1

New Quality Metrics For Web Search Results

Panagiotis Takis Metaxas*, Lilia Ivanova, and Eni Mustafaraj

Wellesley College, Wellesley, MA 02481, USA,
pmetaxas@wellesley.edu
http://cs.wellesley.edu/~pmetaxas/

Summary. Web search results enjoy an increasing importance in our daily lives. But what can be said about their quality, especially when querying a controversial issue? The traditional information retrieval metrics of *precision* and *recall* do not provide much insight in the case of web information retrieval. In this paper we examine new ways of evaluating quality in search results: *coverage* and *independence*. We give examples on how these new metrics can be calculated and what their values reveal regarding the two major search engines, Google and Yahoo. We have found evidence of low coverage for commercial and medical controversial queries, and high coverage for a political query that is highly contested. Given the fact that search engines are unwilling to tune their search results manually, except in a few cases that have become the source of bad publicity, low coverage and independence reveal the efforts of dedicated groups to manipulate the search results.

Key words: Web search, quality metrics, precision, recall coverage, independence, adversarial information retrieval, web spam

1.1 Introduction

The web has changed the way millions of people are being informed and make decisions. Most of them use search engines to access web information. Since people use search engines daily to make all kinds of financial, medical, political or religious decisions, quality of search results is of great importance. In the last ten years the two major search engines, Google and Yahoo, have gained the lion's share in the search market [9].

But does higher market share implies higher search quality? Performance of information retrieval methods is traditionally measured in terms of precision (fraction of results that are relevant to a query) and recall (fraction of relevant items included in the results) [7]. It is well known, however, that web searchers

* Corresponding author. Part of this work was supported by a Brachman-Hoffman grant.

rarely look past the top-10 results [11]. The web has enormous size. More that 50 billion pages are reportedly indexed by search engines, and this represents just a portion of the static web. At the same time, search results on important issues are being heavily spammed [6, 8]. Therefore, high precision is easy to achieve but does not convey useful information, while recall cannot be computed accurately because of the enormous size of the web.

The problem of measuring the quality of search results becomes more interesting when searching for controversial issues. A controversial issue is one that has several possible relevant "answers", depending on one's point of view. Are the results users receive characterized by a reasonably comprehensive *coverage*? In other words, are the various opinions equally represented in the search results?

While search engines are trying to provide unbiased results, Search Engine Optimization (SEO) companies and web spammers are actively trying to force a search engine to list their own sites high on its search results. They do so using a variety of techniques, such as creating "link farms" [6, 9]. How *independent* are the top-10 results? For example, is it possible for a successful group of spammers to claim, not only the top spot in the top-10 search results, but a large group of them?

In this paper we take on the problem of defining coverage and independence of web search results. As far as we know, even though several papers have tried to define search quality (e.g., [1]) the metrics we introduce have never been addressed in the past. In the process we study the structure and density of the web neighborhood that supports each of the web search results according to each search engine, and we observe some interesting characteristics of these neighborhoods and of the way the search engines operate.

1.2 Web search results of controversial issues

To address these questions we decided to do a sequence of web searches on highly contested issues using the two most popular search engines, Google and Yahoo. A controversial (and thus, contested) issue is one that has several possible relevant "answers", depending on one's point of view. The subjectivity of the answers is what separates a controversial issue from a simple ambiguity, which is a separate, well defined search problem. Searching for "jaguar", for example, one expects fo find results that stem from the ambiguous use of the term (a car, an operating system, an animal, etc.) We are interested in queries that are not ambiguous, but have subjective answers. For each of the queries we selected, one can expect that there are at least three possible answers: a "pro", a "con" and a "bal" (short, for "balanced") answer.

1.2.1 Coverage

We argue that for a controversial issue, equal and comprehensive coverage in the top-N results is to have an equal number of pro, con and bal results. We will simply refer to this quality as "coverage" and we will define it below. For example, in the top-10 search results that search engines are giving back by default, equal and comprehensive coverage would be to have 3-4 results (or, on average, 3.3 results) from each category.

Let's assume we have k different categories for a complete coverage (above we have $k = 3$) and we have N results in the top-N slots. Let's further assume that category i received r_i results. We define as *coverage bias* the following quantity B:

$$B = \sum_{1 \leq i \leq k} |r_i - \frac{N}{k}| \tag{1.1}$$

In other words, B is the distance of r_i from the expected number of results N/k. We note that bias B is bounded by:

$$B_{min} = 0 \leq B \leq N + (k-2)N/k = B_{max}$$

Specifically for top-10 search results, when we have equal number of results from each category, we have minimum bias $B_{min} = 0$. At the other end, when one category takes all top-10 spots, the bias is maximized at $B_{max} = 13.3$.

The further bias B is from 0, the worst the coverage is. Therefore we can define as **coverage** C, the lack of bias, that is,

$$C = \frac{B_{max} - B}{B_{max}} \tag{1.2}$$

Coverage C, therefore, has a value between 0 (one-sided coverage) and 1 (equal and comprehensive coverage). When coverage is in the bottom third of this range, we call it *low coverage*, in the middle third *medium coverage*, and on the top third *high coverage*. For example, if the three categories receive 4, 3 and 3 (respectively) of the top-10 results , then bias is small, $B = 0.66 + 0.33 + 0.33 = 1.3$, and coverage is high: $C = \frac{13.3-1.3}{13.3} = 0.9$

1.2.2 Independence

The second metric we introduce is *independence* in search results. To define independence we need first to examine the various ways in which search results can be dependent. We see four kinds of dependent results:

- *URL dependency* is the situation when multiple entries in the top-N results are actually coming from the same site URL, e.g., they correspond to different pages of the very same site, as it is defined by the domain URL. For example, in Table 1.5, results numbered 5 and 6 have URL dependency. In the Tables of this paper we mark one of the two URL dependencies with a (u).
- *Redirection dependency* is the situation when two different site URLs resolve onto the very same location. Redirection is often used by web spammers who try to increase the visibility of a target site by creating many

other sites that will point to the target site [6]. For example, in Table 1.5, results numbered 3 and 4 have redirection dependency. In the Tables we mark one of the two redirection dependencies with an (r).

- *Content dependency* is the situation when the contents of two or more pages included in the search results are essentially the same. The contents may be surrounded by images and menus that are different and are stored on different web sites. This is also a trick used by spammers who try to increase visibility to a target site by creating entries in blogs or "news" sites that lack their own content [6]. For example, in Table 1.11, results numbered 4 and 8 have content dependency. In the Tables we mark one of the two content dependencies with a (c).
- *Link dependency* is the situation when the supporting link structure in the web graph is substantially similar in two or more sites. Link dependency reveals "link farms' of spammers [6]. This type of dependency has been studied extensively in the literature, and it is considered a major tool that the Search Engine Optimization industry is using to acquire high PageRank [3]. In this paper we will focus on the most basic similarity structure, namely the *circular link dependency* between two or more sites. For example, in Table 1.11, results numbered 3 and 10 have circular link dependency, as do results numbered 6 and 7. See Figure 1.1. In the Tables we mark one of the two link dependencies with an (l).

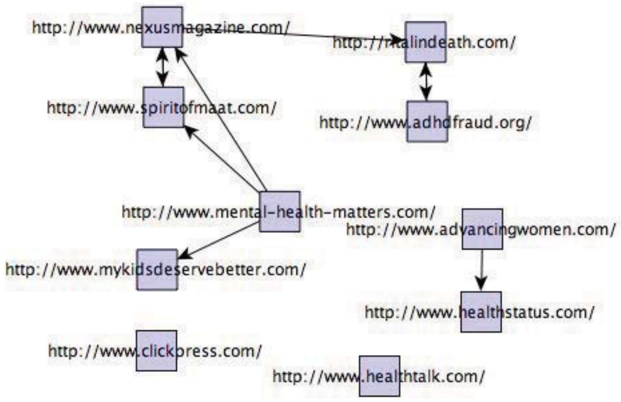
http://www.nexusmagazine.com/
http://ritalindeath.com/
http://www.spiritofmaat.com/
http://www.adhdfraud.org/
http://www.mental-health-matters.com/
http://www.advancingwomen.com/
http://www.mykidsdeservebetter.com/
http://www.healthstatus.com/
http://www.clickpress.com/
http://www.healthtalk.com/

Fig. 1.1. Link dependencies between the top-10 sites of Table 1.11. The backGraph algorithm reveals that there is considerable linking between 8 of these sites, effectively influencing the search results. In particular, the pairs of sites in the upper part have circular link dependencies. These four sites occupy locations 1 and 6 in Google's results and 3, 6, 7, 10 in Yahoo's results.

Let's consider the search results of some controversial query. Let's assume that out of the top-N results, u results are URL dependent, r results are redirection dependent, c results are content dependent, and l results are link dependent. We define as **independence** of the search results the ratio of the non-dependent sites over N, or

$$I = \frac{N - (u + r + c + l)}{N} \tag{1.3}$$

Independence essentially measures the percentage of independent results in a collection of top-N search results. Note that in the formula above we penalize URL dependencies for all but one (their "representative") of the identical URLs. We do the same for all other types of dependencies. Of course, we do not penalize results having multiple dependencies more than once.

Like coverage, independence is also a number between 0 (lack of independence) and 1 (all independent results). As with coverage, the higher the independence, the higher the quality of search results.

We should clarify that, even though it is a very important issue, we are not examining here the correctness or factual accuracy of search results. It is known that misinformation is a serious problem on the internet [12, 2, 5]. But evaluating correctness or accuracy requires significant amounts of time by qualified experts and there is no easy way to be done automatically by computer. We are mainly interested in exploring quality measures that can be computed using automatic or semi-automatic algorithms that will help search engines increase the quality of their search results.

1.3 Experimental Results

An examination of a daily newspaper will reveal many controversial issues for which web users may search, and our colleagues have recommended many others. We chose to follow search results to three queries coming from the commercial, medical and political arena:

- Q1 (commercial): *Human growth hormone (HGH) benefits*
- Q2 (medical): *Is ADHD a real disease?*
- Q3 (political): *Morality of abortions*

Q1 aims at information related to one of the more popular steroids, the human growth hormone. Even though there is ample medical evidence that not only does it not help but may even harm those taking it (see, e.g., [14]), there is a strong industry that is promising (unsupported of any medical evidence) miracles: to reverse aging, reduce body fat, and increase lean muscle, among other things. Companies selling HGH products have strong financial reasons to increase their coverage in the top-10 results, and consequently decrease the coverage of the information that discredits them. For most people, this is not a high visibility issue, however, and it is unlikely that the average web user would know of this controversy.

Q2 aims at information related to a psychiatric condition, Attention-Deficit Hyperactivity Disorder (ADHD - as it is known in the USA) or Hyperkinetic Disorder (as it is known in Europe). While the vast majority of medical doctors accept ADHD as a real disease, there is a small but active vocal opposition by a few parents, doctors and religious organizations (e.g., the Church of Scientology), that dispute it. For many people unfamiliar with the condition, this is also not a high visibility issue.

On the other hand, Q3 is a high-visibility and controversial issue that is gaining extra attention in the USA during election periods. There are strong and vocal groups both favoring and condemning abortion, and every candidate for political office has to take a position on the issue. It is expected, therefore that it is one of the issues that will generate a lot of battles for placement in the top-10 search results.

We note that, while we did not choose the queries based on Wikipedia articles, we observed that all three queries have disputed entries or heated discussions in Wikipedia, an evidence of these battles being carried out in the social networks sphere. See [15, 16, 17].

The question we wanted to address is: Do we get good coverage and independent responses when searching on these controversial issues using the two major search engines, Google and Yahoo?

To evaluate our research question we first conducted the three searches using the Google and Yahoo search APIs in August and September, 2007. We analyzed the search results and computed the coverage and independence of each search.

1.3.1 The Supporting Web Graph

It has been shown [10] that the link structure of pages is evolving at a very fast pace, faster than the pace of change in page contents. We wanted to check this observation with regards to the web site back links. Do they also change fast? To do that, we followed the link structure of the supporting web graph twice, in August and September 2007, with our results reported here.

We examined several questions regarding the comparison of the results between the two search engines. The results of our work appear in the subsections below. First, we explain how we computed the supporting web graphs of each search result.

We define as the *supporting web subgraph* G of depth d for some URL U according to search engine S the graph that is computed using S's back links for d iterations.

To compute the supporting web subgraphs of each search result, we created a java program, `backGraph`, that, given a particular URL U as input, it first collects the set of links $L(U)$ of sites pointing to U according to search engine S. Since we cannot search the whole web from scratch to determine these links, we used the Search APIs provided by Yahoo [13] and Google [4]. This collection, $L(U)$, corresponds to back link depth $d = 1$. In graph theoretic terms, $L(U)$ is expected to be a star of nodes (URLs) with edges corresponding to hyperlinks from other documents pointing to U, the center of the star.

We continued computing the sets $L^2(U) = L(L(U))$ for depth $d = 2$ and $L^3(U)$ for depth $d = 3$ using each of the two APIs. We stop at depth $d = 3$ following [8] that shows is to be sufficient for revealing the structure of the supporting web graph. Going further would strengthen our results but would also require significantly larger computational time.

More formally, the algorithm and the parameters we used are as follows:

Supporting Web Graph Algorithm

```
Input:
   s[i] = each of top-10 results' URL
   d = Depth of back link search (d=3)
   B = Number of backlinks to record (B=100)
   SE = {YahooAPI, GoogleAPI}

Algorithm:
  S = {s[i]}
  Using depth-first-search for depth d do:
    Find the set U of sites linking to sites in S
       using the SE for up to B backlinks/site
    S = S + U

Output:
   Graph recorded in S for each API
```

One may expect that $L(L(L(U)))$ would create a tree of sites pointing to U in no more than 3 links. It turns out that the graph created is not a tree but a directed graph G with a bi-connected core (called BCC in [8]). It has been shown that this graph reveals the deliberate link support that activists and spammers are using to promote a particular web page so that this page scores high on a search engine's query results. We then evaluated the size of each graph $|G|$ by calculating the number of nodes (sites) $|V|$ and edges (links) $|E|$ as well as the size of its BCC. In this paper, we do not discuss the BCC data.

1.3.2 Overall Results

Table 1.1 shows the coverage and independence results of the three queries. We observe that in the commercially important Q1 and the medically important Q2, coverage from both search engines is very low. Recall that Q1 and Q2 have low visibility. Results from both search engines show medium to high independence for these queries. On the other hand, for the highly visible and politically important Q3, coverage is medium-to-high and independence is

Table 1.1. Results for coverage and independence metrics. $C(G)$ and $C(Y)$ are the coverage scores in the Google and the Yahoo results, respectively. $I(G)$ and $I(Y)$ are the independence scores for the Google and Yahoo results, respectively.

Query	bal	pro	con	$C(G)$	bal	pro	con	$C(Y)$	$I(G)$	$I(Y)$
Q1: HGH benefits	0	10	0	0.0	0	10	0	0.0	0.8	0.7
Q2: ADHD real disease	0	1	9	0.1	0	1	9	0.1	0.7	0.7
Q3: Morality of Abortion	5	3	1	0.7	3	2	4	0.8	0.9	0.9

high for both search engines, with Yahoo scoring higher in coverage than Google.

In the late 1990's search engine algorithms used to differ significantly. It has been argued that these days, most search engines use very similar search algorithms. Our results provide some evidence for this.

The size of the support web subgraph is one area where the search results differ significantly between the search engines. Yahoo's supporting web graph size is significantly larger than Google's. Given the popularity and reputation of Google, one might have guessed the opposite.

It has been reported that the search engines APIs do not faithfully report their back links. We found evidence that the reported sizes of the supporting web graphs by Google are affected by some filtering of the back links. This is strongly evidenced by the fact that several supporting web graphs for Google have trivial sizes (e.g., result 7 in Q1 and result 9 in Q2), and this can be explained with the existence of result filtering or some kind of PageRank manipulation.

Another question we addressed is, how much did the top-10 results changed for each search engine in the period of the study. To answer this question we evaluated three similarity measurements between the top-10 search results for each query and for each search engine in the two time periods we retrieved our data (August and September, 2007). We did the same for each query and each time period for the search results by each search engine. As similarity measures we used the *percent of overlap*, the *G-measure* and the *M-measure*.

The first measure is the percent of overlap, O, (which we will simply call *overlap*) between two ordered lists. The greater the overlap, the more items two lists have in common. Overlap varies between 0.0 (no common pages) and 1.0 (permutation of the same pages). Note that the overlap ignores any difference in the ordering of the items in the lists, and it just looks for common presence.

The second measure is the *G-measure* G [18] which tries to address the similarity of item location in two lists. It looks at the absolute difference between the locations of an item appearing in the two lists. For the same overlap, the larger the G-measure, the more nearby positions the items occupy. Like overlap, it ranges betwen 0.0 (no item in common) and 1.0 (identically ordered lists).

Even though G gives a good measurement of ranking similarity, it is not meaningful when the overlap between the top-10 results is very small, because

Table 1.2. Intra-SE comparison: Persistence of top-10 search results between August and September, 2007, for each search engine.

Search Eng.	Query	O	G	M	Analysis
Google	Q1: (HGH)	0.7	0.82	0.91	Largely unchanged top & mid entries
	Q2: (ADHD)	0.5	0.53	0.41	Medium overlap in lower entries
	Q3: (Abortion)	0.7	0.65	0.68	Mostly unchanged top & mid entries
Yahoo	Q1: (HGH)	0.6	0.60	0.77	Largely unchanged top & mid entries
	Q2: (ADHD)	0.3	0.33	0.51	Small overlap in top positions
	Q3: (Abortion)	0.5	0.55	0.63	Mostly unchanged top/mid entries

Table 1.3. Inter-SE comparison: Google vs Yahoo result similarity of top-10 search results per observation period.

Query	Time Period	O	G	M	Analysis
Q1: (HGH)	Aug. 2007	0.2	0.20	0.21	Little overlap, but in higher position
	Sept. 2007	0.1	0.16	0.20	no convergence over time
Q2: (ADHD)	Aug. 2007	0.4	0.27	0.10	Medium O. b/w top/low positions,
	Sept. 2007	0.3	0.27	0.18	increased convergence over time
Q3: (Abortion)	Aug. 2007	0.3	0.33	0.50	Small overlap in high positions,
	Sept. 2007	0.3	0.38	0.55	significant convergence over time

the non-overlapping elements have a major effect it. Also, having two items occupying the last two location in the top-10 will give the same G when the two items are occupying the top two locations. In [19] a new measure, the *M-measure* M, was proposed to capture the intuition that identical or near identical rankings among the top documents is more valuable to the user than are rankings among the lower placed documents. For the same overlap, the greater the M, the smaller the distance of common items near the top. Like the other two measurements, it ranges betwen 0.0 (no item in common) and 1.0 (identically ordered lists).

Since the three measures, O, G and M provide complementary information about the similarity of two lists, we used all three in the analysis below. We present our results for overlap, the G-measure and the M-measure in tables 1.2 and 1.3. In the last column we note our observations based on their values.

In terms of intra-SE comparison, we observe the following:

- Google's results were largely persistent over the time period studied; Yahoo showed more change in its results than Google.
- For both search engines, Q2 (on ADHD) was the query that had the greater change, especially in Yahoo's results.

In terms of inter-SE comparison:

- The results showed low to very-low agreement for queries Q1 and Q2, but significant agreement in Q3 (on abortion).

Next, we discuss results for each query in some detail.

Table 1.4. Top-10 results of the Google search engine when given the query Q1 = *HGH benefits* on August, 2007. See also Table 1.5.

#	Google top-10 results on August, 2007
1	http://www.i-care.net/hgh-benefits.html
2	http://www.alwaysyoung.com/hgh/benefits/benefits.html
3	http://www.ghsales.com/ghsales2/hgh_growth_hormone _benefits.html
4	http://www.humangrowthhormonesales.com/ghsales2/index.html
5	http://www.hgh-human-growth-hormone.org/
6	http://www.hgh-human-growth-hormone.org/benefits-of-hgh.htm
7	http://www.csmngt.com/human_growth_hormone.htm
8	http://www.associatedcontent.com/article/38893/human_growth_hormone_hgh_benefits_risks.html
9	http://www.hgharticles.com/
10	http://www.godswaynutrition.com/products/growthhormone.html

Table 1.5. Supporting web graph sizes, as reported by backGraph, for each site in the top-10 results of the Google API for query Q1 (see Table 1.4). The two columns correspond to backGraph sizes measured in August, 2007 and September, 2007. For each entry we have calculated the size of the backGraph as $(|V|, |E|)$ revealed by the Google API and the change between these two dates. In 4 cases we see significant difference (above 20%) in their sizes over time.

#	August, 2007	September, 2007	Change (%)	Notes
1	(415,426)	(346,353)	-16.6	G1=Y8 (r)
2	(161,167)	(133,137)	-17.4	G2=Y1
3	(1245,1314)	(981,1027)	-21.2	G3=G4 (r)
4	(1245,1314)	(981,1027)	-21.2	G4=G3 (r)
5	(71,77)	(71,77)	0.0	G5=G6 (u)
6	(71,77)	(71,77)	0.0	G6=G5 (u)
7	(1,0)	(1,0)	0.0	filtered
8	(1749,2086)	(2905,3386)	+66.1	
9	(1388, 1585)	(463, 506)	-66.7	G9=Y2
10	(502 ,506)	(419 ,423)	-16.6	

Specific observations for Q1 Results

The size of support web subgraphs reported by Yahoo is 7 times greater than that reported by Google (total of 49886 nodes vs 6848 nodes). In the period of the two months we monitored the sizes of the supporting web subgraphs, we saw small-to-medium percentage variation in the size of the Google graphs and wide range for the Yahoo graphs. Despite that, they seem to almost agree on the top result (G2 = Y1) as well as in other results (e.g., G1=Y8, G9=Y2).

All of the top-10 results for both Google and Yahoo seem to be coming from companies that sell steroids online (the "pro" case). There are no results from medical authorities or research articles that refer to the medical or legal problems from the use of steroids (the "con" case). There are also no "balanced" views represented in the top-10 results. Coverage, therefore, is very low. We believe that these results reveal the work of very successful commercial SEOs who have dedicated lots of resources to gain from the lucrative HGH industry.

Table 1.6. Top-10 results of the Yahoo search engine when given the query Q1 = *HGH benefits* on August, 2007. See also Table 1.7.

#	Yahoo results
1	http://www.alwaysyoung.com/hgh/benefits/benefits.html
2	http://www.hgharticles.com/hgh_benefits.html
3	http://www.hgh-pro.com/pro-blenhgh.html
4	http://www.hghhomeopathic.com/HGH.html
5	http://www.hgharticles.com/
6	http://www.hghnstuff.com/faq-benefits-hgh.htm
7	http://linkspiders.com/HGH/benefits%20of%20hgh.htm
8	http://eyecare.freeyellow.com/hgh-benefits.html
9	http://www.hgh-pro.com/homeopathichgh.html
10	http://www.hgh.com/Descriptions/sec.aspx

Table 1.7. Supporting web graph sizes, as reported by backGraph, for each site in the top-10 results of the Yahoo API for query Q1 (see Table 1.6). The two columns correspond to backGraph sizes measured in August, 2007 and September, 2007. For each entry we have calculated the size of the backGraph as $(|V|, |E|)$ revealed by the Yahoo API and the change between these two dates. In 7 cases we see significant difference (above 20%) in their sizes over time.

#	August, 2007	September, 2007	Change (%)	Notes
1	(13151,16690)	(7829,9294)	-40.5	Y1=G2
2	(2933,3871)	(3380,4634)	+15.2	Y2=G9
3	(9587,11741)	(6376,7727)	-33.5	Y3=Y9 (u)
4	(2137,2402)	(3125,3551)	+46.2	
5	(2933,3871)	(3380, 4634)	+15.2	Y5 = Y2 (u)
6	(3063, 3444)	(5575, 6877)	+82.0	
7	(3194, 3665)	(491, 495)	-84.6	
8	(1041, 1204)	(1990, 2295)	+91.2	Y8=G1 (r)
9	(6376,7727)	(6376,7727)	0.0	Y9=Y3 (u)
10	(5471, 6460)	(7418, 9114)	+35.6	

Specific observations for Q2 Results

As we mentioned, Q2 coverage results are largely similar to Q1 results for both search engines. The size of support web subgraphs reported by Yahoo is, again, far greater than that reported by Google (total of 88737 nodes vs 19901 nodes, or 4 times greater).

These results reveal low coverage, with only one of the top-10 entries differing from the overall "con" direction of the results. There are no "balanced" results included. Interestingly, for both search engines, the "pro" result occupies position 5! The results for Yahoo reveal low link independence as four of the top-10 results are forming two circular link farms (Figure 1.1).

We believe that when a controversial issue is below the horizon of current news awareness, such as the ADHD issue at the time of the search, those that care about an issue can be successful in getting the top spots in the relevant queries. The situation seems to be a bit different for issues that have enormous visibility and equally determined groups of supporters, such as the next query.

Table 1.8. Top-10 results of the Google search engine when given the query Q2 = *Is ADHD a real disease*, on August, 2007. See also Table 1.9.

#	Google results
1	http://www.spiritofmaat.com/archive/oct1/drfred.htm
2	http://www.clickpress.com/releases/Detailed/2728005cp.shtml
3	http://www.wildestcolts.com/mentalhealth/stimulants.html
4	http://www.wildestcolts.com/safeEducation/real.html
5	http://web4health.info/en/answers/adhd-real-disorder.htm
6	http://www.adhdfraud.org/
7	http://www.adhdfraud.org/commentary/5-27-01-1.htm
8	http://www.mykidsdeservebetter.com/adhd/disease.asp
9	http://www.virtualvienna.net/community/modules.php?name=News &file=article&sid=295
10	http://www.escolar.com/Escolar-Parenting_Articles/Escolar-is-adhd-a-real-disease.php

Table 1.9. Supporting web graph sizes, as reported by backGraph, for each site in the top-10 results of the Google API for query Q2 (see Table 1.8). The two columns correspond to backGraph sizes measured in August, 2007 and September, 2007. For each entry we have calculated the size of the backGraph as $(|V|, |E|)$ revealed by the Google API and the change between these two dates. In 7 cases we see significant difference (above 20%) in their sizes over time.

#	August, 2007	September, 2007	Change (%)	Notes
1	(1928,2135)	(1423,1614)	-25.2	G1=Y6
2	(1223,1297)	(912,991)	-25.4	G2=Y9
3	(496,515)	(873,925)	+76.0	G3=G4 (u)
4	(496,515)	(873,925)	+76.0	G4=G3 (u)
5	(2545,2759)	(1688,1790)	-33.7	pro
6	(3280,3912)	(3955,4791)	+20.58	G6=Y3
7	(3280,3912)	(3955,4791)	+20.58	G7 = G6 (u)
8	(1590,1708)	(1557,1848)	-2.1	G8 = Y2
9	(1,0)	(1,0)	0.0	filtered (c)
10	(5062,5764)	(incomplete)	N/C	(c)

Table 1.10. Top-10 results of the Yahoo search engine when given the query Q2 = *Is ADHD a real disease*, on August, 2007. See also Table 1.11.

#	Yahoo results
1	http://www.mental-health-matters.com/articles/article.php?artID=849
2	http://www.mykidsdeservebetter.com/adhd/disease.asp
3	http://www.adhdfraud.org/
4	http://www.healthstatus.com/articles/Is_ADHD_A_Real_Disease.html
5	http://www.healthtalk.com/adhd/diseasebasics.cfm
6	http://www.spiritofmaat.com/archive/oct1/drfred.htm
7	http://www.nexusmagazine.com/articles/ADHDisbogus.html
8	http://www.advancingwomen.com/diabetes/is_adhd_a_real_disease.php
9	http://www.clickpress.com/releases/Detailed/2728005cp.shtml
10	http://ritalindeath.com/Against-ADHD-Diagnosis.htm

Specific observations for Q3 Results

Yahoo's sizes of support web subgraphs is roughly equal to Google's, (total of 125376 nodes vs 113463 nodes for the first nine results) in big contrast with the results in Q1 and Q2. Interestingly, both engines also agree on the top spot (G1=Y1), which represents a "balanced" opinion of the query. They

Table 1.11. Supporting web graph sizes, as reported by backGraph, for each site in the top-10 results of the Yahoo API for query Q2 (see Table 1.10). The two columns correspond to backGraph sizes measured in August, 2007 and September, 2007. For each entry we have calculated the size of the backGraph as $(|V|, |E|)$ revealed by the Yahoo API and the change between these two dates. In 4 cases we see significant difference (above 20%) in their sizes over time.

#	August, 2007	September, 2007	Change (%)	Notes
1	(10263,13422)	(12321,16184)	-16.7	
2	(7610,9556)	(5208,6202)	+46.12	Y2=G8
3	(7567,9093)	(9791,12232)	-22.7	Y3=G6 (l)
4	(13631,17032)	(12590,15645)	+8.3	
5	(7011,8488)	(6704,7993)	+4.6	pro
6	(6425,8651)	(4242,5148)	+51.5	Y6=G1 (l)
7	(8424,10808)	(8266,10184)	+2.0	(l)
8	(19291,25469)	(21150,27145)	-8.8	(c)
9	(5675,6644)	(4920,5775)	+15.37	Y9=G2
10	(2840,3440)	(2359,2604)	+20.4	(l)

Table 1.12. Top-10 results of the Google search engine when given the query Q3 = *Morality of abortion*, on August, 2007. See also Table 1.13.

#	Google results
1	http://www.efn.org/ bsharvy/abortion.html
2	http://atheism.about.com/od/abortioncontraception/p/Religions.htm
3	http://atheism.about.com/od/abortioncontraception/p/AtheistsAbort.htm
4	http://ethics.sandiego.edu/Applied/Abortion/index.asp
5	http://rwor.org/a/038/morality-right-to-abortion.htm
6	http://www.answers.com/topic/abortion-debate
7	http://ocw.mit.edu/NR/rdonlyres/054E18A6-DC9A-460E-826E-9EEC31A573E1/0/abortion.pdf
8	http://www.nrlc.org/news/2002/NRL06/pres.html
9	http://www.manitowoc.uwc.edu/staff/awhite/mark_b97.htm
10	http://www.keele.ac.uk/depts/la/ documents/rfletcherFlagsubamd.pdf

both include "pro-choice" and "pro-life" results, while devoting the remaining top-10 entries on opinion gatherers (such as about.com and wikipedia.org).

Over the two month period we observe much smaller variation of the sizes of the supporting web graph, especially for Yahoo. It is not the case that this was due to lack of interest, because the abortion issue has always been at the top of political agendas in the USA. We conjecture that for such highly sensitive issues, the search engines "tune" their results so that they will present wider coverage of opinions. We have seen this happening in the past in a variety of well publicized queries (such as "miserable failure" [21]).

1.3.3 Supporting Web Graph Overlaps

Another question we studied is how much overlap exists in the backGraphs produced by the two search engines on the same result. Given that Yahoo reports far more links than Google, one might expect that there is a major overlap between their backlink graphs. We have found this not to be the case. See Figure 1.2 for a graphical representation of a typical example. The two search engines seem to report a largely different set of links in their results:

Table 1.13. Supporting web graph sizes, as reported by backGraph, for each site in the top-10 results of the Google API for query Q3 (see Table 1.12). The two columns correspond to backGraph sizes measured in August, 2007 and September, 2007. For each entry we have calculated the size of the backGraph as $(|V|, |E|)$ revealed by the Google API and the change between these two dates. In 3 cases we see significant difference (above 20%) in their sizes over time.

#	August, 2007	September, 2007	Change (%)	Notes
1	(5942,6710)	(5597,6350)	-5.9	G1=Y1
2	(3922,4610)	(3429,4138)	-12.6	G2=G3 (u)
3	(3922,4610)	(3429,4138)	-12.6	G3=G2 (u)
4	(12045,14659)	(10433,13102)	-13.38	G4=Y8
5	(1667,2095)	(2687,3213)	+61.19	
6	(7440,8657)	(5187,6276)	-30.28	G2=Y6
7	(3006,3373)	(2902,3211)	-3.5	
8	(6422,8119)	(5810,7295)	-9.5	
9	(2079,2353)	(1331,1560)	-36.0	
10	(incomplete)	(incomplete)	N/C	

Table 1.14. Top-10 results of the Yahoo search engine when given the query Q3 = *Morality of abortion*, on August, 2007. See also Table 1.15.

#	Yahoo results
1	http://www.efn.org/ bsharvy/abortion.html
2	http://jbe.gold.ac.uk/5/barnh981.htm
3	http://www.rit.org/editorials/abortion/moralwar.html
4	http://en.wikipedia.org/wiki/Morality_and_legality_of_abortion
5	http://www.abort73.com/HTML/I-H-2-morality.html
6	http://atheism.about.com/od/abortioncontraception/p/Religions.htm
7	http://www.ashby2004.com/abortion.html
8	http://ethics.sandiego.edu/Applied/Abortion/
9	http://www.rit.org/editorials/abortion/morality.html
10	http://gospelway.com/morality/index.php

Between G1's 5942 sites and Y1's 17224 sites, there is a mere overlap of 470 sites!

1.4 Conclusion and Open Problems

With this paper, which is the extended and augmented version of [20], we have started an effort to evaluate quality of web search results in terms of two important metrics, coverage and independence. We have found that when searching controversial issues, both of these quality metrics can be low. Given the fact that search engines are unwilling to tune their search results manually, except in a few cases that have become the source of bad publicity [21], low coverage and independence reveal the efforts of dedicated groups to manipulate the search results. Automatizing this computation would be a challenging but rewarding step towards quality analysis of search results.

We also found that Yahoo's API reports a much greater number of back links than Google's API. We do not believe that this means that Google

Table 1.15. Supporting web graph sizes, as reported by backGraph, for each site in the top-10 results of the Yahoo API for query Q3 (see Table 1.14). The two columns correspond to backGraph sizes measured in August, 2007 and September, 2007. For each entry we have calculated the size of the backGraph as $(|V|, |E|)$ revealed by the Yahoo API and the change between these two dates. In no cases we see significant difference (above 20%) in their sizes over time.

#	August, 2007	September, 2007	Change (%)	Notes
1	(17224,21732)	(15367,19843)	-10.8	Y1=G1
2	(17491,22553)	(16137,21632)	-7.74	
3	(13434,17522)	(13752,17656)	+2.37	Y3=Y9 (u)
4	(25672,35167)	(25672,35167)	0.0	
5	(10282,14156)	(12243,16284)	+19.1	
6	(6274, 8169)	(6274, 8169)	0.0	Y6=G2
7	(7370,9919)	(7370,9919)	0.0	
8	(13877,18248)	(13877,18248)	0.0	
9	(13752,17656)	(13752,17656)	0.0	Y9=Y3 (u)
10	(2195,2721)	(incomplete)	N/C	

indexes a smaller portion of the web, but that, as others have suggested [22], Google limits the size of its reported backlinks.

On the other hand, between search engines there is a non-trivial degree of overlapping results in the top-10 results for all the queries, selected (reportedly) among millions of qualifying results. This suggests the employment of similar algorithms and heuristics by the two search engines. More research is needed to measure the effect of this conclusion.

Queries on controversial issues will continue to play an important role in web search. Search engines need tools to help them provide high quality results that include high coverage and high independence between results. Evolution of search results of controversial and highly contested queries over time is also needed, as it may reveal information about the part of the web that is being manipulated by spammers, activists and SEOs. Future research will hopefully shed some light on the extend of this manipulation.

Acknowledgments.

Part of this work was supported by a Brachman-Hoffman grant. We would like to thank Sociology Professor Rosanna Hertz for providing expert evaluation on the coverage of query Q3's results.

References

1. Amento, B., Terveen, L., and Hill, W. (2000). Does authority mean quality? Predicting expert quality ratings of web documents. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.
2. Berenson, A. (2000). On hair-trigger wall street, a stock plunges on fake news. New York Times.

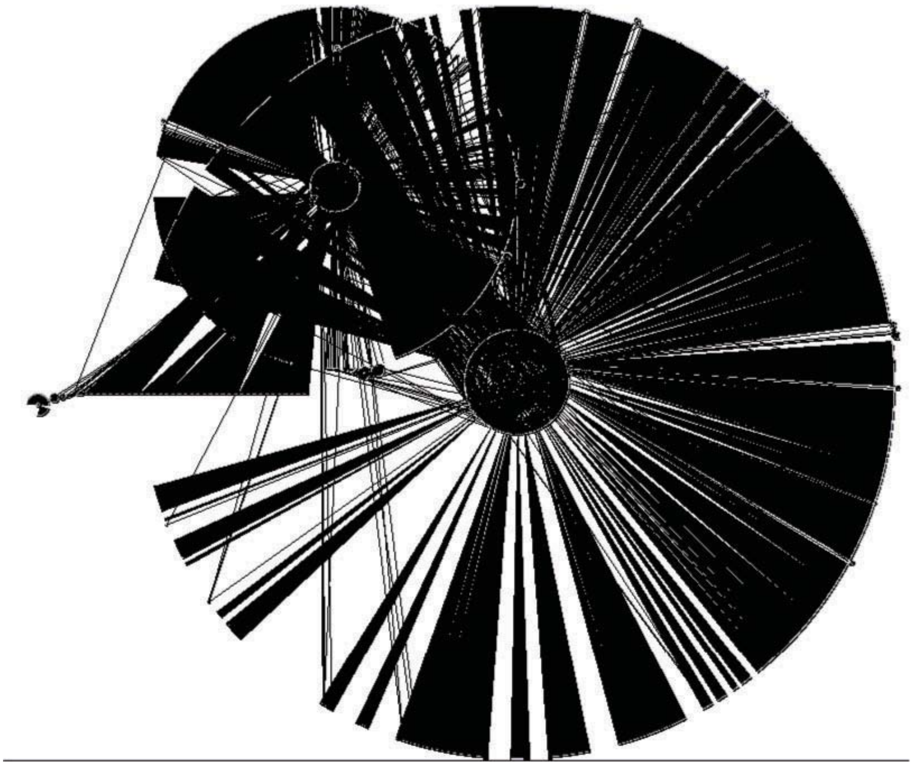

Fig. 1.2. The graph induced by the backlinks of Google's 1^{st} result and Yahoo's 1^{st} result for Q3 (G1=Y1). The graph has been drawn to emphasize the two BCCs. The upper left group is composed by the majority of Google's 5942 sites, the center right large group is composed by most of Yahoo's 17224 sites, while the group in the middle left side (appears as a horizontal line) is mostly composed of 470 sites in the intersection of the two groups.

3. Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.
4. Google (2003). The Google API, google, inc. `http://code.google.com/apis/`
5. Graham, L. and Metaxas, P. T. (2003). "Of course it's true; i saw it on the internet!": Critical thinking in the internet era. *Commun. ACM*, 46(5):70–75.
6. Gyuongyi, Z. and Garcia-Molina, H. (2005). Web spam taxonomy. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan.
7. Manning, C., Raghavan, P., and Schultze, H. (2008). *Introduction to Information Retrieval.* Cambridge Press, Cambridge, UK, (forthcoming) edition.
8. Metaxas, P. T. and Destefano, J. (2005). Web spam, propaganda and trust. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan.
9. Moran, M. and Hunt, B. (2006). *Search Engine Marketing.* IBM Press, New Jersey, USA.

10. Ntoulas, A., Cho, J., and Olston, C. (2004). What's new on the web? the evolution of the web from a search engine perspective. In *Proceedings of the WWW 2004 Conference*, New York, NY.
11. Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12.
12. Vedder, A. (2001). *Misinformation through the internet: Epistemology and ethics*. Intersentia, Antwerpen, Gronigen, Oxford.
13. Yahoo (2006). The Yahoo search API, Yahoo, inc. `http://developer.yahoo.com/search/`
14. Wellness letter, UC Berkeley, June, 2003. Retrieved October 10, 2008. `http://www.berkeleywellness.com/html/ds/dsGrowthHormone.php`
15. Wikipedia entry on Growth hormone. Retrieved October 10, 2008. `http://en.wikipedia.org/wiki/Hgh`
16. Wikipedia entry on ADHD. Retrieved October 10, 2008. `http://en.wikipedia.org/wiki/Hyperkinetic_conduct_disorder`
17. Wikipedia entry on Morality and Legality of Abortion. Retrieved October 10, 2008. `http://en.wikipedia.org/wiki/Morality_legality_of_abortion`
18. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM J. on Discrete Math. 17(1), 134-160, (2003).
19. Bar-Ilan, J., Mat-Hassan, M., Levene, M.: Methods for comparing rankings of search engine results. Computer Networks, 50(10):14481463, (2006).
20. Metaxas, P.T., and Ivanova, L.: Coverage and Independence - Defining Quality in Web Search Results. In: Proceedings of the International Conference on Web Information Systems and Technologies (WEBIST). Madeira, Portugal (2007)
21. Online article entitled "Google Kills Bushs Miserable Failure Search & Other Google Bombs". Retrieved October 10, 2008. `http://searchengineland.com/google-kills-bushs-miserable-failure-search-other-google-bombs-10363.php`
22. McCown, F., and Nelson, M.L.: Agreeing to Disagreeing: Search Engines and their Public Interfaces. In the Proc. of ACM JCDL'07, Vancouver, Canada (2007)