The Information Society, 26: 364–374, 2010

ISSN: 0197-2243 print / 1087-6537 online
DOI: 10.1080/01972243.2010.511560

Routledge
Taylor & Francis Group

PERSPECTIVE

The Politics of Search: A Decade Retrospective

Laura A. Granka

Department of Communication, Stanford University, Stanford, California, USA
User Experience Research, Google, Inc., Mountain View, California, USA

In "Shaping the Web: Why the Politics of Search Engines Matters," Introna and Nissenbaum (2000) introduced scholars to the political, as well as technical, issues central to the development of online search engines. Since that time, scholars have critically evaluated the role that search engines play in structuring the scope of online information access for the rest of society, with an emphasis on the implications for a democratic and diverse Web. This article describes the thought behind search engine regulation, online diversity, and information bias, and it places these issues within the context of the technical and societal changes that have occurred in the online search industry. The author assesses which of the initial concerns expressed about online search engines remain relevant today and discusses how technical changes demand a new approach to measuring online diversity and democracy. The author concludes with a proposal to direct the research and thought in online search going forward.

Keywords democracy, diversity, information retrieval, Internet search, transparency

In "Shaping the Web: Why the Politics of Search Engines Matters," Introna and Nissenbaum (2000) introduced scholars to the political, as well as technical, issues central to the development of online search engines. Their piece encouraged scholars to critically evaluate the decisions made by search engines—particularly with respect to indexing and ranking—and assess how these choices affect the structure and scope of online information access for the rest of society. The issues addressed introduced scholars to the thought behind search engine regulation, online diversity, and information bias, and the subsequent implications for a democratic and diverse Web. A decade later, the search industry is in quite a different place—technically, socially, and politically—and needs a new research agenda from which to shape critical scholarly thought. Some of the initial hypotheses predicted a decade ago have remained quite valid and pertinent; others need to be reassessed. This article describes the current state of online search engines from both a technical and societal perspective. From this, we are able to assess which of the initial concerns expressed about online search engines remain relevant today, and what open questions remain. We conclude with a proposal to direct the research and thought in online search going forward.

Received 25 January 2010; accepted 31 January 2010.

The author thanks Fred Turner for his invaluable comments and discussions.

Address correspondence to Laura A. Granka, Department of Communication, Building 120, Room 110, Stanford University, Stanford, CA 94305-2050, USA. E-mail: granka@stanford.edu; Web site http://laura.granka.com

A RETROSPECTIVE: PREDICTIONS 10 YEARS AGO

A decade ago, the Internet was frequently viewed through a utopian lens, with scholars predicting that this increased ability to share, access, and produce content would reduce barriers to information access (Agre 2002; Barlow 1996; Bennett & Entman 2002; Ferdinand 2000; Gillmore 2004; Powell 2002). Viewed from this perspective, a key advantage of the Web, and subsequently of search engines, was providing more voices an opportunity to be heard: Scholars espoused that online search results should reflect the authorship diversity and viewpoint diversity latent in the online space. As such, the Web was approached in stark contrast to traditional media outlets, where content is created and distributed by a sole media owner. Introna and Nissenbaum were among the first to anticipate, and effectively articulate, specific problems with a market-driven evolution of search engines. Since that time, a number of scholars have extended these arguments to address the broader role that search engines play in distributing and shaping knowledge (Hargittai 2007; Hinman 2009;

Lev-On 2008), the potential for search engines to suppress or bias information (Diaz 2008; Elkin-Koren 2001; Gerhart 2004; Vaughan & Thelwall 2004), the new legal or policy opportunities created by search engines (Gasser 2006; Goldman 2006; Grimmelman 2007), or the broader societal responsibilities that search engines might assume (Hargittai 2007; Pasquale 2006; Van Couvering 2004). Underlying most of this work is a desire to prevent online information from merely mimicking the power structure of the conglomerates that dominate the media landscape. The search engine, subsequently, is seen as an idealized vehicle that can differentiate the Web from the consolidation that has plagued ownership and content in traditional print and broadcast media.

Introna and Nissenbaum (2000) were among the first to urge that online information is too important and too special a commodity to be shaped by market forces alone. They doubted that certain market requirements—namely, choice and competition—could be met in the case of online search, and also predicted that a market for online information would cause information quality and diversity to devolve—into the pandering of mass tastes. To alleviate, and potentially correct, these weaknesses of a search market, Introna and Nissenbaum advocated for (1) public oversight of search engine operation and (2) algorithm transparency.

To date, Web search engines have largely evolved without either of the two correctives originally championed by Introna and Nissenbaum. Search engine operation has instead more closely followed the trajectory of an online information marketplace, with competition and consumer choice dominating. While not what some scholars had originally hoped, the present state of search does enable us to evaluate what is (and isn't) working within the market paradigm, and to assess what advantages algorithm transparency and public oversight might otherwise provide.

This article draws from recent research in the fields of communication and computer and information sciences to extend and re-posit prior predictions about the need for public oversight and transparency in search engine operation. We are fortunate to have a diverse body of research addressing both the technical workings of search engines (e.g., ranking algorithms, Web infrastructure) and the cultural and societal effects. By drawing from these developments made throughout the past decade, we can now assess how close we have come toward achieving the ultimate ideal espoused by Introna and Nissenbaum—a democratic and diverse online information environment.

THE CASE FOR ALGORITHM TRANSPARENCY

Search engines have come under much scrutiny for their perceived role as the primary gatekeepers of online content (see Gasser 2006; Hargittai 2007; Introna & Nissenbaum 2000). For every given search term, a search engine engages with its automated algorithm[1] to decide what information to present to the searcher. For this reason, search engines are seen as wielding a great deal of power in what people know about a given topic—similar to how traditional news editors and journalists shape what audiences know about a given news topic. This directed flow of information—from an elite body to the public—is known as gatekeeping, and is traditionally applied to mass media contexts to understand the decision-making processes that drive news production in traditional broadcast and print media (Beard & Olsen 1999; Clayman & Reisner 1998; Shoemaker et al. 2001; Schoemaker 1996; Whitney & Becker 1982; White 1950). In the context of online information, search engines are also seen to function like the news editor, allowing the search ranking algorithm to determine what content to display for a particular topic. Just as criticisms of bias have been made against news organizations, similarly search engines have been challenged for their selection of sources, and for potentially not representing democratic and diverse interests.

Ultimately, some degree of gatekeeping is necessary to parse through the massive quantities of available information—the key issues are who does the gatekeeping, and how ethically these decisions are made. As we show later, the mechanisms whereby content is selected for inclusion in a user's search result set is fundamentally different than in traditional media—search engines universally apply an algorithm, whereas traditional news media makes case-by-case decisions.

Source Diversity

Fundamental to Introna and Nissenbaum's thesis was the argument of source—what sources a search engine promotes in their rankings, and which sources are hidden from view. Like all forms of media, search engines have to make choices about what content to distribute and show to the public—in the case of search engines, it is about what results to show for a given query, driven by the automatic choices inherent to the search algorithm. Search algorithms are technical decisions, rules, and cues that are created to broadly apply across all user-issued queries and Web content, thus avoiding decisions made on case-by-case bases. Introna and Nissenbaum advocated for algorithm transparency—meaning that search engines should disclose exactly how their indexing and ranking of online information works—for the sake of both consumers and producers of online information. Otherwise, they argued, searchers will be naive to ranking biases, and content creators will have limited knowledge of how search engines are rank their site.

While reasonable and ethical arguments, it should also be acknowledged that the basic principles guiding search

result ranking are in fact publicly available in academic literature and freely available online (Manning, Raghavan, & Schütze 2008; Baeza-Yates & Ribeiro-Neto 1999; Singhal 2001; Croft, D. Metzler, T. Strohman 2009; Langville & Meyer 2006). Subdisciplines in computer science are devoted to improving online information retrieval and ranking algorithms, making the general principles behind search engine ranking widely recognized (Singhal 2001; 2008). Individual search companies, while disclosing generalities about their ranking scheme, traditionally keep confidential the specific weights and scores attributed to each element in a ranking algorithm (Manber 2008).

The Case against Transparency: Quality and Competition

Search engines keep confidential the specifics of their ranking for two primary reasons. The first is for quality: Complete algorithm transparency would subject search engines to a great deal more spam and malicious attacks than what is already quite prevalent (Cutts 2008; Manber 2008). Prior events have proven that people have used their general knowledge of Web search ranking (learned from publicly available information on technology blogs and in academic literature) to adversely manipulate Web ranking. An entire economy of Web spammers, search engine optimizers (SEOs), and "Google bombers" has emerged (Zittrain 2009). Bar Ilan (2007) studied the development of Google bombs as a cultural phenomenon, and the ability of site owners to outwit the search algorithm. Search engine optimization conferences have cropped up, advertising ways for site owners to boost themselves (often unfairly) in search rankings. Conceptually, complete disclosure of an algorithm implies that for any given search, an individual would have the ability to assess how and why those specific results were returned. On principle, any extra information appears to be a clear win for consumers—more information to improve or troubleshoot the Web search experience seems like it would generally be a good thing. However, as this article addresses, the number of times that searchers will seek out, and derive benefit from, this extra information will be outweighed by the new opportunities afforded to spammers.

In addition to quality control, ranking specifics are also kept confidential because complete public disclosure of algorithms would dissolve healthy competition in the search marketplace. Currently, the quality of results returned for a particular query is the key distinguishing feature from one search engine to another. While search engines also differ in terms of speed or appearance, the unique results that a search engine retrieves for a query are the most important factor in creating a diverse and democratic online information space. If ranking algorithms were shared, such that multiple search engines replicated one ranking algorithm, consumers and site owners would be at an obvious disadvantage. Consumers would have less information choice in the marketplace (multiple search engines would return the same sites), and website owners would now be relegated to the same rank across multiple engines, with no increased opportunity to be found on one search engine versus another.

Past evidence has proven that disclosing criteria used in result ranking is prone to ill-intentioned uses of that information (Sullivan 2008). The larger question is whether further knowledge of search engine ranking specifics could benefit legitimate site owners and searchers enough to outweigh the harm. Introna and Nissenbaum so strongly urged for transparency because they assumed that knowing the details on how a site is ranked would empower site owners; they also assumed benefits to the information consumer, who would then be able to clearly understand why certain sites are returned for their queries. The following section describes what is currently known about Web ranking, to understand how complete disclosure in ranking specifics might conceivably provide additional benefit to consumers.

SPECIFICS OF WEB RANKING

In 2008, there were 1 trillion documents on the Web (Alpert & Hajaj 2008), and that number has most certainly multiplied. For any given query (the words a user types into the search box), there may be millions of webpages that contain the users' query words. Because the average person doesn't have the time to personally read or even skim millions of webpages, a search engine has to first identify relevant content, and second, rank order the most pertinent results. The three primary components that make up the rules in a search engine algorithm are linguistic cues, popularity cues, and user behavior cues.

Linguistic Cues

Term frequency. Perhaps most fundamental to search ranking is parsing the words (the query) that a user types into the search box and matching these words with the terms contained in online documents. Search engine algorithms attempt to infer user intent from a query, and then identify the webpages on which the user's query terms seem most important. For example, if a user searches for [Iraq], it is assumed that a webpage with only one mention of "Iraq" is less helpful than a webpage including more mentions of the word. Search engines use this assumption to leverage word uniqueness and frequency when determining ranking and relevance (Singhal 2001; Manning, Raghavan, & Schutze 2008; Baeza-Yates, Ribeiro-Neto 1999). Search engines identify how frequently a term appears within the context of a webpage versus how frequently the term

appears overall, across the entire Web. (This technique is similar to determining the relative uniqueness of a word in the overall English language. *The Oxford English Dictionary* does this by looking at the frequency of word occurrence in a set corpus—comprised of speeches, articles, books, and novels (OED, http://www.askoxford.com/oec). In the case of the Web, simply relying on the sheer counts of term occurrences inherently biases longer webpages, so search algorithms need to control for that by normalizing the term frequency against the overall document length (Manning, Raghavan, & Schutze 2008).

Other important linguistic cues that are not discussed in depth in this article include the identification of synonyms (e.g., "vocalist" versus "singer") and relevant phrases (e.g., differentiating a *hot dog* from a hot *dog*). (Singhal 2008; Manning, Raghavan, & Schutze 2008).

Web metadata. In addition to term frequency, the digital structure of a webpage affords unique properties that can be leveraged to facilitate result ranking. The code that makes up webpages is marked with metadata, which identifies certain properties of the document. For example, the title of a webpage is coded with a *title* tag (e.g.: <title>This is my webpage</title>), allowing search engines to identify which parts of the webpage are titles, headings, links, or images, all of which signify importance with respect to query terms. (Note: The <title> text does not show up to the reader and viewer of the webpage—only the information between the bracket tags is displayed.) Search engines use this data to (1) determine how to present results (e.g., the words identified as the page title will show up on the search result page) and (2) assess how important or prominent certain words are on a page. Another example of a metadata tag is anchor text, which tells the search engine how a site is describing another webpage. For example, one could refer to the Nalgene website with the words "indestructible water bottles," which would allow the search engine to have even more idea about what the Nalgene Web site might be about. (It would appear like indestructible water bottles .) Search engines recognize these terms that people associate with webpages and often incorporate them as cues in search algorithms. As expected, any of these metrics are prone to bias, and search engines typically take a lot of care to ensure that pages are not unnecessary spammed and unduly promoted (Cutts 2008).

PageRank Cues

PageRank (Brin & Page 1998) is the most frequently cited, and perhaps most grossly simplified measure of how online search works. PageRank signals are often the most criticized component of a search algorithm, as critics overestimate its importance in the overall Web ranking structure. PageRank is generally understood to be a popularity measure—the number of links to a site is essentially equivalent to "votes" for that site. Prior to PageRank, Kleinberg (1999) used the network structure of the Web to identify the most authoritative sources of information by analyzing links and hubs of content. Kleinberg looked at the link structure of websites, determining which sites link to which sources and what are the major sites that connect information together. The fundamental premise of PageRank still incorporates link structure, but accounts for not simply the sheer volume of links, but also the relative "authoritativeness" of the sites. To be more effective, and to circumvent spammers, the PageRank algorithm now uses over 200 signals (Singhal 2008).

User Cues: Feedback Loop from User to Algorithm

Perhaps one of the most fundamental differences between content selection in online search and traditional media is the search engine's reliance on searcher behavior to develop and shape the ranking algorithm. While traditional news media outlets do want to satisfy their readers and viewers, it is much more difficult for them to modify their selection criteria in real time, as search engines do. In online search, search engines have immediate feedback about what queries are issued, what content is selected, and what sites are accessed. Evaluating searcher behavior helps search engines understand what results are most helpful, and this information can subsequently be used to modify a ranking algorithm. In the information retrieval community, incorporating metrics about user behavior in attempts to improve Web ranking is called implicit feedback, with user clicks, reading time on webpages, and patterns of query reformulation serving as the most frequently leveraged behaviors (Fox et al. 2005; Kelly & Teevan 2003; Kelly 2005; Joachims et al. 2007; Radlinski et al. 2008). User click behavior is the most fundamental metric that search engines use to infer relevance. For nearly all searches, a user must click on at least one search result to obtain their desired information.[2] As such, clicks on webpages can be used to indicate what information is most central and useful to a user's online search experience (Huffman & Hochster 2007; Huffman 2008). For any query that has been issued more than once, search engines average what results are most frequently clicked on, across all users who have issued that query. Leveraging aggregate click data can help search engines identify which results people find most useful, and this knowledge might be used to promote or demote certain sites. Certain algorithms can learn to rerank results based on what is most frequently click on: For example, if the third search result is clicked on more frequently than the first two results, it is assumed to be more relevant and useful to users, and may be promoted up in the rankings (Joachims, et al. 2007).

Another common user behavior metric is reading time spent on individual webpages (Kelly & Teevan 2003; Kelly 2005). Not only is it important what results a user clicks, but also how much time a user spends on a specific page. If a user spends only a short time on a specific page, the page may be deemed unsatisfactory or less useful than if a user spends more time reading a page (Fox et al. 2005). While any metric is of course prone to error (an individual may accidentally close a window, or keep webpages open for hours in separate browser tabs), on aggregate, all of this noise washes out and enables these user behavior measures to become fairly robust (Joachims et al. 2007; Radlinski et al. 2008). Furthermore, a number of eyetracking studies of search behavior have been conducted, demonstrating that users do sometimes view results that are ranked lower than the selected result (Granka et al. 2002; Joachims et al. 2007; White & Morris 2007).

Search engines also use patterns of query reformulation to better infer user interest from a specific search session (Radlinski et al. 2008). It is often difficult for a search engine to know exactly what a user wants, particularly in the case of broad, single-word queries (e.g., "television" or "Australia"). In these situations, the subsequent query choices that a user makes, and the words a searcher adds to their query, is used to learn what that user may have originally intended from their first query. Other metrics that comprise the academic literature, but are of less practical use in industry, are page scrolling, and printing or bookmarking webpages (to measure interest and content retention) (Kelly & Teevan 2003).

As described, online searchers actively, though often unknowingly, participate and shape the content that is shown in search results. User feedback signals can be likened to a democratic system of "voting with your click" for the sites that are found helpful, useful, and informative. In this respect, both consumers and creators of search algorithms contribute to result ranking. However, scholars have also made the contrary argument that this represents a deficiency in the online information marketplace—once a site is listed in the top few results, it may continually be promoted and clicked in its position at the top of the page, enabling popular sites to rise (or at least sustain) their popularity (Introna & Nissenbaum 2000; Hindman 2003). While ostensibly logical, this assumption underestimates the significance of fundamental technical factors of Web ranking (such as term frequency and webpage metadata), and also ignores both a searcher's ability to exercise discretion and critically evaluate results. For instance, as described earlier, patterns of user click behavior are often used to rerank results. Even if a site is promoted at the top of a search result page, if users spend relatively little time reading that site compared to others, it may lose its hold at the top position.

Controlled experiments have shown that user click behavior changes based on a searcher's perceptions of quality, meaning that a searcher is not likely to blindly satisfice by selecting the first result if there are more relevant options (Joachim et al. 2005; Joachims et al. 2007; Pan et al. 2008). In these experiments, the order of search results was reversed (i.e., the 10th ranked result was put in the top position, the 9th ranked result in the 2nd position, etc.), and the researchers sought to understand whether the distribution of clicks in the experimental conditions differed from the normal ranking. The results revealed that when result order was reversed (with lower quality information at the top of the page), on average, users spent more time evaluating results on the page, clicked on more results, clicked on a lower ranked document (in this condition, a "lower ranked document" was actually of higher quality), and were more likely to reformulate their query (Joachims et al. 2007; Pan et al. 2008). This evidence is encouraging, as it shows that online searchers exercise some degree of selectivity in their quest to find the most useful information to meet their needs.

Potential Benefits of Ranking Disclosure

Now, assume a searcher or a site owner has taken the time to inform themselves of the foundational principles of search engine ranking—meaning, for each query, a searcher broadly knows why certain results appear. Will having the additional information of specific weights and attributes used in each search engine's algorithm notably affect a searcher's subsequent search behavior or a website owner's site management? Perhaps if an individual is proficient enough in understanding Web ranking to distinguish the "quality" of one ranking algorithm from another, then knowing the specifics of each engine's algorithm might encourage one to selectively use one search engine over another. However, the likelihood of this extra information being useful enough to change search behavior is unclear. The average searcher does not have a working knowledge of computer science, and possibly not even a strong desire to learn about it. In many cases, the available principles of information retrieval are likely to suffice for those individuals who care to learn.

This premise of this article is not to disagree that in principle, algorithm transparency is admirable and should be striven form. Taken at face value, algorithm disclosure could certainly avert unethical business operations. Instead, this article seeks to articulate, based on current knowledge of how online search engines operate, what transparency would look like in practice, and what benefit this could have for the average searcher.

Algorithm Transparency and Abuses of Power

The one foreseeable benefit in knowing how a specific result set is ranked is to identify instances when a search engine may be promoting specifics sites for profit at the expense of quality or relevance for the searcher. All or most search engines rely on advertising for revenue, and currently, most search engines explicitly identify these advertisements as such on their search result page (typically with the term "sponsored links"). Any argument for regulation or transparency of search engine algorithms should be less about the *principle* of transparency and whether an algorithm produces "diverse" results, but rather about regulating (i.e., preventing) potential abuses of power. For instance, in attempts to generate more profits, a search engine could resort to unethical behaviors by partaking in acts such as disguising advertisements for search results, or ranking wealthier sites higher if they pay more, all in attempts to generate higher profits.

In the instance that search engines unfairly promote certain sties to make a profit, it would be to a searcher's advantage to know if the search engine is exercising bias toward a paying content owner or sponsorship, thus limiting the diversity and democracy inherent to the information. If any aspect of search engine algorithms were to be regulated, the most important part is identifying when the search engine deviates from their organic algorithm to instead promote profit-making content. Partial algorithm disclosure or regulation could be useful if it ensures that search engines do not include paid results in their ranking at the expense of more relevant organically ranked results.

THE SEARCH MARKET: DIVERSE AND DEMOCRATIC?

Consumer Choice in Online Search

In order for an online information marketplace to properly function in the context of online search, certain conditions about online user choice and behavior must be met. One typical assumption, however unfounded, is that online searchers do not extensively evaluate many results when making their decisions about what pages to click. Introna and Nissenbaum (2000) argued against an online marketplace for search, speculating that searchers are simply not interested in reading multiple sources—searchers would click the first useful result and be done.

While this sort of quick search behavior may be common, it is entirely too simplistic an argument. As mentioned, experimental research has shown that a users' online search behavior will vary significantly based on the type of search being conducted, as well as the quality of results a user is presented with (Joachims et al. 2007; Guan & Cuttrell 2007; Lorigo et al. 2006). Using an eyetracker to measure individuals' online eye movements, researchers are able to assess how many results users evaluate, how quickly they scan the results, and in what order these results are viewed. While, on average, three to four results are scanned, this number differs based on the complexity of the task that a user has set out to complete, as well as the cost of making a decision online. Users spend more time critically evaluating sources when they know they have to spend money on their decision—for instance, making a purchase or planning a trip—than for facts or trivia, like what the weather will be, or the population of Canada. Thus, the number of sources critically considered by the user is highly dependent on the task.

Additionally, users have different search behaviors when viewing content of varying degrees of quality. As previously described, the experiment conducted by Joachims et al. (2007) generated significantly different viewing behaviors for users who were presented with reverse-ranked search results. In this condition, searchers spent more time viewing results and, on average, selected a lower ranked result than those did in the normal condition. Through behavioral data like this, one can see that consumers of online information can be relatively proficient at discerning and estimating the quality of search results.

These findings encourage the development of a stable online information marketplace—particularly one in which the search engines that provide the most relevant and highest quality information will invariably be the most visited. Knowing that users notice differences in result quality should encourage search engines to operate according to the principles of a marketplace, serving the best possible results. If search engines operate ethically, there should be no need for public intervention or regulation; only in instances when search engines abuse power to generate more revenue is there any risk of an information marketplace degrading.

Is the Structure of Old Media Recreated in New Media?

Most scholars critical of search engine behavior have looked at Web behavior on a large-scale aggregate level and have found that the patterns of media dominance and ownership that are present offline are merely reproduced online (Van Couvering 2004; 2007). This means that wealthier site owners have the capacity to create larger sites and therefore attract a larger audience. Existing research has identified the most frequently viewed sites and blogs and the most common search queries (Hindman 2003; 2007; 2008; Tancer 2008). Hindman (2003) has shown that the most popular sites viewed, while few in quantity, comprise over 90% of Web traffic. Similarly, a small number of queries comprise the majority of Web

traffic (Hindman 2007). Similarly, scholars have argued that the elements of ranking algorithms (such as PageRank and user behavior cues) also recreate "old media" structures, in that they simply allow the "rich to get richer," similar to the dominance of major conglomerates in the traditional media marketplace.

Broader arguments against the search marketplace discuss an apparent lack of competition and choice between different search engines (Introna & Nissenbaum 2000; Van Couvering 2007). When confronted with recent research, however, these claims seem more hypothetical than factual: Users turn to other search engines if they are unsatisfied (Heath & White 2008; White & Dumais 2009), and more than 60% of searchers use more than one search engine (Fallows 2008). While there is clearly competition between existing search engines, particularly with respect to international market share, it is admittedly more difficult for newer players to emerge in the search space. Based on economies of scale, the overall startup cost in creating a fast and efficient search engine is quite high—companies need many computers, servers, and a great deal of processing power to index the Web and serve traffic (Varian 2007). Once this infrastructure is in place, the incremental cost of serving additional queries is quite small, explaining the number of competitors that have emerged in the search space.

Market of Markets

When probed more deeply, the "rich get richer" argument against search-engine operation is an insufficient judgment. Most of the research addressing this issue is only based on data analyzed at the aggregate scale. While aggregate analyses are informative on some level, this perspective does not assess the true utility of a search engine, which is in surfacing information for non-mass interests and long-tail queries (Anderson 2004). To effectively understand the democratic implications of search engines, it is important to go beyond the aggregate level (which essentially only measures mass opinion and mass preference) and to look instead at the "market of markets" argument that Introna and Nissenbaum (2000) briefly alluded to. Each query creates a new economy, both financially, in terms of advertising potential (advertising is based on query keywords), and informatively, in terms of content disclosure. The advantage that online search engines have over traditional media is an ability to house and surface the long-tail information that goes beyond the mass tastes of the public (Anderson 2004). By looking at patterns of overall popularity across websites and queries, scholars repeatedly ignore the additional diversity online because, quite simply, more information can be found. An even more significant research oversight is the diversity that might exist within a particular search market (in the case of search, a market would be an individual query). Aggregate analyses of Web traffic and Web behavior only reveal the tastes of mass publics, and because we are not expecting search engines to change innate public opinion, we should seek out more precise measures.

DIVERSITY IN ONLINE SEARCH

The main challenge with addressing diversity in Web search is that the criteria with which to measure its presence, as well as to evaluate the benefits derived from it, have been historically ill-defined in the context of Web search. With the exception of deviant cases, such as censorship of search in totalitarian states (see Vaughan & Thelwall 2004), researchers have failed to tangibly identify specific cases, situations, and problems that are the direct result of bias and diversity in result ranking. To date, most scholars have limited their understanding of "bias" and "diversity" in online search results to the aggregate level—meaning, on average, what sites are most clicked on and most popular. On a theoretical level, content bias and diversity are legitimate and important issues and have been addressed from a policy perspective (Gasser 2006; Goldman 2006; Grimmelman 2007). However, "correctives" have been offered without a clear definition of the problem, or even explaining what could be solved with more diverse information. Some have urged that search engines should assume the responsibilities of a public forum, providing searchers with the opportunity for chance exposure (Lev-On 2008). Others have used this rationale to suggest the forced inclusion of randomized results (of lower ranks) just for the sake of "diversity" (Diaz 2009; Pandey et al. 2005).

Perhaps a better way to define diversity is on a per-query level, according to the "market of markets" paradigm, instead of on an aggregate scale. In this case, diversity would consist of two dimensions: diversity in site ownership, and diversity in the information content. Specifically, "site ownership" diversity would recognize the ownership structure of the sites that are retrieved for a given query, enabling us to draw parallels about the concentration and structure of online and offline media ownership. The latter measure of diversity would assess the incremental difference and value offered in each subsequent result, offering a better indication of the actual utility and information value provided by the many available sources.

MEASURING DIVERSITY

Source Diversity within a Result Set

Instead of aggregate popularity and total volume of traffic, scholars should evaluate whether the results for individual queries also in fact recreate on a micro-scale

the structure of offline media. For example, for a random sample of queries, are the largest conglomerate websites always listed first? Which site domains are promoted in rankings—commercial, educational, governmental, or nonprofit sites? The Internet can surface new, obscure, less prominent sources and content for specific queries that may not readily fall into the domain of mass appeal, and aggregate analyses currently overlook the opportunity to identify diversity within a particular query market. The query [obama] is much different from one that asks [obama health care plan], and the results for each query will be quite different. It will be useful for scholars to understand how diversity is represented in each particular context; perhaps for broad queries, major sites will dominate in rankings, and for more specific queries, lesser known outlets will have an opportunity to emerge. Future research should address source diversity for a given query and should be able to assess the range and ownership of the sources present in the top 10-20 results for a given query. Are these top sites also major media conglomerates in the offline realm? Are they dominant websites? Are they .com, .edu, .gov?

Content Diversity: Added Value of Results

Aggregate measures of diversity are merely an assessment of mass tastes, and do nothing to say whether these major sites are in fact providing lower quality information to its readers than what another, lesser known source might provide. In fact, one can argue that sites like the Mayo Clinic have substantial resources to research and create informative health information, so their information may be of higher quality, more factual, and less opinionated than, say, an individual doctor's personal health blog. While we don't know exactly what information a searcher wants in this sort of situation, we should look more closely at the notion of source diversity and determine what one could actually accomplish or change by enforcing diversity. From the viewpoint of a searcher,[3] it is superficial to simply say that mere quantity—more sources—directly correlates to a higher information value. Additional sources will not benefit the searcher unless we are able to quantify whether these additional sources offer new and valuable information.

One only needs to conduct a simple search like [diabetes treatment] to see that a number of the retrieved results offer information that is redundant with information contained on higher ranked sites (though it may be presented in slightly different ways). Future research should determine the utility curve of the absolute added value of an additional search result, through a standardized content analysis. For instance, if one were to complete a search for [diabetes], does the 11th ranked result offer information that is significantly different from any of the prior 10? At what point are there diminishing returns with respect to new content? Future research should be able to answer the question: For the top 10 or 20 results retrieved for a given query, how different is the information quality offered for each site? What utility does a searcher derive from each additional search result?

Medium (Corpora) Diversity

It is also important to recognize that search engines have recently included different forms of media into their Web ranking, including news results (for current issues), video results, academic results, image results, or local information. Evaluating the different types of information that a search engine retrieves for a specific query may reflect that any given query can have diverse and varied interpretations. Future research should measure the number of different media sources included in a result set, as this might indicate a search engine's attempt to satisfy the diverse interpretations of a given query.

Changes in Ranking Over Time

Another way to assess the presence of diversity is to analyze the changes in popularity and site ranking for specific queries or topics. For example, over time, for a query like [windsurfing] or [diabetes], how likely are fluctuations to occur within the top 10 or 20 results? The ranking change rate on a per-query basis would be particularly revealing to determine how easy it is for a single site to maintain dominance in its own unique market.

CONCLUSION

Up until now, researchers have attempted to understand the relationship between search engines and online diversity by measuring macro-behaviors: the overall distribution of traffic on the Web, where that traffic comes from, and what are the most popular search engine queries. Descriptive analyses such as these provide a useful baseline for understanding the innate preferences of online consumers. However, as this article has argued, these analyses do not deeply inform us of search engine best practices, particularly regarding information diversity. Aggregate traffic merely reflects mass tastes, and, as this article has shown, cannot be immediately extended to search engines. Instead, a micro-level analysis, centered on the level of a searcher's specific search query, would more appropriately assess the degree of diversity in online search. Search engine diversity cannot be measured by simply counting the most popular queries issued to the search engine as a whole; instead, each query should be evaluated separately within its own market, treated as a unique opportunity to provide information to online searchers.

As Lawrence Lessig argues in *Code and Other Laws of Cyberspace* (1999), code is the real power. On the Web, democratic decisions and governance come from the code that is written. In the context of online information, the code (algorithms) behind search engine ranking functions like a gatekeeper of content, and as structured, search engines are inherently dependent on the quality of the Web to do so. Research that effectively understands the implications of search engine ranking needs to tease apart the effects that are specific to search engine operation from those that just reflect the state of the Web—or at the very least, avoid jumping to such causal assumptions.

Because search is still such an explicit process, a user has to be highly motivated to even begin the process, and there is no "inadvertent audience" like there may be in the case of other media (such as chancing upon a newscast while waiting for your favorite sitcom; Iyengar & McGrady 2007). Thus, when one is arguing for diversity or democracy online, it is important to realize that search engines do a lot to ensure that interested individuals have the ability to acquire whatever they need, but are not useful for bringing new knowledge to those who lack the desire to ask for it, similar to the argument Sunstein exemplifies with the "Daily Me" (2008). Perhaps this selective exposure is the heart of the critique against online search, but is too easily misplaced upon the algorithms themselves.

In sum, it is incongruent to address issues about search engine democracy and diversity on an aggregate basis. These analyses will naturally mirror the state of available content on the Web and the innate preferences of online consumers, rather than isolating the potential for democracy or diversity (or lack thereof) due to search engines. Conspicuously lacking in the literature is any research done on the per-query level, isolating and clearly defining the variable of diversity. A "market of markets" analysis, investigating diversity and democracy on the level of the individual query, will help to achieve this.

NOTES

1. The study of algorithms has most recently been popularized through the contexts of Web search and information retrieval. In fact, this is a more recent extension of algorithms. "Algorithm" is a broad term referring to a set list of instructions and processes required complete a task. Algorithms are necessary in computing, enabling processes and tasks to be automated quite easily. For an in-depth discussion of algorithms, and their development from traditional mathematics to extensions in computing, see Chabert et al. (1997), *A History of Algorithms*. For a more focused look at search and information retrieval algorithms, see Manning et al. (2008).

2. Exceptions are those tasks where most search engines present the relevant information directly on the page, such as searches like "weather san francisco" where the current temperature may automatically be displayed.

3. We can also look at source diversity from the perspective of the content creators, though that is not within the scope of this article. Defending the importance of source diversity from the perspective of content producers is more about ensuring equal opportunity to an audience, and that is a separate issue.

REFERENCES

Agre, P. E. 2002. Real-time politics: The Internet and the political process. *The Information Society* 18(5):311–31.

Alpert, J., and N. Hajaj. 2008.*We knew the web was big...* Official Google blog post, July 25. http:// googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (accessed March 26, 2009).

Anderson, C. 2004. The long tail. *Wired*. Issue 12.10. October. http://www.wired.com/wired/archive/12.10/tail.html

Baeza-Yates, R., and B. Ribeiro-Neto. 1999. *Modern information retrieval*. Reading, MA: Addison-Wesley.

Barlow, J. P. 1996. *A declaration of the independence of cyberspace.* http://w2.eff.org/Censorship/Internet_censorship_bills/barlow_0296.declaration.

Bar-Ilan, J. 2007. Google bombing from a time perspective. *Journal of Computer-Mediated Communication* 12(3):article 8. http://jcmc.indiana.edu/vol12/issue3/bar-ilan.html

Beard, F., and R. Olsen. 1999. Webmasters as mass media gatekeepers: A qualitative exploratory study. *Internet Research: Electronic Networking Applications and Policy* 9(3):200–11.

Bennett, L., and R. M. Entman, eds. 2002. *Mediated politics: Communication in the future of democracy*. Cambridge: Cambridge University Press.

Brin, S., and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the seventh international conference on World Wide Web*, 107–17. April. Brisbane, Australia.

Chabert, J. L. et al. 1997. *A history of algorithms.* Berlin: Springer-Verlag.

Clayman, S., and A. Reisner. 1998. Gatekeeping in action: Editorial conferences and assessments of newsworthiness. *American Sociological Review* 63(2):178–99.

Croft, W. B., D. Metzler, and T. Strohman. 2009. *Search engines: Information retrieval in practice*. Reading, MA: Addison-Wesley.

Guan, Z., and E. Cuttrell. 2007. *An eye tracking study of the effect of target rank on web search*. Proceedings of the SIGCHI conference on human factors in computing systems, San Jose, CA.

Cutts, M. 2008. *Using data to fight webspam*. Google Blog post. June 28. http://googleblog.blogspot.com/2008/06/using-data-to-fight-webspam.html

Diaz, A. 2008. Through the Google Googles: Sociopolitical bias in search engine design. In *Web search: Multidisciplinary perspectives,* ed. A. Spink and M. Zimmer, 11–34. Berlin: Springer-Verlag.

Elkin-Koren, N. 2001. Let the crawlers crawl: On virtual gatekeepers and the right to exclude indexing. *Dayton Law Review* 26:179–209.

Fallows, D. 2008. Search engine use. *Pew*. August. http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx

Ferdinand, P. 2000. *The Internet, democracy, and democratization*. London: Cass.

Fox, S., K. Karnawat, M. Mydland, S. Dumais, and T. White. 2005. Evaluating implicit measures to improve Web search. *ACM Transactions on Information Systems* 23(2):147–68.

Gasser, U. 2006. Regulating search engines: Taking stock and looking ahead. *Yale Journal of Law and Technology* 9:124.

Gerhart, S. 2004. Do Web search engines suppress controversy? *First Monday* 9:1–5. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/ view/1111/1031

Gllimore, D. 2004. *We the media*. Creative Commons License. http://www.authorama.com/book/we-the-media.html

Goldman, E. 2006. Search engine bias and the demise of search engine utopianism. *Yale Journal of Law and Technology* 9: 188.

Grimmelman, J. 2007. The structure of search engine law. *Iowa Law Review* 93(1):1–63.

Guan, Z., and E. Cuttrell. 2007. *An eye tracking study of the effect of target rank on web search*. Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, CA.

Hargittai, E. 2007. The social, political, economic, and cultural dimensions of search engines: an introduction. *Journal of Computer-Mediated Communication* 12(3):article 1.

Heath, A. P., and R. White. 2008. Defection detection: Predicting search engine switching. *Proceedings of the 17th International Conference on World Wide Web, Beijing, China*: 1173–1174.

Hindman, M. 2007. A mile wide and an inch deep: Measuring media diversity online and offline. In *Media diversity and localism*, ed. P. Napoli, 327–47. Mahwah, NJ: Erlbaum.

Hindman, M. 2008. *The myth of digital democracy*. Princeton, NJ: Princeton University Press.

Hindman, M., and J. Tsioutsiouliklis. 2003. Googlearchy: How a few heavily linked sites dominate politics on the Web. Paper, annual meeting of Midwest Political Science Association, Chicago, IL, April 3–6, 2003.

Hinman, L. M. 2009. Searching ethics: The role of search engines in the construction and distribution of knowledge. In *Web search: Multidisciplinary perspectives*, ed. A. Spink and M. Zimmer, 67–76. Berlin: Springer-Verlag.

Huffman, S. 2008. *Search evaluation at Google*. Official Google Blog. http://googleblog.blogspot.com/2008/09/search-evaluation-at-google.html

Huffman, S., and M. Hochster. 2007. How well does result relevance predict satisfaction. *Proceedings of the 30th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 567–574.

Introna, L., and H. Nissenbaum. 2000. Shaping the Web: Why the politics of search engines matters. *The Information Society* 16(3):1–17.

Iyengar, S. I., and J. McGrady. 2007. *Media politics: A citizens' guide*. New York: W. W. Norton.

Joachims, T., L. Granka, Bing Pan, H. Hembrooke, F. Radlinski, and G. Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems (TOIS)* 25(2), http://portal.acm.org/citation.cfm?id=12291818jmp=at&coll=portal&dl=ACM&CFID=://tois.acm.org.topdownloads.html&CKTOKEN=tois.acm.org/top-downloads.html#CIT

Kelly, D. 2005. Implicit feedback: Using behavior to infer relevance. In *New directions in cognitive information retrieval*, ed. A. Spink and C. Cole, 169–86. Dordrecht, the Netherlands: Springer.

Kelly, D., and J. Teevan. 2003. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum* 37:18–28.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5): 604–32.

Langville, A., and C. D. Meyer. 2006. *Google's PageRank and beyond: The science of search engine rankings*. Princeton, NJ: Princeton University Press.

Lessig, L. 1999. *Code and other laws of cyberspace.* New York: Basic Books.

Lev-On, A. 2008. The democratizing effects of search engine use: on chance exposures and organizational hubs. In *Web search: Multidisciplinary perspectives*, A. Spink and M. Zimmer, 135–50. Berlin: Springer-Verlag.

Lippmann, W. 1922. *Public opinion*. New York: Harcourt Brace.

Lorigo, L., Bing Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay. 2006. The influence of task and gender on search and evaluation behavior using Google. *Information Processing and Management: An International Journal* 42(4):1123–31.

Manber, U. 2008. *Introduction to Google search quality*. Official Google Blog. http://googleblog. blogspot.com/2008/05/introduction-to-google-search-quality.html

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Pan, B., H. Hembrooke, T. Joachims, G. Gay, and L. Granka. 2007. In Google we trust: Users' decisions on rank, position and relevancy. *Journal of Computer-Mediated Communication* [special issue on The Social, Political, Economic and Cultural Dimensions of Search Engines] 12(3), http://jcmc.indiana.edu/vol12/issue3/pan.html.

Pandey, S., S. Roy, C. Olston, J. Cho, and S. Chakrabarti. 2005. Shuffling a stacked deck: the case for partially randomized ranking of search engine results. *Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30–September 2*, 781–92.

Powell, M. 2002. *Broadband migration III: New directions in wireless policy*. Remarks of Michael K. Powell Chairman Federal Communications Commission at the Silicon Flatirons Telecommunications Program University of Colorado at Boulder. http://www.fcc.gov/Speeches/Powell/2002/spmkp212.html

Pasquale, F. 2006. *Rankings, reductionism, and responsibility*. Seton Hall Public Law Research Paper No. 888327. SSRN: http://ssrn.com/abstract=888327

Radlinski, F., R. Kleinberg, and T. Joachims. 2008. *Learning diverse rankings with multi-armed bandits*. International Conference on Machine Learning, Helsinki, Finland.

Schoemaker, P. J. 1991. *Gatekeeping*. Newbury Park, CA: Sage.

———. 1996. Media gatekeeping. In *An integrated approach to communication theory and research*, ed. M. Salwen and D. Stacks, 79–92. Mahwah, NJ: Lawrence Erlbaum Associates.

Singhal, A. 2001. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin* 24(4):35–43.

Singhal, A. 2008. *Introduction to Google ranking*. Official Google Blog. Retrieved from http://googleblog.blogspot.com/2008/07/introduction-to-google-ranking.html

Sullivan, D. 2008. *What is search engine spam? The video edition.* October 21. http://searchengineland.com/what-is-search-engine-spam-the-video-edition-15202

Sunstein, C. 2007. *Republic 2.0*. Princeton, NJ: Princeton University Press.

Tancer, B. 2008. *Click: What millions of people are doing online and why it matters*. New York: Hyperion.

Van Couvering, E. 2004. New media? The political economy of Internet search engines. Paper, Communication Technology Policy Section at the Conference of the International Association of Communications Researchers, Porto Alegre, Brazil, July 25–30.

Van Couvering, E. 2007. Is relevance relevant? Market, science, and war: Discourses of search engine quality. *Journal of Computer-Mediated Communication* 12(3):article 1.

Varian, H. 2007. *Economics of Internet Search.* Angelo Costa Lecture. http://www.sims.berkeley.edu/∼hal/Papers/2007/costa-lecture.pdf

Vaughan, L., and M. Thelwall. 2004. Search engine coverage bias: evidence and possible causes. *Information Processing and Management* 40:693–707.

White, R. W., and S. T. Dumais. 2009. Characterizing and predicting search engine switching behavior. *Proceedings of CIKM, November*, 87–96.

White, R. W., and D. Morris. 2007. Investigating the querying and browsing behavior of advanced search engine users. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, the Netherlands, July 23–27, 255–62.