Academic search engine spam and Google Scholar's resilience against it

Joeran Beel[1,2] and Bela Gipp[1,2]

[1] University of California, Berkeley, USA
School of Information
[2] Otto-von-Guericke University, Magdeburg, Germany
Department of Computer Science, ITI, VLBA-Lab
{beel, gipp}@sciplore.org

This paper was refereed by the Journal of Electronic Publishing's peer reviewers.

Abstract. In a previous paper we provided guidelines for scholars on optimizing research articles for academic search engines such as Google Scholar. Feedback in the academic community to these guidelines was diverse. Some were concerned researchers could use our guidelines to manipulate rankings of scientific articles and promote what we call 'academic search engine spam'. To find out whether these concerns are justified, we conducted several tests on Google Scholar. The results show that academic search engine spam is indeed—and with little effort—possible: We increased rankings of academic articles on Google Scholar by manipulating their citation counts; Google Scholar indexed invisible text we added to some articles, making papers appear for keyword searches the articles were not relevant for; Google Scholar indexed some nonsensical articles we randomly created with the paper generator SciGen; and Google Scholar linked to manipulated versions of research papers that contained a Viagra advertisement. At the end of this paper, we discuss whether academic search engine spam could become a serious threat to Web-based academic search engines.
Keywords: academic search engine spam, search engines, academic search engines, citation spam, spamdexing, Google Scholar,

1 Introduction

Web-based academic search engines such as *CiteSeer(X)*, *Google Scholar*, *Microsoft Academic Search* and *SciPlore* have introduced a new era of search for academic articles. In contrast to classic digital libraries such as *IEEE Xplore*, *ACM Digital Library*, or *PubMed*, Web-based academic search engines index PDF files of academic articles from any publisher that may be found on the Web.
Indexing academic PDFs from the Web not only allows easy and free access to academic articles and publisher-independent search, it also changes the way academics can make their articles available to the academic community.

With classic digital libraries, researchers have no influence on getting their articles indexed. They either have published in a publication indexed by a digital library, and then their article is available in that digital library, or they have not, and then the article is not available in that digital library. In contrast, researchers can influence whether their articles are indexed by Web-based academic search engines: they simply have to put their articles on a website to get them indexed. Researchers should have an interest in having their articles indexed by as many academic search engines and digital libraries as possible, because this increases the articles' visibility in the academic community. In addition, authors should not only be concerned about the fact that their articles are indexed, but also *where* they are ranked in the result list. As with all search results, those that are listed first, the top-ranked articles, are more likely to be read and cited.
Furthermore, citation counts obtained from Google Scholar are sometimes used to evaluate the impact of articles and their authors. Accordingly, scientists want all articles that cite their articles to be included in Google Scholar and they want to ensure that citations are identified correctly. In addition, researchers and institutions using citation data from Google Scholar should know how robust and complete the data is that they use for their analyses.
In recent studies we researched the ranking algorithm of Google Scholar [Beel and Gipp (2009c), Beel and Gipp (2009a), Beel and Gipp (2009b)] and gave advice to researchers on how to optimize their scholarly literature for Google Scholar [Beel et al. (2010)]. We called this method 'Academic Search Engine Optimization' (ASEO) and defined it as

> "[…] the creation, publication, and modification of scholarly literature in a way that makes it easier for academic search engines to both crawl it and index it." [Beel et al. (2010)]

The idea of academic search engine optimization is controversial in the academic community. Some researchers agree that scholars should be concerned about it, and respond positively in various blogs and discussion groups:

> "In my opinion, being interested in how (academic) search engines function and how scientific papers are indexed and, of course, responding to these… well… circumstances of the scientific citing business is just natural." [Groß (2010)]

> "ASEO sounds good to me. I think it's a good idea." [Ian (2010)]

> "Search engine optimization (SEO) has a golden age in this internet era, but to use it in academic research, it sounds quite strange for me. After reading this publication […] my opinion changed." [Meskó (2010)]

> "This definitely needs publishing." [Reviewer (2010)]

Others argue against ASEO. Some of the critical feedback included statements like:

> "I'm not a big fan of this area of research […]. I know it's in the call for papers, but I think that's a mistake." [Reviewer4 (2009)]

> "[This] paper seems to encourage scientific paper authors to learn Google scholar's ranking method and write papers accordingly to boost ranking [which is not] acceptable to scientific communities which are supposed to advocate true technical quality/impact instead of ranking." [Reviewer2 (2009)]

> "[…] on first impressions [Academic Search Engine Optimization] sounds like the stupidest idea I've ever heard." [Gunn (2010)]

In our last paper [Beel et al. (2010)] we concluded:

> "Academic Search Engine Optimization (ASEO) should not be seen as a guide how to cheat with search engines. It is about helping academic search engines to understand the content of research papers, and thus how to make this content more available."

However, the concern that scientists might be tempted to 'over-optimize' their articles is at least worthy of investigation. Therefore, we researched whether academic search engine spam can be performed, how it might be done, and how effective it is. For us, academic search engine spam (ASES) is the creation, modification, or publication of academic articles as PDF files and resources related to the articles, specially constructed to increase the articles' or authors' reputations or ranking in academic search engines. Or, in short, the abuse of academic search engine optimization techniques.
Initial results were published in a poster [Beel and Gipp (2010)]. The final results of our research are presented in this paper.

2 Research objective

The main objective of this study was to analyze the resilience of Google Scholar against spam and to find out whether the following is possible:

> Performing citation spam to increase rankings, reputation, and visibility of authors and their articles.
> Performing content spam to make papers appear in more search results, increasing their rankings and increasing authors' publication lists.

Placing advertisement in PDFs.

In addition, we present our first ideas on how to detect and prevent academic search engine spam. The results will help to answer the following questions in further studies:

How reliable are Google Scholar's citation counts, and should they be used to evaluate researcher and article impact?
To what extent can the ranking of Google Scholar be trusted?
To what extent can the linked content on Google Scholar be trusted?

3 Related work

To our knowledge, no studies are available on the existence of spam in academic search engines or on how academic search engine spam could be recognized and prevented. However, indexing and ranking methods of Web-based academic search engines such as Google Scholar are similar to those of classic Web search engines such as Google Web Search. Therefore, a look at related work in the field of classic Web spam may help in understanding academic search engine spam.
Most Web search engines rank Web pages based on two factors, namely the Web page content and the amount (and quality) of links that point to the Web page. Accordingly, Web spammers try to manipulate one or both of these factors to improve the ranking of their websites for a specific set of keywords. This practice is commonly known as 'link spam' and 'content spam'.
Link spammers have various options for creating fraudulent links. They can create dummy Web sites that link to the website they want to push (link farms), exchange links with other webmasters, buy links on third party Web pages, and post links to their websites in blogs or other resources. Many researchers detected link spam [Gyöngyi and Garcia-Molina (2005), Benczur et al. (2005), Drost and Scheffer (2005), Fetterly et al. (2004), Benczúr et al. (2006), Saito et al. (2007), Wu and Chellapilla (2007), Gan and Suel (2007)].
Content spammers try to make their websites appear more relevant for certain keyword searches than they actually are. This can be accomplished by taking content of other websites and combining different (stolen) texts as 'new content', or by stuffing many keywords

in a Web page's title, meta tags[1], ALT-tags of images, and body text, or creating doorway pages, and placing invisible text on a Web page. 'Invisible text' usually means text in the same color as the background or in layers behind the visible text. Again, much research has been performed to identify content spam [Urvoy et al. (2006), Nathenson (1998), Geng et al. (2008), Castillo et al. (2007)].
A third type of Web spam is duplicate spam. Here, spammers try to get duplicates of their websites indexed (and highly ranked). Figure 1 shows an example in which the three first results for a search query point eventually to the same document. The chance that a Web surfer would read the document is higher than if only one of the top results had pointed to this paper[2]. Google provides guidelines for webmasters on how to avoid unintentional duplicate content spam[3]. Similar guidelines do not exist for Google Scholar.

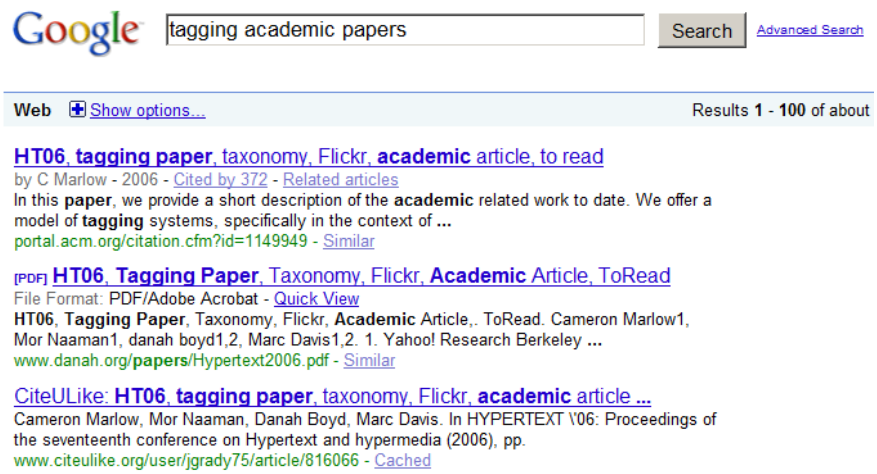


Figure 1: Example of duplicates on Google's result list (search query: 'tagging academic papers')

Although Web spammers are continuously adjusting their methods and developing new techniques (e.g. scraper sites, page hijacking, social

[1] Meta tags are rarely used by spammers since most search engines ignore meta tags due to spam issues
[2] We do not claim that the author of the example paper did duplicate spam. It is likely that Google was not able to identify the different pages as duplicates. However, this illustrates what duplicate spam might look like.
[3] http://googlewebmastercentral.blogspot.com/2009/02/specify-your-canonical.html

media spam, Wikipedia spam, and gadget spam), overall, search engines are capable of fighting Web spam quite well.
Since academic search engines rank scientific articles in a similar way as Web search engines rank Web pages, academic spam can be divided into the same categories as Web spam: content spam, duplicate spam, and link spam; however, in the case of academic papers 'link spam' is equal to 'citation spam.'

4 Motivation

Researchers could be tempted to do academic search engine spam for several reasons: reputation, visibility, and ill will. We discuss these reasons below.

4.1 Reputation

One reason researchers might perform academic search engine spam may be to increase citation counts of their articles and hence enhance their reputations. Citation counts are commonly used to evaluate the impact and performance of researchers and their articles. In the past, citation counts were amassed by organizations such as *ISI's Web of Science*. Direct manipulation of *Web of Science* would be difficult, as *ISI* checks citations in 10,000 journals from the reference lists in those journals from 1900 to the present (and throws out duplicate references in a single article). Nevertheless, some researchers are said to manipulate their citation counts with citation circles, inappropriate self-citations, etc.
Nowadays, citation counts from Web-based academic search engines are also used for impact evaluations. Software like *Publish or Perish*[4] and *Scholarometer*[5] calculate performance metrics such as impact factor and h-index [Hirsch (2005)], based on Google Scholar's citation counts, to assist in analyzing the impact of researchers and articles. These impact measures may be used to support hiring and grants decisions.
We do not know to what extent these tools are used to evaluate the performance of scientists. But several universities recommend *Publish or Perish* as an alternative to *Web of Science* [Harzing (2010)] and many scholarly papers use citation data from Google Scholar for their

[4] http://harzing.com/pop.htm
[5] http://scholarometer.indiana.edu

analysis [Yang and Meho (2006), Harzing and Van der Wal (2008), Kloda (2007), Bakkalbasi et al. (2006), Noruzi (2005), Meho and Yang (2007), Bar-Ilan (2007), Harzing and van der Wai (2008), Kousha and Thelwall (2008), Jacso (2008), Meho and Yang (2006), Moussa and Touzani (2009)]. Some evaluations even take into consideration download counts or the number of readers [Patterson (2009), Taraborelli (2010)].

We believe that this kind of data will play an important role in impact evaluations in the future. And the more these tools are used, the higher the temptation for researchers to manipulate citation counts.

To increase their reputations and publication lists, researchers might also try to create fake papers and get Google Scholar to index these papers. A 'fake paper' could be any document that was solely created for the purpose of manipulating citation counts, etc.

Researchers could try to modify articles by authors who are known in a field, so that the articles reference the researchers' articles or appear to be co-authored by the nefarious researchers. Then it would look as if an authority cited the manipulating researcher's article or as if the authority co-authored with the manipulating author.

Researchers are not the only ones who are evaluated by citation counts; organizations such as universities or journals are evaluated the same way and might therefore consider performing academic search engine spam to increase their citation counts. One publisher has already been caught putting pressure on authors to cite more articles from its publications to increase the impact factor of the publishers' journals [Havemann (2009)].

4.2 Visibility

Researchers could duplicate one of their own articles with enough slight changes and publish it on the Web to make the article appear new to Google Scholar. If Google Scholar indexed it, the duplicate would appear on Google Scholar as separate search result. Users would be more likely read one of these articles than if only one result pointed to the researcher's work. The downside of this approach would be that real citations would be divided among the various duplicates of the article.

Most academic search engines offer features such as showing articles cited by an article, or showing related articles to a given article. Citation spam could bring more articles from manipulating researchers

onto more of these lists. To do so, an author could modify an already published article by inserting many additional references to papers related to the modified paper. Authors of the cited papers would pay attention to the modified article when they examine who is citing them, and readers of the cited articles would more likely pay attention to the citing article when they are searching for related work.

4.3 Ill-Will

Researchers might try academic search engine spamming just for fun, or to damage others authors' reputations by 'pushing' their article rankings so obviously that the other authors are identified as spammers by academic search engines and their articles are removed from the index. On first glance, this idea might seem absurd. However, a similar practice, called 'competitor click fraud', is common in paid search results. Here, companies generate clicks on a competitor's advertisement to exhaust their budget [Wilbur and Zhu (2009), Soubusta (2008), Podobnik et al. (2006), Hadjinian et al. (2006), Gandhi et al. (2006)].
A similar technique, deoptimization, is applied by so-called 'webcare' teams. These teams try to keep negative remarks and negative publicity about a company from showing up high on search-engine results. As a consequence, only positive websites appear high in the result list.

4.4 Classic spam in academic articles

Classic non-academic spammers could place advertisement in manipulated academic articles to generate revenue or create malicious PDF files to either attack readers' computers or attack the search engines' servers themselves. Just recently, Google and other companies were attacked by hackers with malicious PDF files [Müll (2010)].

5 Methodology

There are three basic approaches to academic search engine spam.

> When creating an article, an author might place invisible text in it. This way, the article later might appear more relevant for certain keyword searches than it actually is.

> A researcher could modify his own or someone else's article and upload it to the Web. Modifications could include the addition of additional references, keywords, or advertisements. A manipulating researcher could create complete fake papers that cite his or her own articles, to increase rankings, reputation, and visibility.

Over the past year, we performed several experiments on Google Scholar. We placed invisible text in an article we published, modified existing articles, and created several fake articles to test the resilience of Google Scholar. The articles were uploaded to various websites so Google Scholar could index them. Articles were uploaded to our private homepage, http://beel.org; our project website, http://sciplore.org; the university website, http://ovgu.de; and to the social network websites http://mendeley.com, http://academia.edu and http://researchgate.net.
This paper should not be seen as a thorough experiment on how exactly Google Scholar may be spammed. It is rather a case study and proof-of-concept in which we perform various tests of how to spam Google Scholar.

6 Results

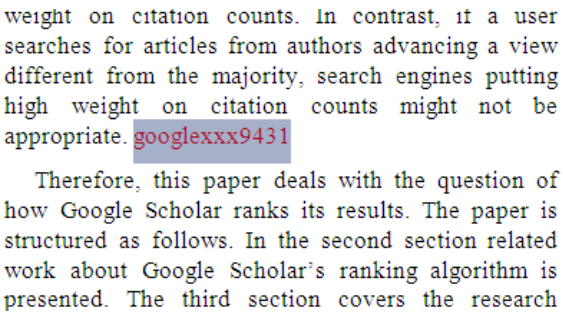
6.1 Websites Google Scholar crawled

Google Scholar did not index our PDF files from mendeley.com and researchgate.com, although other PDFs from those websites are indexed by Google Scholar. PDFs from sciplore.org, beel.org and academia.edu were indexed as well as PDFs from the university's Web space.

6.2 Spamming while writing a real article

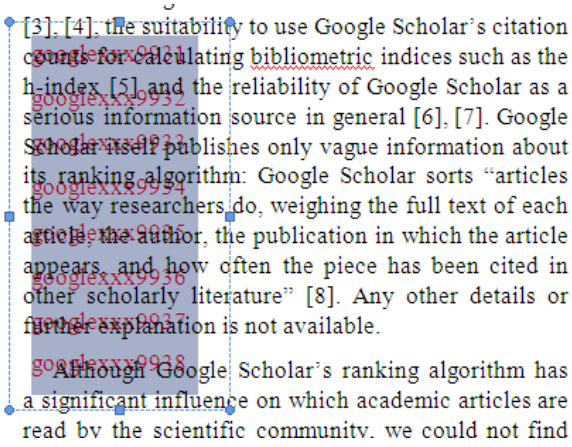
While writing one of our real papers [Beel and Gipp (2009b)], and before it was published, we added words in white color to the first page (see Figure). In addition, we added several words in a layer behind the original text (see Figure). Finally, a vector graphic, a type of picture that can be searched and is machine readable, was inserted. This vector graphic was also placed behind the original text, and contained white text in a tiny font size (see Figure).

The paper then was submitted and accepted for a conference, published by *IEEE*, and included in *IEEE Xplore*. We did not let *IEEE* know what we were doing, and the invisible text was not discovered. About two months after publication the paper was crawled and indexed by Google Scholar, which included the invisible text. That means users of Google Scholar may find our article when they search for keywords that appear only in the invisible text.


weight on citation counts. In contrast, if a user searches for articles from authors advancing a view different from the majority, search engines putting high weight on citation counts might not be appropriate. googlexxx9431

Therefore, this paper deals with the question of how Google Scholar ranks its results. The paper is structured as follows. In the second section related work about Google Scholar's ranking algorithm is presented. The third section covers the research

Figure 2: White text on white background (highlighted for illustration)


[3], [4], the suitability to use Google Scholar's citation counts for calculating bibliometric indices such as the h-index [5] and the reliability of Google Scholar as a serious information source in general [6], [7]. Google Scholar itself publishes only vague information about its ranking algorithm: Google Scholar sorts "articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature" [8]. Any other details or further explanation is not available.

Although Google Scholar's ranking algorithm has a significant influence on which academic articles are read by the scientific community, we could not find

Figure 3: Text in a hidden layer behind the original text (highlighted for illustration)

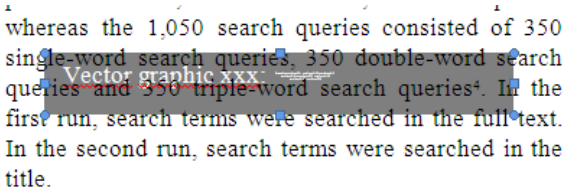

whereas the 1,050 search queries consisted of 350 single-word search queries, 350 double-word search queries and 350 triple-word search queries[4]. In the first run, search terms were searched in the full text. In the second run, search terms were searched in the title.

Figure 4: The tiny white text right of the 'Vector graphic xxx:' is a vector graphic (highlighted for illustration)

6.3 Modifying an already published article

6.3.1 Content modifications

We modified some articles we had already published and added additional keywords (both visible and invisible) throughout the document. Google indexed all modified PDFs and grouped them with the original ones. That means users of Google Scholar may find these modified articles when they search for the additional keywords. In other words, researchers can make their articles appear for keyword searches the original article would not be considered relevant for.
New keywords were also added to the PDF metadata (title and keyword field). However, Google Scholar did not index the additional metadata.

6.3.2 Bibliography modifications

In several existing articles we added new references to the bibliography. Some pointed to articles that were more recent than the original article. These modified articles were uploaded to the Web, and Google Scholar indexed all additional references. As a consequence, citation counts and rankings of the cited articles increased.
That means researchers could easily increase citation counts and rankings of their articles by modifying existing article (and not necessarily their own). This way a researcher could also increase visibility of his articles. He could modify one of his own articles, add references to the bibliography, and the newly cited authors would then probably pay attention to the article.

6.3.3 Adding advertisements

We modified one article [Beel and Gipp (2009b)] and placed *Viagra* advertisement in it, including a clickable link to the corresponding website (see Figure). After a few weeks Google Scholar indexed the PDF file and grouped it with the already indexed files.
That means users of Google Scholar interested in the full text of our research article [Beel and Gipp (2009b)], might download the manipulated PDF containing the Viagra advertisement and we—if we were real spammers—could generate revenue from the researchers visiting the advertised website.

Figure 5: *Viagra* advertisement placed on the first page of an article with a link to a website selling *Viagra*

6.4 Publishing completely new papers

So far, we had modified only existing papers. Google Scholar already knew the articles' metadata—title and author, for instance—when it was indexing the manipulated PDFs.

We also made Google Scholar index papers that were never officially published.

6.4.1 Publishing nonsensical papers

Using the random paper generator *SciGen* [Stribling et al. (2005)], we created six random research papers. These papers consisted of completely nonsensical text and bibliography. Only one real reference was added. We created a homepage for a non-existent researcher and offered the six created papers on this homepage for download. The homepage was uploaded to the Web space OvGU.de, and linked by one of our own homepages, so the Google Scholar crawler could find it. Although Google Web Search indexed the homepage and PDFs after three weeks, Google Scholar did not initially index the PDF files.

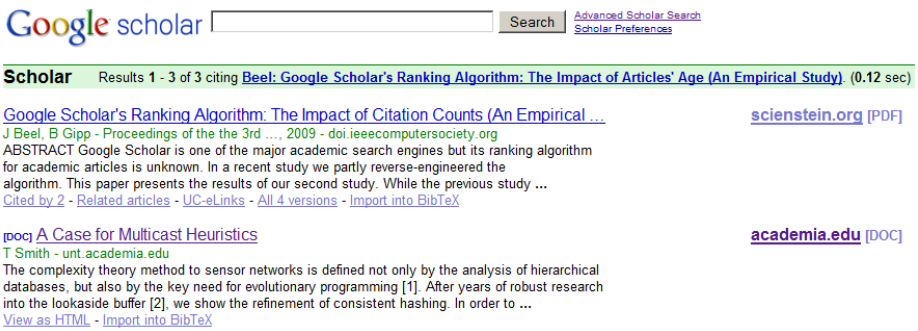


Figure 6: The randomly created article 'A Case for Multicast Heuristic' with nonsensical text and uploaded to Academia.edu is indexed by Google Scholar and increased the citation count and ranking of our 'real' article.

We then uploaded one of the papers to *Academia.edu*. After two months Google Scholar indexed the paper from *Academia.edu* (see Figure 6) and from the university website as well, and ranking of the cited articles increased.

Apparently, Google Scholar has different trust levels for different websites. It indexes unknown articles from the trusted websites, but indexes only known articles from untrusted websites. In this case, academia.edu seems to be considered trustworthy. Each article on that platform is indexed by Google Scholar. It appears that once an article is indexed from *Academia.edu*, other PDFs of that article are indexed, even from websites Google Scholar does not consider trustworthy.

6.4.2 Nonsensical text as real book

Recently created print-on-demand publishers such as *Lulu*, *Createspace*, and *Grin* can publish a book, including ISBN, free, within minutes. We analyzed whether a group of fake articles published as a real book would be indexed by Google Scholar.

We created fourteen new fake articles with *SciGen* [Stribling et al. (2005)]. We replaced the nonsense bibliography of each article with real references. We bundled the fourteen articles in a single document and published this document as a book with the publisher *Grin* [Beel (2009)]. After a few weeks, the book was indexed by Google Books, and some weeks later by Google Scholar. All fourteen articles can be found on Google Scholar and their citations are displayed on Google Scholar too. That means citation counts and rankings of around a hundred articles increased because the fourteen fake papers cited these

articles. Also the (non-existent) authors are now listed in Google Scholar.

6.4.3 Publishing new articles based on real articles (duplicate spam)

In 2009 we published an article about how data retrieved from mind maps could enhance search applications [Beel et al. (2009)]. It was titled 'Information retrieval on mind maps—what could it be good for?' We took this article, changed the title to 'Mind Maps and Information Retrieval' and replaced some references. The body text was not changed. After uploading the article to the Web, Google Scholar indexed it as a completely new article.

That means when users of Google Scholar search for 'mind maps' and 'information retrieval' the result set displays not only the original article, but the modified one as well (see Figure). Accordingly, the probability that users will read the article increases.

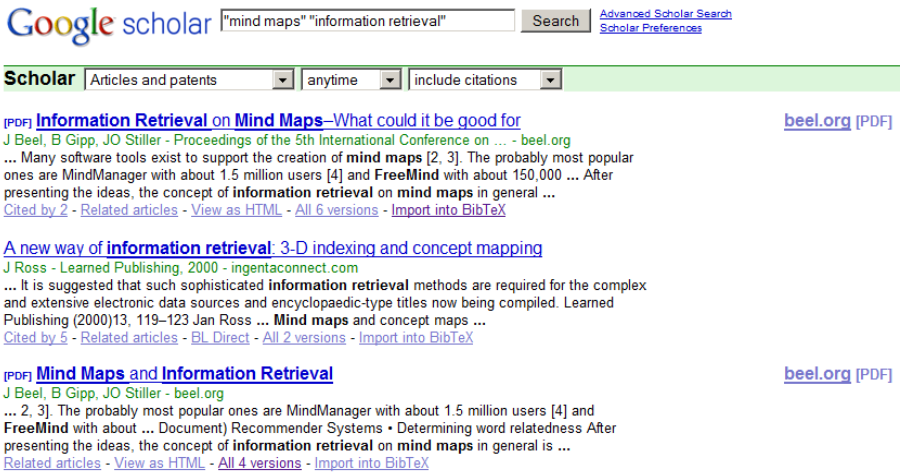
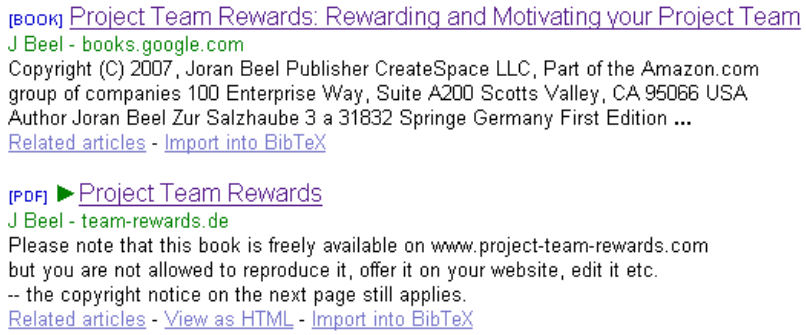


Figure 7: Duplicates with identical content but different title are listed as separate search results

Something similar happened with a book we published about rewarding project teams [Beel (2007)]. Google Scholar indexed the original print version, which is also available on Google Books. When we posted the PDF on the book's website, http://team-rewards.de, Google Scholar indexed it as a new article. Differences between the documents, each about 100 pages, are minimal. However, as Figure shows, Google Scholar has misidentified the title. The correct title is on Google Books: 'Project Team Rewards: Rewarding and Motivating your Project

Team'. The PDF's title was incorrectly identified as 'Project Team Rewards'.


[BOOK] Project Team Rewards: Rewarding and Motivating your Project Team
J Beel - books.google.com
Copyright (C) 2007, Joran Beel Publisher CreateSpace LLC, Part of the Amazon.com group of companies 100 Enterprise Way, Suite A200 Scotts Valley, CA 95066 USA Author Joran Beel Zur Salzhaube 3 a 31832 Springe Germany First Edition ...
Related articles - Import into BibTeX

[PDF] Project Team Rewards
J Beel - team-rewards.de
Please note that this book is freely available on www.project-team-rewards.com but you are not allowed to reproduce it, offer it on your website, edit it etc. -- the copyright notice on the next page still applies.
Related articles - View as HTML - Import into BibTeX

Figure 8: Multiple indexing of the same document

As a consequence of this misidentification, both documents are displayed for searches for the term 'project team rewards' or other similar terms. In addition, the cited articles all received two citations because the original book and the PDF from the website were indexed separately.
Based on these results, it seems that Google Scholar is using only a document's title to distinguish documents. If titles differ, documents are considered different.

6.5 Miscellaneous

In our research we saw some issues that might be relevant in evaluating Google Scholar's ability to handle spam and its reliability for citations counts.

6.5.1 Value of citations

Google Scholar indexes documents other than peer-reviewed articles. For instance, Google Scholar has indexed 4,530 *PowerPoint* presentations[6] and 397,000 *Microsoft Word* documents. It has indexed a Master thesis proposal from one of our students and probably many proposals more. Citations in all these documents are counted[7]. It is apparent that a citation from a *PowerPoint* presentation or thesis proposal has less value than a citation in a peer reviewed academic

[6] The amount of indexed files of a certain type (e.g. ppt) are identifiable via the search query "filetype:ppt"
[7] We took a sample of 10 presentations and Microsoft Word documents that contained citations and all citations in these files were counted.

article. However, Google does not distinguish on its website between these different origins of citations[8].

6.5.2 Wikipedia articles on third party websites

Google Scholar indexes *Wikipedia* articles when the article is available as PDF on a third party website. For instance, the *Wikipedia* article on climate change[9] is also available as a PDF on the website http://unicontrol-inc.com (with a different title). Google Scholar has indexed this PDF (see Figure) and counted its references.

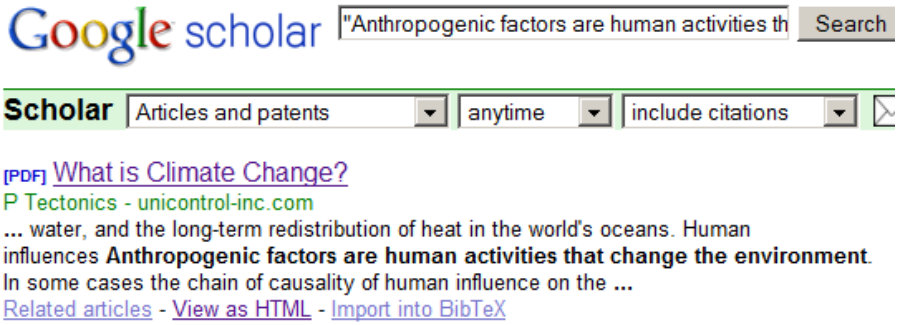


Figure 9: Indexed Wikipedia article from third party website

That means, again, that not all citations on Google Scholar are what we call 'full-value' citations. More importantly, researchers could easily perform academic search engine spam just by citing their papers in *Wikipedia* articles, creating a PDF of the *Wikipedia* article, and uploading the PDF to the Web.

6.5.3 PDF duplicates / PDF hijacking

Google Scholar indexes identical PDF files that have different URLs separately, even if they are on the same server. In case of our article 'Google Scholar's Ranking Algorithm: An Introductory Overview', four PDFs on the domain beel.org (see Figure) were all indexed. Google even considers the same PDF with same URL—once with and once without *www*—as different.

That means a spammer could upload the same PDF several times to the same Web page and all PDFs would be displayed on Google Scholar. Consequently, the probability that a user downloads the manipulated PDF would increase.

---

[8] It could be that Google Scholar weights citations differently when using them for ranking articles. However, third parties parsing Google Scholar cannot identify any distinctions.

[9] http://en.wikipedia.org/wiki/Climate_change

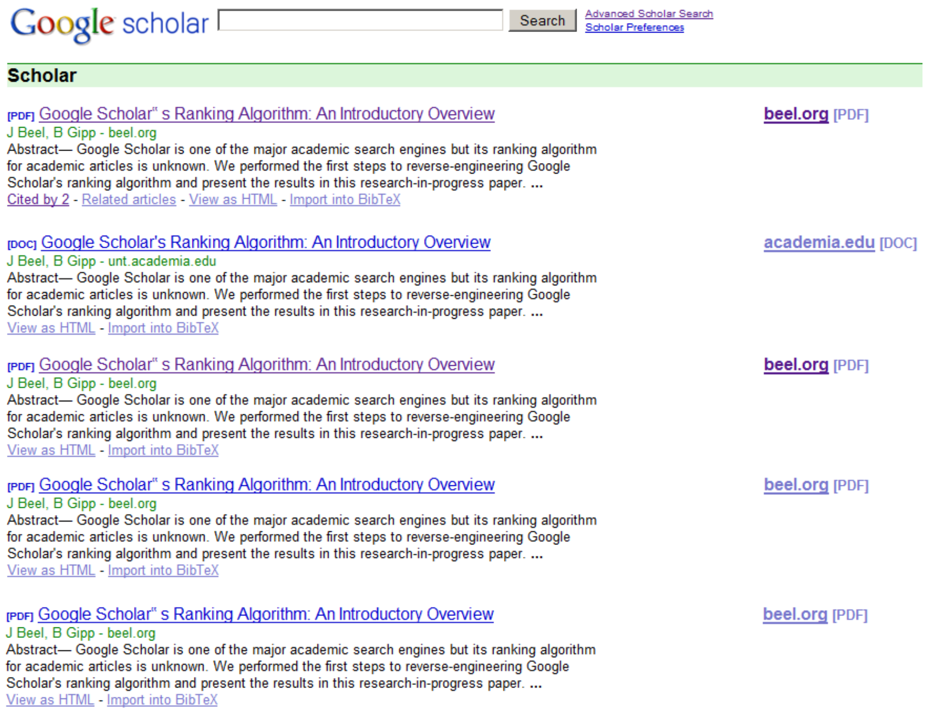


Figure 10: Identical PDFs from the domain beel.org grouped as separate versions

The ranking of grouped PDFs depends mainly on the file date—newer files are listed higher. That means spammers publishing modified versions of an article most likely will see their manipulated PDF as the primary download link for an article. This was also the case in our test with the manipulated PDF containing Viagra advertisement. The manipulated PDF is the most current PDF and displayed as primary download link (see Figure 11).

Figure 11: Ranking of multiple PDF files

A similar practice is known from Web spam. 'Page hijacking' describes the practice that spammers create Web pages (with advertisements, malicious code, etc.) similar to a popular website. Under some circumstances Google identifies the duplicate as the original Web page and displays the duplicates' website as the primary search result.

6.5.4 Misidentification of journal name

By coincidence we realized that it is possible to manipulate the journal name Google Scholar anticipates as the publishing journal of an article. One of our papers [Gipp and Beel (2009)] includes a vector graphic on the second page that illustrates how recommendations are made on our website http://sciplore.org. This vector graphic includes bibliographic information, among others 'Epidemiology, vol. 19, no. 3' (see Figure 12 for a screenshot of that PDF and the vector graphic).

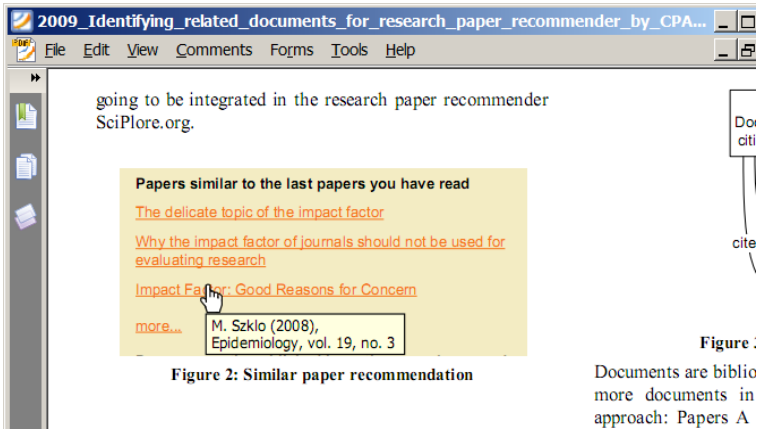


Figure 12: PDF with a vector graphic showing a popular journal name (Epidemiology)

Interestingly, Google Scholar used this bibliographic information as the name of the journal our article was published in (although it was not). A search on Google Scholar for our article shows the article as being published in *Epidemiology*, a reputable journal by the publisher *JSTOR* (see Figure 13).

Figure 13: Misidentification of journal name

Apparently, Google Scholar is using text within an article to identify the article's publishing journal. This could be used by spammers to make their papers appear as if they were being published in reputable journals.

7 Discussion

We discussed academic search engine spam with several colleagues. Some congratulated us on our work; others considered it to be meaningless or even negative for the academic community. Apparently, opinions vary strongly about academic search engine spam. Therefore, we believe that academic search engine optimization and the potential threat of abusing it should be discussed.

We have heard the argument that academic spam might be a less serious threat to academic search engines than Web spam is to Web search engines. First, the effort required for academic search engine spam is high in contrast to the effort required for normal Web spam. Creating spam Web pages, including the registration of new domains, can be done almost automatically within seconds. In contrast, creating modified PDFs or publishing articles with print-on-demand publishers requires significantly more time.

Second, the benefit of spam for researchers is not as immediate and measurable as it is for other Web spammers. While a Web spammer can expect a certain amount of money for each additional visitor, a researcher can hardly specify the benefit of additional citations and readers.

Finally, and most importantly, researchers are not anonymous. In Web search, a website's domain might be banned by the search engine if the site is identified as spam but the spammer could register a new domain within seconds (with a fake identity, if necessary). In contrast, researchers need to think about their reputation. If a researcher doing

academic search engine spam were exposed, the academic search engine would ban all his articles permanently, and his reputation in the academic community would likely be permanently damaged.
However, although the vast majority of researchers are honest, it is widely known that there are some researchers performing unethical and even illegal actions to increase their reputation (see, e.g., [Judson (2004)] for examples). Therefore, it must be assumed that some researchers are willed to do academic search engine spam, despite the risks.
Also journals and conferences might be tempted to do academic search engine spam. Most, if not all, journal and conference rankings consider citation counts as the major or even only factor for calculating the ranking. By citation spam, journals and conferences could dramatically increase their rankings, and therefore, most likely, their revenue.
Journals might also be tempted to perform academic search engine spam to attract more visitors to their websites. The publisher *SAGE* states that 60% of all online readers come via Google and Google Scholar to their journals [SAGE (2010)]. This percentage may increase in a few years. Therefore, academic search engine spam could bring thousands of new visitors and potential revenue. Small and currently unknown journals and conferences might be especially willing to take the risk. If they are discovered, they could found another journal or conference and try again.
Maybe most importantly, 'normal' Web spammers probably will place their spam in modified research articles as soon as they learn that it is possible. Google Scholar provides a new platform to them with hardly any barriers to distributing their spam. There is no reason to assume that normal spammers would not take advantage of this.
Most publishers seem *not* to be aware of the possibility of academic search engine optimization and academic search engine spam. We scrutinized publishing policies of three major publishers in the field of computer science (*IEEE*, *Springer*, and *ACM*) and could not find any rules or policies that address things like including invisible text.
Some publishers are aware of the benefits of academic search engine optimization. The publisher *SAGE*, for instance, suggests the following practice for authors:

> "Search engines look at the abstract page of your article […] Try to repeat the key descriptive phrases [but] don't overplay it, focus on just 3 or 4 key phrases in your abstract." [SAGE (2010)]

> "Ensure the main key phrase for your topic is in your article title." [SAGE (2010)]

This advice does not differ significantly from the guidelines we provided in [Beel et al. (2010)]. However, some journals' recommendations cross what we would consider legitimate ASEO. For instance, the *Journal of Information Assurance and Security* (JIAS) gave the following 'recommendation' to us after a paper we submitted in 2009 was accepted:

> "Please […] improve the introduction/related research section by including all the past related papers published in JIAS."

Also the *International Journal of Web Information Systems* recommended that we "add references from papers previously published in International Journal of Web Information Systems" after one of our papers was accepted in 2010.
To us, the intention of these recommendations seem primarily to be to increase citation counts of the journal and hence to improve metrics such as the impact factor[10].

8 Conclusion

As long as Google Scholar applies only very rudimentary or no mechanisms to detect and prevent spam, citation counts should be used with care to evaluate articles' and researchers' impact. Similarly, researchers should be aware that rankings and linked content might be manipulated. Overall, Google Scholar is a great tool that may help

[10] Due to these recommendations we decided to withdraw the submitted papers.

researchers find relevant articles. However, Google Scholar is a Web-based academic search engine and as with all Web-based search engines, the linked content should not be trusted blindly.
To academic search engines we suggest applying at least the most common spam detecting techniques known from Web search engines. They include analyzing documents for invisible text and either ignoring this text or ignoring the entire document. Also, very small fonts, especially in vector graphics, should not be indexed. With common spam detection methods the PDFs could also be analyzed for 'normal' spam. If an authoritative article directly from the publisher is available, only citations from this article should be counted, and not form other versions of the article found on the Web. It is also questionable whether counting citations from *PowerPoint* slides and *Microsoft Word* documents is sensible.
In addition, documents should be analyzed for 'sense making'. The documents we created with *SciGen* and published with *Grin* and on *Academia.edu* consisted of completely nonsensical text, but still they were indexed. Articles with identical or nearly identical text but different titles should not be listed as separate search results but should be grouped. Also, identical PDFs, especially when they are from the same domain, should not be listed as separate versions.
Finally, we suggest that publishers change their policies: over-optimization of articles should be a violation of their policies and lead to appropriate consequences. However, the academic community needs to decide what actions are appropriate and when academic search engine optimization ends and academic search engine spam begins.

9 Summary

Our study on the resilience of Google Scholar delivers surprising results: Google Scholar is far easier to spam than the classic Google Search for Web pages. While Google Web Search is applying various methods to detect spam and there is lots of research on detecting spam in Web search, Google Scholar applies only very rudimentary mechanisms—if any—to detect spam.
Google Scholar indexed invisible text in all kind of articles. A researcher could put invisible keywords in his article before, or even after, publication and increase the ranking and visibility of this article on Google Scholar.

Google Scholar counted references that were added to modified versions of already published articles. That means authors could add references in their articles after official publication. If these altered articles were published on the Web, Google would index them. This way, researchers could increase citation counts and rankings of the cited articles. They could also bring attention to their articles because the cited authors might investigate who has cited them. Researchers could also modify articles from other authors and add references to their own articles. This way, scholars could create the impression that an authority in their field cited their articles and increase citation counts as well.

Google Scholar also indexed fake articles uploaded to trusted sources such as *Academia.edu* and articles that were published as book with a print-on-demand publisher such as *Grin*[11]. This gives researchers another way to manipulate citation counts and extend their publication lists. An author could create a fake article with his or her name and the name of a popular researcher as co-author. This method could also be used to publish a real article again but with a different title, so the different variations would appear as separate items in the result lists (duplicate spam).

Google Scholar is indexing file formats other than PDFs, such as *PowerPoint* presentations (.ppt) and *Microsoft Word* documents (.doc), and counting references that were made in these files. Although we did not test it, one might assume that it would be easy to create *PowerPoint* presentations and doc files citing a specific article just with the intention of pushing the article's ranking. Google Scholar is also indexing non-peer-reviewed academic documents such as thesis proposals or *Wikipedia* articles offered on third party websites.

It was also easy to perform duplicate spam. With changed titles, basically identical PDFs were identified as separate articles. In addition, Google Scholar seems to rank new PDFs higher than older PDFs. That means manipulated PDFs most likely would appear as the primary download link.

By coincidence we realized that Google Scholar assigned a paper to a journal named in the full text of the article. We did not investigate this

[11] It has to be mentioned that we published this book under our real names. However, it would have been just as easy to publish it with a fake identity (though this would have violated the terms of service of the print on demand publisher).

further, but it might be possible to make an article seem to have been published in a reputable journal although it never was.
Finally, Google Scholar indexed modified versions of articles that contained advertisements. Certainly, researchers would not add advertisement to their own articles. But it is imaginable that normal spammers could download thousands of academic PDFs, automatically place their advertisement in these PDFs, and upload them to the web. Google Scholar would index them, and users of Google Scholar interested in an article's full text might download these modified articles and see the advertisement.
Some might argue that academic search engine spam is a less serious threat to academic search engines than classic Web spam is to Web search engines. However, the potential benefits of academic search engine spam might be too tempting for some researchers. In addition, we see little reason why normal Web spammers should not place their advertisement in academic articles.
To prevent academic search engine spam, Google Scholar (and other Web-based academic search engines) should apply at least the common spam detection techniques known from Web spam detection, analyze text for sense-making, and not count all citations.

Note

We would like to note that the intention of this paper was not to expose Google Scholar. The intention was to stimulate a discussion about academic search engine optimization and the threat of academic search engine spam. We chose Google Scholar as the subject of our study because Google Scholar probably is the best and largest academic search engine indexing PDFs from the Web. Currently, we are developing our own academic search engine, *SciPlore* (http://sciplore.org). As yet, *SciPlore* has no protections against spam either. A very brief investigation of *CiteSeer* and *Microsoft Academic Search* indicates that they do not detect academic search engine spam either.

Acknowledgements

We thank Bert van Heerde from *Insyde* for his valuable feedback.

About the authors

Joeran Beel

Joeran Beel is a visiting scholar at UC Berkeley and PhD student in computer science at OvGU Magdeburg (Germany). He obtained an *MSc in Business Information Systems* at OvGU Magdeburg and graduated with distinction and was cited as the

computer science department's best student. In addition, he obtained an *MSc in Project Management* at Lancaster University Management School (UK).
In recent years he published several papers about academic search engines. He is a co-founder of the academic search engine *SciPlore* (www.sciplore.org) and the machine readable digital library *Mr. dLib* (www.mr-dlib.org). He won several awards for his research, including one from German's Chancellor Gerhard Schröder. Contact details may be found on his homepage (www.beel.org).

Bela Gipp

At the age of 15 Bela Gipp won several prizes at Jugend Forscht, Germany's premier national youth science competition. The German Chancellor honored him after he took first prize in the state-level round for the third time. Scholarships allowed him to study in Australia, England, and China, and at UC Berkeley, in California. After obtaining his master's in the field of computer science and an MBA, he started working for SAP AG in a research joint venture.
Currently he is a visiting researcher at UC Berkeley, where he works on his doctoral research. Central topics are network theory and bibliometric analysis. Besides his theoretical research, he develops open-source software for scientists as a founder of SciPlore.org. Publications can be found on his website www.gipp.com.

References

[Alcala et al. (2004)] Alcala, F., J. Beel, A. Frenkel, B. Gipp, J. Lülf, and H. Höpfner (2004). UbiLoc: A System for Locating Mobile Devices using Mobile Devices. In K. Kyamakya (Ed.), *Proceedings of 1st Workshop on Positioning, Navigation and Communication 2004 (WPNC 04)*, pp. 43–48. University of Hanover. Also available on http://beel.org.

[Bakkalbasi et al. (2006)] Bakkalbasi, N., K. Bauer, J. Glover, and L. Wang (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries 3*.

[Bar-Ilan (2007)] Bar-Ilan, J. (2007). Which h-index? - A comparison of WoS, Scopus and Google Scholar. *Scientometrics 74*(2), 257–271.

[Beel (2007)] Beel, J. (2007, November). *Project Team Rewards - Rewarding and Motivating your Project Team*. Scotts Valley (USA): Createspace. ISBN 978-1434816269. Also available on http://project-team-rewards.com.

[Beel (2009)] Beel, J. (Ed.) (2009). *Computer Science: New Generations*. Munich (Germany): Grin Verlag. ISBN 978-3-640-32875-8.

[Beel and Gipp (2009a)] Beel, J. and B. Gipp (2009a, July). Google Scholar's Ranking Algorithm: An Introductory Overview. In B. Larsen and J. Leta (Eds.), *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, Volume 1,

Rio de Janeiro (Brazil), pp. 230–241. International Society for Scientometrics and Informetrics. ISSN 2175-1935. Also available on http://www.sciplore.org.
[Beel and Gipp (2009b)] Beel, J. and B. Gipp (2009b, April). Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study). In S. Latifi (Ed.), *Proceedings of the 6th International Conference on Information Technology: New Generations (ITNG'09)*, Las Vegas (USA), pp. 160–164. IEEE. ISBN 978-1424437702. Also available on http://www.sciplore.org.
[Beel and Gipp (2009c)] Beel, J. and B. Gipp (2009c, April). Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study). In A. Flory and M. Collard (Eds.), *Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science (RCIS'09)*, Fez (Morocco), pp. 439–446. IEEE. ISBN 978-1-4244-2865-6. Also available on http://www.sciplore.org.
[Beel and Gipp (2010)] Beel, J. and B. Gipp (2010, June). On the Robustness of Google Scholar Against Spam. In *Proceedings of the 21th ACM Conference on Hyptertext and Hypermedia*, Toronto (CA), pp. 297–298. ACM. Also available on http://www.sciplore.org.
[Beel et al. (2009)] Beel, J., B. Gipp, and J. O. Stiller (2009, November). Information Retrieval on Mind Maps – What could it be good for? In *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09)*, Washington (USA), pp. 1–4. IEEE. ISBN 978-963-9799-76-9. Also available on http://www.sciplore.org.
[Beel et al. (2010)] Beel, J., B. Gipp, and E. Wilde (2010, January). Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co. *Journal of Scholarly Publishing 41*(2), 176–190. University of Toronto Press. Also available on http://www.sciplore.org.
[Benczúr et al. (2006)] Benczúr, A., K. Csalogány, and T. Sarlós (2006). Link-based similarity search to fight web spam. *Adversarial Information Retrieval on the Web (AIRWEB), Seattle, Washington, USA*.
[Benczur et al. (2005)] Benczur, A., K. Csalogány, T. Sarlós, and M. Uher (2005). SpamRank – Fully Automatic Link Spam Detection. In *Adversarial Information Retrieval on the Web (AiRWEB'05)*.
[Castillo et al. (2007)] Castillo, C., D. Donato, A. Gionis, V. Murdock, and F. Silvestri (2007). Know your neighbors: Web spam

detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 430. ACM.
[Drost and Scheffer (2005)] Drost, I. and T. Scheffer (2005). Thwarting the nigritude ultramarine: Learning to identify link spam. *Lecture Notes in Computer Science 3720*, 96.
[Fetterly et al. (2004)] Fetterly, D., M. Manasse, and M. Najork (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. pp. 1–6.
[Gan and Suel (2007)] Gan, Q. and T. Suel (2007). Improving web spam classifiers using link structure. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp. 20. ACM.
[Gandhi et al. (2006)] Gandhi, M., M. Jakobsson, and J. Ratkiewicz (2006). Badvertisements: Stealthy click-fraud with unwitting accessories. *Journal of Digital Forensic Practice 1*(2), 131–142.
[Geng et al. (2008)] Geng, G., C. Wang, and Q. Li (2008). Improving Spamdexing Detection Via a Two-Stage Classification Strategy. pp. 356.
[Gipp and Beel (2009)] Gipp, B. and J. Beel (2009, October). Identifying Related Documents For Research Paper Recommender By CPA And COA. In S. I. Ao, C. Douglas, W. S. Grundfest, and J. Burgstone (Eds.), *International Conference on Education and Information Technology (ICEIT'09)*, Volume 1 of *Lecture Notes in Engineering and Computer Science*, Berkeley (USA), pp. 636–639. International Association of Engineers (IAENG): Newswood Limited. ISBN 978-988-17012-6-8. Also available on http://www.sciplore.org.
[Groß (2010)] Groß, M. (2010, February). academic search engine optimization. Blog.
[Gunn (2010)] Gunn, W. (2010, January). Feedback on ASEO. Blog Comment.
[Gyöngyi and Garcia-Molina (2005)] Gyöngyi, Z. and H. Garcia-Molina (2005). Link spam alliances. In *Proceedings of the 31st international conference on Very large data bases*, pp. 528. VLDB Endowment.
[Hadjinian et al. (2006)] Hadjinian, D. et al. (2006). Clicking away the competition: The legal ramifications of click fraud for companies that offer pay per click advertising services. *Shidler JL Com. & Tech. 3*, 5–16.

[Harzing and van der Wai (2008)] Harzing, A. and R. van der Wai (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics(ESEP) 8*(1), 61–73.
[Harzing and Van der Wal (2008)] Harzing, A. and R. Van der Wal (2008). A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *Journal of the American Society for Information Science 60*(1), 41–46.
[Harzing (2010)] Harzing, A. W. (2010). Publish or Perish in the news. Website.
[Havemann (2009)] Havemann, F. (2009). *Einführung in die Bibliometrie*. Humboldt University of Berlin.
[Hirsch (2005)] Hirsch, J. E. (2005, November). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences 102*(46), 16569–16572.
[Ian (2010)] Ian (2010, January). Feedback on ASEO. Blog Comment.
[Jacso (2008)] Jacso, P. (2008). Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for FW Lancaster. *Library Trends 56*(4), 784–815.
[Judson (2004)] Judson, H. (2004). *The great betrayal: fraud in science*. Houghton Mifflin Harcourt (HMH).
[Kloda (2007)] Kloda, L. (2007). Use Google Scholar, Scopus and Web of Science for comprehensive citation tracking. *Evidence Based Library and Information Practice 2*(3), 87.
[Kousha and Thelwall (2008)] Kousha, K. and M. Thelwall (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics 74*(2), 273–294.
[Meho and Yang (2006)] Meho, L. and K. Yang (2006). A new era in citation and bibliometric analyses: Web of Science, Scopus, and Google Scholar. *Arxiv preprint cs/0612132*.
[Meho and Yang (2007)] Meho, L. and K. Yang (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology 58*(13), 2105–25.
[Meskó (2010)] Meskó, B. (2010, January). Academic Search Engine Optimization in Google Scholar. Blog.
[Müll (2010)] Müll, D. (2010, January). Der Cyberkrieg hat begonnen: 34 US-Unternehmen teils stark von Attacken getroffen. News Website.

[Moussa and Touzani (2009)] Moussa, S. and M. Touzani (2009). Ranking marketing journals using the Google Scholar-based hg-index. *Journal of Informetrics*.
[Nathenson (1998)] Nathenson, I. (1998). Internet infoglut and invisible ink: Spamdexing search engines with meta tags. *Harv. J. Law & Tec 12*, 43–683.
[Noruzi (2005)] Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *Libri 55*(4), 170–180.
[Patterson (2009)] Patterson, M. (2009, September). Article Level Metrics. Blog.
[Podobnik et al. (2006)] Podobnik, V., K. Trzec, and G. Jezic (2006). An auction-based semantic service discovery model for e-commerce applications. *Lecture Notes in Computer Science 4277*, 97.
[Reviewer (2010)] Reviewer, J. (2010, March). Your Submission: Academic Search Engine Spam. Private Communication (Email).
[Reviewer2 (2009)] Reviewer2, C. (2009, March). Feedback I on Reverse Engineering Google Scholar's Ranking Algorithm (Conference Reviewer). Private Communication (Email).
[Reviewer4 (2009)] Reviewer4, C. (2009, March). Feedback II on Reverse Engineering Google Scholar's Ranking Algorithm (Conference Reviewer). Private Communication (Email).
[SAGE (2010)] SAGE (2010). Authors - How to help readers find your article online. Website.
[Saito et al. (2007)] Saito, H., M. Toyoda, M. Kitsuregawa, and K. Aihara (2007). A large-scale study of link spam detection by graph algorithms. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp. 48. ACM.
[Soubusta (2008)] Soubusta, S. (2008). On Click Fraud. *Information Wissenschaft und Praxis 59*(2), 136.
[Stribling et al. (2005)] Stribling, J., M. Krohn, and D. Aguayo (2005). Scigen-an automatic cs paper generator.
[Taraborelli (2010)] Taraborelli, D. (2010, September). ReaderMeter: Crowdsourcing research impact. Blog.
[Urvoy et al. (2006)] Urvoy, T., T. Lavergne, and P. Filoche (2006). Tracking web spam with hidden style similarity. In *AIRWeb 2006*, pp. 25.
[Wilbur and Zhu (2009)] Wilbur, K. and Y. Zhu (2009). Click fraud. *Marketing Science 28*(2), 293–308.

[Wu and Chellapilla (2007)] Wu, B. and K. Chellapilla (2007). Extracting link spam using biased random walks from spam seed sets. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp. 44. ACM.

[Yang and Meho (2006)] Yang, K. and L. Meho (2006). Citation analysis: A comparison of google scholar, scopus, and web of science. *43*.