screen contained the organic results and the right side contained information from its *knowledge graph* [9] with official information about the candidate (see Figure 3). Bing and Yahoo have also adopted a similar interface (e.g., see Figures 1 and 2). We should point out that, even though Google's electoral search results has been studied over time, to our knowledge, Bing's and Yahoo's performance to defending against spam has not been studied in depth in the past.

Recently, some researchers have raised the possibility that Google might secretly decide to manipulate its own results

 ISBN: 978-1-61208-563-0

[10]. That is, they worry that Google might be tempted to use its ranking algorithm to support one political candidate over another [11]. In particular, [10] has measured the possible influence that manipulated search results can have on unsuspecting audiences. They have found that, even though the effect of manipulation may not be large, it can have a significant effect in close elections.

While such claims created a lot of interest from news organizations, the realization that people's opinion can be influenced by search results is hardly new. Every advertiser is well aware of the importance of their ranking, and a whole industry, called *"search engine optimization" (SEO)*, has tried to increase product placement through blog posting and even Web spamming. The SEO industry is reportedly worth tens of billions of dollars [5]. SEOs organize conferences and training workshops selling expertise on how one can do exactly this type of manipulation. While the work of [10] is focusing on a particular SE, Google, and has called for federal regulation of its search results, one needs to examine all major SEs for biased behavior. We argue that such a concern is rather overstated: Google is the major SE and it would have everything to lose by manipulating its rankings. Data collection such as the one done for this paper could reveal enough evidence of its manipulation and it could be done by anyone with basic programming skills in scrapping. Further, many people inside Google would know it and the likelihood of a whistleblower is rather high.

Our paper's contributions are as follows: We investigate to what degree the SE query results related to searches of electoral candidates' names showed any signs of alteration by anyone (spammers or SEs) during the six months prior to the 2016 US congressional election. This was an election that saw the rise of "fake news" sites, and so it is doubly important to see to what degree "fake news" stories infiltrated search results. We also examined the number of times that "fake news" sites (that is, sites that have been characterized as hosting "fake news" stories by [12]), appeared in the top-10 search results for the examined SEs.

Our results indicate that the three most commonly used SEs, Google, Yahoo and Bing, had strikingly different degree of success defending against spamming. Google gave consistent preference to reliable and official sources in the top-6 search results when queried with the name of any electoral candidate, and did not allow much variation in the ranking of the top-10 results. Bing and Yahoo, on the other hand, do not have as good a record. Their search results showed little effort of consistency and the number of "fake news" sites appearing in their results were higher. This was especially obvious in the search autocomplete box "suggest" options presented to the user forming the query.

The rest of the paper is organized as follows: The next Section II describes our data collection and preparation for analysis, Section III explains our methods, while Section IV describes our results. Finally, Section V contains our conclusion and future work.

II. Data Collection and Preparation

According to the Pew Foundation [13], Google, Bing and Yahoo have a combined market share of 98.34% with the greater portion going to Google (79.88%). It is safe to assume that if there was a successful attempt to manipulate SE results before the US elections, one could detect its success by monitoring the query results for suspicious variations during that period in these three SEs.

For the six months prior to the November 2016 US presidential and congressional elections, we collected query results using as query strings the names of the two major presidential candidates Donald Trump and Hillary Clinton, for Bernie Sanders (because at the beginning of the data collection he was still a contestant for the Democratic nomination) and of 340 congressional candidates, 150 of them Republican, 142 of them Democratic, and 64 of them of smaller parties or unaffiliated (58 independent or libertarian, and 6 local parties). Of these candidates, 74 were running for the 34 seats in the Senate, so we examined every senatorial candidate. We also examined 279 candidates for the House. The latter were a subset of the more than 2000 candidates running for the 435 seats in the House. To avoid overloading the SEs with over 2000 query requests, we selected the candidates for the first six states, in alphabetical order: AL, AK, AZ, AR, CA, CO. We have no reason to believe that the search results in the remaining states would have been any different. The candidate names were chosen from a website specializing in monitoring the electoral candidates [14].

We used the following method to collect the data: between June 2 and November 8, 2016, on a roughly bi-weekly basis (44 data collection dates in total), Google, Bing, and Yahoo were queried and scraped for the top 10 results using `requests` and `urllib` python libraries, and matching using regular expressions for the top 10 results tags. Each of the search results were then aggregated into files for each candidate, as follows: for each collection date, each candidates file contains a list of websites that appeared in the top 10 search results, in the numerical order they appeared. For consistency in the overall data collection, websites that did not appear in the top-10 results for a certain date were assigned a rank of 0 for that date.

We point out an caveat in our data collection. In the middle of the search results scraping, Bing changed its formatting a couple of times (at least) and our algorithms collected fewer top-10 results in a consistent fashion. The analysis we present here is based on the earlier dates and may be incorrect for Bing overall.

III. Method and Analysis

Processing the data involved the following steps: for each candidate, we created a table with the top-10 links per date for each of the 44 data collection dates. We wanted to know the particular domain that a link was pointing to, instead of the specific link within the site. To account for that, we extracted the site domain of each link. For example, all articles from the New York Times were represented in our data tables by the site domain `nytimes.com`.

For each domain in the table, we calculated the number of times it appeared over the whole collection period. To control for data sparsity, we introduced the measure of *website appearance percentage* (WAP), defined as the minimum percentage of times a website appeared in the top-10 results over the period of data collection. For the data reported below, we used WAP values of 33%, 50%, 66% and 75%.

We also compute a domain's *mode*, defined as the top-10 location it appeared for a WAP percentage of the time.

 ISBN: 978-1-61208-563-0

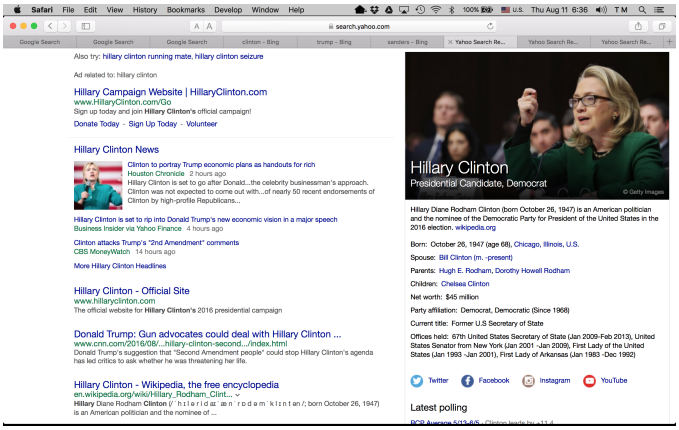

Figure 1. Yahoo sample search results for Hillary Clinton on Aug. 11, 2016. The first item is an ad, followed by recent news. Organic results start with her official site, cnn and wikipedia. The knowledge graph's information appears on the RHS.

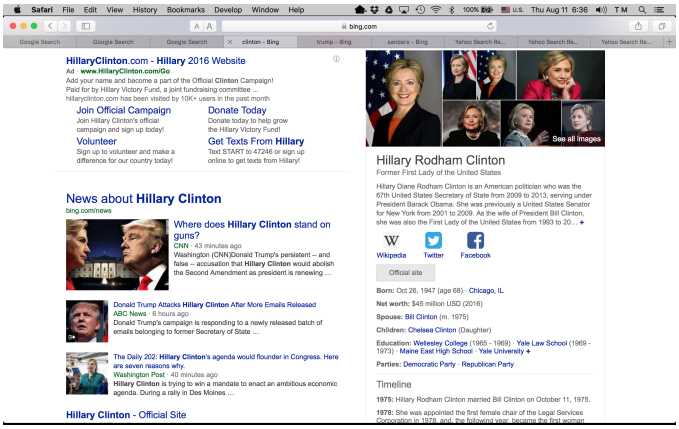

Figure 2. Bing sample search results for Hillary Clinton on Aug. 11, 2016.

For less than a particular WAP, the mode is not defined. One way to get a sense of the usefulness of defining the mode is to think in terms of predicting future search results for some query. Consider the following example: In Google's complete collection for "Hillary Clinton" the URL item `hillaryclinton.com`, her official campaign site, had a mode of 1 with a WAP of 0.75. That means that, at least 75% of the time, searching for "Hillary Clinton" on Google resulted in her official campaign page being first in the top 10 results. It also means that next time we could do the same search, at least 3 out of 4 chances is that `hillaryclinton.com` will appear in the 1st position. Similarly, for the same search, `wikipedia.com`'s mode was 2 and `twitter.com` was 3.

In summary, the higher the WAP we could define for the modes of domains of search results, the more predictable the ranking of search results will be in a future search, and the less likely the search results were altered by propagandists of the SE itself.

IV. Results

Given the plethora of news, blogs and political analysis around the time of elections, if SEs was using a dynamic ranking method, such as straightforward PageRank, to compute

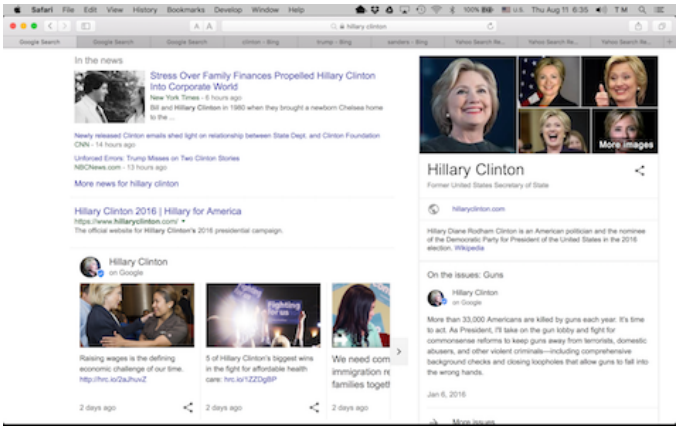

Figure 3. Google sample search results for Hillary Clinton on Aug. 11, 2016. Note that all three SEs have adopted the "knowledge graph" on the RHS of each search, which makes them even more visible occupying a large portion "above the fold". For prominent candidates, the LHS may contain an ad followed by recent news about the candidate. Our research measures the changes in the "organic results" typically appearing under news.

the top-10 results, it would be surprising to have mode defined at all. It is more reasonable to think that, as news was being produced and gaining prominence in online sources, the location of every URL item would change considerably over time. On the other end, if SEs were using a static list of predefined top-10 results to respond to search queries, all 10 modes would be defined over the data collection.

Of course PageRank is one of the factors that search engines are using to rank their results. The greater the contribution of PageRank in the final ranking, the less often mode is defined. Intuitively, when for a search query we have a large number of modes defined, say 6–10 modes per candidate collection, we can deduce that the query results are not updated dynamically over time as much. But if a small number of modes were defined, say 0-4 modes, the query results are rather dynamically altered, possibly driven by PageRank or spammers. Finally, when 5 modes are only defined, one would say that dynamic and static ranking methods are equally at work.

In any case, we should also point out that it is important to consider *which* modes are defined. For example, if all modes at the top-5 locations are defined, the user will be shown practically the same results above the fold. For users who do not look below the fold, it would appear that search results are not changing.

To see the significance of the number of modes defined in some search result, consider the following scenario. Assume that for some search engine, query results over time for some candidate define 8 modes. That means that, since ranking does not change a lot, the next time we are querying the search engine for the candidate we may expect to see the same 8 URLs in the same location of the search results. So, we can say that we can predict most of the search results. In particular, it is very likely that the above-fold search results (often referred as the top-5) will be the substantially same in the future.

Let us compare that scenario with another scenario of a search engine or candidate whose search defines only 3 modes. In this case, the search result URLs will be significantly

 ISBN: 978-1-61208-563-0

different. If Web spammers are trying to influence the search results this latter search engine could fall victim. To be fair, if the 3 modes defined are the top-3, it would indicate an attempt by a search engine to defend its search results in the locations above the fold that are more important, while leaving the rest to the algorithm.

Our mode of location results for our data collections are as follows:

1) Google mode averages for different WAPs:

WAP	75%	66%	50%	33%
Overall	5.37	5.78	6.64	7.50
Democrats	5.32	5.78	6.66	7.67
Republicans	5.44	5.83	6.67	7.45
Senate	5.18	5.55	6.54	7.38
House	5.43	5.85	6.69	7.56

2) Yahoo mode averages for different WAPs:

WAP	75%	66%	50%	33%
Overall	2.67	2.92	3.28	3.21
Democrats	2.57	2.85	3.17	3.13
Republicans	2.75	2.93	3.35	3.26
Senate	2.68	2.93	3.31	3.18
House	2.65	2.89	3.26	3.19

3) Bing mode averages for different WAPs: As we mentioned, due to problems with scraping Bing our results are not complete and so we will not present them here. For the period we have complete results we can report that Bing's mode averages are even lower than Yahoo's.

Our results indicate remarkably different number of modes (and therefore, ranking methods) between Google and the other search engines, even for the more restrictive WAP of 75%. For Google, it is possible to define the mode of each Google search for an average of 5.37 top-10 locations (median = 6) in all of our 340 searches. For WAP 50% (i.e., at least half the time) almost 7 modes are defines (6.64 to be exact). On the other hand, in Yahoo, it is only possible to define the mode for an average of 2.67 locations (median = 3) for WAP 75% and 3.28 for WAP 50%. Finally, in Bing it appears that the average is even less than Yahoo's. In other words, users who queried Google about a congressional candidate, they saw little variation above-fold in their organic search results over time. Users who queried Yahoo saw much greater variation but not in the first 3 organic locations, while users who used Bing saw almost always different results, except maybe in the first location.

But which modes are defined? Not surprising, for all SEs, the most common predicted modes were the top results, with Bing's being the top 2, Google's being the top 7, and Yahoo's being the top 5. This shows that Google contains more websites that are consistently appearing in the top-10 results than Yahoo and Bing.

If spammers (or the search engines themselves) were trying to manipulate their congressional search results, they were not succeeding in Google. It is much more likely that they were successful in Yahoo and, especially in Bing.

Next, we will address the question whether search engines showed any preference to a particular party. Were there any differences in the modes for Democratic and Republican candidates?

The answer is no, all three SEs treated the candidates of both parties in a similar way: Compared to the overall average of 6.64 (for WAP = 50%), Google's mode of Democratic candidates was 6.66 while for Republicans was 6.67. Similarly for Yahoo (3.17 and 3.35, respectively) and Bing.

Remarkably, the averages for Senate candidates vs. House candidates are similarly consistent. Google's Senatorial candidates have average of 6.54 and House candidates 6.69. Yahoo's averages are 3.31 and 3.26, respectively.

Thus, while Google showed little alteration to its organic search results compared to Yahoo and Bing, all three search engines treated all candidate, Democratic or Republican, for Senate or for House, consistently.

A. *"Fake news" stories*

Finally, we counted the occurrences of items from sites that were characterized as "fake news" sites appearing in [12]. We discovered that over all SEs and over all the searches, there were 85 "fake news" sites in Democratic candidate search results, 139 for Republicans, and 27 for independents. Thus, there are more appearances of "fake news" sites in search results of Republican candidates. We should clarify that we have done no analysis as of this writing on whether the "fake news" items were positive or negative for the candidate, or whether the stories were true or false. Doing so is beyond the scope of this paper.

We then counted the number of "fake news" occurrences per search engine. For Google, there were 68 unique stories from sites characterized as "fake news" sites that appeared over the six months period we studied. By contrast, there were 83 for Bing, and 95 for Yahoo. Thus, Google is less prone to listing "fake news" sites in its top-10 results, followed by Bing and last Yahoo. Again, we should clarify that we have done no analysis as of this writing to see whether the "fake news" items appearing on top-10 results were true or false (given that not every story that appears in a "fake news" site may be false).

V. CONCLUSIONS

Our paper studied the extent to which Google, Bing, and Yahoo were prone to Web spamming during the last 2016 congressional elections. While we cannot be sure whether anyone tried to manipulate search results ranking for congressional candidates, we can tell whether they were successful in altering the ranking of the search results, if they tried.

Our results indicate that, by and large, there were no variations of top-6 websites in Google, and only a few "fake news" stories that appeared over the six months period we studied in the top-10 results. On the other hand, there was significant variation in the search results for Yahoo and almost constant change in Bing. If spammers were trying to manipulate search results in Yahoo and Bing, they were more successful. We also found that all three search engines treated similarly Democratic and Republican candidates, and Senatorial and House candidates.

Even though the market share for Yahoo and Bing is small, spammers can introduce biased information into the search results, affecting the perception of candidates for users

 ISBN: 978-1-61208-563-0

who used these two SEs. Further evidence of the ease at which Bing and Yahoo are manipulated by spammers can be found [15]. Thus, we call for a thorough effort by Bing and Yahoo to increase their defense against Web spammers. Further investigations are needed for Yahoo and Bing data, as well as for the nature and type of any spam that were introduced to various demographics of candidates.

REFERENCES

[1] C. Castillo and B. Davison, *Adversarial Web Search*, ser. Foundation and trends in information and technology. Now Publishers, 2011. [Online]. Available: http://bit.ly/2pdYxrZ

[2] P. Metaxas, "Web Spam, Social Propaganda and the Evolution of Search Engine Rankings," *Lecture Notes BIP, Springer-Verlag*, 2010. [Online]. Available: http://bit.ly/ffYsuC

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[4] "Google Keeps Tweaking Its Search Engine," 2007, URL: http://nyti.ms/2n5JjGx [accessed: 2007-06-03].

[5] "The SEO industry is worth $65 billion; will it ever stop growing?" 2016, URL: http://selnd.com/2p0PyvV [accessed: 2017-02-01].

[6] "Propaganda, Misinformation, "Fake News", and what to do about it." 2017, URL: http://bit.ly/2qQkwpn [accessed: 2017-05-11].

[7] P. Metaxas, "Network Manipulation (with application to political issues)," in *Proceedings of the Interdisciplinary Workshop on Information and Decision in Social Networks, Cambridge, MA, May 31 - June 1, 2011*. MIT Media Lab, May 2011, URL: http://bit.ly/NRmNox/ [accessed: 2017-02-15].

[8] P. T. Metaxas and E. Mustafaraj, "The battle for the 2008 us congressional elections on the web," in *In the Proceedings of the 2009 WebScience: Society On-Line Conference*, March 2009.

[9] "Knowledge Graph – Inside Search – Google." 2017, URL: http://bit.ly/2r9QOvG [accessed: 2017-05-11].

[10] R. Epstein and R. E. Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," *PNAS*, pp. E4512–E4521, 2015.

[11] "Could Google rankings skew an election? New group aims to find out." 2017, URL: http://wapo.st/2r7WSFg [accessed: 2017-05-11].

[12] "List of False, Misleading, Clickbait-y, and/or Satirical "News" Sources," 2016, URL: http://bit.ly/2pe1ZmC [accessed: 2017-01-02].

[13] "Search engine use over time," 2012, URL: http://pewrsr.ch/2q2htgx [accessed: 2016-02-15].

[14] "2016 US Congressional Elections Candidates," 2016, URL: http://www.politics1.com/p2016.htm [accessed: 2016-06-02].

[15] "Dr. Epstein, You Dont Understand How Search Engines Work." 2017, URL: http://bit.ly/2pE2Ifl [accessed: 2017-05-11].

 ISBN: 978-1-61208-563-0