Revealing the Relationship Network Behind Link Spam

Apostolis Zarras
Ruhr-University Bochum
apostolis.zarras@rub.de

Antonis Papadogiannakis
FORTH-ICS
papadog@ics.forth.gr

Sotiris Ioannidis
FORTH-ICS
sotiris@ics.forth.gr

Thorsten Holz
Ruhr-University Bochum
thorsten.holz@rub.de

***Abstract*—Accessing the large volume of information that is available on the Web is more important than ever before. Search engines are the primary means to help users find the content they need. To suggest the most closely related and the most popular web pages for a user's query, search engines assign a ranking to each web page, which typically increases with the number and ranking of other websites that link to this page. However, link spammers have developed several techniques to exploit this algorithm and improve the ranking of their web pages. These techniques are commonly based on underground forums for collaborative link exchange; building a relationship network among spammers to favor their web pages in search engine results. In this study, we provide a systematic analysis of the spam link exchange performed through 15 *Search Engine Optimization* (SEO) forums. We design a system, which is able to capture the activity of link spammers in SEO forums, identify spam link exchange, and visualize the link spam ecosystem. The outcomes of this study shed light on a different aspect of link spamming that is the collaboration among spammers.**

I. INTRODUCTION

The World Wide Web offers an abundance of information accessible to anyone. To identify the most useful information among the vast amount of available web pages, users rely primarily on search engines. Search engines typically classify a huge number of web pages and present the ones that seem most relevant to user queries ranked by their estimated relevance and their popularity. The users typically visit the highest-ranked web pages and ignore the rest [17]. To attract more users, it is therefore important for each web page to rank high in the search engine results. While honest web pages achieve this ranking due to the quality of their content, dishonest ones try to mislead search engines to rank them higher than they deserve. We refer to the attempts of these dishonest pages that try to deceive search engines as *link spam*.

Link spam is used for several reasons, ranging from money-related activities to malware propagation. Therefore, it is becoming more popular and more sophisticated as the rapid increase of the Internet users leads to higher revenue for spammers [12,18,24]. Unfortunately, this has a negative impact to user's experience: link spam is annoying as users often cannot find what they are searching for, while constitutes a security problem due to possible malicious content on spam pages. Additionally, it causes headaches to the search engines themselves. Search engines must exert significant effort to filter link spam and satisfy users expectations. Thus, over the years, many different anti-spam techniques have been developed [3,9,13,20]. However, adapting to such techniques, spammers always improve their strategies to evade detection.

To determine the reputation and popularity of a web page, search engines commonly rely on the number and ranking of the other web pages that link to it. The more websites linking to a page p, and the more popular these websites are, the higher is the ranking that the page p will receive from a search engine [4]. Although this is a reasonable way to define page ranking, it can also be exploited by spammers to boost the ranking of their pages by increasing the number of links to it. So called *Search Engine Optimizations* (SEO) forums are often used by spammers for this reason. Despite the fact that search engines approve SEO as a way used by web page owners to achieve a better recognition of their websites [7], *blackhat SEO* techniques considered as unfair means of boosting the web page ranking. For simplicity, in the rest of the paper we refer to the blackhat SEO as SEO.

SEO forums bring together page owners who want to improve the ranking of their web pages. However, these forums also attract spammers who swap ideas on spamming methods and exchange links. In this work, we study how the information found in SEO forums is related to link spam. We build a system that collects and analyzes the links posted in public threads or sent via private messages. We noticed that many spammers, to avoid detection, tend to exchange links only through private messages. Hence, to collect data from such spammers, our system uses *honey accounts* that behave like typical spammers; post on public threads asking for links exchange with other websites. Next, we analyze the harvested links with respect to their spam affiliation, their frequency, and their occurrences among different users and different forums. Additionally, we crawl the web pages found in SEO forums posts or messages to examine their structure and extract their links, which are then matched against the other web pages found on monitored forums. Finally, to visualize the developed relationships among link spammers, we use graph structures based on the extracted information from the observed exchanges of links.

We examined 15 popular SEO forums and after a three-month period we collected 97,658 web pages that participated in link spam. Overall, we discovered two major categories of spammers with distinct features. Each category behaves in a completely different manner and thus, we need different approaches to reveal their spam web pages. In addition, a deeper analysis of the collected data revealed few clear differences in the type of link exchange and in the relationship network between URLs exposed in public threads and those sent via private messages. These results improve our knowledge on the web spam ecosystem and shed light on the activities performed in underground forums related to link spam.

In summary, we make the following main contributions:

- We collect a corpus of spam links from SEO forums. Our approach is the first one that uses *honey accounts* to harvest spam links from private messages.
- We analyze the links found in SEO forums and validate link exchange by *crawling the respective web pages*. Instead of solely relying on collected data from SEO forums, our analysis correlates this information with data extracted from the actual web pages.
- We present an in-depth analysis of data gathered from 15 popular SEO forums. The outcomes reveal the different approaches and strategies used by advanced and inexperienced web spammers.

II. Web Spam

The intention of web spamming is to increase the ranking of spam web pages by misguiding the ranking algorithms used by search engines. According to Gyongyi et al. [8], the term web spamming refers to any deliberate human action that is meant to trigger an unjustifiably favorable relevance or importance for a web page, considering the page's true value. There are several reasons for creating web spam, such as increased ad revenue, phishing attacks, profit from illegal activities and malware distribution. It is obvious that the highest-ranked web pages get clicked much more often by Internet users and this is something that drew miscreants' attention. As a matter of fact, users tend to trust search engines as the primary mean of finding information in a fast and effective way, and they rarely question the returned results. Therefore, as spammers target end users through search engines, they often try to boost their own web pages to appear higher in the search engines returned results with the aid of term and link manipulation techniques [8, 10, 13]:

Term Spam: To mislead search engines, spammers repeat specific terms on their spam web pages to trick the engines into deciding that the page is closely associated with these terms. Such terms often do not build a useful sentence: to increase the numbers of queries that are associated with a spam web page, attackers dump a large number of unrelated terms on the web page, which are often copied from dictionaries. This is an effective technique to lead rare queries to a spam web page because they are not included in many benign web pages, and a spam page with such terms is highly ranked. Additionally, misleading meta-keywords or anchor text can boost the page's ranking as well.

Link Spam: This method aims to modify the structure of the web graph by increasing the number of the backlinks targeting the spam pages. To this end, spammers often post backlinks to their web pages on guest books, wikis, messages boards, blogs, or in web directories. Such procedures are easy to handle and do not need extra effort or money. Moreover, as the cost of web servers is very cheap, spammers can leverage these servers to build a link structure. These servers can be used in different ways to push the ranking of the spam pages. Some of the servers could provide useful information by copying the content of other benign web pages, and linking also to the spam page. On top of that, spammers build *link farms* where they can interact with other spammers and exchange URLs. As means of communication, spammers usually contact each others through SEO forums. This procedure is called *link exchange* and is the main focus of our study.

There are three different kinds of link exchange: (i) the one-way link exchange, where only one website a links to another website b, (ii) the two-way link exchange, where two web pages a and b are both linking to each other, and (iii) the three-way link exchange, where the web pages do not link directly to each other, but they use a third web page to build a circle of links. For instance, the website a links to c, and the website c links to b. This type of link exchange is harder to be detected in practice. In this work, we study the two- and three-way link exchange. When a node a participates in more than one N-way exchanges, it is considered to participate in a link farm. For instance, in case a links to b and b also links to a, and similarly a links to c and c to a, a is a part of a link farm.

Usually, web spammers are miscreants that own malicious web pages, such as phishing or drive-by-download pages, or even pages that participate in ad frauds and thus, they try to monetize them. As we already mentioned, more traffic to a website increases its value in the underground black market, which translates to more money for the site's owner. Hence, among all the other tricks the cybercriminals utilize in order to attract more users to their web pages, they also leverage link exchange as a tool to achieve their nefarious tasks. However, it is worth noting that not all web spammers own a malicious website. Some times, holders of blogs, personal web pages, or any other kind of benign web pages use similar techniques to increase their web page ranking for several other reasons. In this chapter, we define web spam to be *any* link exchange among websites, even if this action involves benign websites.

III. Data Collection

In this section, we describe in detail the goals of our work, and provide an overview of our measurement methodology.

A. Study Objectives

Identifying link spam is a continuous process for search engines, and typically this process is hidden from the average end users. According to Wall [19], link spammers form alliances in order to exchange links among their web pages, resulting in *global* link farms. The most common channel used by individuals to communicate with each other is SEO forums. Thus, it is almost impossible to discover these farms with traditional anti-spam techniques. In this work, we study the spammers and spam websites that use SEO forums and expose their alliances. To this end, we developed an infrastructure that allow us (i) to identify web pages that use SEO forums for improving their page ranking, and (ii) to visualize these pages and the formed relationships among them. We use this system to study the link spam ecosystem through the information that is available or can be collected at the popular SEO forums, and measure the extent of utilization of SEO techniques and their impact on the Web. Keep in mind that with this work we do not want to provide yet another detection system, but we want to study how advanced and inexperienced spammers collaborate and create link farms.

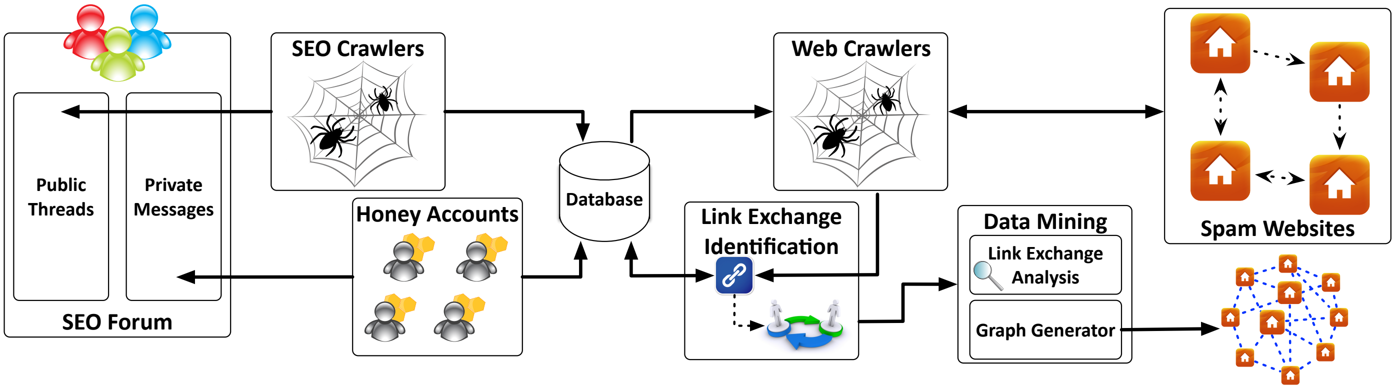

Figure 1: Architecture of our approach.

Our system is based on the idea that there is a mutual exchange of URLs among link spammers. More precisely, when a spammer A wants to increase the number of URLs linking to her web page, she is willing to add the URL of another spammer B in her web page, if the latter adds the web page of the A in her own site as well. Having that knowledge in advance, we can accurately discover web pages that try to increase their ranking with fraudulent means. Therefore, we have created a set of crawlers that gather URLs from SEO forums' public threads (under the sections of *Link Exchanges*), and a set of honey accounts that send messages to other users requesting link exchange with honey web pages, while hooking the responses through private messages.

We use the notion of *link spam* throughout the paper. For this work we consider as link spam any web page that unethically tries to increase its page ranking by participating in global link farms. In addition, we consider as link exchange any mutual link transaction between two pages with similar or disparate web contents. The combinations of all these link exchanges form link spam relationship networks. In conclusion, the main goal of our work is to study the ecosystem behind these networks, and try to expose and categorize the behavior of different spammers.

B. Data Collection Architecture

Figure 1 depicts the general architecture of our approach. The core elements of our infrastructure are: (i) the SEO forum crawlers, (ii) the honey accounts army, and (iii) the web page crawlers. While the SEO forum crawlers are responsible for harvesting the URLs from SEO forums, the honey accounts try to lure link spammers to expose information that is not publicly available and would only reveal to other peers through private messages. Finally, the web page crawlers examine each page found in the set of collected links and validate the actual link exchange. The data collected by these crawlers are forwarded to the components that are responsible for analyzing and correlating them in order to reveal the relationship network among spammers. Finally, the link exchange and the spammers' relations visualized by creating the respective relationship network graphs.

SEO Crawlers. The SEO crawlers search underground forums for URLs that participate in link exchange. These forums have predefined places (sub-forums) where the users can exchange URLs. Therefore, the crawlers target these sub-forums to find and extract the URLs. For this procedure it is important to consider the HTML structure of the forums. Our experimental results reveal that all of our examined forums use one of the following platforms: (i) vBulletin, (ii) phpBB, or (iii) MyBB. This fact allow us to create crawlers that have identical behavior on more than one SEO forum. Consequently, the crawlers can extract all the necessary information from the forums (such as links, post authors, usernames) with small modifications in their configurations. Similarly, the URL extraction from the forums needs to be handled carefully. There exist posts that are not spam related and the included links are not posted for spam purposes. For instance, we observed a plethora of links to popular websites. For that reason, SEO crawlers use a whitelist to decide if a link should be extracted and stored in the database. Additionally, we filter all the links found in users' signatures. Moreover, the users frequently quote posts of other users. This leads to double posted links in one thread and is appearing the same link to be posted by more than one users. To find the user who had originally posted the link, our crawlers percolate all the *quoted* elements from the posts and keep only these that differ from the previous. This way, we obtain a clean mapping between the link and the user who originally posted it.

Honey Accounts. We have witnessed that a significant fraction of the link exchange is performed through private messages. Hence, in order to gain access to that kind of information we need an approach that lures link spammers to expose themselves. From our prior knowledge, we know that a spammer is willing to reveal a certain type of information only to a fellow spammer. Thus, it is necessary for our approach to create fake accounts (i.e., Sybils) and make them to behave as if they are real spammers. For this purpose, these Sybil accounts can post requests for link exchange and harvest the responses sent by private messages. In addition, they have the capability to reply back to other users when the received private messages do not contain any exchanged URL. As a matter of fact, the responses are different each time they reply back, which make it more difficult to categorize these accounts as Sybils. We know that creating honey accounts to retrieve internal information constitutes a short-term solution and can not be used as an anti-spam technique. Nevertheless, this approach provides us with valuable information that it could not have been retrieved with different means.

Web Crawlers. The web crawlers map the link structure of the networks behind the harvested URLs. They follow all the outgoing links up to a defined depth and store the extracted URLs in a database for further analysis. To have more accurate results we used instrumented browsers as crawlers. When a crawler visits a new web page, it searches for every link on the page and checks if this link satisfies some predefined conditions. Initially, the link is reduced to its hostname. If the hostname of the link and the current web page match, it is ignored because it is a navigational link. Our experimental results revealed that most of the analyzed spam pages host their outlinks on their main page. In case of an outlink is found, the crawler will check if the linked page is already crawled. Finally, if the link is not crawled, it will be appended to the crawler's queue and the pair of source and destination URL will be stored in the database.

Link Exchange Analysis. With all the information gathered in a central database, we correlate the relationships among the crawled web pages. More precisely, we discover the connections among different entities and observe the real interactions in link farms (*link exchange identification*). For instance, if a user claims that she will add the web page of another user and this is never happened, this relationship is classified as *broken*. This is a major difference from previous works that handle all the posted URLs in SEO forums as accomplished [5]. Our approach can also recognize two- and three-way link exchange. Although a two-way link exchange is quite straightforward, a three-way requires more sophisticated techniques. To do so, we correlate all the links contained in public threads or in private messages; we consider all these links as a cluster and try to find relationships among them. If there is a relationship that includes more than two URLs, we conclude that there is at least one three-way link exchange in this cluster.

Graph Generator. This component is responsible for graphically represent the spammers' relationships. A graphic representation can provide a clear view of the spam web pages, i.e., the pages that participated in a large number of link exchange. As we can recognize the major players on link exchange, we can easily extract viable conclusions about the procedure that these spammers follow, such as if they require link exchange for one or more pages, whether they prefer to advertise the link exchange with public posts or private messages, if they use a common template. Our system provides different levels of graph representations, such as relationships among users or web pages, two- or three-way link exchange, and link exchange through public posts or private messages.

IV. SEO Forums Analysis

In our study, we analyzed 15 SEO forums that contain sub-forums for link exchange and gathered data for a three-month period. These forums are among the top websites where users can search for SEO boosting techniques and they number hundreds of thousands active users. Each of these forums includes sections that describe how to boost web pages to appear higher in search engines returned results as well as sections that offer link exchange among their users. In our research, we only focused on the sections related to link exchange. To this end, we analyzed in total 9,617 threads and extracted 25,338 unique URLs generated by 7,923 users. Our results indicate that there is a ratio of 3.6 replies per thread. This means that for each web page that tries to boost its page rank, there is an average of three other web pages that are willing to contribute to this goal. It is worth noting that during our measurements we analyzed all the reply messages in public threads and found that 26.79% of them did not contain any actual URL but a reference to a private message.

Table I: Percentage of URLs and users that appear on one up to three different forums.

Number of forums	URLs	Users
1	95.53%	97.99%
2	4.04%	1.71%
3	0.43%	0.30%

A. Spammers Behavior

Initially, we examined the behavior of users in SEO forums based on the number of posts they make and on the number of URLs they send on a single thread. We saw that the majority of users (53.94%) that participate in a thread make one post including one URL in each thread. The percentage is getting lower as the number of posts and URLs increasing. On average, each user generates 1.08 posts and 1.61 URLs per thread. This reveals that most of the users that post publicly own maximum one website. We also noticed that the users who own many websites usually present all URLs in a single post. The rest of the posts mainly contain information such as the category and ranking of the web pages.

Forums have very strict rules for spamming. Users are allowed to freely post in the forums, however, they are forbidden to spam. As spam is considered, among the others, the replication of the same content in different threads. The accounts that caught spamming are permanently banned. According to this, we measured the frequency of each posted URL in all threads. We observed that the 75.61% of all the posted URLs occurred only one time, while the 98.11% was found in up to five different posts. This lead us to the conclusion that users try to avoid excessive spam inside these forums.

Next, we measured the total contribution of a user in a forum. We assessed the activity of users by counting the total number of posts they make. The forums usually have different tiers for their users depending on their activity. The higher tier an account has, the more benefits it gains. We noticed that 78.13% of the users made less than 20 posts. The amount of users participated in this category is so large because most of the forums allow posts in the link exchange sub-forums only when a user has a defined number of posts, which varies from 5 to 20 posts. We observed that the users who only want to exchange links are not very active on other parts of the forums.

Table I shows the percentage of URLs and users that appear on up to three different forums. Most of the users are active only in one forum (97.99%). This percentage, however, may not be very accurate as users may not use the same username in each forum. Taking into consideration that most of the unique URLs (95.53%) are also posted just in one forum, the percentage of users in the different forums discovered by our approach has to be close to the absolute number. Interestingly, we did not find any user or URL to be presented in more than three different forums.

As we previously mentioned, a fraction of users reveal their web pages only via private messages. In order to harvest these pages we created honey accounts. These accounts sent requests for link exchange to threads where the initial post does not include the exchanged URL, and reply to private messages. We wanted to retrieve a wide variety of URLs and thus, we followed two different strategies. In the first strategy, we created web pages and enriched them with content from various categories. These pages had low page ranking and therefore, we were able to attract other pages with similar ranking. In the second strategy, we advertised highly-ranked pages that we did not own and spurious claimed that we offered them for link exchange with other highly-ranked web pages. This method attracted owners of higher ranked websites. We overall retrieved 718 unique URLs from both strategies. We compared these URLs against the ones we collected from crawling the public threads and found an overlap in only three URLs. These results showed us that there are two disjoint set of link spammers: spammers that exchange their web pages through publicly available posts, and spammers that exchange their pages only via messages. The users who send their URLs only via private messages were proved to be more suspicious of being detected for link spamming.

B. *Spam Pages Categorization*

Another interesting aspect on comprehending the link spam network is to measure the page rank of the web pages that request link exchange. This will help us to understand if the highly-ranked web pages behave in a similar way with the low-ranked pages. To this end, we used Google's toolbar queries to get the page rank of the pages we found in SEO forums. Figure 2 depicts the outcomes of our analysis. The page rank `0` contains web pages, which have yet to be ranked. We discovered that the majority of the spam pages have page rank `2`, while more than 90% of spam pages belong to a page rank lower than or equal to `4`. Additionally, pages with rank greater than `6` are only the 0.88% of the total amount of spam pages. Consequently, we believe that owners of highly-ranked pages behave completely different compared to owners of low-ranked pages. Obviously, highly-ranked pages usually do not participate in link exchange networks to such an extent as the low ranked, either because they do not need it, or because they find different ways to increase their ranking, such as paying search engines to pitch their pages in better positions, or paying for more effective and targeted advertising methods. On the other hand, low-ranked pages, which are usually blogs or personal websites, seek cheaper solutions to increase their page ranking. Finally, there is the category of malicious web pages that cannot increase their ranking in any legal means, so they are forced to look for illegal ways, such as link exchange, to achieve that. Note that identifying malicious web pages is outside the scope of this paper.

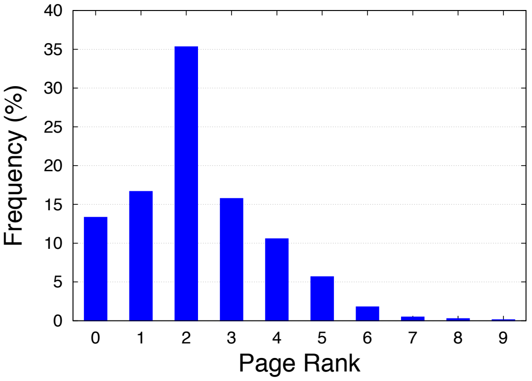

Figure 2: PageRank of web pages that request for link exchange in SEO forums.

In the following experiment we categorize the web pages based on their contents and spot these categories that are more prone to link exchange than others. In essence, every web page has a theme that defines its content. Usually, the owners are trying to exchange links with other pages that belong to a similar category. To get an overview of the different provided themes, the threads are analyzed for the main subject of the related web pages. The results in Table II illustrate the top themes requested for link exchange. We observe that there are many different types of web pages found in the forums, and most of them have a very close popularity. Only the web pages regarding *travel* and *health* protrude. The 27.68% of all the web pages in the examined threads have several other themes that are rarely requested.

Table II: Distribution of the requested themes.

Theme	Percentage
Travel	16.33%
Health	10.34%
Finances and Business	7.33%
Adult Content	7.31%
Shopping	6.67%
Technology and IT	6.66%
Online Games	6.34%
Internet and Web Design	6.00%
Entertainment	5.34%
Other Themes	27.68%

Table III: Breakdown of countries hosting web spam.

Country	Percentage
United States	71.21%
United Kingdom	7.62%
Germany	2.46%
Netherlands	2.31%
Canada	1.77%
France	1.69%
Bahamas	1.68%
Japan	1.12%
India	1.05%
Other Countries	9.09%

Finally, we tracked the countries where the spam web pages hosted. In Table III we can see the breakdown of the spam websites into the respective countries that hosted them. We observed that the vast majority of web pages were hosted to English speaking countries. United States are ranked first hosting 71.21% of the total amount of spam pages, while the United Kingdom follows in the second place with only 7.62%. The remaining countries shared the rest 21.17% of the web spam hosting.

V. Link Exchange

Through analyzing the SEO forums' threads, we discovered that spammers request web pages with identical page ranking for link exchange. Likewise, the replies to link exchange threads propose an exchange with a similar or higher ranked page. Usually, most of the replies deal with two-way and only a small fraction of users require implicitly three-way link exchange. Moving on the messages gathered from the honey accounts, we discovered that users with a higher page rank in their websites prefer three-way link exchange, compared to users with a lower page rank that prefer two-way exchanges or many times they do not care about the type of link exchange at all.

Table IV: Statistics of requested link exchange types.

Type	Public Posts	Private Messages
Two-way link exchange	65.47%	87.56%
Three-way link exchange	0.81%	8.47%
Link farm	1.45%	3.93%
Not defined	32.27%	0.04%

Previous studies [5] focused only on the available information provided by public forum threads. In contrast, we went one step further: we crawled the actual web pages to validate the exchange and found a number of link exchange in SEO forums that were not defined in the actual websites, mainly in the links found in public threads. We classified as *not defined* all the spam pages that we do not have a clear picture about the category of the link exchange they belong to, or we are not sure if the link was actually exchanged between the two pages. There are two possible reasons for not defined links: (i) the web pages did not actually exchange the link, or removed it at some point, or (ii) they exchanged the link through a private message and thus, it was not included in our dataset.

Table IV shows the classification of link exchange requested types in both public posts and private messages. We see that a significant amount of link exchange (32.27%) is not defined in public posts. This mostly happens because a portion of spammers (26.79%) disclose the requested exchange type and the requested URL only through private messages. On the other hand, when we analyze the private messages, the not defined link exchange types decrease to just 0.04%. We did not expect private messages with not defined exchange type, but a deeper analysis revealed that very few spammers misbehave and remove the outlink from their web pages, transforming the two-way into one-way link exchange. Thus, the large percentage of not defined link exchange type we observed in public posts is moved into the two- and the three-way link exchange as well as to the link farm in the case of private messages. More precisely, the two-way link exchange increased from 65.47% to 87.56%, the three-way link exchange from 0.81% to 8.47%, and the link farm requests from 1.45% to 3.93%. The most interesting increase in ratio is the one happened in the three-way link exchange. As we already mentioned, spammers with higher ranked web pages require a link exchange with other higher ranked pages, or a three-way link exchange. These spammers do not jeopardize to publicly reveal their websites and they only contact other spammers through private messages. As a result, we notice this increase in the three-way link exchange ratio when we move from public posts to private messages.

Overall, SEO crawlers and honey accounts collected 26,053 unique URLs. These URLs were the initial seeds for our web crawlers. During the three-month period we crawled more than 10 million web pages. A deeper look in the results revealed 97,658 web pages that participate in link exchange. The vast majority of these pages where part of two-way link exchange, however we spotted 274 situations where the web pages participated in three-way link exchange.

Next, we further investigated these 97,658 web pages. Surprisingly, we discovered 842 IP addresses that were serving 5,916 different domains. In addition, we analyzed the whois records to possibly link together domains. We detected that the 5.24% of all the unique domains were registered by the same entities. Furthermore, the information in 11.69% of the domains were protected by private domain registrations.

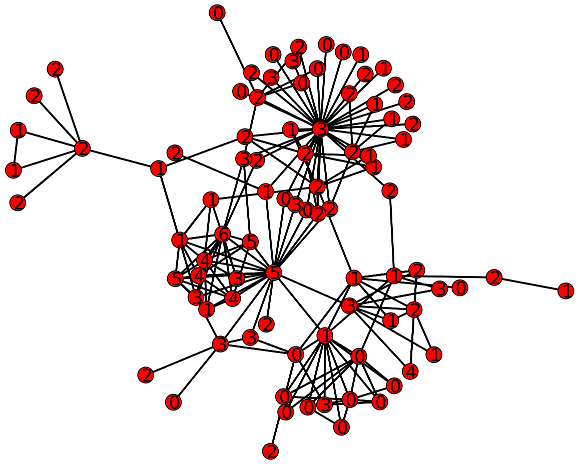

Figure 3: A two-way link exchange network. The numbers on nodes indicate the web page ranking.

VI. Relationship Network Graph

We analyzed our data to find clusters of spam web pages. Each cluster denotes a link exchange relationship among the involved pages. For each cluster we produced a *relationship network graph* where each node participates in a two- or three-way link exchange. During our study, we found 41 clusters with more than 50 nodes, 983 clusters that consist from 10 to 50 nodes, and 2,758 clusters that contain up to 10 nodes.

Then, we generated relationship network graphs from our dataset based only on messages found in public threads in the SEO forums we monitored. The edges in these graphs represent two-way link exchange among spam web pages. All the nodes are pages that we retrieved from the SEO forums. The number within each node represents the page rank of the respective web page. A close observation on these graphs can reveal how inexperienced link spammers interact with each other. Some of the pages in these clusters have exchanged links with only a small number of other spam pages, while others have exchanged up to 25 links. We mostly see web pages with small page ranks in clusters like the one presented in Figure 3. One explanation for this is that the inexperienced spammers ally with other spammers of similar level.

Figure 4 shows a relationship network graph for a cluster of web pages that appeared in both public threads and private messages. Similar to Figure 3, each edge represents a link exchange and each node a page with its respective page rank. The circles depict pages found in public threads, while the rhombuses pages collected from private messages. As we already mentioned, URLs sent by private messages usually contain pages with higher page rank compared to those who are publicly posted. Additionally, we notice the creation of clusters where the URLs sent by private messages collaborate with each other, similarly to the publicly posted pages. We also observe some pages that we saw them only in private messages to collaborate with pages that appeared in public threads.

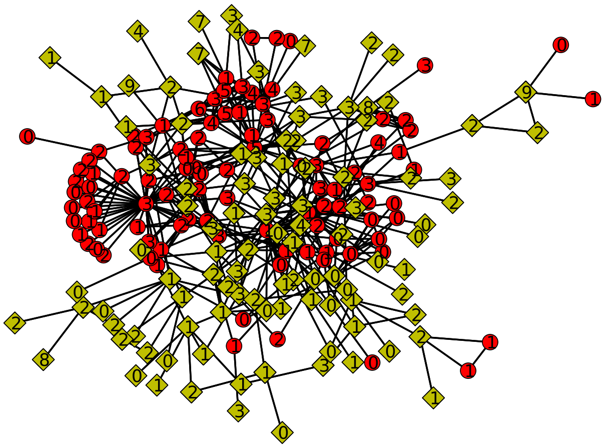

Figure 4: A link exchange network including web pages from both public threads and private messages.

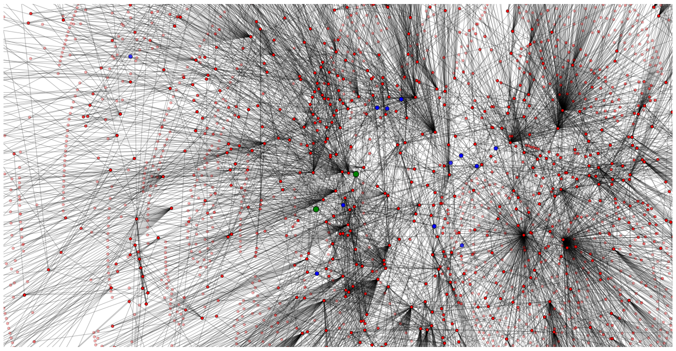

Figure 5: A webspam ecosystem.

In the previous experiments, we chose to be conservative with the generation of the relationship network graphs and only considered as nodes of the graph web pages that appeared in SEO forums. A more liberal approach that count all the outlinks from a spam web page as possible spam web pages could lead to bigger graphs (ecosystems).

To extend our dataset with more link exchange, we started with the web page pairs that were participating in two-way links exchanges, as initial nodes in the extended graph. Then, we started to recursively crawl these web pages to retrieve URLs that fulfill the two-way link exchange requirements. It is worth noting that a small portion of the discovered nodes were already in our database. This proves that even without the prior knowledge of all the participating nodes in a link exchange, it is possible to retrieve nodes with a similar behavior. Figure 5 displays how such a link spam ecosystem looks like. With green nodes we represent the initial pair of web pages we collected from SEO forums and they participate in link exchange, with blue nodes we define the pages that were already presented in our database, and with red nodes all the web pages that web crawlers revealed and they are part of link exchange network.

VII. Summary of Findings

In this work, we studied 15 SEO forums that contain subforums for link exchange and we tried to understand how their members behave. Our study reveals that spammers who use link exchange behave in similar manner, and hence, we are able to extract their heuristics and classify them into categories. The analysis results reveal two main categories of link spammers. The first, which counts the majority of the investigated members, consists of spammers that own low ranked websites. These spammers usually post their websites publicly and thus, it is easy to identify them. Additionally, they belong to the *hit-and-go* group, which means that they are not active in SEO forums and contribute only with a limited number of posts just to be able to participate in link exchange. On the other hand, we have the more experienced link spammers. These spammers, do not post publicly and they communicate with the other members by sending private messages. They own a sufficient amount of websites including highly-ranked domains. These spammers are more difficult to be identified and therefore, advanced techniques should be used to lure them to expose their websites.

Regarding the link exchange, we notice that the first category prefers a two-way link exchange. We assume that these spammers have limited knowledge of SEO optimization techniques and presume that by having more links targeting their websites can mislead the search engines. In contrast, the advanced spammers, are aware of how the page ranking system operates and thus, they prefer the three-way link exchange. They know that many ranking systems do not count the backlinks if there is a two-way link exchange involved. Hence, they create *dummy* websites that exchange with other users to achieve their goals.

VIII. Discussion

We believe that our analysis provides an accurate insight on the behavior of current techniques used by link spammers. This is because we collect and carefully analyze a large volume of data, while we also correlate different data sources to validate the link exchange in SEO forums. As we cannot have a direct access to the complete information stored in the databases of the SEO forums, we make a best effort approach to collect as much data as possible, either by public sources, or by trying to convince other users to send spam links to honey accounts. Therefore, we are not able to collect and analyze all link exchanges through these forums. However, we believe that our approach provides us with a representative and adequate sample of the spammers' activity.

Our approach utilizes honey accounts to harvest data that are not publicly available. Although these accounts have a certain level of intelligence, they could be identified if SEO forums deploy more advanced detection techniques. Additionally, there exist cases in which these Sybil accounts do not know how to act. This happens when the algorithm behind them cannot successfully recognize and thus, categorize the text in public posts. This can also happen when it comes to private messages' replies. In these cases, a manual input is required. Consequently, we do not recommend Sybil accounts as a long term solution. Nevertheless, in our study they were a necessary "evil" in order to uncover disclosed information, which we could not access by any other means.

IX. Related Work

Web spam, in which link spam belongs to, as a phenomenon is nearly as old as the Web itself and thus general aspects of web spam have been discussed in a large number of studies over the last years. Previous works focused on a wide variety of issues including economic aspects of web spam [11, 16], cloaking and redirection techniques used by web spammers [8, 21], and content analysis of spam web pages [14, 15]. However, there are relatively few examples of empirical studies that identify the means by which spammers communicate with each other, most likely due to the private nature of this communication.

Many anti-spam methods such as TrustRank [9], BadRank [20] and SpamRank [3] have been proposed to detect link spam or denote its influence on page ranking. Adali et al. [1] demonstrated that generating pages with links targeting a single page is the most effective means of link spam, while Zhang et al. [25] showed how to make PageRank [4] robust against attacks. Finally, Fetterly et al. [6] investigated the cases where web pages are mosaics of textual chunks copied from legitimate pages and presented methods for detecting them. Our work is complementary to these studies, since we are focusing on the link structure of web spam.

Using honey accounts is an aged old idea on conducting interactive studies. Such approaches appeared for example in studies that investigate spam appearances in instant messaging systems: HoneyIM [22] is a system that uses decoy accounts in user's contact lists to detect content sent by instant messaging malware. Similar, HoneyBuddy [2] is an active architecture that constantly adds "friends" to its decoy accounts and monitors a variety of instant messaging users for sign of contamination. On the other hand, systems such as Camouflage [23] do not try to hide their *"honey"*-based behavior, but instead they advertise it, in order to protect users from infections while visiting malicious web pages. Our system is based on the same basic principles. We create active decoy accounts and try to tempt link spammers to reveal disclosed information, which otherwise would remain hidden.

X. Conclusion

In this paper we presented a large-scale study of the relationship networks that exist behind link spam. The key idea that motivated our data collection and analysis is that link spammers tend to generate relationships with each other, by performing link exchange, in order to unethically boost the ranking of their web pages. They usually utilize SEO forums to get in contact with other co-spammers. In essence, we wanted to expose these formed alliances and thus we systematically collected spam links from SEO forums, analyzed them, and validated the link exchange by crawling the respective web pages. Also, we enhanced a typical forum crawler with honey accounts, which collect data from private communications. In addition, we visualized the link spam network using a graph representation of the revealed link exchange and relationships found among spammers. The outcomes of our experiments indicate that there is a medium-sized but quite active community that seeks ways to unethically improve the ranking of its websites and therefore, it creates tens of new threads and posts every day, with ultimate goal to form alliances that will deceive search engines.

References

[1] S. Adali, T. Liu, and M. Magdon-Ismail. Optimal Link Bombs Are Uncoordinated. In *Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[2] S. Antonatos, I. Polakis, T. Petsas, and E. P. Markatos. A Systematic Characterization of IM Threats Using Honeypots. In *ISOC Network and Distributed System Security Symposium (NDSS)*, 2010.

[3] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. SpamRank–Fully Automatic Link Spam Detection. In *Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[4] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1), 1998.

[5] Z. Cheng, B. Gao, C. Sun, Y. Jiang, and T.-Y. Liu. Let Web Spammers Expose Themselves. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.

[6] D. Fetterly, M. Manasse, and M. Najork. Detecting Phrase-Level Duplication on the World Wide Web. In *Annual International Conference on Research and Development in Information Retrieval*, 2005.

[7] Google. Search Engine Optimization - Starter Guide. http://goo.gl/OmdtX, May 2013.

[8] Z. Gyongyi and H. Garcia-Molina. Web Spam Taxonomy. In *Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[9] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web Spam With TrustRank. In *International Conference on Very Large Data Bases (VLDB)*, 2004.

[10] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in Web Search Engines. *ACM SIGIR Forum*, 36(2):11–22, 2002.

[11] B. J. Jansen. Adversarial Information Retrieval Aspects of Sponsored Search. In *Adversarial Information Retrieval on the Web (AIRWeb)*, 2006.

[12] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing Your Enemy: Understanding and Detecting Malicious Web Advertising. In *ACM Conference on Computer and Communications Security (CCS)*, 2012.

[13] P. T. Metaxas and J. DeStefano. Web Spam, Propaganda and Trust. In *Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[14] G. Mishne, D. Carmel, and R. Lempel. Blocking Blog Spam With Language Model Disagreement. In *Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[15] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting Spam Web Pages Through Content Analysis. In *International Conference on World Wide Web (WWW)*, 2006.

[16] R. R. Sarukkai. How Much Is a Keyword Worth? In *International Conference on World Wide Web (WWW)*, 2005.

[17] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a Very Large Web Search Engine Query Log. *ACM SIGIR Forum*, 33(1):6–12, 1999.

[18] B. Stone-Gross, R. Stevens, A. Zarras, R. Kemmerer, C. Kruegel, and G. Vigna. Understanding Fraudulent Activities in Online Ad Exchanges. In *ACM SIGCOMM Conference on Internet Measurement (IMC)*, 2011.

[19] A. M. Wall. *Search Engine Optimization Book*. State College: Aaron Matthew Wall, 2005.

[20] B. Wu and B. D. Davison. Identifying Link Farm Spam Pages. In *International Conference on World Wide Web (WWW)*, 2005.

[21] B. Wu and B. D. Davison. Detecting Semantic Cloaking on the Web. In *International Conference on World Wide Web (WWW)*, 2006.

[22] M. Xie, Z. Wu, and H. Wang. HoneyIM: Fast Detection and Suppression of Instant Messaging Malware in Enterprise-Like Networks. In *Annual Computer Security Applications Conference (ACSAC)*, 2007.

[23] A. Zarras. The Art of False Alarms in the Game of Deception: Leveraging Fake Honeypots for Enhanced Security. In *International Carnahan Conference on Security Technology (ICCST)*, 2014.

[24] A. Zarras, A. Kapravelos, G. Stringhini, T. Holz, C. Kruegel, and G. Vigna. The Dark Alleys of Madison Avenue: Understanding Malicious Advertisements. In *ACM SIGCOMM Conference on Internet Measurement (IMC)*, 2014.

[25] H. Zhang, A. Goel, R. Govindan, K. Mason, and B. Van Roy. Making Eigenvector-Based Reputation Systems Robust to Collusion. In *International Workshop on Algorithms and Models for the Web-Graph*, 2004.