

Impact of Search Engine Editorial Policies on Small Business and Freedom of Information

Ted Goldsmith

[Search Engine Honesty](#)

Azinet LLC

April, 2007 REV 7/07

Introduction

Search engines have become an essential part of the way we use the Internet to communicate. Although nearly anyone can now develop and operate a web site, public access to all those sites is dependent on search engines. In contrast to the proliferation of web sites¹, market forces act to consolidate search. At present, only three major search engines (Google, Yahoo Search, and Microsoft Search) provide more than 90 percent of all U.S. searches² and Google's share is more than 60 percent and rapidly increasing. Continuing consolidation is likely. The same companies are also extremely strong worldwide and support many languages.

Courts have determined that search engine results pages are publications and that therefore search engines have editorial free-speech rights similar to those of a newspaper to edit, bias, censor³, manipulate, and otherwise alter search engine results⁴ in almost any desired manner. The modern information processing technology used by search engines can be used to implement any desired editorial policy and all the major search engines have policies that direct how they process and display search results. As will be

¹ **Web Site:** A collection of web pages hosted at a single domain name (e.g. somedomainname.com).

² **Search Traffic:** Nielsen//Netratings measured searches performed by U.S. home and work web surfers for April 2007: Google and its partners (AOL Search) using Google data 60.6 percent of searches, Yahoo 21.9 percent, MSN Search 9.0 percent, Ask 1.8 percent, others (total) 8.5 percent; for July 2006: Google and its partners (AOL Search) using Google data 55.5 percent of searches, Yahoo 23.8 percent, MSN Search 9.6 percent, Ask 2.6 percent, others (total) 11.1 percent.

³ **Censoring – Semantic Note:** The term *censoring* is normally used to refer to government or other official deletion or suppression of information provided by others. *Redaction* has a similar connotation. In connection with traditional publications, the exercise of editorial authority to elect not to publish certain information would probably be referred to as "editing." However, in a traditional publication, such election is typically a small part of the editorial process. Reporters do not write stories expecting that there is a significant chance the story will be entirely rejected.

In search engines, editorial policies are primarily exerted by restricting or preventing access to information *provided by others* that would otherwise be available. In this document, "censor", for lack of a better word, means deleting reference to or display of web site information in implementation of an editorial policy. "Suppression" means negatively altering the display of certain information such as by altering page position. Deletion or suppression of entire web sites is common.

⁴ **Search Engine Results:** Specifically non-sponsored or unpaid results listings (also known as *organic* results) as opposed to paid or sponsored "results" (advertised web sites).

described, the current situation has a severe negative effect on small business and can be expected to greatly impact our general freedom of information access.

Purpose of Search Engine Editorial Policies

Search engines all have similar mechanisms for operating including complex software constructs or algorithms that govern how they acquire and index data from web pages using robot “spiders” and how they rank different pages in search results. See [Search Engine Mechanics](#) for a detailed description of these systems. Implementing these systems involves judgments that could be considered part of an editorial policy. Most search users who give it any thought realize that such algorithms and judgments must exist in order to implement a search engine but assume these algorithms and judgments are applied equally and fairly to all web sites in an essentially mechanical manner. The spidering and ranking algorithms indeed contain a great amount of general criteria that are applied equally to all web sites; however, search engines also exercise editorial policies against specific individual web sites.

It has recently become widely known that Google is censoring search results displayed to their Chinese language users ([Google.cn](#)) to conceal the existence of specific individual sites that the Chinese government considers objectionable. It is much less well known that all the major search engines also censor English language sites displayed to U.S. users using similar techniques.

A major need for editorial control involves reduction of web spam⁵ and web site deception⁶. People pick a search engine based on the perceived comprehensiveness of search results (ability to find relevant pages), freshness (how often the search engine visits pages and updates its index to reflect new information), and quality of results (probability that a given result page is useful as opposed to “spam”). A poll conducted by *Search Engine Honesty* indicates that the *last factor is the most important for 60 percent of users*. It is therefore no surprise that search engines are trying hard to improve the average quality of their results by censoring or suppressing spam sites and sites employing deceptive techniques.

The *Google Annual Report 2006* says under “Risk Factors”: “There is an ongoing and increasing effort by ‘index spammers’ to develop ways to manipulate our web search results. For example, people have attempted to link a group of web sites together to manipulate web search results. ... If our efforts to combat these and other forms of index spamming are unsuccessful, our reputation for delivering relevant information could be

⁵ **Web Spam:** A web site listed in search results that turns out to have little or no useful content. Spam sites irritate search users while consuming search engine resources. Spam is now an epidemic that especially threatens the smaller search engines. Since new and used domain names are inexpensive, there is a continuing flood of new spam sites using steadily improving technology.

⁶ **Deceptive Web Sites:** Sites that use deceptive practices to gain an unfair ranking advantage over competing sites.

diminished. This would result in a decline in user traffic, which would damage our business.”

All the majors (Google, Yahoo Search, MSN Search) admit on “webmaster guidelines” pages to censoring access by their users to sites that employ deceptive practices to unfairly increase their search engine exposure. Because, like the Chinese government, search engines take the position that any site that they have deleted “has done something wrong” they prefer the terms “banning” or “penalization” to “censoring.” However, functionally, there is no difference. A site that has been censored (or “banned”, “penalized”, “blackballed”, “blacklisted”, “de-listed”, or “removed from our index”) cannot be found no matter how relevant its pages are to a search. Censored sites are manually, on a site by site basis, removed from and barred from a search engine’s index. (Search engines can also use *site-unique bias*⁷ to suppress access to individual web sites.) Search engines see use of deceptive practices, also known as “black hat” search engine optimization, as a pro-active attack on the search engine’s integrity and function and therefore employ “punitive” procedures against small businesses using such practices.

None of the majors admit to any censoring or suppression on pages likely to be seen by their users (people doing searches).

Note carefully that there is a difference between outright censoring, in which all (sometimes all but one) of a site’s pages are removed from the index, and a “rank” problem where it is only less likely that a site will be found. It is easy to determine if a site has been deleted. (See [Is Your Site Banned?](#)) It is much harder to detect even gross bias against an individual site in a search engine’s ranking algorithm or depth-of-crawl algorithm. Search engines frequently make general changes in their algorithms causing changes in the ranking of any particular site.

Search engines do delete sites for using deceptive practices but they also de-list sites for inconvenient practices and may also remove sites for competitive reasons or other editorial reasons. Generally speaking, small sites (less than 100 pages hosted on a domain name) do not appear to be currently at risk for being banned except for clearly deceptive practices. It is also possible for any site to fail to appear on a search engine because of technical site configuration issues. A very small and insignificant site could conceivably be missed by a major search engine, especially if it had very few incoming links from other sites.

Nearly 60 percent of our poll respondents said that search engines should provide the option of seeing censored results and 90 percent said that search users should at least be advised that some sites had been intentionally deleted.

⁷ **Site-Unique Bias:** The use of a search engine ranking or depth-of-crawl algorithm that contains or refers to a list of domain names, trademarked names, site-unique phrases, or other site-unique or business-unique information in order to negatively or positively bias search result ranking or page indexing of individual, hand-selected web sites.

Deceptive Practices

Deceptive practices involve features of a web site specifically designed to “trick” search engines. Such practices are designed to take advantage of weaknesses in a search engine's system in order to get an unfair advantage in search engine exposure. The following is a list of common deceptive practices:

- Using invisible text (same color as the background) to feed different text to the search engine spider from that seen by a viewer; using tiny type at the bottom of a page for the same purpose; “stuffing” keywords in “ALT” or “Keywords” tags (usually not seen by viewers); many other similar techniques.
- Programming a web server to detect when it has been accessed by a search engine's spider and feeding the robot different information than would be received by a viewer (“cloaking”).
- Use of multiple “doorway pages” that are each designed to be optimum for a particular search engine.
- Links in locations or on pages that would seldom or never be seen by human visitors for the purpose of gaming link popularity.
- Gross duplication of data such as hosting the same site on multiple domain names. (See [The Redundancy Explosion](#).)

Deceptive practices are typically aimed at improving the search results rank a site would have for particular keywords relative to a “legitimate” site on the same subject. This problem is made more difficult by the fact that search engines are extremely reluctant to define “legitimate” web site design practices in any detail. “Deceptive” is therefore a gray area and many small-business sites are enticed by unscrupulous “black hat” search engine optimization operators into using techniques that may cause temporary rank improvement but eventually result in penalization.

A second goal may be to increase site traffic generally by using hidden keywords for popular but off-topic subjects. This could be useful if the site is displaying pay-by-impression advertising or advertising a subject of very general interest (e.g. CocaCola).

Search engines *may* be willing to describe the particular deceptive practice causing censoring if requested by a webmaster. In addition, Google is reported to be testing a system whereby webmasters of sites censored for a deceptive practice would be advised by means of email to webmaster@hostname.com that their site has been censored and the reason for the action. All the major engines have procedures whereby webmasters that notice that their site has been censored, determine the nature of the deceptive practice, and fix it, can apply for reinstatement (See [Webmaster Guidelines](#).)

Our impression is that sanctions for grossly deceptive practices are more or less fairly applied. A large-business, Fortune 500 website engaging in clearly deceptive practices will likely be censored as well as a minor site. A major car manufacturer's site was recently briefly censored by Google, apparently for using doorway pages. Reinstatement (Google's term is “reinclusion”) is likely to be much slower for a small-business site that

has ceased the deceptive practice. Google is widely reported to have “punishment” or “timeout” periods preceding site reinstatement.

Inconvenient Practices

Inconvenient practices involve features of web sites which, while not deceptive, nonetheless represent a problem for a search engine. More specifically, the automated, software-driven processing at the engine does not handle these features in a way that is satisfactory for the search engine’s management. It is easier to manually delete thousands of entire sites than fix the problems with the software. Notice that in this case the site isn’t “doing anything wrong”; the problem is actually at the search engine. If the search engine design were changed such that another feature became a problem, then sites having that feature would also be censored. Sites that have been operating for five years or more have been suddenly banned by search engines. (See [Case Studies.](#))

Censoring for inconvenient practices is much less fairly applied. Banning of large-business sites for convenience, competitive, or editorial reasons is rarely, if ever done. If Google censored Amazon for convenience, competitive, or editorial reasons there would be hell to pay. Suits would be filed, Congressional investigations would be held, PR campaigns would be executed. A small-business owner doesn’t have these advantages.

If a small-business site has been censored for an inconvenient practice, it may be very difficult to determine which aspect of the site is causing the problem. Search engines are understandably very reluctant to disclose, especially in writing, that they have suppressed public access to an entire site for their own convenience. They are even more reluctant to disclose that a site has been suppressed for criteria that are conspicuously not being applied to other sites.

Here are some potentially inconvenient practices and features:

-Large number of pages – Sites with a large number of pages may be a problem for some search engines. If the engine indexes the entire site, a large amount of search engine resources (disk space, bandwidth) could be consumed by a site that might not be very important. Normally, we would expect the depth-of-crawl algorithm to handle this by indexing a relatively smaller number of pages in sites receiving relatively less traffic or otherwise having less [merit](#). There is increasing evidence that the major search engines do indeed ban small-business sites merely for having a large number of pages. It is also true that all medium and large sites have (by definition) a “large number of pages.”

-Links. Sites that have a large number of outgoing links such as directories or sites with a large “links” page may tend to upset the link popularity scheme for some search engines. Sites that have forums, message boards, guestbooks, blogs, or other features that allow users to publish a link may also be seen as interfering with the link popularity concept. Google’s *PageRank* site ranking algorithm is less susceptible to these problems because it automatically penalizes pages for outgoing links while rewarding them for incoming links. Some site owners claim that their sites have been censored merely for having a

links page. Buying or selling of links, or requiring links in return for a service could improve a site's link popularity.

Google says in a form email sent to web sites that inquire why they have been banned:

“Certain actions such as buying or selling links to increase a site's PageRank value or cloaking - writing text in such a way that it can be seen by search engines but not by users - can result in penalization.”

The largest single buyer of links is probably Amazon, which has a very successful affiliate programs in which web sites get a commission for duplicating Amazon provided data and linking to Amazon pages. Needless to say, Google has not “penalized” Amazon. Google reports (3/06) indexing 144 *Million* pages at Amazon.com! Yahoo sells links from their directory. Google reports indexing 233 *Million* pages at Yahoo.com. AOL/Netscape's Open Directory Project (ODP) specifically requires users of ODP data to install massive link farms pointing to ODP, the sort of practice that Google's annual report mentioned. Google has not instituted punitive procedures against AOL. Google runs a copy of the Open Directory on its own site (directory.google.com). Google reports (3/06) they index 12.4 *Million* pages in their own directory, each page of which links to AOL's ODP.

Censoring for Competition Suppression and Editorial Reasons

Directories or other collections of links compete directly with search engines. Our [case studies](#) strongly suggest that search engines censor small-business directory sites in order to suppress competition. Search engines also engage in many other business activities such as selling of things, provision of email, message board, video, photo, and mapping services, etc. As long as it is legal to do so, it is unreasonable to expect that they would not suppress larger competitive small-businesses in search results. Suppressing of other larger small-business sites that compete with a search engine or compete with a business partner of a search engine is also likely. Censoring a single small-business competitor would certainly have no effect on the bottom line of a major search engine. Censoring thousands of such sites would obviously have a beneficial effect.

There is currently no convincing evidence that any of the major search engines bans small sites (less than 100 pages) for editorial or competitive reasons. There are plenty of small “Google sucks” sites out there that have not been banned by Google.

The National Security Argument

There is a National Security Argument that goes to the effect of: “We are punishing you but we can't say why we are punishing you or give you an opportunity to defend yourself because doing so might disclose information that could be used by the enemy.” Search engine people use a version of this argument to justify their refusal to disclose why a

particular site has been censored. The idea is that a spammer might have found and exploited a previously undisclosed weakness in a search engine. If the search engine discloses the reason the site has been banned, it might add some confirmation that the deceptive technique works. The spammer might spread the word or be more likely to use the technique on another site.

This argument may have had some validity ten years ago but is currently ridiculous. Search engines have been around for a long time (by Internet standards). Search technology is well developed. Abuse methods are now well known; you just read about most of them. A spammer that has implausibly found a new weakness has other ways to measure the effectiveness of its method.

A much more plausible explanation is that search engines need to conceal the fact that the site is being banned for a practice that others are being allowed to continue (buying links, duplication of data, directories, links pages, guestbooks, message boards, etc.) or that the site is being censored for competitive or arbitrary editorial reasons. Search engines are using the “national security argument” to conceal their own discriminatory practices.

Google has announced a plan to notify *some* webmasters of censored sites (presumably the ones that have been banned for clearly deceptive practices) that their site has been censored and the reason for the action. The notification will be automatic and *not* at the request of the webmaster. Our understanding is that Google will generally continue to refuse to disclose the reason for censoring to webmasters that *do request* such notification. This allows Google to disclose that a site has been censored for a deceptive practice while continuing to deny that it is censoring other sites for reasons other than deceptive practices.

The Google Sandbox

Many webmasters report that Google has a “sandbox” in which websites are confined “for being bad” but after they have ceased the “bad” behavior, requested reinstatement, and have been reincluded in Google's index. The site is no longer completely censored and can be found in the index, but has an abnormally low rank, much lower than it had before being banned. New websites are also said to be sent to the sandbox for some period of time. Google people have generally denied the existence of a sandbox although some agree that there is a “sandbox *effect*.” (Notice an interesting continuation of the parent-child psychology here. Webmasters are often willing to see themselves as children being “punished” for “being bad” by being sent to the “sandbox” for a “timeout.” They are very willing to assume that if they are “banned” or “in the sandbox”, they are “doing something wrong.”)

The sandbox effect *could* be partly explained by the [site popularity](#) factor. A site banned by Google would lose traffic and therefore lose site popularity. When the site was restored to the index, it would still have reduced traffic, and site popularity and therefore poorer rank initially. Gradually, traffic and therefore site popularity and rank would improve. Voila, the sandbox effect. This effect would be much more noticeable with

Google than with the other search engines because Google generally contributes more to a site's traffic. Therefore a site banned by Google would lose more traffic and site popularity. If this were the case, we would expect to see at least somewhat proportional reaction from the other major search engines.

However, using the age of a site (domain name) as a factor in ranking is an obvious anti-spam technique. Since spam sites are continuously being discovered and banned by search engines, spammers have a continuing need for new unsullied domain names. Therefore, a site that has been operating for a long time is much less likely to be a spam site and it makes sense for search engines to rank newer sites lower. Spammers counter by buying previously-owned domain names released by failed non-spam web sites in order to get their "history" and residual incoming links. (Legitimate site owners can be inadvertently banned because of purchasing or inadvertently duplicating a domain name previously used by a spammer.) Search engines have apparently countered by monitoring domain name use. If a name used to host a web site disappears (web site down) and then reappears it might now be owned by a spammer.

Some webmasters report that their Google referred traffic suddenly dropped drastically and never recovered, and that there was not a great correlation with other search engine traffic. They further report a similar drop in PageRank reported by Google. This suggests that Google is "manually" adjusting PageRank (site-unique bias) for some sites.

Other Censoring Issues

The people in a censoring department, be it in China or at a major search engine, are probably relatively poorly paid. They sit at a monitor all day adding sites to the censored sites list. If they exceed their quota, they might get a bonus. These people may spend time randomly surfing around looking for sites that meet the criteria specified for censoring but they are more likely to be reviewing sites that have been *nominated* for censoring. They are not paid to make subtle and complex value judgments.

All the major search engines have a system where anybody can nominate a site for censoring by filling out an online "spam report" form. People can nominate the sites of their competitors or any site they don't like for any reason. The larger the traffic a site has and the more people that see the site, the more likely it is that someone will nominate it.

The search engines can also have mechanical means to nominate sites. For example, tracking data⁸ and site popularity data can be analyzed in order to produce nominations.

Censoring for competitive or convenience reasons is important to a search engine only if the target site has significant traffic or has a large number of pages indexed. If the site is

⁸ **Tracking data:** Data showing the web usage (pages visited, time spent on each page, time-of-day visits were performed, subject matter of pages visited, etc.) of individual web users. Usage patterns of legitimate users differ from click-fraud participants and can also be used to identify spam sites. Tracking data can be obtained anytime a user visits a web site owned by a search engine or one of its advertising partners.

“naturally” getting very few referrals (clicks) from the search engine, then the beneficial effect of banning that site would be minor. Many small-business site owners report getting banned only after they achieved considerable size and popularity.

Because censoring can be performed from anywhere, the censoring department is an obvious choice for outsourcing to an offshore location that has lower labor costs.

One problem is that censoring could obviously be used for all sorts of nefarious purposes. A censor could decide to rent himself out to the highest bidder. If advertising is going for \$2 per click, imagine what it would be worth to censor a competitor’s site, even temporarily, just before Christmas. Maybe one of the censors just doesn’t like Democrats. The possibilities are endless. Even if search engine management has not, at a corporate level, used its censoring power for such purposes, what steps have they taken to prevent abuse by individual censors or groups of censors? If censoring is done offshore, there are additional concerns.

All of the precautions cost money. Censored sites could be reviewed at a second level but it would cost more. Complaints from sites that happen to notice that they have been censored could be reviewed by people other than the group that did the original censoring, but it would cost more. Since search engines consider all of these practices to be trade secrets, we have no way of knowing what precautions they employ. Since search engines consider that censoring is an exercise of their free-speech rights and that they have no obligation to webmasters, it appears unlikely that they would pay for extra precautions to protect web sites.

Google has recently changed their policy to re-review and consider reinstating only those sites that stipulate in writing, in advance, that they are guilty of a deceptive practice and have discontinued that practice. Google will no longer consider reinstating a small-business site that has been banned for competitive, editorial, or convenience reasons, a site banned for reasons not specified in their guidelines, a site banned by mistake, or a site that does not know why it has been banned. One can easily infer that this change was implemented because of massive complaint volume from sites banned for non-deceptive reasons.

Search Engine Bias and Site-Unique Bias

Recall that in order to incorporate site factors such as site popularity and age-of-site into their ranking algorithms, search engines necessarily must maintain a database of *sites* as well as a database of *pages*. We have mainly been discussing outright censoring (banning) in which a web site is completely excluded from access by a search engine’s users. Outright censoring is [easily detected](#) by a site owner. All the major search engines admit to banning on a site-by-site basis. Major search engines will, upon request, generally confirm or deny that a particular site has been censored while often concealing the reason for such action.

However, all the majors claim or at least strongly imply that their ranking and depth-of-crawl algorithms are fairly applied equally to everybody. They say that if your site is not censored, any sudden change in rank is due to a general change that the search engine made in its algorithm. The same algorithm is applied to all sites handled by that search engine.

The preceding paragraph is literally true. The algorithms are indeed applied to all sites. However, this does *not* mean that there cannot be gross bias against particular individual sites incorporated *into* an algorithm. For example, a ranking algorithm could easily contain a “rule” that says: “Check if this site is on our ‘bad sites list’, if so, subtract 264 from its merit ranking value; if it is on our ‘good sites list’, add 29 to merit.” This sort of site-unique bias would be more subtle than outright censoring and more difficult for a site owner to prove, even if it was so severe that it was effectively impossible to find the site in a search.

Site-unique bias is easy to incorporate into the site database that a search engine must already maintain. Much more complex site-unique bias schemes are obviously possible. We can define *site-unique bias* as a case where a search engine ranking algorithm or depth-of-crawl algorithm contains or refers to site-unique information such as domain names, trademarks, words, or phrases unique to a particular site, or other information identifying particular sites in order to apply bias to the ranking or crawling of individual, hand-selected sites. Many site owners claim that their drop in ranking at a single search engine is so catastrophic that it must be the result of site-unique bias. (See [Kinderstart Case](#) for a convincing instance of site-unique bias.) Because site-unique bias is more difficult to prove than outright censoring, it represents an opportunity for search engines to avoid some of the hassles surrounding censoring such as the need for “re-review” of complaining censored sites. (A bias scheme that is uniformly applied to all sites (site popularity, age-of-site) is not considered to be site-unique bias.)

Search engine bias can also be used against specific ideas as opposed to against specific sites. Most people would consider it reasonable to rank pages and sites containing pornographic words lower than pages not containing pornographic words, especially if the search terms did not contain pornographic words or phrases. Most would also consider it reasonable to rank pages containing phrases such as “site under construction” below otherwise similar pages. Any search algorithm presumably expresses the results of many such judgments. The same technology could also easily be used to rank pages containing “Democratic candidate” below pages containing “Republican candidate”, an action most Democrats would consider “unreasonable.” Because algorithms are considered trade secrets, it is difficult (but not impossible) to determine if search engines are employing “unreasonable” bias. As far as we can determine, political search bias is not currently illegal.

Anti-Competitive Impact of Censoring on Small Businesses

Search engine censoring departments are careful not to censor or apply major negative site-unique bias to any site that appears to be owned by a relatively large business unless the site employs practices *clearly* intended to deceive such as hidden text or doorway pages. Large businesses have lawyers, publicists, and other ways of “pushing back” if their site is censored for the convenience of the search engine or for competitive or editorial reasons and the benefit to the search engine of censoring any one site is generally relatively small. Therefore, essentially all censoring for these purposes is done against small businesses. For example, our studies found no case in which a large business using Open Directory data had been censored by any major search engine while small businesses using Open Directory data or otherwise containing directories are very frequently censored, especially by Google. (See [Search Engine Censoring of Sites Using Open Directory Data](#), [The SeekOn Case](#), and [The Kinderstart Case](#).) Amazon and other large companies are allowed to “buy links” or use “non-original” content where small companies are frequently censored for doing so. Search engine censoring therefore works to suppress small businesses in favor of large businesses. This is especially unfortunate because otherwise the web represents a major opportunity for smaller businesses.

Issues with Search Engine Editorial Policies

Is there really anything legally wrong (or even morally wrong) with a search engine having an editorial policy? Many people would say no. Maybe a search engine has just as much right as a newspaper or magazine to determine what information they pass on to their users. A search engine is investing in its index just as a newspaper or magazine is paying for each square inch of page space. Maybe a search engine should have equally complete control over what they put in their index. We don’t expect to see articles favorable to *Newsweek* or other competitor in *Time*. We expect publications to have an editorial “slant” or bias. There is not even any requirement or expectation that a newspaper or magazine disclose its editorial policies.

It should also be obvious from the discussion of [“merit” ranking](#), algorithm bias, site databases, and the existence of censoring departments that search engines have the built-in capability to execute any desired editorial policy. If a search engine decided to be “right-wing” or “left-leaning”, or to suppress competition, the necessary infrastructure already exists. There is no technical reason and apparently currently no legal reason why the merit algorithms couldn’t easily be adjusted to favor sites about Republicans over those about Democrats or institute any other desired point of view.

However, the existence of the capability for search engine editorial policies and strong evidence suggesting that such policies are being implemented is very disturbing for several reasons. In many ways a search engine is *not* like a newspaper or other publication.

Unaware Audience: Anybody who has been born and raised in a free country knows all about editorial bias in media including newspapers, magazines, radio, and TV. “Filtering truth” when obtaining information from these sources is second nature. However, most people do not think of editorial bias, censoring, or

other filtering as applicable to search engines. Search engines are seen as mechanical devices and therefore incapable of bias. A search for a pornographic word or phrase returns millions of hits adding to the false impression that no censorship or editorial policy is in place. Search engines do everything they can to enhance this impression. In our opinion, this is deceptive and dishonest.

Unfulfilled Need: Most people use search engines precisely because they are trying to get access to the largest and most diverse body of information possible with the least amount of editorial filtering possible. If you don't mind having your information filtered through some editorial filter, there are many much better sources of information. The growth of the Internet itself was largely fueled by the public's desire for unedited, uncensored information. Search services that are more openly editorial (such as Ask.com) have been relatively unsuccessful.

Absence of Diversity: In the United States alone, there are quite a few TV networks, radio networks, newspapers, and magazines. Worldwide there are many more. But there are only three major search engines *worldwide*. A **very small group** of people is setting the editorial policies for these search engines. This small group is controlling a very important source of information for a very large number of people. (See [The Web Czars](#).) This has important implications for the political process in the United States and elsewhere where elections could be influenced by search engine bias. Even worse, if the most comprehensive search is needed, Google is rapidly becoming a sole source.

Control Without Responsibility: Publishers, while having complete editorial *control* over their publications, are also *responsible* for what they publish. If someone libels a person in a newspaper article, that person can sue the newspaper. If the publication contains illegal material such as child pornography, or promotes illegal activity, it can be shut down. Rules for publishers have been developed during a period spanning more than 500 years.

At the same time it is understood that an organization that merely acts as a conduit or connection service for information (Telephone Company, Internet Service Provider (ISP), Post Office) is not responsible for the content of the information they convey. You can't sue the Post Office because you were taken by mail fraud. You can't blame the phone company if you get a threatening call. ISPs have (so far) been able to avoid any responsibility for information they convey or store (e.g. child pornography) as long as they *do not* edit or filter the information.

However, organizations that act as information conduits (connection services) are not allowed to pick and choose the information they carry. The phone company is not allowed to decide, using secret internal criteria, who can have a phone or what they can say on the phone. Any restrictions regarding who can or cannot have or use a phone (or other connection service) must be very well documented, very

public, and very fairly enforced. If it were not this way, the phone companies would be running the country.

The major search engines currently have it both ways. They are able to editorially and using undisclosed criteria “cherry-pick” the information their users are allowed to see while denying any responsibility for the content of that same information. Will they be able to continue to do this indefinitely or will a court or legislative decision eventually force either more responsibility for content or more fair handling of information providers? We will have to wait and see. The Internet and search engines are new technology. Law and regulation have not caught up yet.

Essential Nature of Search: Imagine what would happen to most businesses if their phone service suddenly disappeared. What if the phone company could arbitrarily refuse to reconnect them for no stated reason? Maybe the phone company blackmails the business into buying “advertising” in order to be reconnected. Now imagine that there were only three major phone companies and one of them controlled more than 60 percent of all phone traffic. The main way for more than half of the people to reach you is through this company. If *Time* never publishes a favorable story about a business, they can certainly live with that. For most businesses, being disconnected by the phone company would be a terminal event. For many, being disconnected by Google is equally terminal. This will be increasingly true in the future.

Service Nature of Search

Many people clearly and increasingly use search as a connection service. Many searches are for company names or other company-unique information. If someone searches for a unique trademarked company name and that company’s site does not appear near the top of Google results the searcher can reasonably conclude that the company has gone out of business or is so backward that it does not have a web site. A search for the corner gas station produces a response. Certainly a search for any company that does any significant business would also produce a result.

Also, in providing a “clickable link” to destination web sites, search engines clearly provide a connection service. It is far easier to merely click on a link to connect to the web site than correctly typing www.searchenginehonesty.com as would be needed if, for example, you read about the site in the newspaper or other print publication and for some reason wanted to connect without using search.

Private Nature of Communication: Publications are public. However, search engine results are private. John Q. Public presumably does not want the fact that he is searching for “hot blond chicks” to become public and expects privacy in the same manner that he would expect privacy in a telephone connection or other connection service. People making phone calls expect not only that the content of

the conversation is private, but also that the time and date of the call and person or business called is private, unless obtained under court order. People using search services to connect to web sites have the same expectations.

Individual Nature of Communication: Publications are designed for a mass audience. However, like information conveyed in telephone conversations, search engine results are specifically designed for the single individual that conducted the particular search. Search results are a service not a publication.

Source of Information: A telephone company may provide the wires, software, and other infrastructure for processing and handling your voice but you are providing the information content when you talk on the phone. The phone company does not own your communicated information and is not allowed to use it for their own purposes even though it has access and technically could easily eavesdrop or record your conversation. Similarly, search engines don't actually provide original information in results data but only mechanically process and handle information provided jointly by web sites and searchers. Web sites and searchers provide this information for the express purpose of obtaining connection. The “publisher” is actually the web site. The search engine is a connection service. If anybody has “free-speech” rights it should be the web site.

So the 64 billion dollar question is this: Is a search engine more like a newspaper or more like a telephone company? Is a search engine providing a publication or a connection service? In minor court cases fought by large-cap search engines against tiny small businesses (e.g. [Kinderstart v. Google](#)), search engines have so far been able to maintain the idea that search results are publications and that therefore they have free-speech editorial rights over search results. We can expect to see this question argued very extensively in the courts as well as in the court of public opinion in the next few years. For an illustration of search engine censoring issues see [The Googlecomm Fable](#).

Google Defamation of Small-Business Sites

PageRank (PR) is Google's [merit factor](#) used (along with search term relevance) in determining the ranking of pages in Google search results. Where other search engines internally develop similar merit factors for sites and pages they index, they do not publish the merit factors. Google publishes the PageRank (as a number between zero (minimum page “value”) and ten (maximum) with a corresponding length green bar) of sites listed in their [Google Web Directory](#), which is a clone of the [Netscape Open Directory](#). (The only significant difference between Google's directory and the AOL/Netscape Open Directory is the addition of PageRank.)

Users can also download a free Google toolbar that displays PageRank of any page being displayed in the user's browser. Users can use the PageRank to assess the “quality” and “importance” of the site and page they are viewing. Generally, even very low-traffic, very minor sites have a PageRank of at least 2. Google's home page has a PageRank of 10. Other very popular sites (Yahoo, MSN, Excite, AOL) have a PageRank of 9.

Internally, Google certainly uses a PageRank numerical value that has more than 11 gradations for ranking search results. The displayed, 0 - 10, PageRank is thought to be a logarithmic representation of the internal PageRank, which is said to be named after Google cofounder Larry Page as opposed to being named for its page ranking function. As described above, PageRank is applied to sites as well as pages.

Google tells their users (Google Technology <http://www.google.com/technology/index.html> 7/2006):

"PageRank relies on the uniquely *democratic nature of the web by using its vast link structure* as an indicator of an *individual page's value*. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves 'important' weigh more heavily and help to make other pages 'important.'

Important, high-quality sites receive a *higher PageRank*, which Google remembers each time it conducts a search. Of course, important pages mean nothing to you if they don't match your query. So, Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query."

"A Google search is an easy, **honest and objective** way to find *high-quality* websites with information relevant to your search."

"PageRank is the *importance* Google assigns to a page based on an **automatic calculation of factors** such as the *link structure of the web*."

(Google Toolbar Help http://toolbar.google.com/button_help.html 7/2006):

Importance ranking. The Google Web Directory starts with a collection of websites selected by Open Directory volunteer editors. Google then applies its patented PageRank *technology* to rank the sites based on their *importance*. Horizontal bars, which are displayed next to each web page, indicate the *importance* of the page, *as determined by PageRank*. This distinctive approach to ranking web sites *enables the highest quality pages to appear first as top results* for any Google directory category". (Google Web Directory Help <http://www.google.com/dirhelp.html> 7/2006)

[emphasis added]

Google's description for their users certainly unequivocally and emphatically states that PageRank is "honest", "objective", and the result of an "automatic calculation", that is completely dependent on external factors such as the "democratic nature of the web" and whether the site is "highly-regarded by others" as indicated by links to the site from other sites. However, as we have seen, there is substantial evidence that Google is using its own, internally determined, subjective, and manually applied site-unique bias to suppress PageRank for individual hand-picked sites.

Also, for pages on sites that are censored (banned) by Google, Google's directory listing and toolbar indicate a PageRank of zero (minimum). Google's toolbar indicates "not

ranked by Google” as opposed to zero on those pages that have *not* been evaluated by Google's system. A zero therefore means to a Google user that the page *has been* evaluated by Google’s automated, honest, and objective technology and found to be meritless and of minimum “importance” and “quality.” In actuality, a zero in the case of a banned site means that the site has been manually banned by the Google censoring department based on undisclosed subjective criteria and has little or nothing to do with external factors such as how the site is regarded by others or any other objective criteria.

It therefore appears that a zero or otherwise artificially depressed PR for pages on a manually censored or biased site, in combination with Google's description of the automated, honest, and objective nature of PageRank, represents a knowingly false derogatory statement (defamation) by Google regarding the suppressed web site. Google is “adding insult to injury.”

[Kinderstart](#) sued Google, in part, based on the idea that Google is engaging in defamation and libel of web sites that have been banned (e.g. [SeekOn](#)) or been subject to “blockage”, arbitrary assignment of PR=0, or other arbitrarily imposed reduction in PageRank (e.g. KinderStart).

See the [Open Directory Case Study](#) for data showing that Google sets PageRank to zero for sites that have been banned by Google. Google admits to the practice of manually banning sites based on undisclosed criteria.

Since Google apparently only censors or suppresses small businesses, the defamation issue currently only impacts small businesses.

It seems that Google is phasing out their Open Directory clone. It has not been updated since 2005.

Analysis

The major search engines have been very successful in convincing most search users that search results are “fair”, “honest”, and “objective”, while simultaneously convincing courts that search results are merely “opinions” of an editorial entity. If practices such as those described here were to become common knowledge among search users, or worse yet, the majority of users began to see search engines as editorial entities, all the major search engines might be adversely affected.

People select editorial entities (newspapers, TV news, etc.) based at least somewhat on the editorial slant of the publication. Conservatives are more likely to read the Wall Street Journal, watch Fox News, and so forth. If people began to choose search engines as they choose newspapers or TV news sources, then organizations such as Fox News and the New York Times would be encouraged to set up their own search engines based on their own already widely respected editorial policies. This would likely adversely affect the existing search triad.

Suppose Google entered into a relationship with Barnes and Noble and then censored all of Amazon's pages (and their partner site pages) from the Google index, apparently a completely legal⁹ and legitimate business maneuver. The resultant publicity would act to destroy the public illusion that Google was not an editorial entity and lead to the scenario described above. This promotes a situation in which censoring and major site-unique bias actions by search engines are only taken against small businesses.

Similarly, in order to conceal the double standard regarding their treatment of small-business sites vs. large-business sites, and the general existence of editorial policies, search engines are unable to publish detailed guidelines for webmasters. This produces widespread confusion among small-business webmasters regarding the rules, most specifically concerning what constitutes "black hat" or "white hat" optimization. Search engines therefore represent a continuing and critical uncertainty for small businesses.

Search engines (especially Google), aided by the dearth of published rules, have also been very successful in convincing small-business web site owners that catastrophic rank problems or complete disappearance of their site from search results is the consequence of some site configuration error or inadvertent violation by the webmaster of some arcane and unpublished rule. Site owners have an enormous incentive to believe that their problem has such an easily fixable cause as opposed to the idea that their site was banned because it was seen as competition or otherwise has a design feature only allowed in large-business sites, an unfixable situation. It can literally take years to determine the truth through a trial-and-error process.

The massive and continuing onslaught of spam limits the resources search engines can reasonably be expected to expend on a per-site basis to determine whether a small-business site is or is not spam and leads to a shotgun approach. See [*Does Google Ban Large Sites by Small Businesses?*](#) for more discussion.

To an unprecedented extent, a search engine's infrastructure is hidden from users. They all have similar simple entry pages into which one enters search terms. The result pages look similar. A search engine could be constructed for .01 percent of Google's capital investment and the only significant difference between this engine and Google would be in the quality of results, an almost completely subjective parameter. Results of the cheap alternative might well be adequate for entertainment purposes. Google at about 60 percent of searches is not a monopoly. Google can point to hundreds of search engines that superficially resemble Google. However, Google has many advantages including massive infrastructure investment that allow it to deliver qualitatively better results than any competitor. People who need high-quality search for research, business, education, or other more serious applications are therefore in an increasingly sole-source situation. See [*Is Google an Unbreakable Monopoly?*](#) for more discussion.

⁹ The author is not a lawyer. Communications law is a very complex area.

The major search engines would likely be the greatest corporate victims of any loss of *net neutrality*¹⁰. The first move any ISP would make would be to put up their own search engine and suppress access to the major search engines. The search engines and their supporters are now in the position of having to claim that ISP censoring or suppression is bad but, somehow, search engine censoring and suppression is OK! Functionally, the effect on end-users or businesses is very similar.

Conclusions

Search engines are legally seen as editorial entities but the reality is that they perform connection services that are increasingly essential to successful operation of any public web site. Web sites in turn are increasingly essential to the operation of most businesses.

Search engine editorial policies heavily discriminate against small businesses.

In view of their editorial-entity status, search engines can legally and properly institute editorial policies that incorporate political bias or any other bias that would be acceptable in other publishing such as newspapers and TV networks. Because of the very small number of significant search engines, because of Google's general dominance, and because of Google's functional monopoly in the "high-quality" search area, this represents a dangerous loss of editorial diversity. If search engines are to continue to be considered editorial entities, the public should be educated to that effect and the appalling lack of editorial diversity must be addressed. If search engines are eventually determined to be connection services then they should be constrained by regulations similar to those that apply to other connection services.

Unless this situation is corrected by legislative or judicial action, small business and freedom of information will be very adversely affected.

Any solution to the net neutrality problem should address these search engine issues.

Additional Information

Azinet operates a web site for discussion of search engine issues at <http://www.searchenginehonesty.com/>. See this site for additional information including the following reports:

The Kinderstart Case

Search Engine Censoring of Sites Using Open Directory Data

Is Google an Unbreakable Monopoly?

Search Engine Mechanics

Search Engine Webmaster Guidelines

¹⁰ *Net Neutrality Issue*: Should cable companies, telephone companies, and other Internet Service Providers be allowed to suppress or degrade their customer's Internet communications with specific hand-picked destinations relative to other destinations?

