Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation

Ronald E. Robertson
Northeastern University
rer@ccs.neu.edu

Shan Jiang
Northeastern University
sjiang@ccs.neu.edu

David Lazer
Northeastern University
d.lazer@neu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

Autocomplete algorithms, by design, steer inquiry. When a user provides a root input, such as a search query, these algorithms dynamically retrieve, curate, and present a list of related inputs, such as search suggestions. Although ubiquitous in online platforms, a lack of research addressing the ephemerality of their outputs and the opacity of their functioning raises concerns of transparency and accountability on where inquiry is steered. Here, we introduce recursive algorithm interrogation (RAI), a breadth-first search method for auditing autocomplete by recursively submitting a root query and its child suggestions to create a network of algorithmic associations. We used RAI to conduct a longitudinal audit of autocomplete on Google and Bing using a focused set of root queries – the names of 38 US governors who were up for reelection – during the summer of 2018. Comparing across search engines, we found a higher turnover rate among longer and lower ranked suggestions on both search engines, a higher prevalence of social media websites in Google's suggestions, a higher prevalence of words classified as a swear or a negative emotion in Bing's suggestions, and periodic shocks that spanned across most of our root queries. We open source our code for conducting RAI and discuss how it could be applied to other platforms, topics, and settings.

CCS CONCEPTS

• **Information systems → Content ranking**;

KEYWORDS

search queries; autocomplete; suggestions; algorithm auditing

ACM Reference Format:
Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3292522.3326047

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326047

1 INTRODUCTION

When people seek out information beyond their social circles, they are generally limited to one of two options. They can either seek it out on their own, or seek out the assistance of a librarian – a specialist who manages bodies of information – and begin the complex, exploratory, and iterative process of formulating, communicating, and negotiating a *query* [3, 31, 48]. Historically, people would rely on libraries for this type of assistance [7], but today, libraries have been largely replaced by web indices, and librarians by algorithms. To illustrate, recent surveys have found that 82.3% of American adults use search engines one or more times a day [12], but only 48% of Americans over the age of 15 have visited a public library in the last year [25].

Among the most widely used and trusted information-seeking mediums of today are web search engines [4, 13, 15, 16, 19, 37, 49], and particularly analogous to the process of query negotiation with a librarian, is the interactive assistance that search engines offer users as they input their queries [23, 35]. On Google and Bing, which together account for 84.8% of all search engine traffic [36], this assistance comes in the form of algorithmically generated search suggestions. Much like how librarians save people time by guiding them to the right sections of a library, Google estimates that its suggestions reduce "typing by about 25 percent" and save "over 200 years of typing time per day" [18].

Given the immense amount of traffic that autocomplete algorithms receive [36], the influence that they have on the queries that people view and click (through filtering and ranking) [24, 34], and the inherent ephemerality of their output, it is crucial that we develop methods for mapping not only the explicit suggestions that they provide, but also the implicit associations underlying those suggestions [11, 27, 39, 52]. Such maps could potentially increase the interpretability of their outputs, help platform moderators proactively identify the inappropriate associations that autocomplete has a history of making, and enable researchers and the public to hold them accountable for those associations [39].

In this paper we present methods for preserving, mapping, and analyzing the autocomplete search suggestions that people are exposed to while conducting web searches. To construct these maps, we introduce *recursive algorithm interrogation* (RAI), a breadth first search method for auditing autocomplete by recursively submitting a root input and its child suggestions to recover their underlying associations. The data resulting from this process can be modeled as a weighted and directed tree-like network that we refer to as *suggestion networks*; where nodes are suggested items, links are

algorithmic associations, and link weights are derived from each suggestion's ranking.

We conducted an exploratory experiment using RAI in which we held web identity constant by submitting root queries from a single server with a fixed location and user-agent. Using the names of 38 US Governors as root queries, we initiated RAI on Google and Bing's autocomplete algorithms in parallel twice a day, at 9am and 6pm, for approximately 10 weeks between June and August 2018. We then used a combination of descriptive statistics, NLP, and information theoretic measures to compare the suggestion networks produced by each search engine. We found similarities in the structural and linguistic bounds of their suggestion networks, but substantial differences in their content and temporal dynamics. For example, Google was twice as likely as Bing to suggest social media (especially YouTube), and Bing was more likely to make a suggestion that contained a word classified as a swear or a negative emotion. We also found a higher turnover rate among longer and lower ranked suggestions, and periodic shocks that spanned across most of our root queries and which could potentially indicate algorithm updates.

Overall, our work makes the following contributions:

- We introduced RAI, a generalizable method for mapping the associations generated by autocomplete algorithms.
- We used RAI to conduct the first parallel and longitudinal audit of Google and Bing's autocomplete algorithms.
- Our results suggest that RAI could be an effective and valid tool for expanding a set of root queries.
- We found patterns of periodic shocks that affected the suggestions produced for most root queries, potentially identifying algorithm updates.
- We open source our tools for conducting RAI and constructing association networks, with the hope that they will spur further research in other topic areas.

Outline. The rest of the study is organized as follows. First, we examine documentation on and prior audits of autocomplete (§ 2). Next, we introduce the components of RAI (§ 3), and describe the results of our exploratory audit using RAI and US Governors names as root inputs (§ 4). Finally, we discuss our limitations (§ 5) and findings (§ 6).

2 BACKGROUND

In this section, we review official documentation for Google and Bing's autocomplete algorithms and the limited number of audits conducted on them.

Autocomplete Documentation. While both Google and Bing offer some documentation for how their autocomplete algorithms work, the data and decisions governing these algorithms are largely opaque. The documentation for Google's autocomplete system states that their suggestions are based on factors including the terms you type, the popularity and freshness of those terms, your search and browsing histories, and trending topics in your area [17]. Documentation for Bing's autocomplete functionality is less explicit, though it also appears to be based on the characters entered and the popularity of terms [6]. Both search engines also provide factors which might lead one to not see certain suggestions, including suggestions that (1) contain disparaging or sensitive terms, (2) violate a policy regarding sex, hate, violence, and/or dangerous speech, (3) are not novel enough, or (4) are not popular enough.

Given the ephemerality of their output and the opacity and lack of regulatory oversight on how autocomplete works [27], the influence that these systems have over user inquiry raises concerns of transparency and accountability. Among these are concerns about the data that is used to train them, the stereotypes and biases that they might implicitly or explicitly evoke, and the specific censorship rules that govern them. Within this line of concern, and motivating the root selection for our audit, Google has previously been criticized for it's suggestions related to political actors. For example, in mid-2016, people found that searches for "lying" returned "lying ted" as a suggestion, in reference to Republican candidate Ted Cruz, but "crooked" did not return "crooked hillary" as a suggestion, in reference to Democratic candidate Hillary Clinton [29]. Similarly, Google has come under international criticism and entered a wide array of legal disputes related to their autocomplete algorithm generating a "combination of words that [are] capable of conveying a deceitful or misleading message" [27].

Autocomplete Audits. One principled method for outlining patterns of algorithmic decision making is known as the *algorithm audit* [45]. This framework involves feeding an algorithm a set inputs while systematically varying who is asking, what they're asking, where they're asking from, and when they're asking. Examples of how this method has been applied include studies of political bias, personalization, and localization on Google [20, 26, 28, 42], racial and gender discrimination in the gig economy [14, 22], and dynamic pricing on Amazon [8, 21, 33]. Prior audits of autocomplete, however, are relatively scarce, and include informal Search Engine Optimization (SEO) blogs [46, 53], investigative computational journalism reports [10, 11], and a peer-reviewed article from critical discourse researchers [1].

The SEO industry posts were focused on how to manipulate Google's autocomplete to make positive suggestions for a client's name rise in the ranks, and how to make negative suggestions disappear [46, 53]. This research demonstrated that, in 2013, one could directly influence search suggestions by creating a crowdsourcing task where participants were instructed to enter a specific search query and click on the first result listed [46]. Furthermore, it demonstrated that it took approximately one week for the changes to take place, indicating significant lag in Google's autocomplete updates at the time.

The computational journalism reports provided a more rigorous examination, and focused on the topics of censorship and defamation. In the first study, Diakopoulos focused on the suggestions returned by Google and Bing for various sex and violence-related words [11] – topics that Google explicitly states it excludes from autocomplete [17]. He found that certain words were censored on Google – returning no suggestions – while others were not, and the differences were somewhat expected from Google's autocomplete FAQ, but were also somewhat arbitrary depending on how the query was formulated. In the defamation study, Diakopoulos looked at whether entering the names of public figures and corporations on Google's autocomplete produced suggestions that could be

considered defamatory, finding that name disambiguation makes it hard to tease apart the associations between such queries and their suggestions [10].

In the critical discourse paper, Baker and Potts (2013) "interrogated" (inspiring the name of our method) Google in April 2011 by submitting a series of carefully crafted questions and documenting the outputs [1]. After selecting 12 identity groups (*e.g.,* Black, Muslim, Gay), the authors mapped these groups into 2,690 question fragments by adding a *wh-* starting string (*e.g.,* why do, what do, where do) or an auxiliary fronting (*e.g.,* should, are, do) to each group. Overall, Baker and Potts found that their method elicited questions that made relatively distinct associations about each group, including physical characteristics for Jewish and Black people, and negative stereotypes about Gay people. Although they did not consider how the presence of these stereotypes in autocomplete might affect users, they noted that, at the time, no method existed for users to flag inappropriate suggestions, a feature which Google has since added.

Although illuminating, most of these studies were never peer-reviewed or automated and scaled up, leaving open questions on how to audit autocomplete. While other studies have also been conducted on autocomplete, their primary focus was on user engagement and improving relevance, a different, more internally focused type of audit [2, 24, 34]. To the best of our knowledge, none of the studies conducted thus far on autocomplete have examined the structures and associations that emerge when collecting suggestions recursively.

3 METHOD

In this section we describe the three components of RAI: selecting root inputs, conducting a breadth first search, and trimming the resulting suggestion networks. In our application of RAI, we explicitly did *not* study personalization or user generated queries. Instead, we used a fixed set of *root queries* and held web identity constant by submitting queries from a single server with a fixed location in the Northeastern US. This approach enabled us to map the autocomplete search suggestions for a fixed set of queries – from the perspective of a fixed user with no history – and measure their change over time. That is, our audit is not focused on human behavior, but machine behavior: the algorithm is our subject [41].

Root Selection. To seed RAI, we used the names of 38 US governors (26 Republican, 7 Democratic, 2 Independent, and 1 Democrat-Farmer-Labor) as our *root queries*. US Governors are popularly elected officials who serve four-year terms as "chief executive officers of the fifty states and five commonwealths and territories" [38]. We selected these roots for two primary reasons. First, because Google has previously been criticized for it's suggestions related to political actors [29]. Second, we used these root queries because 15 of the governors were up for reelection in 2018. These elections were spread across six days during our data collection window: 2018-06-05 (5), 2018-06-12 (3), 2018-06-26 (3), 2018-08-02 (1), 2018-08-04 (1), and 2018-08-07 (2). These events, tied directly to the root input, provided opportunities to measure the impact of external shocks on Google and Bing's autocomplete algorithms.

Breadth First Search. To conduct a breadth first search on Google and Bing we first identified two URLs that we could leverage as APIs. We then designed a program to submit a single *root query* to each search engine in parallel, add each root's suggestions to its respective queue, and then recursively repeat this process until the queue was extinguished or until the process reached a maximum *depth* – the number of steps from the root – that we set at 8 for this study.[1] For example, using the name of the Massachusetts governor, "charlie baker," as our root, Google returned a set of suggestions, including "charlie baker email," "charlie baker height," "charlie baker twitter," "charlie baker salary," and "charlie baker approval rating." We then recorded the rank, depth, and time of collection for each of these, and then submitted each of them to Google, and so forth, generating a tree-like directed network structure that we refer to as *suggestion networks*, where the nodes are n-grams and the links indicate which node suggested which. While building these networks, we did not collect duplicate edges resulting from cycles. If we observed a suggestion that we had already seen, we drew the link but did not add the suggestion to the queue. If a given node was linked to twice or more, we kept its depth and rank from the first occurrence.

Suggestion Network Trimming. Given that the goal of our audit was to examine the autocomplete associations relevant to the root queries we had selected, we explored our data for cases where the suggestions obviously deviated from the root input. We identified two cases that resulted in what we call *emergent roots*: a suggestion that (a) is not relevant but somehow related to the original root query (*e.g.,* "governor of california jerry brown biography" → "biography"), and (b) initiates a new and unrelated branching process (*e.g.,* Figure 1).

The first case of emergent roots involves a type of conceptual network teleportation. For example, in a network built for the root query "scott walker," we found an edge from "scott walker wisconsin recount" to "recount." The query "recount" then began spawning its own suggestion network, including morphological and informational suggestions – such as, "recounting," "recounted," and "recount definition" – that no longer had direct relevance to the original root query (Figure 1).

The second case involves root disambiguation, and occurs when there is a well-known person with the same name as, or a name similar to, the root query. For example, for the roots "matt mead" and "kate brown," the autocomplete algorithms began suggesting morphological variants, such as "matt meadows" or "matt meader" and "kate abramson" or "kate brannan." The breadth first search would then continue on with these new names, each functioning as its own emergent root, producing a large number of suggestions unrelated to our root query.

Both of these cases were more prevalent on Bing, where the output of RAI surged at depth 5 and onward due to emergent roots (Figure 2).[2] To trim our networks of these, and reduce the amount of noise in the suggestion networks we were building, we used a

[1] As a practical matter, we limited the depth because of the amount of time that it took to reach greater depths, and because submitting queries too quickly increased the risk of being rate limited.

[2] Note that the high variability at greater depths is due to a small number of networks that produced nodes at those depths.

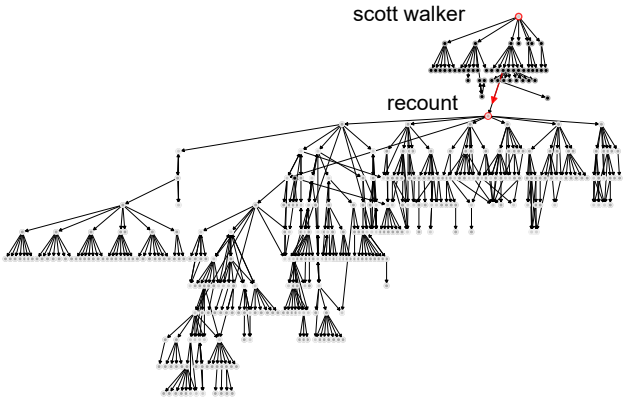

Figure 1: An example of our trimming on a suggestion network for the root "scott walker." The red edge leads from "scott walker wisconsin recount" to "recount," an emergent root that we trimmed from our suggestion networks in order to maintain their relevance to the root queries we selected.

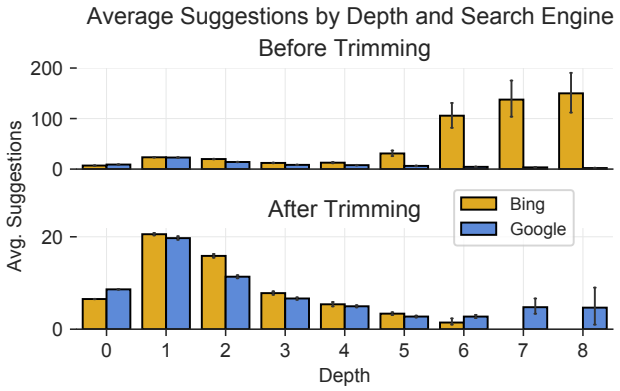

Figure 2: Before trimming, the suggestion networks produced by Bing were substantially larger at greater depths. However, after trimming the two cases of emergent roots we identified, the suggestion networks provided by Google and Bing were well aligned.

simple and conservative rule: trim any edges in which the target node did not contain both the first and last name of the root query. The result was largely a reduction in Bing's suggestions networks at depth 5 and onward, removing emergent roots, and leaving a distribution that more closely matched Google's (Figure 2).

4 RESULTS

Here we examine the suggestion networks that we collected using RAI and our root queries. More specifically, we describe and compare their structural and linguistic features, examine their change over time, and explore a method for reducing them down to a web of associated n-grams.

4.1 Structural Features

To characterize the structure of our suggestion networks, we calculated several canonical network science metrics. Out-degree (k_{out}), the number of out-bound links from a given node, has a clear interpretation (*i.e.,* the number of suggestions it produced). However, in-degree (k_{in}) has a less straightforward meaning. Excluding the roots, all nodes must have an in-degree of at least one (*i.e.,* they must have been suggested), so $k_{in} > 1$ is the result of converging suggestions, meaning that a node is relevant to multiple queries. These merge points could indicate a potentially important or ubiquitous association.

Most networks had right-skewed in- and out-degree distributions (Figure 3), indicating that most of the queries we submitted produced relatively few suggestions. This relationship varied by search engine, with Google never producing more than ten suggestions ($\mu = 1.2, SD = 2.1$), and Bing never producing more than eight ($\mu = 1.1, SD = 1.9$).[3] Using a Spearman's correlation, we found that as depth increased, the number of suggestions decreased for both search engines, but less so for Google than Bing (Google $\rho = -0.11^{***}$; Bing $\rho = -0.34^{***}$).[4] Conversely, the number of merge points increased with depth for both search engines (Google $\rho = 0.52^{***}$; Bing $\rho = 0.43^{***}$).

4.2 Linguistic Features

We explored the linguistic features of suggestions networks in three respects: basic query characteristics, lexicon classifications, and the mentions of social media.

Query Characteristics. Considering all queries (*i.e.,* all roots and their child suggestions), the average query in our dataset consisted of 4.6 words ($SD = 1.2$), and this feature was not substantially different across Google and Bing (Figure 4). This finding aligns with prior work on real users in 2012, which found that the queries conducted on Google and Bing were on average 4.3 words [9], suggesting that RAI may sample queries from a similar distribution. However, our findings diverge from a 2010 report which indicated that 54.5% of queries conducted on Google were greater than three words [32]; 80.4% of the queries we conducted were greater than three words. This difference is likely a result of our root queries all being two words long, and the tendency for the autocomplete algorithms of both search engines to add rather than remove words when providing suggestions.

Indeed, in terms of the change from query (source node) to suggestion (target node), the difference in the number of words (target - source) was somewhat normally distributed around one (Figure 4). Overall, relative to their parent query, 87.7% of suggestions were longer by one or more words, while 11.8% of suggestions had the same number of words (*e.g.,* "asa hutchinson bio" → "asa hutchinson biography"), and the remaining 0.5% of suggestions decreased by one or two words (*e.g.,* "nathan deal new laws" → "nathan deal news").

[3] This maximum does not appear to be temporally stable – research conducted in 2017 found a maximum of only 4 suggestions for Google [43].

[4] Throughout the rest of the paper use the standard significance notation for P values, *** P < 0.001, ** P < 0.01, * P < 0.05

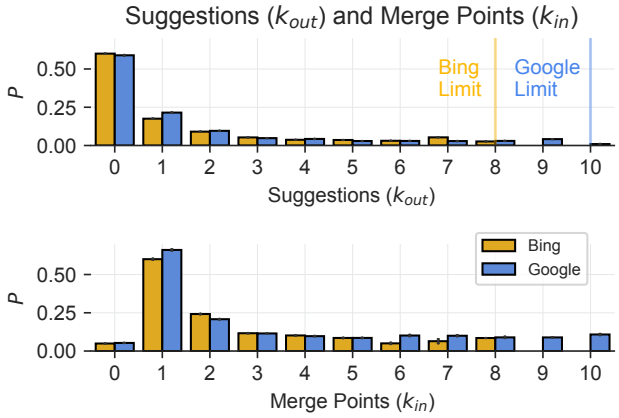

Figure 3: Mean probability of a query producing k_{out} suggestions (top) and being suggested k_{in} times (bottom). The blue and yellow vertical lines represent the maximum number of suggestions that Google and Bing provided, respectively.

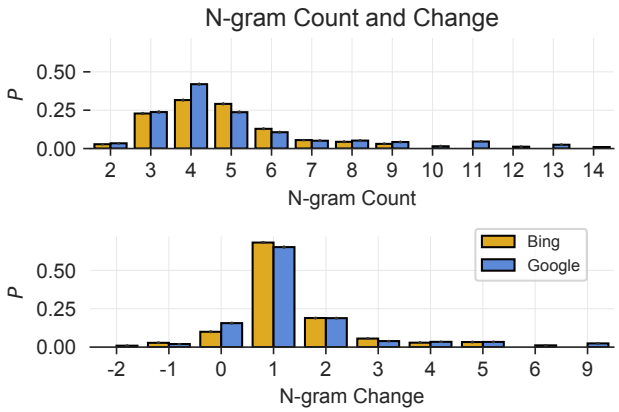

Figure 4: Most suggestions contained four or more words (top), and the majority of suggestions increased by one word or more relative to the query that produced them (bottom).

With respect to the interaction between the linguistic and structural features of suggestion networks, we found that longer queries are not only more likely to appear at greater depths, but they are also less likely to produce more suggestions. More specifically, query length (the number of words in a query) was positively correlated with depth (Google: $\rho = 0.61^{***}$; Bing: $\rho = 0.69^{***}$) and negatively correlated with k_{out} (Google $\rho = -0.33^{***}$; Bing $\rho = -0.50^{***}$).

Lexicon Classifications. To attach qualitative context to the suggestions that we collected, while maintaining the automated nature of our method, we parsed our suggestions for n-grams that matched with words in the Linguistic Inquiry and Word Count lexicon (LIWC). LIWC is a widely used lexicon that was compiled by social scientists for classifying the psychological meaning of words [40, 47].

We observed several cases where suggestions were classified as containing "swear" words. For example, Bing returned the suggestion "gun control **idiot** jerry brown signs bill" at depth four and rank one for the root "jerry brown" (its immediate parent query was "jerry brown gun control"). This suggestion was short-lived, however, appearing at 9am and disappearing at 6pm on June 27, 2018. In contrast, the only Google suggestion that was classified as "swear" was "gina raimondo **freak**onomics" for the root "gina raimondo" at depth zero, rank nine. However, as people familiar with the popular economics book *Freakonomics* will recognize, this is clearly a misclassification by LIWC, and upon further investigation, we found that the suggestion likely resulted from Governor Raimondo appearing on a podcast of the same name. While the Jerry Brown examples demonstrate how suggestions can cast politicians in a negative light, the Gina Raimondo example stresses the importance of examining suggestions in the context of real world events.

While the cases for swear words are somewhat anecdotal, search engines frequently suggested words with negative emotions (i.e., "negemo" for LIWC) to their users. For example, Bing produced the suggestions: "governor kate brown of oregon"→"kate brown governor of oregon **bad** law", "scott walker wisconsin economy"→"wisconsin economy **failing** under scott walker", and "andrew cuomo bar exam"→["andrew cuomo **failing** the bar exam 4 times", "andrew cuomo **failing** the bar exam"]. However, suggestions classified as containing negative emotions typically only survived for less than 10 days. In total, Google produced suggestions containing negative emotions for three (7.8%) root queries and Bing for six (15.8%) root queries.

Social Media Suggestions. Given the growing concerns about how web search can funnel users towards social media (*e.g.,* Google's embedded Twitter and YouTube results [42]), we examined the rate at which major social media platforms – specifically, Facebook, Twitter, Instagram, LinkedIn, Reddit, and YouTube – were suggested in autocomplete. Overall, we found that Google's suggestions included references to social media twice as often (10.6%) as Bing's (4.6%). We also found differences by root, with Google including a suggestion to Twitter for all 38 roots, while Bing only did so for 32 roots. Suggestions mentioning Facebook were more equal (Google: 25, Bing: 23). Conversely, Bing mentioned Instagram for seven roots, while Google only did so for two. There were also substantial differences among suggestions that mentioned YouTube, where Google referenced it's sister platform for 14 roots, while Bing only mentioned YouTube for four roots.

There are several caveats here worth mentioning. Although we trimmed suggestions that did not explicitly mention the root, there are some disambiguation challenges that this did not resolve. For example, while both Google and Bing linked Dan Malloy, the governor of Connecticut, to Instagram, these suggestions typically also mentioned "surf" or "surfer." We looked into this and found that the 63 year old governor was not in fact an avid surfer, but there is another Dan Malloy who is. Similarly, Bing provided a suggestion for California governor Jerry Brown that mentioned Facebook, "jerry brown news3lv facebook," but this was in reference to the

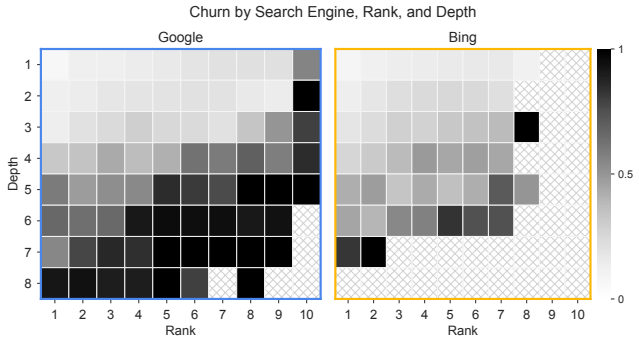

Figure 5: As rank and distance from the root increases, the suggestions of Google and Bing become less stable.

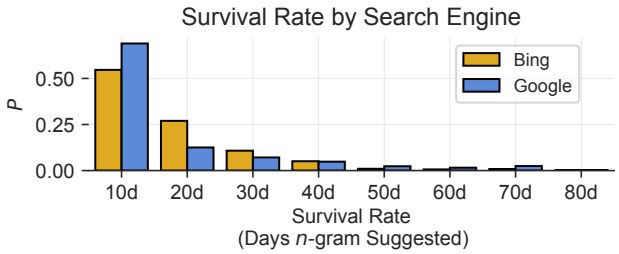

Figure 6: Most queries only appeared for a handful of days, indicating a generally high turnover rate among the suggestions made by search engines, as characterized by our method.

meteorologist Jerry Brown for News 3 Las Vegas. These findings highlight the challenges in automated methods for name disambiguation, especially in autocomplete [10], and again emphasize the need to consider context when evaluating suggestions.

4.3 Temporal Features

We explored the temporal change of suggestion networks in three ways, first by quantifying the *churn* of nodes, second by examining the *survival rate* of nodes, and third by calculating normalized mutual information over time. We found that the greatest variability among the suggestions produced by Google and Bing occurs among the lowest ranked nodes, that most nodes have a relatively short life lifespan, and some evidence that Google and Bing's autocomplete algorithms periodically shift in the suggestions they deliver. These shifts might occur due to real life events directly or peripherally involving the roots (*e.g.,* their primary elections), but also might occur when the algorithms are being updated.

Churn. Given the impact of depth on suggestion networks' structural and linguistic features, we calculated churn by examining each root query's suggestion network, and asking how often the n-grams at each rank and depth changed between that network and the one that we collected on the next crawl. More formally, let $S_{e,r,d}(t,q)$ represent the set of suggestions at time $t \in T$ for root query $q \in Q$ from search engine e at rank $r \in R$ and depth $d \in D$. Churn $c_{e,r,d}(t,q)$ is defined as the Jaccard Index between time t and $t+1$, i.e.,

$$c_{e,r,d}(t,q) = 1 - \frac{|S_{e,r,d}(t+1,q) \cap S_{e,r,d}(t,q)|}{|S_{e,r,d}(t+1,q) \cup S_{e,r,d}(t,q)|}. \quad (1)$$

Then, we aggregate $c_{e,r,d}(t,q)$ over time t and root query q to get average churn $\bar{c}_{e,r,d}$ from search engine e at each rank r and depth d, i.e.,

$$\bar{c}_{e,r,d} = \frac{1}{|Q|}\frac{1}{|T|}\sum_{q\in Q}\sum_{t\in T} c_{e,r,d}(t,q). \quad (2)$$

We found that the churn for Google and Bing followed a similar pattern: turnover was higher for lower ranked suggestions and suggestions that appeared at greater depths (Figure 5). We also examined correlations between churn and depth (Google: $\rho = 0.43^{***}$; Bing: $\rho = 0.54^{***}$), churn and rank (Google: $\rho = 0.10^{***}$; Bing: $\rho = 0.20^{***}$), and churn and query length (Google: $\rho = 0.29^{***}$; Bing: $\rho = 0.48^{***}$). While these correlations point towards the importance of depth and query length, it is also important to consider them in light of our findings from § 4.1, where we found that query length was positively correlated with depth (longer queries at greater depths) and out-degree was negatively correlated with depth (fewer suggestions at greater depths).

Taking the relationships between query length, depth, and churn into account, we examined the relationship between churn and rank at each depth and found the greatest correlation at depth zero (Google: $\rho = 0.39^{***}$; Bing: $\rho = 0.47^{***}$). Together, these findings indicate that (1) longer queries are more volatile over time, and (2) highly ranked suggestions tend to be more stable over time. The latter suggests a potential cumulative advantage for the associations that make it into highly ranked positions, in terms of user attention, because of the rank-biased way that users examine suggestions [24, 34]. This stability may also help to prevent attempts to game or manipulate the rankings, but our results are descriptive and we can only speculate on this matter.

Survival Rates. To examine the survival rate of suggestions, we tracked the number of days that each unique suggestion persisted across our data collection window. Using this count, we found that the majority (Google: 70%; Bing: 54.6%) of suggestions appeared for ten days or less (Figure 6), while less than 1% of suggestions (Google: 0.2%; Bing: 0.1%) appeared for the entire duration of our crawl. Given our previous findings related to query length, we also examined the relationship between query length and survival rate, but found no significant differences. We also examined the differences between the two search engines with respect to the LIWC classifications by ranking each LIWC category according to its mean survival rate. We found that these rankings were not correlated ($\rho = 0.17, P = 0.16$), indicating that the search engines' autocomplete algorithms deviate in the types of associations that they make for US politicians, according to LIWC classifications. These results indicate that for the suggestions that do churn, their life span is relatively short, potentially due to the presence of trending topics (or "freshness") as indicated in the autocomplete documentation.

External Shocks. To explore how suggestions might react to real world events, we examined the rate of change for each root's

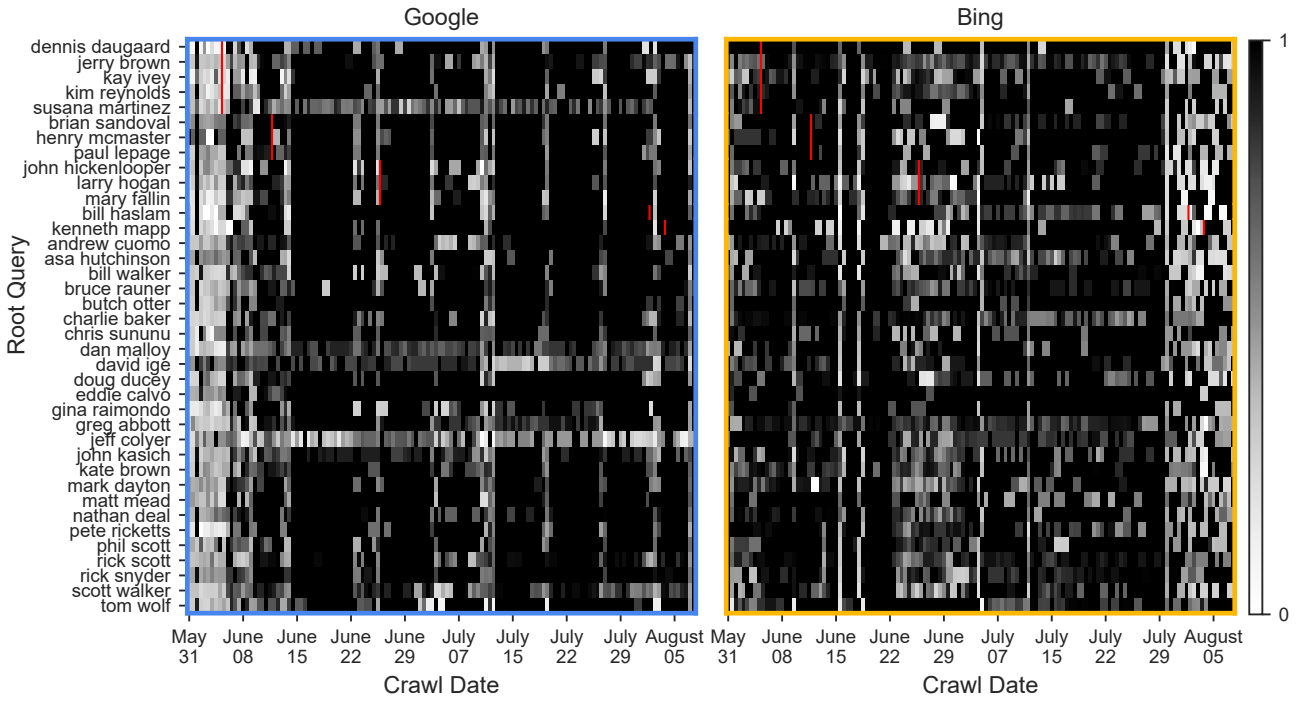

Figure 7: Normalized mutual information (NMI) over time and by search engine. A value of 1 indicates that there was no change in a root's information set since the previous time step, while a value of 0 indicates a complete change. Roots are clustered and arranged vertically by primary, and the red vertical lines represent the date of the primary for each clustering of governors.

suggestion network over time using normalized mutual information (NMI), a measure commonly used when comparing the overlap between two information sets [54]. At a high level, NMI measures the difference between two distributions and returns a value that scales from 0 to 1, where 0 means the distributions were completely different, and 1 means that they were the same. Here we explicitly ignore the ranking and depth of suggestions, as our primary interest is measuring any type of change in the information being retrieved for a given root query. While our analysis here was exploratory, our expectation was that external events would provide some sort of shock to the system, perhaps due to an increase in search volume.

We calculated NMI by first extracting all of the words present in each suggestion network at each time step. We excluded the root queries during this tokenization process to prevent them from inflating similarity. Then, for each root, we moved across each time step in our data collection window and measured the NMI for that root query between time t and time $t+1$. More formally, following the notation we used for churn (Equation 1) and using $H(A)$ to denote the entropy of A, and $I(A;B)$ to denote the mutual information between A and B, we calculated NMI as:

$$NMI_{e,r,d}(t,q) = \frac{2 \times I(S_{e,r,d}(t,q); S_{e,r,d}(t+1,q))}{[H(S_{e,r,d}(t,q)) + H(S_{e,r,d}(t+1,q))]} \quad (3)$$

We then plotted this relationship over time, marking the gubernatorial primary dates of the 15 governors who had a primary during our data collection period (Figure 7). While we did not find a clear decrease in NMI among the roots during their respective elections (which would have indicated a surge of novel suggestions) it did reveal relatively periodic shocks that appear to affect all roots, especially for Google. While both Google and Bing showed variations on these root-wide surges in NMI, the timings did not align well across the search engines, suggesting that some of these changes may be due to internal factors that affect each search engine differently; such as an algorithm update being pushed. These surges could also be tied to fluctuations in search activity more broadly, and different search engines have different users with different needs and habits.

4.4 Association Networks

To explore the underlying associations being made for politicians' names by Google and Bing, we first reduced each suggestion down to the new information that it contained relative to its root. That is, if the query "asa hutchinson" produced the suggestion "asa hutchinson biography," we reduced that suggestion to "biography." We executed this reduction with memory, so if at the next depth, "asa hutchinson biography" (now just "biography") linked to "asa hutchinson bio," the resulting edge was from "biography" to "bio." In effect, this procedure reduced the redundant information and enabled us to look at what we refer to as n-gram *association networks*.

After completing this process, we aggregated all of the resulting association networks into a single network for each search engine (Figure 8). The resulting networks reflect all of the associations made across our data collection window, giving us a new way of examining the associations that persist across all politicians, and

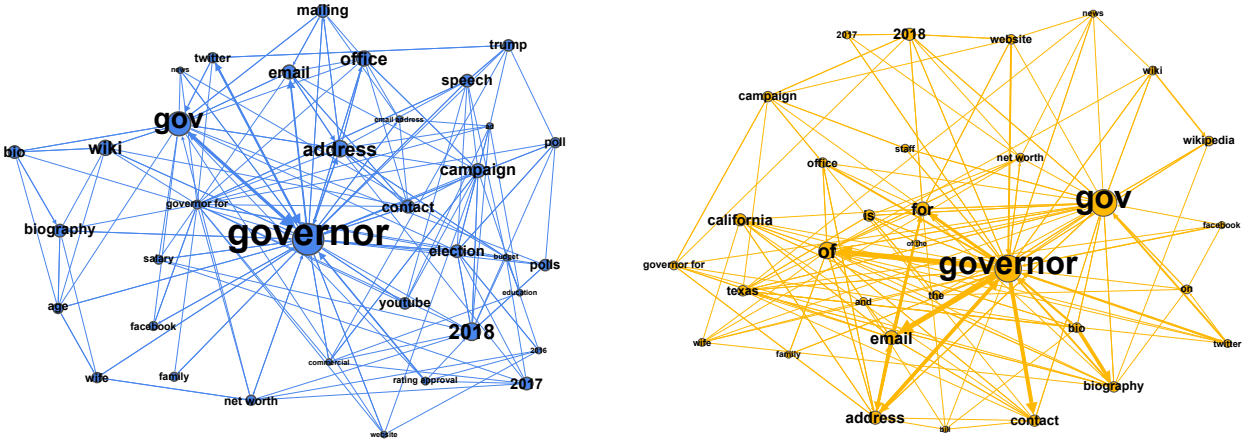

Figure 8: Association networks for Google (blue) and Bing (yellow) aggregated across root queries. The nodes represent n-grams, and the directed edges indicate a suggestion between two n-grams. The size of nodes and their labels are relative to their k_{in}, and edge size are representative of their weight – the total count for that suggestion across all crawls and roots. For visualization purposes we set a threshold of $k_{in} \geq 15$.

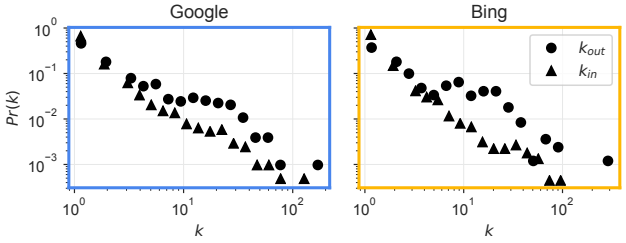

Figure 9: The in- and out-degree distributions for the aggregated association networks produced by both search engines were heavy tailed and fairly similar.

how these associations differ by search engine. Structurally, using the same standard network metrics as before (in- and out-degree), we found that Google and Bing both had similar heavy-tailed degree distributions (Figure 9). This structural similarity is related to the maximum number of suggestions that Bing and Google provided at the time of our audit, and is also likely limited by our trimming method.

Examining these associations networks more closely, we found moderate correlations between the n-gram associations produced by Google and Bing. More specifically, we used the in- and out-degree to rank all of the n-grams we observed, and found that both were moderately correlated (in-degree: $\rho = 0.49^{***}$; out-degree: $\rho = 0.38^{***}$). This means that there were significant differences in the associations being made for politicians by Google and Bing. The top-10 n-grams with the highest in-degree for Google were, rank-ordered: ['governor', 'email', 'twitter', 'gov', 'address', 'net worth', 'wife', 'rating approval', 'contact', 'facebook']. The same for Bing were: ['governor', 'gov', 'of', 'gov.', 'address', 'email', 'contact', 'twitter', 'for', 'bio'].

5 LIMITATIONS

Here we discuss the limitations of our approach, including disambiguation and context issues, the fixed and artificial nature of our queries, and the absence of user behavior and personalization.

Disambiguation. As noted in prior work, disambiguation is a problem inherent to audits of human names in autocomplete [10]. While our method for trimming suggestions (§ 3) proved to be a useful method for removing irrelevant queries generated by morphologically similar names, it did not solve the disambiguation problem, as we saw in our results on the presence of social media websites. However, the number of cases where this occurred was minimal, and it likely did not have an impact on our main findings. Although one could add an n-gram to their root queries to help disambiguate the search intent, this could further distance the root from a real user query, and further limit the number and breadth of suggestions returned. Future research should examine additional disambiguation techniques to trim suggestion networks.

Context. While we attempted to study autocomplete in isolation – by holding location, user, and queries constant – autocomplete is inherently tied to the social world. Suggestions occur, in part, because people are searching for them, and people search for things for myriad reasons. For example, suggestions might contain phrases that are ironic or satirical, but without context, could be classified and interpreted as negative. One potential way to capture this context is to expand our method to capture the search results returned for each suggestion, though without cooperation from a search engine, the amount of crawling required to collect this additional data is not feasible. On the other hand, it is likely that these mischaracterizations would wash out with a large enough sample.

Query Selection. Providing Google or Bing with a root input other than a name produces different types of suggestions, and adding characters to the end of the names we submitted (*e.g.,* name + "a") also produces a different set of suggestions than those generated by submitting the names alone, but there is typically overlap (*e.g.,*

name + "age" often appears in both). However, as a practical matter, appending each letter of the English alphabet to the end of a set of root queries would expand the amount of data collected exponentially. More importantly, data collection on that scale would likely not be allowed by Google and Bing, as it would require a constant, high-speed stream in order to complete RAI for all root + alphabet combinations in a reasonable time window.

Similarly, we did not collect suggestions in the setting that they are typically encountered: as the user types. When typing a query into Google or Bing, each new character entered results in a flash of suggestions based on the characters entered up to that point. The structures that we collected are a coarse grain version of what information scientists refer to as *tries*; a tree of possible suggestions that is updated as each new character is inputted. Again, as a practical matter, collecting a trie for each root query is not feasible without unfettered access to a search engine's autocomplete. Despite this limitation, prior work has shown that 53% of user engagement with autocomplete occurs after the user has typed the last character of a word [34], which is essentially what we simulated with RAI.

Personalization. We reemphasize that our study was aimed at studying the behavior of Google and Bing's autocomplete algorithms under fixed conditions that eliminated user-based personalization and held localization constant. This approach made the suggestions we collected comparable across search engines, root queries, and time, but does not address the heterogeneity of real user behavior. That is, users have widely different search strategies [44], and in practice, these strategies would systematically affect the way that they formulate queries and therefore the suggestions that they are exposed to. One could incorporate real user queries – through a browser extension, for example – and measure differences across query categories, but then time and location would become confounding variables. Localization, on the other hand, could be incorporated into our study by submitting root queries from geographically dispersed servers. Prior work has shown that real user queries systematically vary by location, and this variance is correlated with demographic variables, suggesting that it could be used as a proxy to understand personalization [5, 51].

6 DISCUSSION

In this paper, we introduced RAI, a method for preserving and mapping an algorithm's associations around a given root input. Using the names of 38 US governors who were up for re-election in 2018 as root inputs, we applied RAI to conduct an exploratory and non-personalized audit of Google and Bing's autocomplete algorithm in parallel over the course of two months. Our results demonstrate consistency in the structural and linguistic bounds of the two search engines' autocomplete algorithms, and shed light on differences in their content and temporal dynamics.

We found that Google and Bing's suggestions generally have a high degree of churn that is mediated by (1) the rank of the suggestions, with highly ranked positions being more stable than lower ranked positions, and (2) the length of the query being submitted, with shorter queries producing suggestions that were more stable than suggestions produced by longer queries. (Figure 5 and Figure 6). Given the attention biases in how users interact with autocomplete (*e.g.,* order effects) [24, 34], this stability at high ranks may limit exposure to volatile suggestions while not entirely excluding them. That is, highly ranked suggestions may be limited to stable trends around a given query, while lower ranked positions offer a more exploratory view of fleeting trends or breaking news. Limiting user exposure to volatile suggestions in this way may provide useful friction for combating inaccurate, offensive, or misleading information, and future research should explore this.

Our findings on the enhanced presence of social media in Google's suggestions, especially YouTube, has implications that could be seen as concerning. For example, YouTube is owned by the same parent company as Google, and regulators have previously levied record-breaking fines on Google for favoring its own products and services. Similarly, YouTube has recently come under heavy criticism for politically radicalizing its viewers [50], and in light of that, steering users who are searching politicians' names towards the video platform does not seem ideal. Future research should examine this link more in-depth, perhaps by examining the search results returned for queries that mention YouTube.

The periodic temporal dynamics that we picked up on may identify algorithm updates. These show up in Figure 7, where we observed approximately weekly shocks across roots, especially for Google. This finding ties in to prior work, which found that Google's suggestions took about a week to update [46]. These periods of accelerated turnover could be a particularly important time to conduct a more in-depth and qualitative analysis in future research.

The association networks that we derived from our data could potentially be used to assist human moderators by providing them with a macro view of the associations that an algorithm is making. Such a perspective could potentially reduce their viewing burden and enable them to more readily spot policy violations. Future research should examine methods for improving the filtering and visualization of these networks.

We hope our results and tools will provide a foundation for future research mapping information pathways on other platforms. For example, RAI could be applied to map YouTube recommendations, similar to the AlgoTransparency project [30]. Scholars who research fairness, accountability, transparency, and ethics in algorithms may wish to apply our method to study the associations that Google and Bing make with respect to gender, race, or other groups [39]. For example, do the suggestions for female or male names systematically differ? Although using more question-like root queries – as Baker and Potts did [1] – is an appealing idea, we explored a few such examples (*e.g.,* "why do <group>"), and found that they often returned zero suggestions on Google and Bing, suggesting that they are already being intentionally blocked. In hopes of spurring such research, we make our tools freely available at https://github.com/gitronald/suggests.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments and Alexandra Olteanu, Piotr Sapieżyński, and others for invaluable discussions and comments on this work. This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Paul Baker and Amanda Potts. 2013. 'Why Do White People Have Thin Lips?' Google and the Perpetuation of Stereotypes via Auto-Complete Search Forms. *Critical Discourse Studies* 10, 2 (May 2013), 187–204. https://doi.org/10.1080/17405904.2012.744320

[2] Ziv Bar-Yossef and Naama Kraus. 2011. Context-sensitive query auto-completion. In *Proceedings of the 20th international conference on World wide web*. ACM, 107–116.

[3] Nicholas J Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5, 1 (1980), 133–143.

[4] Edelman Berland. 2017. 2017 Edelman Trust Barometer. (2017). http://www.edelman.com/trust2017/ Accessed: 2017-03-07.

[5] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the Demographics of Search Users: Social Data Meets Search Queries. In *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*. ACM Press, Rio de Janeiro, Brazil, 131–140. https://doi.org/10.1145/2488388.2488401

[6] Bing. 2018. A deeper look at autosuggest. https://blogs.bing.com/search/2013/03/25/a-deeper-look-at-autosuggest. (2018).

[7] Lionel Casson. [n. d.]. *Libraries in the Ancient World*. Yale University Press.

[8] Le Chen, Alan Mislove, and Christo Wilson. 2016. An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International World Wide Web Conference*.

[9] Chitika. 2012. *Ask.com Has The Most Long-Winded Searchers, Report Says*. Technical Report. Chitika. http://searchengineland.com/ask-com-has-the-most-long-winded-searchers-report-says-109202

[10] Nick Diakopoulos. 2013. Algorithmic defamation: The case of the shameless autocomplete. http://www.nickdiakopoulos.com/2013/08/06/algorithmic-defamation-the-case-of-the-shameless-autocomplete/. (2013).

[11] Nick Diakopoulos. 2013. Sex, Violence, and Autocomplete Algorithms: Methods and Context. http://www.nickdiakopoulos.com/2013/08/01/sex-violence-and-autocomplete-algorithms-methods-and-context/. (2013).

[12] William H. Dutton, Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. [n. d.]. Search and Politics: The Uses and Impacts of Search in Britain, France, Germany, Italy, Poland, Spain, and the United States. ([n. d.]). https://doi.org/10.2139/ssrn.2960697

[13] William H. Dutton, Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. 2017. Search and Politics: The Uses and Impacts of Search in Britain, France, Germany, Italy, Poland, Spain, and the United States. (2017).

[14] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics* 9, 2 (2017), 1–22.

[15] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.

[16] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM: Human-Computer Interaction* 1, 42 (2017). Issue 2.

[17] Google. 2017. Search using autocomplete. https://support.google.com/websearch/answer/106230. (2017). Accessed: 2017-04-01.

[18] Google. 2018. How Google autocomplete works in Search. https://www.blog.google/products/search/how-google-autocomplete-works-search/. (2018).

[19] Jeffrey Gottfried and Elisa Shearer. 2016. News use across social media platforms. Pew Research Center. (2016). http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

[20] Aniko Hannak, Piotr Sapieżyński, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd International World Wide Web Conference*.

[21] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 2014 ACM Conference on Internet Measurement*.

[22] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr.. In *CSCW*. 1914–1933.

[23] Donna Harman. 1988. Towards interactive query expansion. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 321–331.

[24] Kajta Hofmann, Bhaskar Mitra, Filip Radlinski, and Milad Shokouhi. 2014. An eye-tracking study of user interactions with query auto completion. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 549–558.

[25] John B. Horrigan. [n. d.]. Library Usage and Engagement by Americans. ([n. d.]). http://www.pewinternet.org/2016/09/09/library-usage-and-engagement/

[26] Desheng Hu, Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Auditing the Partisanship of Google Search Snippets. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*, Vol. 16. 58.

[27] Stavroula Karapapa and Maurizio Borghi. 2015. Search engine liability for autocomplete suggestions: Personality, privacy and the power of the algorithm. 23, 3 (2015), 261–289. https://doi.org/10.1093/ijlit/eav009

[28] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 ACM Conference on Internet Measurement*.

[29] Search Engine Land. 2016. Google says it's not deliberately filtering "Crooked Hillary" suggested search to favor Clinton. https://searchengineland.com/google-crooked-hillary-251152. (2016). Accessed: 2017-04-01.

[30] Paul Lewis and Erin McCormick. 2018. How an Ex-YouTube Insider Investigated Its Secret Algorithm. *The Guardian* (Feb. 2018).

[31] Don McFadyen. 1975. The psychology of inquiry: reference service and the concept of information/experience. *Journal of Librarianship* 7, 1 (1975), 2–11.

[32] Matt McGee. 2010. Google weighs in on query length: Long tail alive and well. (2010). http://www.smallbusinesssem.com/google-query-length/3273/.

[33] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2012. Detecting price and search discrimination on the Internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*.

[34] Bhaskar Mitra, Milad Shokouhi, Filip Radlinski, and Katja Hofmann. 2014. On user interactions with query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 1055–1058.

[35] Yael Nemeth, Bracha Shapira, and Meirav Taeib-Maimon. 2004. Evaluation of the real and perceived value of automatic and interactive query expansion. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 526–527.

[36] NetMarketShare. 2018. *Search engine market share*. Technical Report. NetMarketShare. https://www.netmarketshare.com/search-engine-market-share.aspx

[37] Nic Newman, David A. L. Levy, and Rasmus Kleis Nielsen. 2017. Reuters Institute Digital News Report 2017. *SSRN Electronic Journal* (2017). https://doi.org/10.2139/ssrn.2619576

[38] NGA. [n. d.]. Governors' Powers & Authority. ([n. d.]). https://www.nga.org/consulting/powers-and-authority/

[39] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

[40] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001).

[41] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477.

[42] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM: Human-Computer Interaction* 2 (2018).

[43] Ronald E Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 955–965.

[44] Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in Web search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. ACM, New York, NY, USA, 13–19.

[45] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Proceedings of "Data and Discrimination: Converting Critical Concerns into Productive Inquiry"*.

[46] Lauren Starling. 2013. How to remove a word from Google autocomplete. (2013). http://www.laurenstarling.org/how-to-remove-a-word-from-google-autocomplete/

[47] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[48] Robert S. Taylor. 2015. Question-negotiation and information seeking in libraries. *College & Research Libraries* 76, 3 (March 2015), 251–267. https://doi.org/10.5860/crl.76.3.251

[49] Francesca Tripodi. 2018. Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices. Data & Society. (May 2018).

[50] Zeynep Tufekci. 2018. Opinion | YouTube, the Great Radicalizer. *The New York Times* (June 2018).

[51] Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 523–530.

[52] Wayne A. Wiegand. [n. d.]. The "Amherst Method": The Origins of the Dewey Decimal Classification Scheme. 33, 2 ([n. d.]), 175–194.

[53] Wiideman. 2010. Beat the autocomplete - A study of Google auto-suggest. (2010). https://www.wiideman.com/research/google-autocomplete/study-results

[54] Pan Zhang. 2015. Evaluating Accuracy of Community Detection Using the Relative Normalized Mutual Information. *Journal of Statistical Mechanics: Theory and Experiment* 2015, 11 (Nov. 2015), P11006.