Protecting PageRank: Helping Search Engines Maintain Result Integrity

Cordelia Ludden
Tufts University

Helena Simson
Tufts University

Sarah Radway
Harvard University

Daniel Votipka
Tufts University

1 Introduction

We are increasingly reliant upon search engines to provide us with news—a survey by Pew Research identified that as of 2020, 68% of Americans get news from search engines such as Google Search [6]. Therefore, the content presented in these search results is of large importance.

Only top-ranked search results are likely to be seen: only 9% of Google users reach the bottom of the first page of results; a mere 0.44% examine the second page [2]. The order of Google Search results is determined by an algorithm called PageRank; PageRank considers many features, ranging from the website's age, to its frequency of keyword occurrence, to the website's domain authority [3].

Thus, the emergence of third-party services offering to manipulate Google PageRank results are of grave concern. In recent years, various websites have appeared that advertise their ability to manipulate the ranking of search results, providing "Removal & Suppression of Negative Search Results" [1] and "negative content removal and reputation repair services" [5]. These services are often branded as reputation management, intended for people or businesses to suppress negative press or reviews. The ability to suppress negative content allows individuals with financial power to decrease the online visibility of their wrong doings or other content harmful to their brands. This creates an inequitable system, and fuels distrust in search results.

A limited number of previous works have looked at specific features associated with PageRank modification. Most relevantly, Leontiadis et al. examines the frequency of search redirection attacks from established websites as a method of increasing the validity of sites associated with spam and fraud [4]. Xing et al. investigated how search result personalization allows for search history pollution attacks to modify Google Search results [7]. However, to the best of our knowledge, there has been no previous work investigating PageRank manipulation at scale.

In this work, we investigate this threat by observing changes in search rankings over time, to identify possible PageRank manipulation. We are currently collecting Google Search result data for 2,465 individuals of political and corporate prominence, who would have elevated concern for their public perception. We are searching these individuals names over the course of 6 months (May-November 2024), and collecting the results. We track changes in the order of results returned for these queries, and use both analysis methods (1) from cybersecurity-based anomaly detection and (2) based in sentiment analysis to determine features organizations can use to identify artificial PageRank manipulation. In this way, search engines can ensure they employ this knowledge to potentially protect against bad actors. We present initial findings here, and discuss future steps forward for this work.

2 Methodology

As this is a work in progress, we present the current methodology which we have implemented for data collection, and outline our full analysis plan.

2.1 Data Collection

Over the course of six months, we are searching the names of individuals with significant interest in their public-facing image. Specifically, we are searching 1965 individuals running for U.S. Congress in 2024, and 1462 individuals that are CEO/CFO/CTOs of Fortune 500 companies. Each day, for each query, we collect the top results appearing in Google, and scrape the associated pages' content. We target search results to a specific location (a city in the district of each political candidate, and Washington DC for the C-level executives), as this is one of the main ways that Google targets results to individuals.

2.2 Evaluation

We categorize changes in page rank into two categories: (1) natural changes (caused by a real world event triggering a change of the top pages) and (2) artificial changes (where

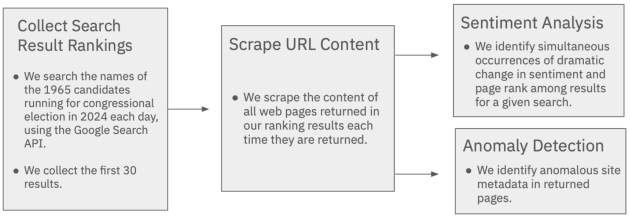

Figure 1: High level overview of analysis methodology

the top pages increased or decreased with no observable real world stimuli). We identify potential artificial changes through a combination of sentiment analysis and identification of anomalous metadata.

2.2.1 Sentiment Analysis

We will be performing extensive sentiment-based evaluations of the collected page content. Specifically, we are implementing a relational approach to sentiment analysis, where we evaluate the page's sentiment with respect to the relevant searched politician. We first create a dataset of political articles about currently serving congressional representatives, who are not up for re-election. A pair of researchers go through each article, assigning them as being in favor of a given candidate or criticizing a given candidate on a 5-point Likert scale. We then train a classifier on these labeled articles and politician names, and apply this classifier to our own examined articles, to produce a 5-point Likert-scale value for the article's sentiment towards a given candidate. Given this sentiment value for each page in our rankings, we are able to contextualize changes in their ranking.

We identify large changes in search results accompanied by large changes in sentiment. When we detect a large change in sentiment in combination with a large change in the content and order of results, we search for a real-world event that occurred, looking through the provided search results, and consulting local news sites to confirm. If there was a real-world event occurring on the given date, this suggests a natural change. However, if there is not a real-world event, this suggests potential PageRank manipulation.

2.2.2 Anomaly Detection

To further assert this manipulation, we perform an anomaly detection-based evaluation, focusing on features associated with PageRank: namely, back link quality and count, external link quality and count, and domain authority. We will be using these features to identify anomalous results, that suggest tampering (in line with early studies of search poisoning [4]). We will use these features to likewise identify what constitutes abnormal results in the wild.

3 Initial Results

While we currently have minimal data, we have a series of initial findings, that we would be interested to receive feedback and suggestions on from both reviewers and the broader SOUPS community.

3.1 Churn

In order to provide a baseline for search result change, we measure both entry/exit churn and rank churn. We collect a week's worth of the first page of search results for election candidates, measured across each run, and then from start to finish across the week. Entry churn/exit churn represents the number of entries that entered/exited the list. We see that the mean entry churn across all candidates is 1.52 (meaning that about one entry enters or leaves the rankings throughout the week). Rank churn represents the absolute change in ranking of entries that remained on the list throughout the week. This value represents the sum of the differences in positions for items that remained in the lists from start to finish. We see that the mean of the summed rank churns across all candidates is 5.24, with the mean per site being 0.64.

This means that while generally the list of top results stays the same, their relative ranking will change to some degree over time. However, we can reasonably expect that an anomalous event is occurring if an item drops more than five positions, as on average, all results are only changing by a combined total of around 5.

3.2 Qualitative Observations

It appears that most high-ranked websites are candidate-affiliated campaign pages, government sites for currently elected officials, or listings of all elected candidates. We collect the ten most common keywords across websites, using NLTK, excluding stop words. Some of the most common keywords included U.S., district, election, news, candidate, and primary, each present in around a third of articles. We believe that as the number of scandals increases as election dates grow nearer, we may see substantial changes in the content of top-ranked pages.

4 Discussion

It can be challenging to tow the line between 'search optimization' and foul play. We hope that this study, when completed, will allow us to move closer to identifying search result optimization, in an effort to increase transparency surrounding those with the ability to afford these services. We believe that our initial steps in setting baseline observations for benign search behavior will allow us to uncover PageRank manipulation at scale in the coming months, particularly as the 2024 election cycle unfolds.

References

[1] Reputation 911. We make you look good online. https://reputation911.com.

[2] Brian Dean. How people use google search (new user behavior study), 2023 August. https://backlinko.com/google-user-behavior.

[3] Mr Anuj Joshi and Priyanka Patel. Google page rank algorithm and it's updates. In *International Conference on Emerging Trends in Science, Engineering and Management, ICETSEM-2018*, 2018.

[4] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 930–941, 2014.

[5] Google Reputation Manager. Google search reputation management repair. https://googlereputationmanager.org.

[6] Elisa Shearer. More than eight-in-ten americans get news from digital devices, January 2021. https://www.pewresearch.org/short-reads/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital#:~:text=About%20two%2Dthirds%20of%20U.S.,%2C%20like%20Google%20(65%25).

[7] Xingyu Xing, Wei Meng, Dan Doozan, Alex C Snoeren, Nick Feamster, and Wenke Lee. Take this personally: Pollution attacks on personalized services. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 671–686, 2013.