IR 13.0 - Technologies
AoIR ASSOCIATION OF INTERNET RESEARCHERS
Salford, UK
18-21, October 2012

Creating Networks Through Search

PageRank, Algorithmic Truth, and Tracing the Web

John Jones
West Virginia University
United States of America
john.jones@mail.wvu.edu

Abstract

This paper analyzes PageRank, a key feature of *Google's* search algorithm, showing how its primary function is not to identify quality Web pages but rather to identify hubs within a network defined by the Internet's link structure. While PageRank's method has been compared to the process of using the wisdom of crowds to determine quality, by relying on network effects to identify hubs, the algorithm does not allow users the independence and diversity necessary for crowdsourcing to be completely effective. For these reasons, *Google* and other search engines cannot be simply understood as information providers, for their role in defining the network structure of the Web makes these search companies the holders of a significant form of network power: programming. However, users can offset this power by becoming switchers who actively connect networks in order to diversify their information sources.

Keywords

search; networks; algorithms; programming and switching; *Google.*

In his description of the features of the network society, Castells (2009) notes that networks are inefficient compared to top-down "command-and-control structures" unless those networks are paired with computing technology that allows users to manage their "complexity" (p. 22). Castells directly credits the creation of the World

Wide Web and the concomitant rise of search for managing the complexity of networked information online, making it useful to the average user (2000, pp. 50-51). While Internet search has taken on various forms (Dennis, Bruza, & McArthur, 2002), from curated, user-contributed directories like *Yahoo!* to the query-based search of *Google*, the exponential growth in complexity of the Internet (Barabási, 2002, pp. 37-38) has rendered directories less effective for Web search. Instead, users increasingly rely on query-based search to find information. Query-based search relies on algorithmic methods that not only find matching documents but seek to determine their relevance to the query and rank them accordingly. This is the method applied by *Google*, typified by its PageRank formula, which was adapted into an algorithm that assigns a rank to Web pages based on an analysis of the links directed toward them (Brin & Page, 1998). Currently, *Google* is not only the most popular Internet search engine, but also the most visited site on the Internet <www.alexa.com/topsites>, a fact that underscores the extent to which search is used as a portal for navigating the Web (Halavais, 2009, p. 33). As such, the methods used by *Google* and other search engines are important objects of research not only for the ways that they provide access to the Web, but also how they mediate the Web for users. Recently, increased scrutiny has been brought on the methods employed by search engines to generate results and how those methods relate to knowledge access and the effects of Internet search on information seeking and culture (Halavais, 2009; Vaidhyanathan, 2011). A clear understanding of the effects of search algorithms is crucial for users to craft their searches in effective ways.

Since the centerpiece of *Google's* algorithm is PageRank, which claims to model users' information-seeking behaviors (Brin & Page, 1998, p. 110), an analysis of the effect of PageRank on information seeking is instructive both for understanding how search constructs the Web and how users can be better informed information seekers. Of course, PageRank is only one part of how *Google* determines search results, but, despite the efforts of researchers (Evans, 2007) and *Google's* increasing acknowledgement of the changes it makes to its algorithms (cf. Huffman, 2012), there are few public, concrete details about *Google's* search algorithms. For this reason, the analysis in this paper is limited. Nor is this analysis intended as a particular critique of *Google* or to

suggest that the effects of choices made by the search company do not have a counterpart in the choices made by other search engines. Rather, the goal of this paper is to analyze the network properties of PageRank and its effect on search users.

While PageRank has been extensively analyzed, such analysis tends to focus either on its mathematical properties (Ding, He, Husbands, Zha, & Simon, 2002; Evans, 2007; Gu, Chen, Chen, & Lu, 2012; Nie, Davison, & Qi, 2006) or the application of the formula to networks besides the Web (Iván & Grolmusz, 2011; Takahashi, et al., 2011). In this paper, I will describe how PageRank mediates the Web by bringing into existence a version of the network that would not exist without it. I will then show how that network privileges unique goals and values, examining how those goals and values prescribe a particular role for users within that network. I will begin by describing PageRank's method of determining the relevance of pages to a query and, following Latour (2004, p. 132), I will argue that this method traces a unique network within the wider Web. I will then ask how this tracing determines the role of the user in assigning relevance to search results. Finally, I will explore the impact these answers have on users' information-seeking behaviors. After this analysis, I will suggest some ways that acknowledging the network properties of algorithms like PageRank in discussions of search could provide new directions for research on this topic. Because of the important role search plays in navigating online information, it is crucial to understand how search algorithms like PageRank privilege certain kinds of information and information-seeking over others, as well as the role that networks play in this process.

PageRank and tracing the Web

Network theorists have argued that networks are not things, but rather are constituted in "discourses" (Castells, 2009, p. 53) and the "frames" and values (Castells, 2009, p. 46) they provide for particular information or interactions. As Latour (2004) puts it, networks are not simply physical entities like "nylon thread" (p. 132); rather, they are "the trace left behind by some moving agent" (p. 132) through both technological and cultural structures. As such, descriptions of networks are necessarily partial and interpretive. Being a network, the Internet is not a distinct thing, but it is a process that connects many different—and constantly changing—resources, both

technological and human. In their discussion of search algorithms, Heineman, Pollice, and Selkow (2009) state that efficient search depends not only on the effective structuring of data, but also on choosing the right algorithm for the particular search task (p. 106). For a search to return results quickly and effectively, it is important to organize the information being searched in a way that allows the algorithm to quickly navigate it for the result that is desired. For this reason, algorithms and data structure are intimately connected, and the choice of one directly influences the choice of the other. Indeed, while *Google* is well known for its search algorithms, it has also been praised for the unique methods it uses to index and structure Web data. O'Reilly (2005) has specifically claimed *Google's* primary value was as a "specialized database," not a search engine. As Halavais (2009) notes in his cultural analysis of search, search algorithms present particular views or biases of information (p. 87), and I argue that this claim not only covers the methodology of search, but also the ways that search conceptualizes the information that is being searched for. Search engines do more than simply provide a window onto the Web, but, in indexing and formatting the Web in particular ways, they value certain forms of participation and certain classes of information over others. In this section I will discuss what the known portions of the PageRank algorithm tell us about how it structures a network, describing both how PageRank creates a trace of the Web and the types of pages that are privileged in this tracing.

Describing PageRank

In making the case for PageRank, Brin and Page (1998) argue that then-current methods like directories and keyword searching were unable to handle the constant growth of the Web. In the first case, while directories like *Yahoo!* were good for some purposes—covering "popular topics," for example—they had trouble scaling and updating with new web content, and in the second case traditional keyword search "return[ed] too many low quality" results (p. 107). Brin and Page's answer to these problems was to use "the link structure of the Web," or the "citation (link) graph," to determine the "quality" of "each Web page" encountered by their algorithm (p. 109). PageRank gives each page it in its index a score based on how many other pages link to

it and how many pages it links to. Further, links that come from pages with a high PageRank are weighted more heavily within the system than links that come from pages that do not have a high PageRank (p. 110). This method is supposed to be "a model of user behavior," where the user is described as

> a 'random surfer' who is given a Web page at random and keeps clicking on links, never hitting 'back' but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. (p. 110)

In the initial presentation of PageRank, then, we are told that it can determine the importance of a Web page by quantifying the way in which pages are linked together, or cite each other, and that this method should replicate how an average user would navigate the Web, following links from high quality pages to other high quality pages.

Network effects and PageRank

Page, et al. (1998), in their discussion of citations, note the work of Goffman who argues that information moves through citation networks via an "epidemic process" (p. 2), and they draw on research that models the Web using graph theory and other forms of network analysis, identifying the Web, for example, as "hubs and authorities" (p. 2). In his discussion of the development of graph theory and network analysis, Barabási (2002) addresses an individual like the "random surfer" the *Google* founders adopted as their conceptual model for PageRank, and Barabási uses this conceptual device—which, following Brin and Page, I will call the random surfer—to describe two features of the Internet that allow algorithms like PageRank to work: "small worlds" (p. 37) and "preferential attachment" (p. 85). In the first case, the Web is a "small world" (p. 37), where "small world" means that any page on the Web is only a few links—around nineteen, at the time of Barabási's analysis—away from any other. This fact allows for the random surfer to navigate the Internet, but it does not mean that it is necessarily easy for this surfer to find a particular page online, as the exponential growth of the Web would make finding a page nineteen links away involve sifting

through millions of documents (p. 37). In addition to the small worlds effect, the Web is governed by another network feature: "preferential attachment" (p. 85). Preferential attachment describes the likelihood that nodes in a network with many links will acquire more links. Or, as Barabási puts it, "[w]hen choosing between two pages, one with twice as many links as the other, about twice as many people link to the more connected page" (p. 85). This process creates the "hubs" that we find on the Web, pages with many more links than other pages, and it is these hubs that PageRank is so effective in ferreting out. Writing a decade ago, Barabási states that pages that are not hubs—those with only 1-2 incoming links—are frequently ignored by search engines (p. 58), a situation that has largely changed as search engines have been able to create more comprehensive Web indexes. However, the fact remains that non-hubs are not preferred by PageRank, which utilizes the links in its index to discover hubs, and those hubs receive preferential rankings in the system.

In short, while it is possible for a random surfer to find relevant pages on the Web by moving from link to link, because this surfer is constrained in time, she or he would never be able to reliably find particular documents due to the sheer size of the Web. Yet, individuals do find pages by "interpreting the links," looking for the most relevant one (Barabási, 2002, pp. 37–38). While the *Google* system described by Brin and Page (1998) attempts to take into account the text associated with links in order to identify page content, in lieu of this interpretation PageRank largely relies on the status of particular page as a hub as a substitution for the situated decision making of the random surfer. In doing so, PageRank traces a particular version of the Web, one characterized not by decisions about relevance but by preferential attachment. To put it another way, what PageRank describes is not necessarily a map of quality Web pages, but rather a map of the growth of preferential attachment defined by links between pages.

Of course, it is not a given that the hub structure of the Web does not map onto quality. However, I would point out two other features of the network that militates against the easy conflation of quality with network structure. First, search engines cannot reliably index the entire Web. Second, and more importantly, by focusing on Internet structure, this tracing or model of the network excludes from considerations of

quality significant forms of action from the actor-network, thus devaluing the user. In the first case, because the Internet is a directed network—meaning that if there is a link from page A to page B, it is possible for a user to navigate from A to B, but not from B to A—this has particular effects on how it is possible to navigate the network. Whereas in a network that allows connections in both directions all pages would be navigable from all other pages, the directed nature of links on the Internet has resulted in it being divided into distinct "continents," or groups of connected nodes (Barabási, 2002, p. 166). Barabási (2002) identifies four such continents on the Internet: the central core, the IN continent, the OUT continent, and the tendrils (p. 166). The central core consists of most major sites, and a page on this continent is able to connect with any other page (p. 167). The IN continent is connected to the core, but it is not possible to navigate from the core back to this collection of nodes, while pages in the core are connected to the OUT continent, but it is not possible to return to the core from it (p. 167). Finally, the tendrils are a series of interconnected nodes that are not connected to any other continents (pp. 167-168). Because the Web has this topology, search engines can only reliably index roughly 24% of Web pages, leaving the rest "unreachable by surfing" (p. 165) or by ranking methods that emulate surfing. It is clear, then, that the random surfer is not reliably reaching all portions of the Web, so any definition of quality based on this model would have to account for this lack of coverage. I wish to deal with the second case—how PageRank accounts for the judgments of Internet users—more fully in the next section.

The user and PageRank

As I have argued, the link structure of the Web is not a concrete thing, but rather a particular tracing of a wider actor-network (Latour, 2004, p. 132). Further, PageRank does not simply present that link structure, but rather it interprets it, placing more importance on some links than others, and using that interpretation to identify hubs within the network; indeed, this hub structure is itself merely a representation of the Internet, seeing as it only covers a small portion of the documents on the network (Barabási, 2002, p. 165). Because it is an actor network, one that puts particular emphasis on the role of users in its creation, we should examine the network PageRank

creates not simply for its technological features but also for the role that individuals play in it.

In their various descriptions of PageRank, Brin and Page (1998) place a high level of importance on the role of the user, stating that PageRank's formula and resulting algorithms replicate users experience in navigating the Web (p. 110). Additionally, although it is not directly stated, there is another role for the user, that of deciding which pages to link to from a particular page, thus making choices about connections that are mined by PageRank to determine quality in Web search. While the authors do not use this phrase, the way in which they frame the "link structure" of the Internet (Page, Brin, Motwani, & Winograd, 1998, p. 1) as a means of determining the value of individual pages is very similar to the idea of using the "wisdom of crowds," or "crowdsourcing"—the process of using aggregate decisions of many people, as opposed to top-down decisions from one or a few, to determine importance—concepts that would later be a key component of what was called "Web 2.0." Indeed, O'Reilly (2005), the author of a seminal document describing Web 2.0, specifically connects PageRank to crowdsourced methods of evaluating information. In this section I wish to complicate the relationship between the user and PageRank, suggesting that, while users do influence the results returned by the algorithm, their influence is significantly constrained. Specifically, I will look at the role that the user plays in the collective action that forms the link structure of the Web as that structure is modeled by PageRank.

According to Surowiecki (2004), there are four conditions that allow for wise crowds:

> diversity of opinion (each person should have some private information, even if it's just an eccentric interpretation of the known facts), independence (people's opinions are not determined by the opinions of those around them), decentralization (people are able to specialize and draw on local knowledge), and aggregation (some mechanism exists for turning private judgments into a collective decision). (p. 10)

Whereas PageRank is wildly successful at aggregating results and does little to interfere with the local knowledge that informs decisions made about links, it has biases that interfere with diversity of opinion and independence. Namely, the way in which

PageRank structures the index of the Web allows for it to harness particular network effects that it then uses to create search results. However, these network effects can limit both the diversity of opinion available in search results as well as the independence of individual users.

In the previous section, I noted that PageRank identifies the hub structure of the Internet by discovering those pages that have a large number of incoming links. The assumption being made by this formula is that these incoming links are indicators of quality, and Page, et al. (1998, p. 2) cite research into citation analysis as a means of justifying the connection between incoming links and relevance. Yet, despite this connection, by focusing more on the existence of links rather than the interpretation of those links, mere citation can lead to two problems. The first is that citation without interpretation can fail to distinguish between negative citations and positive ones. In a well-known case, the owner of an online store that sold sunglasses would regularly harass and otherwise provide poor customer service to certain customers in order to get them to write negative reviews of his site (Segal, 2010). The reason given for this behavior was that the owner found that mentions of any kind, positive or negative, tended to cause his site to rise in *Google* rankings (Segal, 2010). While *Google* has since claimed that they have altered their other algorithms to account for negative reviews, PageRank, in itself, does not allow for diverse opinions, counting each link as a single vote for another page. The effect is, as Surowiecki (2004) puts it, to limit the "private information" (p. 10) that users can contribute to web results. Whereas a scientific paper that is negatively reviewed in a journal, or cited in order to debunk its claims, would never be mistaken as authoritative by human readers of the article, PageRank does not distinguish between positive and negative mentions. It is this feature of PageRank that led to famous "*Google* Bombs" ("*Google* bomb," n.d.) instances where individuals have used the features of PageRank to influence search results for generally non-commercial purposes. Interestingly, while Brin and Page (1998) note as an example of poor search quality that a then-major search engine returned a joke page as its first link for searches of "Bill Clinton" (p. 116), the most famous *Google* Bomb

connected searches for the phrase "miserable failure" to then-president George W. Bush's official White House biography page ("*Google* bomb," n.d.).[1] In short, because the algorithm cannot natively distinguish between positive and negative mentions of a particular source, its effect is to limit all user knowledge that contributes to its calculations to only public knowledge—the links themselves—thus restricting individual interpretations and other private knowledge.

While it may seem that noting the direction of links is itself a significant form of reference, and that individuals can draw on their private knowledge to interpret search results, there is another aspect of PageRank that affects the role of individuals. Even though PageRank counts links in aggregate, individual choices about links are constrained by network effects. As mentioned earlier, pages with large numbers of incoming links are hubs, but, as Barabási (2002) points out, hubs are not random. Rather, hubs are the result of preferential attachment in the growth of networks (p. 91). The presence of preferential attachment indicates that some networks generate hubs as an outcome of their growth process, and hubs develop not simply because of inherent qualities but because of early preferences (cf. Waldrop, 1992, p. 45). As Barabási (2002) notes, individuals, when choosing to create links, are far more likely to create links to pages that already have a great number of links (p. 85). While such linking behavior is not centrally controlled, it is constrained by this feature of networks, and this constraint is a direct product of the way that PageRank traces the Web. The

[1] Examples like these, and the fact that Google must constantly fight spam, suggest another problem with the simple equation of PageRank's results with quality. Page et al. (1998) note that it is possible to try to "manipulat[e]" the PageRank system by "buying" links to a particular page, thus increasing the page's rank by increasing the number of pages that link to it (p. 12). However, in 1998 they argued that this type of manipulation was largely preventable, suggesting that the cost of such a practice—purchasing a number of ads on various sites—would make it prohibitive. They did not anticipate the rise of free services like blogs were users could easily create numerous unique subdomains then have them link to a site, or they could create numerous comments on other blogs that would link to the site in question. *Google* has taken steps to limit the effectiveness of this type of manipulation, both by directly restricting it ("Report spam, paid links, malware, and other problems to google," 2012), and by modifying its algorithms to devalue sites that they identify as so-called link farms (Singhal & Cutts, 2011). Yet, if PageRank was a means of assessing quality and was based on an unconstrained version of crowdsourcing, one would expect that this effort on *Google's* part would not be necessary, as low quality sites would be identified as such and devalued by users. However, because PageRank's true function is to discover hubs, not quality, and because hubs can be created by anyone with sufficient time and resources, the arms race against spammers is a constant concern to the search company.

tendency of PageRank to produce this behavior is hinted at even in the earliest writings about the formula. Page, et al. (1998) argue that, given a limited set of Web pages, PageRank is an excellent predictor of citation counts (p. 13). That is, when not all citations are known, the PageRank of a particular page within the known set of pages is strongly correlated with the number of citations for a given page within a larger or complete set. Or, as the authors put it, "PageRank may even be a better citation count approximation than citation counts themselves" (p. 13). While this result may suggest confidence in PageRank's results—indeed, the authors argue that it is a "powerful justification for using PageRank" (p. 13)—it may be that it is instead merely an indicator of the formula's effectiveness of ferreting out the structure of the network that it examines. Or, to rephrase Page, et al.'s (1998) statement, PageRank is a better predictor of a network's structure than the known structure of the network.

These issues are exacerbated by the tendency of search engines to offer personalized search results, results designed to fit the particular interests of the individual user ("*Google* personalized search," n.d.). Indeed, Page, et al. (1998) suggest searching the saved bookmarks on a user's computer as one means of creating these personalized results (pp. 11–12). While the ostensible purpose of personalized search is to increase the relevance or quality of results returned to searchers, one way of conceptualizing personalized search is that it shifts the location of the search from the wider Web to a more limited network. That is, rather than identifying the hubs of particular topics or other information within the Web, personalized search merely identifies the hubs and other network features of what (it assumes) are the individual's personal or local networks. At first blush, this may seem to be a means of increasing the independence of search results by giving more prominence to user-generated choices. However, because PageRank operates by identifying the hubs in a link network, the effect of personalized search is to limit the network that is being searched to one that is relevant to the searcher, thus effectively limiting independence by excluding what are determined to be irrelevant networks and their associated hubs. Similarly, personalized search could give too much prominence to individual interpretations of facts, thus not allowing for enough outside opinion to temper individual results.

These conclusions do not necessarily argue against the use of PageRank. Indeed, one of the benefits of an algorithm like PageRank is that it identifies Web structure. Barabási (2002) notes that, because of the density of links on the Web and its small world properties, while any page is theoretically within nineteen clicks of any other page, the random surfer has no effective means of determining the shortest path between those two pages (p. 37). However, algorithms like PageRank can effectively surf connections, and, consequently, identify hubs and other major features of the Web (p. 38). Yet, PageRank does not represent a model of the random surfer because it does not provide searchers with the shortest distance between two pages. Rather, it reformats the data it indexes to emphasize the hub structure of the Web that it has defined using the link information it collects. One result of identifying this hub structure is to make those pages more visible for users, thus intensifying the effect of preferential attachment on those pages. For this reason, to the extent that PageRank influences search results, it does not allow for true independence in the establishment of hubs. Drawing attention to a page as a potential hub increases the likelihood that it will draw more links, thus giving it more attention, thus drawing more links, and so on. For these reasons—PageRank places particular, although not absolute, restrictions on the diversity of opinion and independence of its results—the algorithm constrains the role of the user in creating the knowledge that informs its definitions of quality. Within the actor-network, prominence is given to the technical features of the network—the hubs—while the choices of users are constrained by that network and network effects.

Algorithms, "programming" and "switching," and networked search

So far I have discussed the effects of PageRank on the creation and growth of networks of information. The question that remains, is: how do these effects influence the information-seeking activities of *Google's* users? This is a relevant line of inquiry, for as search has become the dominant interface for information seeking in a digital society, "permeating our social lives" as Halavais (2009, p. 31) puts it, a common argument has been that search skills are an essential emerging literacy. Halavais (2009) identifies the focus of many search literacy efforts as being on the role of query formation on the part of users and the continual refinement of those search queries so as

to close in on the object of the search (p. 34). Understanding how to successfully query a search engine is an important part of being able to craft effective searches, but, as I have argued, the unique network effects of PageRank suggest that the role of networks and their structure must be taken into account as part of the search process. That is, in combination with choosing the correct query, it is important to choose the correct network or networks to search and—when possible—to understand the structure of those networks. While Jenkins, Clinton, Purushotma, Robison, and Weigel (2006) do not specifically identify search literacy in their list of the core skills of twenty-first century media education, search nevertheless underlies many of the skills they do identify, particularly "networking," which the authors describe as "the ability to search for, synthesize, and disseminate information" (p. 4).

In this concluding section, I will suggest that we need to think of user interactions with search not simply as interactions with an interface or with search queries, but as interactions with networks. To that end, I address the effects of PageRank in terms of Castells's (2009) "programming" and "switching" (p. 45), describing how the network-creating features of PageRank can influence search results and how users can respond to this influence in order to be more informed and effective information-seekers. Before addressing how users are affected by the network features of search, I will briefly describe some challenges facing researchers interested in search procedures, describe how what we know about *Google* searches can impact users in light of the analysis in this paper, then, following Jenkins et al.'s (2006) description of the role of search in networking, suggest how search users can increase their access to information by becoming effective "switchers" (Castells, 2009, p. 45).

PageRank and creating networks through search

Rather than simply being a passive recorder of the Web, PageRank makes *Google* a wielder of a form of power uniquely suited to networks: programming. Castells (2009) argues that the primary means of exercising power within a network are "programming" and "switching" (p. 45), and he defines these skills respectively as:

> (1) the ability to constitute network(s), and to program/reprogram the network(s) in terms of the goals assigned to the network; and (2) the ability to connect and ensure the cooperation of different networks by sharing common goals and combining resources, while fending off competition from other networks by setting up strategic cooperation. (p. 45, original emphasis)

Because of PageRank's role in tracing or "constituting" a particular network, first identifying the hubs in that network and then reinforcing those hubs through preferential attachment (or undermining them when it appears that they have been created in ways that violate its policies), *Google* has significant programming power. By establishing the goals of the network it wishes to trace—one where a large number of inbound links, or inbound links from important sites, equals quality—PageRank effectively programs users' Web experience so that it privileges these features.

Yet, acknowledging the role of PageRank in programming networks, it is necessary to point out that while PageRank is a highly significant factor informing *Google's* search algorithms, it is not the sole determiner of the company's search results. While a search for a particular term would, in theory, yield a list of pages matching that query ordered by their PageRank, in practice *Google* search results are determined by hundreds of factors (Evans, 2007, p. 23) and the algorithms governing them are altered frequently (Huffman, 2012). Whereas the PageRank algorithm is publicly available, these other factors are either only generally known or are kept completely secret. Evans (2007) notes the following issues that effect what we know about *Google's* ranking methods:

- There are over 200 different factors (or signals) used by *Google* to calculate a page's rank.[2]
- What these factors are is unknown, as is the weighting of each factor towards the final rank.

[2] N.B.: Evans (2007) is referring here to the order in which a Web page is ranked within a set of search results, not that page's particular PageRank.

> • The weighting of each factor used to determine the top ten results may be different from the weighting used for the remainder.
>
> • Different query terms may employ different ranking factors and/or different weights (Bifet et al., 2005).
>
> • *Google* has multiple data-centres distributed across the world, not all of which are in sync at any one time. Thus the ranking algorithm used in one data-centre may change subtly from the ranking algorithm used in another (Cutts, 2006). (Evans, 2007, p. 23).

Vaidhyanathan (2011) notes that we now know the bare outlines of some of those factors, for *Google* localizes and, in some cases, personalizes search results (p. 139). In the first case, the company tailors certain search results to particular regions of the world, sometimes actively filtering search results to comply with local laws (p. 47), and sometimes complying with local views about contested issues like national borders (p. 117). Similarly, personalized search places primacy on those results that *Google's* algorithms calculate to be of most relevance to the user—although, as Evans (2007) points out, the top results in a search may be unaffected. This is not to say that there is not some benefit to localization and personalization for certain types of searches; however, it is not clear that this type of restriction is beneficial for the type of open-ended research that benefits knowledge creation (Vaidhyanathan, 2011, p. 192 ff.).

As I have suggested in my earlier analysis, the effect of localization, personalization, and the other factors used to determine *Google's* search results would be to create smaller, more specialized networks. There is no reason to think that network effects like preferential attachment are not operative within these networks, albeit with different outcomes and creating different hubs. Consider the case of localization, where the results of a search query made in the United States might be quite different from one made in a different country. However, such a search, while not necessarily searching the entire network consisting of the Web pages *Google* indexes, would still be dominated by hubs and other network effects just as with the larger network, although those effects would be constrained to the sub-network. That is, PageRank, in conjunction with other

search criteria like localization and personalization, creates sub-networks that also privilege these biases. Networks that are localized and personalized are even more constrained in their access to information because, in addition to limiting the diversity of opinion and independence of those results (Surowiecki, 2004, p. 10), they would emphasize these constraints by drawing from a smaller network of websites.

Impact of search networks on information-seeking

We know that PageRank continues to play an important role in *Google* searches, despite the fact that it is not the only factor used to calculate search results. As I have described above, this means that individual searches do not interact with a network that is determined by *Google's* entire index of the Web, but rather with smaller sub-networks. The effects of these sub-networks on information-seeking have the potential to be significant. In previous sections, I have argued that the network-creating functions of PageRank do not fully tap into Surowiecki's (2004) "wisdom of crowds" and, rather than simply organizing "quality" information online (Brin & Page, 1998, p. 107), the effect of the algorithm is to create hubs via network effects like preferential attachment.

There are two problems associated with the way that PageRank interacts with *Google's* localization and personalization features. By excluding information through focusing on local or personal sub-networks, search algorithms can effectively "disappear" some topics in the eyes of the searcher, obscuring difficult problems that should be addressed or giving the searcher a false sense of the universality of his or her beliefs. In the first case, Vaidhyanathan (2011) describes how *Google* localizes searches in France and Germany to "actively block anti-Semitic" websites (p. 47). While at first blush it may seem that this is a positive action—who wishes to defend access to such sites?—such blocking may give the impression to searchers in those regions that such sites do not exist or that the thinking they promote is non-influential. The effect of excluding the websites of anti-Semitic groups—or any category of information—from search relegates such groups to their own networks, and those networks—as Barabási (2002) has shown with the different continents of the Web (p. 165)—can grow despite being cut off from searches. However, by being disconnected from Web search, the primary interface for many Web users, it is possible that the growth of such networks

could go unnoticed by those users, thus masking the extent of the problem represented by such thinking and slowing potential interventions. In the second case, personal searches could obscure or deemphasize contrary views if such views are determined by the personalization algorithm to be outside of a user's preferences. For example, a personalized search for a political figure that primarily delivered results that were deemed to support the users' beliefs might reinforce those beliefs by giving the user the impression that those beliefs were more widespread—or unchallenged—than is actually the case. Such search has the potential to turn into an echo chamber, where only what pleases the searcher is contained in the network of search results. In both cases, the ability to see the importance or growth of ideas obscured by these searches would be lost, and this loss would be reinforced by the effects of preferential attachment within the sub-networks created around these searches. Such searches simply reinforce the features of PageRank described in this analysis by enacting them in smaller networks.

In light of these problems, users cannot simply improve search results by crafting better queries or simply understanding the limits of search engines. When access to networks is key, an important skill would be the "networking" ability described by Jenkins et al. (2004). The skill would lie not simply in discovering static knowledge, but discovering the networks can point to or accommodate the knowledge being sought. In this way, effective searching should involve Castells's (2009) "switching," bringing the network of search results into contact with other networks that are able to provide outside perspectives on those results that may not be constrained by the same preferential attachments. Different networks will show their own biases, but querying many different networks should help alleviate the effects of those biases. In the case of monitoring the rise of hate groups, even when access to that material is inaccessible via search networks, users could connect with groups that monitor such activity—as the Southern Poverty Law Center <splcenter.org> does with hate groups in the United States. This switching—connecting these two networks in order to "disrupt dominant" connections that serve to obscure (Castells, 2009, p. 431)—is an important form of information literacy. Similarly, in the case of personal search, users could seek out opposing views, to both inform themselves and understand the relations between different arguments. Indeed, switching is a key response to network power, a

"counterpower" (Castells, 2009, p. 431) to networks, like those created by PageRank, that are incomplete representations of the information they claim to characterize. In this way, switching allows for private information (Surowiecki, 2004, p. 10), thus providing for greater coverage of crowdsourced information, and in theory better information-seeking outcomes.

Conclusion

Like other search algorithms, to operate efficiently the PageRank algorithm relies on constructing the Web in a particular way, as a series of hubs. While this method is able to keep up with the rapid growth of Web pages when compared to Web directories, it sacrifices the individual interpretive decision making of the user and replaces it with an analysis of the network effects present in this network. Further, rather than simply recording the network, it affects the network itself. When a page appears near the top of *Google* search results, this reinforces that page's importance as a hub, thus tending to cause the page to acquire more links through preferential attachment. PageRank, by identifying and reinforcing the hub-structure of the Web and limiting the agency of users in determining value, has a powerful role in creating hubs on the net, effectively programming the primary version of the network that many users interact with. This is also true of sub-networks generated by *Google's* localization and personalization features. While there has been great utility in this structure, users should not simply accept it as being a stand-in for quality, as Brin and Page (1998, p. 107) argue, but should both recognize the function of the algorithm in programming this version of the network and identify those instances where this program can actually conflict with the identification of quality information. While this paper is merely the beginning of a longer inquiry into the effects of algorithms on users, here I have suggested that, given the networked nature of this information, the most effective user response is to engage in switching, or bringing multiple networks into contact in order to counteract the biases of "dominant" connections (Castells, 2009, p. 431) that exist in more powerful networks.

This view of search algorithms could open some new areas for research in this field. First, researchers could examine search results not simply as lists of ranked pages,

but rather as networks generated in response to particular search queries. While we are aware of localization and personalization, we know little about the actual effects of these features. Evans (2007) has attempted to examine search results by seeing how sites with particular search engine optimization (SEO) enhancements are ranked by *Google*. Similarly, an analysis of the network connections—or lack thereof—between items in a list of search results could offer insight into how *Google's* algorithms work and the type of coverage they offer. Comparing the coverage of such networks to those generated by other search terms or affected by localization or personalization could yield valuable information about how these features impact search results. Further, an investigation of other network effects, such as those outlined by O'Reilly (2005) in his description of Web 2.0 would be beneficial for understanding how the networks created by search algorithms affect users. For example, the role of "user contributions" (O'Reilly, 2005) and their affect on the creation of these networks is an important one that should be examined. Finally, a fruitful area for information-literacy research would be the studies of actual users' switching habits and how those users cope with search limitations in information-seeking. Such research is necessary, for this analysis has only introduced the possibility of these effects on search results. We need a greater understanding of how actual searches impact users in order to inform greater search- and information-literacy initiatives.

Acknowledgments

This research was sponsored in part by grants from the West Virginia University Office of the Provost and the Faculty Development Grant Program. The author would like to thank the editors and reviewers of *SPIR* for their helpful feedback during the revision process for this paper.

References

Barabási, Albert-László (2002). *Linked: The new science of networks*. Cambridge, MA: Perseus.

Brin, Sergey, & Page, Lawrence (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30*, 107–117.

Castells, Manuel (2000). *The rise of the network society* (2nd Ed.). Oxford: Blackwell.

Castells, Manuel (2009). *Communication power*. Oxford: Oxford UP.

Dennis, Simon, Bruza, Peter, & McArthur, Robert (2002). Web searching: A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology 53* (2), 120-133.

Ding, Chris, He, Xiaofeng, Husbands, Parry, Zha, Hongyuan, & Simon, Horst D. (2002). Pagerank, hits and a unified framework for link analysis. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.* Tampere, Finland.

Evans, Michael P. (2007). Analysing google rankings through search engine optimization data. *Internet Research 17* (1), 21-37.

Wikipedia. (n.d.). *Google bomb*. Retrieved from http://en.wikipedia.org/wiki/Google_bomb

Wikipedia. (n.d.).*Google personalized search*. Retrieved from http://en.wikipedia.org/wiki/Google_Personalized_Search

Google, (2012) Report spam, paid links, malware, and other problems to google. *support.google.com*. Retrieved from http://support.google.com/webmasters/bin/answer.py?hl=en&answer=93713

Gu, W.L., Chen, W., Chen, J., & Lu, X.Y. (2012). Improved pagerank algorithm. *Jisuanji Xitong Yingyong - Computer Systems and Applications 21* (2), 214-217.

Halavais, Alexander (2009). *Search engine society*. Malden, MA: Polity.

Heineman, George T., Pollice, Gary, & Selkow, Stanley (2009). *Algorithms in a nutshell*. Sebastopol, CA: O'Reilly.

Huffman, Scott. (2012). Search quality highlights: 39 changes for may. *Google Inside Search* , June 7. Retrieved from http://insidesearch.blogspot.com/2012/06/search-quality-highlights-39-changes.html

Iván, Gábor, & Grolmusz, Vince (2011). When the web meets the cell: Using personalized pagerank for analyzing protein interaction networks. *Bioinformatics 27* (3), 405-407.

Jenkins, Henry, Clinton, Katie, Purushotma, Ravi, Robison, Alice J., & Weigel, Margaret (2006). *Confronting the challenges of participatory culture: Media education for the 21st century* John D. and Catherine T. MacArthur Foundation. Retrieved from http://digitallearning.macfound.org/atf/cf/%7B7E45C7E0-A3E0-4B89-AC9C-E807E1B0AE4E%7D/JENKINS_WHITE_PAPER.PDF

Latour, Bruno (2004). *Politics of nature: How to bring the sciences into democracy*. Cambridge, MA: Harvard UP.

Nie, Lan, Davison, Brian D., & Qi, Xiaoguang (2006). *Topical link analysis for web search. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* Seattle, Washington, USA.

O'Reilly, Tim (2005). *What is web 2.0: Design patterns and business models for the next generation of software*. Retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Page, Larry, Brin, Sergi, Motwani, Rajeev, & Winograd, Terry (1998). *The pagerank citation ranking: Bringing order to the web*. Retrieved from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Segal, David (2010). A bully finds a pulpit on the web. *New York Times,* Nov 26. Retrieved from http://www.nytimes.com/2010/11/28/business/28borker.html

Singhal, Amit, & Cutts, Matt. (2011). Finding more high-quality sites in search. *Google Oficial Blog,* Feb 24, Retrieved from http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html

Surowiecki, James (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday.

Takahashi, Yuku, Ohshima, Hiroaki, Yamamoto, Mitsuo, Iwasaki, Hirotoshi, Oyama, Satoshi, & Tanaka, Katsumi (2011). *Evaluating significance of historical entities based on tempo-spatial impacts analysis using Wikipedia link structure*. Paper presented at the *22nd ACM conference on Hypertext and hypermedia*.

Vaidhyanathan, Siva (2011). *The googlization of everything (and why we should worry)*. Berkeley: U of CA Press.

Waldrop, M. Mitchell (1992). *Complexity: The emerging science at the edge of order and chaos*. New York: Simon & Schuster.