A Reference Collection for Web Spam

Carlos Castillo[1,3], Debora Donato[1,3], Luca Becchetti[1], Paolo Boldi[2], Stefano Leonardi[1], Massimo Santini[2] and Sebastiano Vigna[2]

[1] Università di Roma "La Sapienza" Rome, ITALY

[2] Università degli Studi di Milano Milan, ITALY

[3] Yahoo! Research Barcelona Catalunya, SPAIN

Abstract

We describe the `WEBSPAM-UK2006` collection, a large set of Web pages that have been manually annotated with labels indicating if the hosts are include Web spam aspects or not. This is the first publicly available Web spam collection that includes page contents and links, and that has been labelled by a large and diverse set of judges.

1 Introduction

The term "**spam**" has been commonly used in recent years to refer to *unsolicited (and possibly commercial) bulk messages* (U[C]BE). The most common form of electronic spam is **e-mail spam**, but in practice each communication medium creates a new opportunity for sending unsolicited messages. As the request-response paradigm of the HTTP protocol makes it impossible for spammers to actually "send" pages directly to the users, the type of spam that is done on the Web takes a somewhat different form than in other media. What spammers do on the Web is to try to deceive search engines, undermining the trust relation established between search engines and Web users [Gyöngyi and Garcia-Molina, 2005].

Spamdexing (search engine spamming) is defined in [Gyöngyi and Garcia-Molina, 2005] as "any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance for some Web page, considering the page's true value". A **spam page** is a page that is used for spamming or receives a substantial amount of its score from other spam pages. Another definition of spam, given in [Perkins, 2001] is "any attempt to deceive a search engine's relevancy algorithm" or simply "anything that would not be done if search engines did not exist".

These definitions raise many questions: which actions can be considered *deliberate*?, which actions are *unjustifiable*?, what is a page's *true value*? The fact is that there is a large gray area between "ethical" search engine optimization, that is, making sure that a page can be found by search engines, and "unethical" spamdexing, that is, deceiving search engines.

There are pages on the Web that do not try to deceive search engines at all and provide useful contents to Web users; there are pages on the Web that include many artificial aspects that can only be interpreted as attempts to deceive search engines, while not providing useful information at all; finally, there are pages that do not clearly belong to any of these two categories.

The presence of Web spam negatively affects the quality of current search engines. Often, pages that most users would consider of low quality score very high on search engine rankings.

For instance, the authors of [Eiron et al., 2004] report that : "among the top 20 URLs in our 100 million page PageRank calculation (...) 11 were pornographic, and these high positions appear to have all been achieved using the same form of link manipulation".

Actually, for every ranking algorithm, in particular for those that count replicable features of Web pages, there exist some potential manipulations [Page et al., 1998]. This has created an "arms race": between Web site administrators trying to rank high on search engines and search engine administrators trying to provide relevant, credible results. The arms race has in turn created the field of **Adversarial Information Retrieval**, that studies how to adapt information retrieval techniques for contexts in which part of the collection has been maliciously modified to affect ranking algorithms.

This article presents a reference collection designed for Web spam research. We think this collection might become a valuable tool for researchers studying these problem from different perspectives (e.g.: information retrieval, machine learning, computer security, etc.). In particular, it will help in the understanding of how is Web spam in practice, and in the development of new algorithmic techniques for detecting and demoting Web spam content.

As to this point, automated strategies for spam detection and demotion necessarily assume the presence of characteristic or recurring patterns in spam pages. Agreeing on the set of patterns to consider as indicators of possible or likely spam activity is an essential input to a spam detection tool. Our work provides empirical evidence that consensus on what is spam and what is not is high but not total. As we point out further in this paper, this calls for maybe to some extent arbitrary, but less ambiguous criteria to decide what is spam and what is not; so far, **we know it when we see it**[1].

The rest of this paper is organized as follows: the next section describes the collections used in previous works about Web spam. Section 3 describes the process carried to obtain the collection, and Section 4 describes the results of the labelling process. Section 6 presents our conclusions and describes how to obtain the data.

2 State of the art

The lack of a reference collection is one of the problems that has been affecting the research in the field of spam detection and demotion. This often obliges researchers to build their own data sets to perform experiments, with a twofold drawback. First of all the data sets are often generated to constitute a good representative of the phenomenon researchers are investigating and so, in many cases, are biased toward it. Second and more important, techniques cannot be truly compared unless they are tested on the same collection.

The problem of biased data sets is underlined in a number of previous works on spam detection and demotion. In [Davison, 2000] the author bewares of the fact that one of the collection was arbitrary collected having the task in mind. The same problem indeed characterizes the collections used in [Gyöngyi et al., 2004, Benczúr et al., 2005, Benczúr et al., 2006b] despite of the accuracy of the sampling process. The overall crawled data set was divided into buckets, each containing a

[1] Judge Potter Stewart wrote in a famous verdict about hard-core pornography in 1964: "I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that."

different number of sites/pages with scores summing up to 5 percent of the total Pagerank score and, from each of these buckets, 50 sites were randomly extracted. Obviously, the obtained samples are biased toward pages with high PageRank. Nevertheless, this can be a positive aspect since the goal of the authors was to detect spam sites/pages that are in the first rank positions. A different problem characterizes the sample used in [Ntoulas et al., 2006] as the collection was obtained using a Web crawler that already applies some spam filtering techniques.

Another common problem is that the most of the collections were tagged by one single author with an evident bias due to the subjective judgment, this problem affects [Davison, 2000, Gyöngyi et al., 2004, Becchetti et al., 2006]. In [Benczúr et al., 2005] the collection was labeled by all the authors of the paper, but they observed a poor agreement over the pages evaluated by more than two authors. This provides evidence of the difficulty of the labeling process also for humans judges and justifies the choice, for the collection presented here, of having each hosts evaluated by at least two volunteers.

A list of some of the data sets used in previous works is summarized in Table 1 where, for each collection, we report: the specific top-level domain (TLD) crawled, the crawler used, the number of pages, edges and hosts of the original data sets, the number of labeled links /pages/ hosts, and the reference paper.

Table 1: Existing data sets used in the Web spam literature.

TLD	Crawler	Date	Pages	Edges	Hosts	Labelled	Reference
-	-	-	-	-	-	1,536 links	[Davison, 2000]
-	DiscoWeb	1999	7M	-	-	750 links	[Davison, 2000]
.uk	UbiCrawler	2002	18,5M	-	98,452	5,750 hosts	[Becchetti et al., 2006]
-	AltaVista	2003/Aug	∼Bill.	-	31M	1,000 sites	[Gyöngyi et al., 2004]
.ch	Search.ch	2004	20M	-	300K	728 hosts	[Benczúr et al., 2006a]
.de	Polybot	2004/Apr	31.2M	962M	-	1,000 pages	[Benczúr et al., 2005]
-	MSN	2004/Aug	105M	-	-	17,168pages	[Ntoulas et al., 2006]

We want to underline that despite of the overall effort spent in classifying pages, none of these collections is freely available with the only exception of the data sample used in [Becchetti et al., 2006].

3 Web Spam Data Collection Process

Our main goal was to build a reference Web spam collection for testing Web spam detection/demotion algorithms. Since there was no reference data set for testing antispam techniques, our primary goal was to build one with the broad objectives of being:

Large: the collection should include many examples of spam and non-spam content.

Clean: the collection should contain little classification errors.

Uniform: the collection should represent a uniform random sample over a dataset.

Broad: the collection should include as many different Web spam techniques as possible.

Open: the collection should be freely available for researchers.

The process of assembling this collection consists of the following phases: Web crawling, elaboration of Web spam guidelines and classification interface, labelling, and post-processing, which are described in the rest of this section.

3.1 Crawling of base data

We started in May 2006 by collecting a large set of UK pages. These pages were downloaded at the Laboratory of Web Algorithmics[2] at the Università degli Studi di Milano. The crawl was done using the UbiCrawler [Boldi et al., 2004] in breadth-first-search mode for cross-host links (depth-first exploration was adopted for local links), starting from a large seed of over 190,000 URLs in about 150,000 hosts under the `.uk` domain listed in the Open Directory Project2.[3] The crawler was limited to the `.uk` domain and to 8 levels of depth, with no more than 50,000 pages per host: these restrictions were such that indeed only a part of the hosts contained in the seed were actually crawled. The obtained collection includes 77.9 million pages and over 3×10^9 edges, and includes pages from 11,000 hosts.

The collection was stored in the WARC/0.9 format[4] which is a data format proposed by the Internet Archive, the non-profit organization that has carried the most extensive crawls of the Web. WARC is a data format in which each page occupies a record. A record includes a plain text header with the page URL, length and other meta-information, and a body with the verbatim response from the Web servers, including the HTTP header. The collection is distributed in 8 volumes compressed using `gzip`, containing about 55 GB of compressed data per volume.

The links found in the collection form a Web graph, which is stored in the compressed format described in [Boldi and Vigna, 2004], and uses about 2.9 bits per edge for a total size of about 1.2 GB. This is the version obtained by using the default settings, which provides good time/space tradeoffs; there is also a highly compressed version that uses 2.2 bits per edge.

3.2 Elaboration of Web spam guidelines and classification interface

We reviewed the existing literature about Web spam mentioned in Section 2, as well as the guidelines of the Web search engines operated by Google, Yahoo, and MSN search for dealing with Web spam. We assembled a list of **spamming aspects** and collected several examples of pages using deceptive techniques.

The guidelines we provided to the reviewers consist of a list of Web spam aspects, a set of examples and the guidelines from the three search engines listed above. The main question the judges were asked was: **are there aspects of this page that are mostly to attract and/or redirect traffic?**.

The Web-base interface has the twofold task of randomly assigning hosts to 2 different judges per host and to help in the classification. The interface is very simple and consists of a Web application with a layout of three panels, as shown in Figure 1. The left panel presents the user with a "work unit", a list of 20 hosts chosen at random that s/he has to classify. The center panel presents information about the selected host, including in- and out-links, *WHOIS* information about the domain registrar, as well as a list of sample pages. The right panel visualizes the Web site and

[2] `http://law.dsi.unimi.it/`

[3] `http://www.dmoz.org/`

[4] `http://www.niso.org/international/SC4/N595.pdf`

allows browsing. The interface presents the user with the home page of each host downloaded from a local cache, and the current version downloaded from the live Web for the other pages.

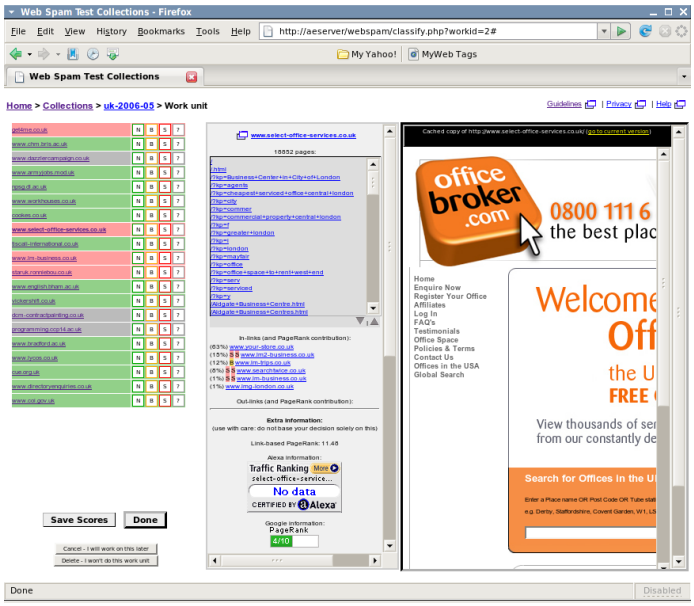

Figure 1: A screenshot of the classification interface.

The interface allows the user to classify each host into `normal`, `spam` or `borderline`, as well as "skipping" a host for cases in which the host is considered impossible to classify.

3.3 Labelling

The volunteers were recruited at the beginning of July 2006 by emails posted on three mailing lists subscribed by a large number of researchers active in the area of Web Analysis and Information Retrieval: it includes `SIGIR-IRList`, `DBWorld` and `WebIR`. A total of 33 volunteers were involved in the project. The volunteers were provided with the guidelines and a password to access the classification interface, and they were asked to classify a minimum of 200 hosts (only 19 of them classified 200 hosts and are listed in the acknowledgments section of the collection)).

The labelling process last 2 weeks, and during it the volunteer issued 6,552 evaluations including "normal", "borderline", "spam" and "can not classify". At least two "normal" and/or "spam" evaluations were obtained for 2,725 hosts. After the two weeks, we searched for hosts in which two judges gave opposite evaluations and asked a third judge to provide one more evaluation.

After the labelling a post-questionnaire was submitted in order to evaluate the total time effort and to obtain a feedback on how to improve the classification process in the future. As reported by the volunteers who responded the questionnaire, they spent about 10 hours on average for classifying 200 hosts (so the entire collection can be considered as the result of about 300 man-hours of work). The job was slower at the beginning but, at the end, each volunteer was able to classify one host every 2-3 minutes. The guidelines were considered very useful, and a common problem raised by the judges was that the evaluation of borderline cases is very subjective. Indeed, many Web sites that use spam techniques also provide some contents, so that it is very difficult to classify them as spammers.

3.4 Post-processing

After collecting the judgments, we anonymized all human judgments about the collection by assigning them numbers in an arbitrary order. In the file containing the labels, human judges are identified by the codes `j1`, `j2`, ..., `j33`. We also added two special "judges":

- Judge "`odp`" labels as normal all the 5496 hosts with at least one page mentioned in the Open Directory Project on May 2006. Not all of them are normal; about 1% of the ODP domains were tagged as spam by at least 2 human judges.
- Judge "`domain`" labels as normal all the 3106 `.uk` hosts ending in `.ac.uk`, `.sch.uk`, `.gov.uk`, `.mod.uk`, `.nhs.uk` or `.police.uk`. These hosts were not assigned to human judges during the labelling phase, to focus their evaluations in the other hosts.

Labels are contained in a plain text file with one line per host, containing the corresponding hostname and all the judgements associated to it.

4 Description of the labels

We obtained 6,552 evaluations. A first observation is the amount of hosts reviewed by each reviewer. Figure 2 describes the distribution of pages reviewed among reviewers. More precisely, for every value x we plot the number of reviewers that reviewed at least x pages. Note that here and in the following we did not take into account pages that were skipped (i.e. their rating was "can not classify"). The picture remains substantially the same if also these pages are considered. We can observe three main trends: less than one third of reviewers reviewed less than 40 pages, while more than two thirds reviewed more than 100 pages. Finally, a relatively large group of reviewers, again roughly one third, reviewed strictly more than 200 pages, that was the amount of work suggested by the organizers.

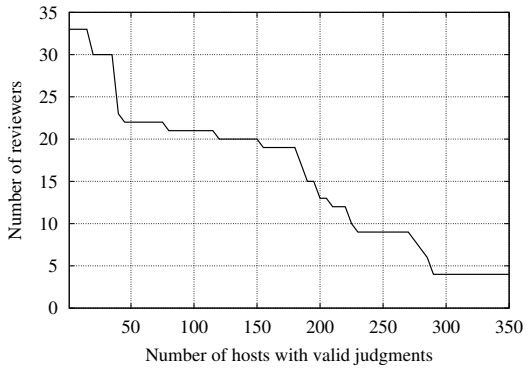

Figure 2: Distribution of the number of pages reviewed by each judge.

4.1 Overall spamicity

The distribution of the labels assigned by the judges is shown in Figure 3. The most common label was "normal", followed by "spam", followed by "borderline".

Label	Frequency	Percentage
Normal	4,046	61.75%
Borderline	709	10.82%
Spam	1,447	22.08%
Can not classify	350	5.34%

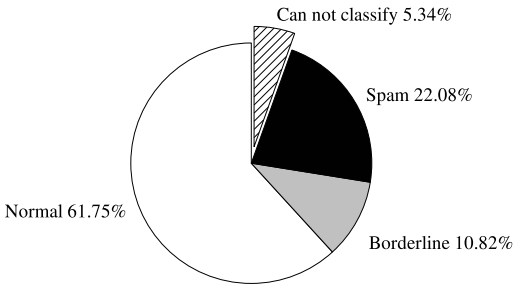

Figure 3: Distribution of the number of pages reviewed by each judge.

We calculated a spamicity measure by assigning 1 point for each "spam" judgment, 0.5 points for each "borderline" judgment, 0 points for each "normal" judgment, and taking the average. Figure 4 shows the spamicity distribution of the hosts in 5 buckets, considering only the hosts that were labelled by at least 2 human judges.

Spamicity	Frequency	Percentage
$[0.0, 0.2]$	1,530	56%
$(0.2, 0.4]$	342	13%
$(0.4, 0.6]$	179	7%
$(0.6, 0.8]$	261	9%
$(0.8, 1.0]$	413	15%

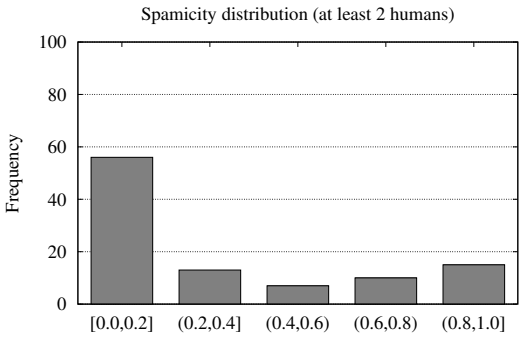

Figure 4: Distribution of the spamicity metric, including only hosts labelled by at least 2 human judges.

We considered that the final label for a host with an average of over 0.5 should be spam, for a host with an average of less than 0.5 normal, and for a host with exactly 0.5 undecided. Using this scheme, we labelled 71% of the hosts as normal, 25% as spam and the remainder 4% as undecided.

4.2 Reviewer overlap

Before measuring the consistency of the judges' subjective judgments, we studied the overlap in the reviewed page sets, that is, the degree to which page sets rated by different reviewers overlap. This aspect has important implications, since having two or more reviewers rate the same set of pages allows us to infer information about the degree of consensus as to what is spam and what is not. This aspect is further discussed below, where we consider agreement in reviewers' ratings.

For every reviewer r, we define the binary vector A_s describing the set of pages rated by r. Let $A_s(j)$ denote its j-th component; $A_s(j) = 1$ if and only if r rated page j, 0 otherwise [5]. Consider two reviewers i and l and the sets S_i and S_l of pages they rated. We define their overlap as $O(i,l) = |S_i \cap S_l|$. Obviously, $O(i,l) = A_i' A_l$. We also define the overlap index between i and l as $OI(i,l) = O(i,l)/||A_i||_2||A_l||_2$. Notice that the overlap index falls in the interval $[0,1]$, the value 1 being achieved when A_i and A_l coincide componentwise.

The average and maximum values of overlap indices are 0.5184 and 0.0268 respectively. The average has been taken considering all possible reviewer pairs (528 in total). In fact, most reviewer pairs have few pages in common, but enough of them have an overlap in the order of a few tenths of pages, as described in Figure 5. The picture gives the distribution of overlap between reviewer pairs. In particular, for every value x of the overlap, we plot the number of reviewer pairs with overlap at least x.

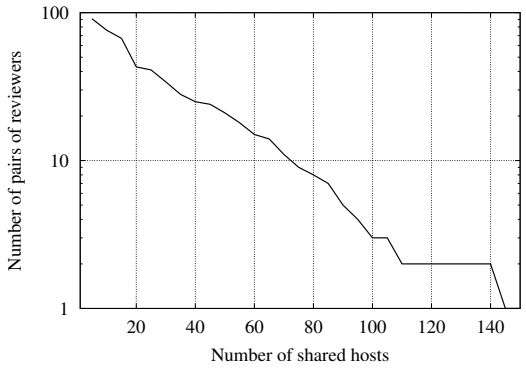

Figure 5: Distribution of the number of pages reviewed by each judge.

In total, 91 reviewer pairs share 5 hosts or more, while 43 pairs have overlap at least 20, a value that can allow to draw some preliminary conclusions, as we do below.

4.3 Disagreement metrics

One question we wanted to address was the following: is there a general consensus on what is spam and what is not? It seems reasonable to assume that the answer to this question should to some extent drive research on spam detection techniques. Still, this answer does not seem to be obvious. In fact, our measurement campaign, though in part providing preliminary results, seems to indicate that there is an only partial consensus on what is spam and what is not.

Kappa statistic. A first choice for quantifying the agreement among judges is to use the kappa statistic [Cohen, 1960], a statistical measure of inter-rater reliability:

$$k = \frac{P - P_e}{1 - P_e}$$

that is defined as the difference between how much agreement is actually present ($P - P_e$) compared to how much agreement would be expected to be present by chance alone ($1 - P_e$). P is the relative

[5] Note that $A_s(j) = 0$ if s reviewed j but expressed no rating (i.e. his/her rating was "can not classify").

agreement among judges and P_e is the probability that agreement is due to chance. In particular we use Fleiss' kappa [Green, 1997, Fleiss, 1971], a variant of Cohen's kappa, that works for any constant number of raters giving categorical ratings to a fixed number of items. For the interpretation of the statistic, we use the scale presented in Table 2.

Table 2: Interpretation of Kappa.

Kappa	Agreement
< 0	Less than chance agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Almost perfect agreement

Considering all three possible valid judgments (normal, spam and borderline) we observe an overall kappa of 0.56 that can be interpreted as a moderate agreement among the judgments. In Table 3, we can observe substantial agreements for the normal and spam categories and a slight agreement for the borderline category. The difficulty in evaluating borderline pages was clearly expressed by a several volunteers in the post-questionnaire. Indeed if we restrict the evaluation of the kappa only to the subset of hosts labeled as normal or spam, we obtain a much higher agreement with an overall value of 0.82, almost perfect according to the scale we are using.

Table 3: Kappa values for the category normal/spam/borderline

Category	Kappa	Interpretation
normal	0.62	Substantial agreement
spam	0.63	Substantial agreement
borderline	0.11	Slight agreement
global	0.56	Moderate agreement

Agreement index. Another way of evaluating the disagreement in this task is to consider a **cost matrix**. The cost matrix is a symmetric square matrix indexed by the possible labels (normal, borderline and spam), in which the entry a, b corresponds to the cost of replacing label a by label b (this is similar to the *PAM matrices* used in bioinformatics). Obviously the diagonal elements of this matrix are zero. One possible cost matrix is the following:

	Normal	Borderline	Spam
Normal	0	0.5	1
Borderline	0.5	0	0.5
Spam	1	0.5	0

This particular cost matrix means that, if in a pair of judgments, one judge considers that a host is normal and the other spam, this is a stronger disagreement than if, for instance, one judge considers that the host is normal and the other considers that it is borderline.

Now, given two reviewers i and l, we define their agreement with respect to pages in $S_i \cap S_l$ as follows: for every $j \in S_i \cap S_l$, the agreement $A_j(i,l)$ of i and l on j is $1 - cost(a,b)$ in which a is the label assigned by judge i and b is the label assigned by judge l.

We define the agreement index between i and l as $AI(i,l) = \sum_{j \in S_i \cap S_l} A_j(i,l)/|S_i \cap S_l| = \sum_{j \in S_i \cap S_l} A_j(i,l)/O(i,l)$. We considered the average, maximum and minimum value of the agreement index for increasing values of the overlap. In particular, for every value x of the overlap we restricted to all pairs with (i,l) of reviewers such that $O(i,l) \geq x$ and took the minimum, maximum and average accordingly. Figure 6 plots these values, for every $x \leq 80$ (a minimum of 10 pairs of judges with that amount of overlapping hosts).

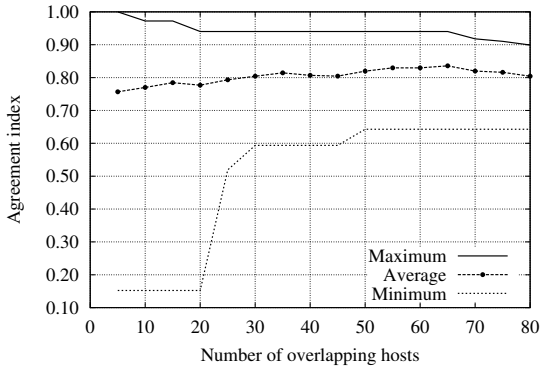

Figure 6: Agreement index as a function of the number of overlapping pairs.

Some comments are in order. First of all, as remarked earlier, most reviewer pairs actually have little or no overlap. This suggests iterating the experiments on a larger scale. In spite of this, relatively many reviewer pairs overlap significantly, at least enough for us to draw some first conclusions as to the degree of their agreement.

The average agreement is never more than about 80% (slightly more than 83% on the 14 reviewers with overlap at least 65) and never below 75%. Also, and to some extent surprisingly, the average agreement does not seem to grow with the overlap. In fact, it starts decreasing for values higher than 65, when the number of reviewer pairs over which the average is taken is still relatively high (between 10 and 15). This result should probably be further checked on larger instances, but it seems to indicate that a non negligible degree of "disagreement" is maybe not the result of statistical noise. Rather, it seems to be inherent to human rating of Web spamming and seems to indicate, to some extent, the lack of a general consensus on what exactly is spam and what not. A consequence of this fact might be the need for a stricter and unambiguous definition of what should be considered spam and what not.

5 Qualitative aspects of spam hosts

Finally, we wanted to evaluate the prevalence of different spamming aspects. For this end, and as a preliminary study, we ran a second round of evaluations by sampling at random 200 hosts that were tagged by at least two judges as Web spam. We wanted to examine the most relevant features found in hosts that were tagged as spam. After inspection of these hosts, we decided to tabulate them using the following (non-exclusive) criteria:

- **Keywords in URL**: The host contains keywords in the URLs, separated by minus, underscore or the plus sign. This is not necessarily a spamming aspect.
- **Keywords in anchor text**: The host contains pages with adjectives or query-looking keywords in the anchor text of links. This is not necessarily a spamming aspect.
- **Multiple sponsored links**: The host contains pages with a large number of sponsored (paid-for) links, or sponsored links constitute most of the clickable elements in the page. This feature is very common among spam sites.
- **Multiple external ad units**: The host contains pages with two or more external ads units (Google, eBay, Amazon, Overture, etc.) or external ad units make most of the clickable elements in the page. This is not necessarily a spamming aspect.
- **Text obtained from Web search**: There are many pages containing titles, URLs and short excerpts from other Web pages. The purpose is to increase the quantity and quality of the page's keywords and/or to increase the score of the page by pointing to reputable sources. This is very frequently a spamming aspect, except when a real search engine is provided (but often spammers just repeat the same links in a large group of their pages).
- **Synthetic text**: The host contains text that does not appear to be natural language, but consists of phrases and words "stitched" together to form meaningless paragraphs. This is almost always a spamming aspect.
- **Parked domains**: The host belongs to a domain that is "parked" by a company that owns the domain name, but there is no Web site associated to the domain name. There is often a form for bidding for the domain name and many links to other domains owned by the same operator.

The results in our sample are shown in Table 4. These numbers indicate general trends and are not conclusive.

Table 4: Aspects found in a sample of 200 spam hosts.

Aspect	Prevalence
Keywords in URL	84%
Keywords in anchor text	80%
Multiple sponsored links	52%
Multiple external ad units	39%
Text obtained from Web search	26%
Synthetic text	10%
Parked domains	4%

6 Conclusions

One way of making the spam judgments more objective is to propose the reviewers a certain procedure they must follow to label hosts or pages. Such a procedure could take the form of a

"checklist" of spam aspects such as the ones studied in Section 5. We considered that for this first collection it would have been premature to state such procedure. However, future collections should include at least a sub-set of labels obtained in such a way.

Web spam is a challenging area in which many things are yet to be discovered. An interesting and challenging problem is to study how to stay ahead of spammers, proposing general methods that can be easily adapted to new types of Web spam. There is a strong economic incentive to score high in search engines, so "[o]ne might try to address speculative Web visibility scams individually (as search engine companies are no doubt doing); however, the bubble is likely to reappear in other guises." [Gori and Witten, 2005].

6.1 Licensing and availability

Labels can be freely downloaded and are available under a Creative Commons *Attribution-NonCommercial-ShareAlike 2.5* license [6]. This license basically states that researchers are free to use the data and that we make no warranties about it. Researchers can use the data for any purpose, even in a commercial environment. The *NonCommercial-ShareAlike* clause applies only for redistributing the data publicly. We advice researchers not to use these labels directly for search engine ranking.

The Web graph in compressed format can be freely downloaded. It was obtained by crawling the Internet following commonly accepted methods, and indexing publicly available documents.

The contents of the pages are available upon request. Due to the large size and the nature of this information, researchers are required to sign a data usage agreement before obtaining the contents of the pages. See `http://www.yr-bcn.es/webspam/` for details. The collection is currently hosted by Yahoo! Research Barcelona.

6.2 Acknowledgements

This collection was possible due to the work of a team of volunteers. We thank them for their time and effort:

Thiago Alves	Antonio Gulli	Tamás Sarlós
Luca Becchetti	Zoltán Gyöngyi	Mike Thelwall
Paolo Boldi	Thomas Lavergn	Belle Tseng
Paul Chirita	Alex Ntoulas	Tanguy Urvoy
Mirel Cosulschi	Josiane-Xavier Parreira	Wenzhong Zhao
Brian Davison	Xiaoguang Qi	
Pascal Filoche	Massimo Santini	

Many of the volunteers also provided valuable feedback about the guidelines and the interface before starting the labelling process. We also thank Ludovic Denoyer and Ricardo Baeza-Yates for their help with the guidelines and the interface.

This research was partially funded by the DELIS project (Dynamically Evolving, Large Scale Information Systems) [7].

[6] http://creativecommons.org/licenses/by-nc-sa/2.5/deed.en

[7] `http://delis.upb.de/`

References

[Becchetti et al., 2006] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. (2006). Using rank propagation and probabilistic counting for link-based spam detection. In *Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD)*, Pennsylvania, USA. ACM Press.

[Benczúr et al., 2006a] Benczúr, A., Csalogány, K., and Sarlós, T. (2006a). Link-based similarity search to fight web spam. In *Adversarial Information Retrieval on the Web (AIRWEB)*, Seattle, Washington, USA.

[Benczúr et al., 2006b] Benczúr, A. A., Bíró, I., Csalogány, K., and Uher, M. (2006b). Detecting nepotistic links by language model disagreement. In *WWW*, pages 939–940.

[Benczúr et al., 2005] Benczúr, A. A., Csalogány, K., Sarlós, T., and Uher, M. (2005). Spamrank: fully automatic link spam detection. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan.

[Boldi et al., 2004] Boldi, P., Codenotti, B., Santini, M., and Vigna, S. (2004). Ubicrawler: a scalable fully distributed web crawler. *Software, Practice and Experience*, 34(8):711–726.

[Boldi and Vigna, 2004] Boldi, P. and Vigna, S. (2004). The webgraph framework I: compression techniques. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 595–602, New York, NY, USA. ACM Press.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Psychological Bulletin*, 20:37–46.

[Davison, 2000] Davison, B. D. (2000). Recognizing nepotistic links on the web. In *Aaai-2000 Workshop On Artificial Intelligence For Web Search*, pages 23–28, Austin, Texas. Aaai Press.

[Eiron et al., 2004] Eiron, N., Curley, K. S., and Tomlin, J. A. (2004). Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA. ACM Press.

[Fleiss, 1971] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

[Gori and Witten, 2005] Gori, M. and Witten, I. (2005). The bubble of web visibility. *Commun. ACM*, 48(3):115–117.

[Green, 1997] Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the Twenty-Second Annual Conference of SAS Users Group*, San Diego, USA.

[Gyöngyi and Garcia-Molina, 2005] Gyöngyi, Z. and Garcia-Molina, H. (2005). Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*.

[Gyöngyi et al., 2004] Gyöngyi, Z., Molina, H. G., and Pedersen, J. (2004). Combating web spam with trustrank. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB)*, pages 576–587, Toronto, Canada. Morgan Kaufmann.

[Ntoulas et al., 2006] Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. (2006). Detecting spam web pages through content analysis. In *Proceedings of the World Wide Web conference*, pages 83–92, Edinburgh, Scotland.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project.

[Perkins, 2001] Perkins, A. (2001). The classification of search engine spam. Available online at http://www.silverdisc.co.uk/articles/spam-classification/.