3-2009

The Battle for the 2008 US Congressional Elections on the Web

P. Takis Metaxas
pmetaxas@wellesley.edu

Eni Mustafaraj
emustafa@wellesley.edu

THE BATTLE FOR THE 2008 US CONGRESSIONAL ELECTIONS ON THE WEB

Panagiotis Takis Metaxas and Eni Mustafaraj
Computer Science Department, Wellesley College, Wellesley, MA02481, USA
pmetaxas@wellesley.edu

Keywords: Search Engines, Web Science, Web search, Web Spam, Social Networks

Abstract: It has been reported that, in the past, political activists have tried to influence web search results. They did that using link-bombing techniques to raise negative web pages with contents close to the their agendas to the top-10 search results. Google has admitted that this happen in the 2006 US Elections, but did it still happen in the all-important 2008 US Congressional Elections? In this paper we try to evaluate whether "gaming" the search engines during the election period is a widespread problem, how serious is it, and how search engines have tried to maintain the integrity of their search results.

1 INTRODUCTION

We live in an increasingly interconnected world, one in which a growing number of people turn to the web to make important medical, financial and political decisions (The Pew Foundation, 2008). As more people use the Web's search engines daily as their primary source for locating information on many important issues, search engines are in the position to influence what is perceived as relevant information through their mechanism of ranking web pages. However, as studies have shown (Metaxas, 2009a), interested groups and individuals can also make use of web spamming mechanisms to trick search engines in ranking their pages higher than those of their rivals. Furthermore, the battle between search engines and groups with their own agenda is not confined in areas where there is an expected financial gain from web transactions. It is spread among many ideological, cultural, and political issues where controversial positions vie for the public support, on issues such as abortion legality and morality, children vaccination risks, creationism vs. evolution, homosexuality, etc. (Hindman et al., 2003)

It has been widely reported in the news that, in 2006, political blogs had been actively trying to influence the US elections by pushing web pages carrying negative content to the top of the relevant search results of the major search engines. This practice of "gaming" the search engines was implemented with link bombing techniques (also known as Google-bombing), in which web site masters and bloggers use the anchor text to associate an obscure, negative term with a public entity (McNichol, 2004). In particular, during the 2006 US midterm congressional election, a concerted effort to manipulate ranking results in order to bring to public attention negative stories about Republican incumbents running for Congress took openly place under the solicitation of the progressive blog, MyDD.com (My Direct Democracy) (Zeller Jr., 2006).

Search engines have admitted that something like that may be happening. Google, who had been the main target of this attack, initially tried to ignore this practice, claiming that interfering with its ranking would effectively alter their powerful algorithms (Mayer, 2005). Eventually, however, it decided that bad publicity was more damaging and reportedly tweaked its algorithms to minimize the impact of many Googlebombs (Moulton and Carattini, 2007; Hansell, 2007). At the time of this writing, Google has not elaborated further on how they hoped to minimize the impact, but the latest occurrence (Sullivan, 2009) suggests that the fixing of the Googlebombs was done by human editors.

Researchers have worried for a long time not only

for the practice of "gaming" search engines, but also for the search engines' responses to such practices because they may also have serious implications for our political life (Hindman et al., 2003; Introna and Nissenbaum, 2000).

But how widespread of a problem is it, and how successful has it been? Since no search engine ever reveals the details of their algorithms or any adjustments on them, we decided to investigate how such "adjustments" affect search results. Unfortunately, it is impossible to measure the effect that link bombing may have had in the 2006 US Congressional elections, since the web is highly evolving and there are no publicly available comprehensive archives kept for extended periods of time to test this claim. The 2008 US Congressional elections, however, provided us with a unique opportunity. At a time when the attention of the public and the mainstream media was focused on the presidential race, we decided to follow the most contested races for the Congress.

The purpose of this research, therefore, was to examine and evaluate the results of any adjustments that Google did to avoid the gaming of its algorithms and the effect that these adjustments might have for the 2008 US Congressional Elections. In the following sections we describe briefly the data gathering process and three initial findings. After completing our analysis of the data collected, we plan to make them available to other interested researchers for further analysis.

2 Data Collection

The 2008 US elections had a presidential component, and a congressional component. The presidential elections were covered in great detail by the world media since they were considered of historical significance: their outcome would be either the first non-white US president, or the first non-male vice president. Due to the media interest and scrutiny, it is unlikely that any spamming efforts would go unnoticed. However, it is the Congressional elections that were allegedly spammed in the previous elections and that, even though they are very important for determining the formation of the legislative branch of government, were largely ignored by the media until the last couple weeks in the race. We decided to collect data from search results to follow the congressional elections.

There are 535 congressional seats in the USA, 100 in the Senate and 435 in the House. However, due to the electoral law, only 470 seats were contested in November 4, 2008. This included all 435 House seats, but only 35 out of the 100 Senate seats. Moreover, because the electoral system gives an advantage to the incumbent (the candidate currently holding the position), the vast majority of seats were not seriously contested. In early June, 2008, we decided to follow 59 races that were reported as highly contested, 11 in the Senate and 48 in the House.

We based our decision on what races to follow on two principal resources. First resource was the Electoral Vote Predictor (www.electoral-vote.com), created by the CS professor Andrew Tanenbaum in 2004, a website that analyzes all polls in order to identify challenging races (and also predict their outcome). The second resource was the Open Secrets website (openSecrets.org), maintained by the nonpartisan group "The Center for Responsive Politics", a website that tracks the amount of money raised by candidates and other groups during election campaigns. As a sanity check for our decision, we also consulted the websites of the two major parties: the Democratic Congressional Campaign Committee (www.dccc.org) and the National Republican Congressional Committee (www.nrcc.org), in order to find out what races were seen as important from their perspective, since it is known that these committees choose to support with financial contributions those candidates that will have a chance to win in contested races.

In early June, 2008, there was wide agreement among the resources we mentioned on 59 contested races, and we decided to follow them. We revisited our decision on October 24, 2008, ten days before the elections. As things had shifted in public opinion, at that time, news organizations reported that there were 77 contested races (Wandsness, 2008). Of those, 27 were marked as "tossup" (i.e., impossible to predict statistically) and 50 of them as "leaning" (i.e., one candidate had an advantage in the polls but still the race was too close to call). The contested races for the Senate had not changed from our monitoring group (still the same 11 races). However, the number of contested races for the House had rose to 66, or 18 more than we were tracking. All of our tracked races were included in this larger set.

2.1 Method of Data Collection

Since June 9, 2008, initially in weekly intervals and after September 15, 2008, in bi-weekly intervals, we automatically issued queries (the names of the candidates) to the Google API and collected the top 20 search results along with the total number of hits for each query. For each result in the ranked list we collected the back-links supporting them, and stored their HTML content.

Although Google urges the use of its API to query

the search index, the results of the API are reportedly not always identical to those returned when using the Google web user interface (WUI) (McCown and Nelson, 2007). We found that the collections of reported back-links is only a sub-sample of the data Google uses for its calculation of PageRank. Another difference we found is the aggregation of results from several sources into a single top-10 results page. In particular, searches for well-known candidates such as the Alaska senator Ted Stevens, or the comedian-turned-politician Al Franken, display results containing three special sections. On the top of the search results page appear fresh news results (from Google News), in the middle appear video results (mostly from YouTube), and at the bottom appear fresh blog posts. However, this does not seem to be a general practice and we found that the API results still represent a broad picture, though not identical to the results that web users using a browser will see. When comparing the two lists of the top 10 results, we find that the average overlap between entries is 0.88, while the average M-measure (that weighs more heavily similarity in the top of the lists), is 0.9.

2.2 Selecting Search Queries

It is reasonable to expect that many citizens use a search engine to be informed on the elections of their representatives. Although people might use various types of queries to search for electoral information, we decided that using the names of the candidates as queries would be a common way. After all, this is what the alleged attack in the 2006 elections reportedly did. An additional reason supporting our decision is the following: In one widely reported instance , the entrance into the electoral fight of a relatively little-known political figure at the time, Sarah Palin, was followed by a huge increase in web searches with her name as the query (Google Trends, 2008). This was preceded by the editing of the relevant entry in Wikipedia, apparently by someone closely connected with the candidate. (Noguchi, 2008)

This fact provides strong evidence that (a) web users will use a search engine to find out about a candidate and (b) those who care about a candidate will make sure there is positive news in the search results about them, especially in Wikipedia. Therefore, we created a list with the names of all candidates. With a script that run automatically once a week, each Monday night, we instructed the Google API to return the top 20 hits for each candidate name in the list, using as queries the candidates' names entered as phrases (using quotes), in order to avoid spurious results.

In addition, we also used the "link:" option of the Google API, to collect the backlinks reported by Google for each of the top-20 URLs. However, we were aware that the number of backlinks reported by Google is only a sample of the whole set of links that Google uses for its calculations of PageRank (McCown and Nelson, 2007). Finally, we also retrieved and stored the HTML content of each of the 20 top URLs, so that in a second phase, we can analyze the polarity of each page (pro candidate, against candidate, or balanced).

Sites	top-1	top-2	top-3	top-4	top-5
campaign	53.5	44.7	36.9	15.6	6.0
congress	38.6	27.9	11.1	3.8	1.5
wikipedia	4.6	21.3	26.7	32.9	14.0
Total	96.7	93.9	74.8	52.3	21.5

Table 1: Percentages of three groups of sites in the top-5 positions for Google search results. The query issued was the name of each candidate in the 59 most contested races for the 2008 US Congressional Elections.

3 Our Findings

3.1 Stable positions at the top-5

When it comes to the top-5 search results, we report two major findings:

(a) Averaged over 24 weeks, almost 70% of the top-5 search results belong to either the official campaign site of a candidate, the official congressional site (in case of an elected official), or the candidate's Wikipedia entry. In the top-2 positions, the dominance of these three categories is almost absolute, reaching in the mid-90s (see Table 1)

(b) These entries remained stable over the time, as measured by standard statistical metrics described in the literature (Bar-Ilan et al., 2006). As the study of the distribution in the top-5 positions shows (see Figure 2), each of the top-5 results moved very little during the observation time.

Moreover, the candidate's Wikipedia's page occupied mostly positions 2, 3 or 4 in the search results (see Fig. 3), with a very small overall reduction in positions as the elections drew near.

3.2 The role of blogs

After the elections of 2004, (Adamic and Glance, 2005) analyzed the linking patterns and discussion topics of political bloggers. One of their results was the identification of the top 40 blogs (20 blogs for

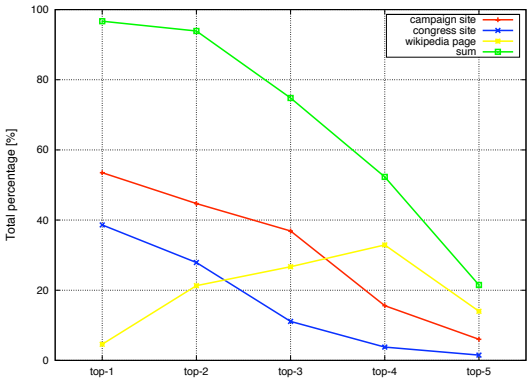

Figure 1: Visual representation of the data in Table 1.

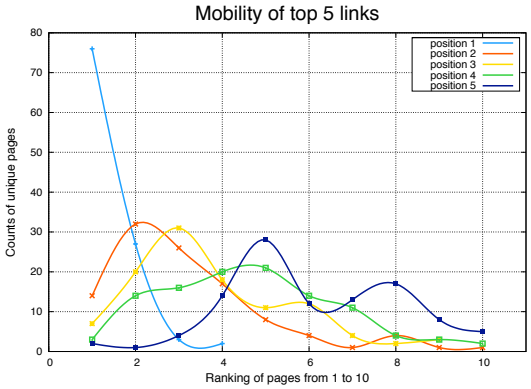

Figure 2: Distribution of top-5 link positions during the period of observation. For 76 candidates, their top result remained in top position during all this time. For 32 of them, their second result stayed in second position, and for 14 moved to the first position. Overall, there was not much movement in the search results for the top-5 items.

each of the conservative and liberal camps), based on the number of links pointing to them. We used this list of blogs to check how often these influential blogs appear among the top 20 results for the candidates.

Our results show that only 14 out of the 40 top political blogs of 2004 are among the collected results, and their cumulative count during the period of 20 weeks amounts only to 0.74% of the total count generated by 2029 different websites during this period. The top 5 blog sites are shown in Table 2.

However, when it comes to back-links of search results, at least 30% of them do belong to blogs. Moreover, the official websites of Democratic can-

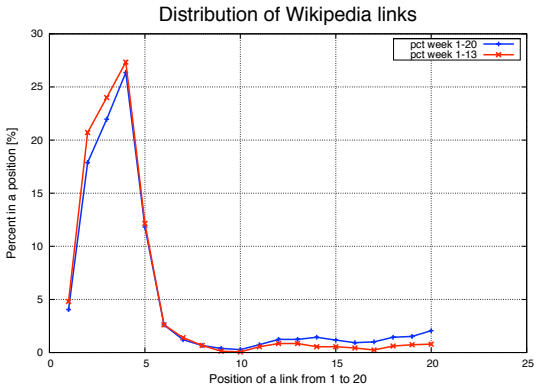

Figure 3: Distribution of the Wikipedia links in the top-20 search results. Wikipedia occupies a prominent place in the search results for congressional candidates. By week 13 it was in 89% of the top-5 positions. By week 20 this had fallen slightly to 82%.

Blog name	Counts	Avg. Position
mydd.com	103	13.3
dailykos.com	72	13
wonkette.com	66	16.8
realclearpolitics.com	24	10.4
michellemalkin.com	14	10.6

Table 2: The reportedly "most influential blog sites" in the 2004 elections appear in diminished positions in the search results of the 2008 elections. Counts are over a 20 week period.

didates have twice as many back-links compared to those of their Republican opponents (with the majority of these back-links coming from blogs). We are currently analyzing the role of back-links from blogs in pushing positive or negative stories about candidates. One interesting case is presented in the Conclusion section, below.

3.3 New Media vs. Traditional Media

Leaving aside the above-mentioned blogs, 25% of links in the top 10 results belong to only 12 websites. Among them, only one belongs to traditional media (The Washington Post), and the rest belong to web-only information sources. Three of the biggest social media sites, YouTube, Facebook and Flickr, are among them. All the rest belong mostly to non-partisan, fact gathering and checking sources such as sourcewatch.org, govtrack.us, ontheissues.org, etc., with two exceptions: therealdemocratstory.com (con-

servative) and actblue.com (liberal).

4 CONCLUSIONS

While there was an open effort to "game" the search engines in the 2006 elections, to the best of our knowledge, no such open effort was announced in 2008. Our results indicate that, if there was some overt orchestration to use anchor text to promote negative items up the search results for candidates, it did not materialize. Google, in particular, had a consistent pattern in the top-5 search results for political candidates. Their official web sites, their campaign web sites and their Wikipedia entries dominated the top-5 results.

But even though this was the prevailing pattern, it had some exceptions. In one case we analyzed, the Republican senatorial candidate for Louisiana had negative results consistently in his top-5 results. The negative sites rose to the top-2 position in mid-August (up from the 8th position in early June) and remained there, behind the Wikipedia entry, until the end. Further analysis of the bi-connected component (Metaxas, 2009b) of this negative result reveals that it was strongly supported by liberal bloggers.

There is no doubt that, in this election, the liberal activists were much more organized in their online strategy than their conservative counterparts. We found twice as many links from liberal sites supporting their candidates that conservative ones. The liberal New Politics Institute, (www.newpolitics.net) a liberal think tank that is credited with organizing online political activism, had reportedly targeted buying time on Cable TV, engaging the liberal bloggers, buying search ads and publishing for the Spanish-speaking population in 2006. In 2008, however, they created a far more sophisticated network by adding to their online political tools mobile phone tools, YouTube videos, and engaging social networks.

Going into the election, of the 22 Senatorial candidates, 7 were incumbents and of the 96 House candidates, 38 were incumbents. The candidates of the Democratic Party were able to win most of the contested seats, picking up 21 more seats in the House, and 8 more seats in the Senate that they previously held. As of the writing of this, one final senatorial seat (Minnesota) is still contested in the courts.

ACKNOWLEDGEMENTS

The authors would like to thank Professor Marion Just, Rebecca Graber and Michaela Wilkes Klein for their valuable contribution in selecting and evaluating the data.

REFERENCES

Adamic, L. and Glance, N. (2005). The political blogosphere and the 2004 u.s. election: Divided they blog. In *Conference on Knowledge Discovery in Data*.

Bar-Ilan, J., Mat-Hassan, M., and Levene, M. (2006). Methods for comparing rankings of search engine results. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 50(10):1448–1463.

Google Trends (2008). Sarah Palin's query trends. http://google.com/trends?q=sarah+palin.

Hansell, S. (2007). Google keeps tweaking its search engine. New York Times, June 3.

Hindman, M., Tsioutsiouliklis, K., and Johnson, J. (2003). Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political Science Association*.

Introna, L. and Nissenbaum, H. (2000). Defining the web: The politics of search engines. *Computer*, 33(1):54–62.

Mayer, M. (2005). Googlebombing "failure". http://googleblog.blogspot.com/2005/09/googlebombing-failure.html, September 16.

McCown, F. and Nelson, M. (2007). Agreeing to disagree: Search engines and their public interfaces. In *Joint Conference on Digital Libraries* .

McNichol, T. (2004). Engineering google results to make a point. New York Times, January 22.

Metaxas, P. T. (2009a). On the evolution of search engine rankings. In *In the Proceedings of the 2009 WEBIST Conference*.

Metaxas, P. T. (2009b). Using propagation of distrust to find untrustworthy web neighborhoods. In *In the Proceedings of the 2009 ICIW Conference*.

Moulton, R. and Carattini, K. (2007). A quick word about googlebombs. http://googlewebmastercentral.blogspot.com/2007/01/quick-word-about-googlebombs.html, January 25.

Noguchi, Y. (2008). Palin's wikipeda entry gets overhaul. http://www.npr.org/templates/story/story.php?storyId=94118849, August 29.

Sullivan, D. (2009). Obama Is "Failure" At Google & "Miserable Failure" At Yahoo. http://searchengineland.com/yahoo-obama-is-a-miserable-failure-16286, January 22.

The Pew Foundation (2008). Pew internet and american life project. http://www.pewinternet.org.

Wandsness, L. (2008). Even in deep-red states, GOP feels the heat. Boston Globe, October 24.

Zeller Jr., T. (2006). Gaming the search engine, in a political season. New York Times, November 6.