Fraudulent Support Telephone Number Identification Based on Co-occurrence Information on the Web

Xin Li, Yiqun Liu, Min Zhang, Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
x-l08@163.com, {yiqunliu,z-m,msp}@tsinghua.edu.cn

Abstract

"Fraudulent support phones" refers to the misleading telephone numbers placed on Web pages or other media that claim to provide services with which they are not associated. Most fraudulent support phone information is found on search engine result pages (SERPs), and such information substantially degrades the search engine user experience. In this paper, we propose an approach to identify fraudulent support telephone numbers on the Web based on the co-occurrence relations between telephone numbers that appear on SERPs. We start from a small set of seed official support phone numbers and seed fraudulent numbers. Then, we construct a co-occurrence graph according to the co-occurrence relationships of the telephone numbers that appear on Web pages. Additionally, we take the page layout information into consideration on the assumption that telephone numbers that appear in nearby page blocks should be regarded as more closely related. Finally, we develop a propagation algorithm to diffuse the trust scores of seed official support phone numbers and the distrust scores of the seed fraudulent numbers on the co-occurrence graph to detect additional fraudulent numbers. Experimental results based on over 1.5 million SERPs produced by a popular Chinese commercial search engine indicate that our approach outperforms TrustRank, Anti-TrustRank and Good-Bad Rank algorithms by achieving an AUC value of over 0.90.

Introduction

The Web has become a major source of information around the world. Therefore, when individuals encounter problems with purchased products, many will search for product support phone information using search engines. However, because a credible editorial process is lacking for many Web sites, swindlers have designed a large number of Web pages that provide fraudulent support telephone numbers in order to obtain personal or financial information for illegal gain. The fraudulent support phone problem has had a substantial negative impact on society [1] and is particularly serious in China. In 2008, individuals in Beijing, Shanghai, Guangdong and Fujian lost more than 600 million CNY by calling fraudulent support telephone numbers. [2] In addition, the loss of trust has resulted in revenue lost for commercial search engines. Several popular Chinese search engines have become involved in scandals due to the existence of fraudulent support phone information on their SERPs. [3]

Despite the severe consequences of this problem, few techniques exist to help commercial search engines identify and warn the users of fraudulent support phone numbers. Most technologies to detect fraudulent telephone numbers focus on detecting voice phishing (or vishing) numbers (Wang et al. 2008; Salem, Hossain, and Kamala 2010). In voice phishing, swindles call users while posing as government officials or customer clients to gain access to the users' personal information. This activity differs from the fraudulent support phone deception because in the latter scenario users dial the misleading numbers to obtain support information on or service for their purchases. Therefore, it is difficult to adopt voice phishing detection techniques, such as voice coding features (Chang and Lee 2010), to identify fraudulent support phones.

Fraudulent support phone detection also differs from general purpose Web spam page detection (Fetterly, Manasse, and Najork 2004; Castillo et al. 2007), which aims to identify activities that try to gain "an unjustifiably favorable relevance or importance for some web page, considering the page's true value" (Gyongyi and Garcia-Molina 2005). Occasionally, fraudulent support phones are located on spam pages. In many other cases, they appear in ordinary pages, particularly Web 2.0 resources (see Figure 1 for examples in which the same fraudulent support telephone number appears in two spam pages and an ordinary Web 2.0 page. Although the fraudulent support number is the same, the pages claim that they provide "official" repair services for a stove, a water heater and an air-conditioner, respectively). Therefore, the detection of spam pages cannot fully solve the problem of fraudulent support phones.

As a result of the lack of technical solutions for fraudulent support phone detection, most commercial search engines turn to business partners to decrease the risk of providing

[1] http://abclocal.go.com/ktrk/story?section=news/consumer&id=9173227

[2] http://www.315chahao.com/weiquanchangshi/686.html

[3] http://www.techinasia.com/cctv-accuses-baidu-of-allowing-fraudulent-websites-in-promoted-links/

Figure 1: Examples of several fraudulent support telephone numbers and their corresponding Web pages. The first two examples are extracted from two spam pages within the same domain, which describe a stove brand and a water heater brand, respectively. The highlighted sections are the declared official phone numbers for each brand, which are the same. The last example is extracted from an ordinary Web 2.0 page located on Ganji.com, which contains the same fraudulent support phone number.

fraudulent support phone information. Therefore, nearly all Chinese search engines encourage reputable support service providers to register their telephone information with the search engine so that this "official" information could be placed at the top of a ranking list. However, the number of products and service providers is so large that it is difficult to compel all of the providers to approach the search engines to register. Additionally, the various forms of user query make it difficult to identify all of the queries with similar intents and connect them with the official information. Thus, it is important to identify fraudulent support telephone numbers and warn search engine users regarding the false information provided on Web pages. Based on the assumption that neither official service providers nor swindlers will place fraudulent support phone numbers together with official numbers on the same page or page block, we propose a framework to detect fraudulent support phones based on their distance from seed official telephone numbers and seed misleading telephone numbers on a page-level/page-block-level co-occurrence graph.

The main contributions of this paper can be summarized as follows:

- To our best knowledge, this is the first attempt to identify fraudulent support telephone numbers, which cause substantial harm to society and in particular to search engine companies and users.
- An approach to identify fraudulent support telephone numbers is proposed based on the co-occurrence relations between telephone numbers on Web pages or page blocks.
- An evaluation dataset is constructed that contains millions of SERPs and a large number of annotated fraudulent/official support telephone instances. [4]

The remainder of this paper is organized as follows: after a discussion of the related work in the next section, we introduce the fraudulent support telephone number identification algorithm in the third section. The fourth section presents an evaluation and discussion of our approach. Finally, the last section concludes the paper.

Related Work

Most approaches to the detection of misleading telephone numbers proposed in the literature focus on the interaction between the swindlers and their victims, e.g., voice phishing. (Maggi 2010) observes that a good share of scammers rely on automated responders to streamline the voice phishing campaigns. This work analyzes the content of the conversation and finds certain recurring, popular terms such as "credit", "press" (a key) and "account". (Maggi, Sisto, and Zanero 2011) develop a data collection system to capture different aspects of phishing campaigns, with a particular focus on the emerging use of the voice channel. Their system analyzes instant messages and suspicious emails and extracts telephone numbers, URLs and popular words from the content, which are correlated using cross-channel relationships between messages to recognize campaigns. However, it is often too late to detect a deceptive telephone call after the call has been answered. If we can recognize fraudulent telephone numbers in advance, we will be able to eliminate them before users call them spontaneously.

Our research focuses on detecting fraudulent support telephone numbers on search engine result pages, which also makes this research relevant to the task of Web spam detection. Many spam detection approaches have been proposed to combat all types of spam page using the information extracted from the content (Ntoulas et al. 2006; Abernethy, Chapelle, and Castillo 2008), user behavior (Liu et al. 2008a; 2008b) and hyperlink structure (Becchetti et al. 2006; Benczúr, Csalogány, and Sarlós 2006). TrustRank and Anti-TrustRank are among the most popular and effective solutions to Web spam page detection on hyperlink graphs. (Gyöngyi, Garcia-Molina, and Pedersen 2004) proposed the TrustRank algorithm, which can semi-automatically separate reputable, reliable pages from spam. The technique starts by selecting a small set of reputable seed pages and discovers other pages that are likely to be reliable using the link structure of the Web. Their results indicate that based on a seed set of less than 200 reliable sites, spam from a significant fraction of the Web can be effectively filtered out. (Krishnan and Raj 2006) introduced the Anti-TrustRank algorithm. Similar to TrustRank, their method selects a manually labeled seed set of pages, and their experiments on the WebGraph dataset show that their approach is effective at detecting spam pages from a small seed set. (Liu et al.

[4] http://www.thuir.cn/group/%7eYQLiu/publications/aaai2014.7z

2013) combine trust and distrust propagations. They propose the Good-Bad Rank algorithm, which propagates trust and distrust simultaneously from both directions. Experimental results demonstrate that their algorithm outperforms the traditional link-based anti-spam algorithms that diffuse only trust or distrust.

However, there are many differences between fraudulent support telephone number identification and Web spam detection. 1. Whereas a large number of Web pages that contain misleading telephone numbers are also spam pages, only a small percentage of spam pages contain misleading telephone numbers. 2. The link relations between Web pages are directional, whereas the co-occurrence relations between telephone numbers on Web pages are not. 3. Telephone numbers that appear on the same Web page with misleading telephone numbers are more likely to be deceptive because swindlers tend to display a large number of misleading telephone numbers on their Web pages. However, Web pages linked by spam pages are not necessarily fraudulent because spam pages often link to a large number of reputable pages in an attempt to improve their PageRanks (Gyongyi and Garcia-Molina 2005). Consequently, Web spam detection approaches cannot be directly applied to our task of identifying fraudulent support telephone numbers. Later in this paper, we compare the performance of our approaches with the TrustRank, Anti-TrustRank and Good-Bad Rank algorithms.

Fraudulent Support Telephone Identification

In this section, we introduce the framework of our approach. Its flowchart is shown in Figure 2.

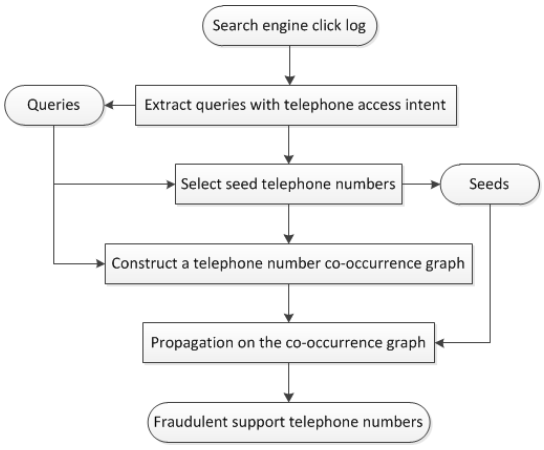

Figure 2: Flowchart of our approach

Extracting queries with telephone access intent

Before applying our methods of identifying fraudulent support telephone numbers, we must obtain as many telephone-number queries as possible so that we can extract telephone numbers from SERPs that are possibly exposed to search users. We use the click logs from a popular Chinese commercial search engine in May 2012 and extract all of the queries that contain the keyword "telephone" from them. We reserve all of the unique queries that were searched for more than 30 times in a month by search engine users. However, not all queries that contain the keyword "telephone" involve the intention to locate a telephone number. Additionally, not all of the queries for telephone numbers explicitly contain the keyword "telephone". For example, users who submit the query "internet telephone" typically are not searching for a certain telephone number. Additionally, the query "how to contact Facebook" will lead to search results that contain Facebook support phone services. To address the first problem, we submit all of the queries to the search engine and only reserve the queries that have telephone numbers in the search results because queries without the intent to locate telephone numbers usually do not produce search results that contain telephone numbers. To address the second challenge, we conduct random walks on the click-through bipartite. However, because the click-through bipartite is highly connected (Cao et al. 2008), random walks may result in irrelevant queries: there may exist paths between two completely unrelated queries or URLs. Therefore, we limit random walks to two steps. Specifically, let the set of original queries be Q. First, we obtain the URL set U, where each URL $u \in U$ is connected to at least one query $q \in Q$ in the click-through bipartite. Then, we expand the set Q into $Q^{'}$, where each query $q^{'} \in Q^{'}$ is connected to at least one URL $u \in U$ in the click-through bipartite. In this way, we extract 157,592 queries intended to locate telephone numbers in total. We use these queries to extract the seed telephone numbers and to construct the co-occurrence graph.

Selecting seed telephone numbers

We select two sets of seed telephone numbers: a set of seed official support telephone numbers and a set of seed fraudulent numbers. The selection of seed official support numbers is based on the fact that certain reputable support service providers have already registered their official telephone information with search engines, as described in the introduction. Thus, if a user searches for the official support telephone number for a product or service that is officially certified by the search engine, the number will appear in the first position in the search result list with an official support telephone number symbol. Therefore, we submit the extracted queries with telephone access intent to the search engine. We reserve the queries with an official support telephone symbol in the result list and extract the corresponding telephone numbers from the first search result. We cluster these queries according to the corresponding telephone numbers. Each cluster represents an official support telephone number with related queries, which constitutes the seed set of official support telephone numbers. In total, we obtain 3,556 telephone numbers in the seed official set.

We adopt two methods to select seed fraudulent telephone numbers. The first method is based on the assumption that a real support telephone number should be related to only one brand of product or service. For example, in Figure 1, the same telephone number appears as the

official support telephone number on the Web pages of two completely irrelevant products, which is obviously deceptive. As described above, one cluster of an official support telephone number and several related queries corresponds to one product or service. Therefore, we submit the reserved queries with the official support telephone number symbol in the search result to the search engine and extract the telephone numbers in the snippets (the few lines of text that appear under every search result). If a telephone number appears in the snippets of queries that belong to two or more clusters, it will be detected as a fraudulent telephone number and added into the seed set. The second method is based on crowdsourcing annotations from the public. Several Web sites in China, such as 00cha[5] and chahaoba[6] provide a platform for individuals to report fraudulent support telephone numbers. When someone has been victimized by a fraudulent telephone call, they may report the telephone number to these Web sites so that others may be made aware of it and exert caution. Thus, we extract telephone numbers from the logs of these Web sites and count the number of times they have been reported. To avoid malicious reporting, we only reserve the telephone numbers that are reported by at least 3 users. Finally, we combine the telephone numbers detected using the two methods to form the seed set of fraudulent telephone numbers, of which are 563 and 166, respectively. To verify credibility of the constructed fraudulent seed set, we randomly sample 100 numbers for manual labeling. We label the numbers according to their descriptions on the Web pages or by dialing the number and making a judgment based on through the conversation. As a result, 96 of the 100 telephone numbers were fraudulent, which indicates the high precision of our method of detecting seed fraudulent numbers.

Constructing a telephone number co-occurrence graph

On the Internet, a Web page may contain several telephone numbers, and a telephone number may appear on several Web pages. If we take telephone numbers as vertices and the co-occurrence relations between telephone numbers as edges, we can construct a co-occurrence graph of telephone numbers. We submit the extracted search-engine queries intended to locate telephone numbers and extract the telephone numbers and their co-occurrence relations from the top 10 result pages of each query. In this way, we construct a co-occurrence graph with the extracted vertices and edges, which covers all of the seed official support telephone numbers and the seed fraudulent numbers.

However, not all of the telephone numbers on the same Web page are closely related. For example, on certain portal Web sites, customer service hotlines or complaint hotlines may be listed at the bottom of each Web page that are mostly irrelevant to the content located in the main body of the page. Therefore, it is unreasonable to add edges between the hotlines and the telephone numbers that appear in the main body of the Web page. Thus, it is necessary to divide a Web page into different regions for different content. We use the **VI**sion-based **P**age **S**egmentation (VIPS) algorithm to segment a Web page into semantically related content blocks based on its visual presentation, as proposed in (Cai et al. 2003). VIPS has been demonstrated to improve web search, link analysis and pseudo-relevance feedback in Web information retrieval in a number of studies (Yu et al. 2003; Cai et al. 2004a; 2004b) that use page layout features to partition the page at the semantic level. Each node in the extracted content structure corresponds to a block of coherent content in the original page. After segmenting a Web page into different blocks, we regard each block as an individual page. We extract the telephone numbers and their co-occurrence relations from each block separately and reconstruct the co-occurrence graph of the numbers, which can effectively separate the telephone numbers on the main body of the Web page from the hotlines in other blocks.

Propagation on the co-occurrence graph

The propagation on the co-occurrence graph of phone numbers is based on the following assumptions:

Assumption 1 (Co-occurrence assumption) *Telephone numbers co-occurring on pages containing official support numbers tend to be non-fraudulent; meanwhile, telephone numbers co-occurring on pages containing fraudulent numbers tend to be fraudulent as well.*

In most cases, official support telephone numbers appear on normal Web pages. The other telephone numbers on these Web pages are likely to be normal as well. However, the Web pages that contain fraudulent telephone numbers are mostly not reputable pages, and there is a greater chance that the other telephone numbers that appear on these Web pages are also fraudulent. Most of the time, official support telephone numbers and fraudulent ones will not appear on a same Web page.

Assumption 2 (Nearby assumption) *A shorter distance between a telephone number and fraudulent telephone numbers on the co-occurrence graph indicates a larger likelihood that a number is fraudulent.*

Based on these assumptions, we design the propagation algorithm as follows. First, we assign the seed official support telephone numbers with a score of 1 and seed fraudulent numbers with -1. Then, we diffuse the scores of the seed official support telephone numbers and the seed fraudulent numbers on the co-occurrence graph with Breadth First Search (BFS). Each time a telephone number is diffused, we assign it the score of its parent vertex multiplied by a certain decay factor. We limit the propagation to a certain number of steps. After the propagation, each telephone number is assigned two scores, one diffused from the seed official support telephone numbers and the other from the seed fraudulent numbers. We add the numbers to obtain the final score. Algorithm 1 describes the propagation approach.

It can be observed from the description of the algorithm that our approach differs from general-purpose Web spam detection methods. We diffuse the trust scores from the seed

[5]http://www.00cha.com/

[6]http://www.chahaoba.com/

Algorithm 1 Propagation on the co-occurrence graph

Require:

The set of seed official support telephone numbers, O;
The set of seed fraudulent telephone numbers, F;
The co-occurrence graph of telephone numbers, $G = (V, E)$;
The decay factors, β_o, β_f;
The propagation step threshold, θ_o, θ_f;

1: **for** each v in V **do**
2: $score(v) = 0$
3: **end for**
4: Construct graph $G_o = (V_o, E_o)$
5: $V_o = V \cup s_o, E_o = E \cup \{(s_o, v)|v \in O\}$
6: Construct graph $G_f = (V_f, E_f)$
7: $V_f = V \cup s_f, E_f = E \cup \{(s_f, v)|v \in F\}$
8: Perform Breadth First Search in G_o from s_o and in G_f from s_f, respectively
9: **for** each v in V_o **do**
10: **if** $depth^o(v) \leq \theta_o$ **then**
11: $score(v) = score(v) + \beta_o^{depth^o(v)-1}$
12: **end if**
13: **end for**
14: **for** each v in V_f **do**
15: **if** $depth^f(v) \leq \theta_f$ **then**
16: $score(v) = score(v) - \beta_f^{depth^f(v)-1}$
17: **end if**
18: **end for**

official support telephone numbers and the seed fraudulent numbers simultaneously. Additionally, Web spam detection methods make full use of the link relations between Web pages in case a spam page links to reputable pages in an attempt to improve its PageRank. In our approach, a telephone number's score is only determined by its parent vertex, which is closest to the seed set. This design is due to the fact that non-fraudulent and fraudulent telephone numbers generally do not co-occur on the same Web page and the shortest distance of a telephone number from the seed set indicates the degree to which it is normal or fraudulent.

Experimental Results

In this section, we report the experimental results of our approach. We present the score distribution and evaluate the precision and recall of the detected fraudulent telephone numbers using our propagation algorithm. We compare our approach with the TrustRank, Anti-TrustRank and Good-Bad Rank algorithms on ROC curves and with respect to AUC value.

Dataset

We use the click logs for one month from a popular commercial search engine, which contain approximately 80 million click records per day. As described in previous sections, we extract 157,592 queries intended to locate telephone numbers in total. We obtain 3,556 telephone numbers in the seed official set and 729 telephone numbers in the seed fraudulent set. We construct two co-occurrence graphs using the telephone numbers and their co-occurrence relations on SERPs, one with VIPS, the other without VIPS. Both graphs contain 2,245,963 vertices because they share the same set of telephone numbers. The graph constructed without VIPS contains 6,248,750 edges, whereas the VIPS graph contains 5,454,758 edges because the VIPS algorithm will remove the edges between the telephone numbers in different blocks within a same Web page. Next, we compare the performance of our propagation algorithm on these two co-occurrence graphs.

Propagation algorithm performance

We apply the propagation algorithm on the two telephone number co-occurrence graphs using the following parameters: $\beta_o = 0.85, \beta_f = 0.9, \theta_o = \theta_f = 20$. (Multiple trial iterations indicate that these parameters obtain the best results, which we believe to be reasonable because the co-occurrence relations between the fraudulent telephone numbers are stronger than those between the official support telephone numbers, and a telephone number that is diffused more than 20 steps from the seed sets has little significance for the final score.)

With the algorithm, 1,879,871 telephone numbers are diffused from the seed set, which represents 83.7% of all of the telephone numbers on the co-occurrence graph. Because the score diffused from seed official telephone numbers $\in (0, 1]$ and that from seed fraudulent telephone numbers $\in [-1, 0)$, the final score of each telephone number is between -1 and 1. We segment the score range into 10 buckets and count the number of telephone numbers in each bucket. We compare the ratios of telephone numbers in different buckets with propagation on the two co-occurrence graphs. As described by Algorithm 1, a lower score indicates a higher likelihood that a telephone number is fraudulent. The percentages of telephone numbers in the first bucket with and without VIPS are 9.5% and 10.0%, respectively. In practical application, we tend to believe that the telephone numbers in the first bucket are fraudulent because they obtain significantly lower scores. To evaluate our approach to the detection of fraudulent telephone numbers, we sample 100 numbers in the first bucket (excluding seed fraudulent telephone numbers) using the two methods for manually labeling. We label the telephone numbers with 3 levels after placing calls, as described below.

- *Fraud*: fraudulent support telephone numbers, where a recording is played or the caller is misled. For example, the person who receives the call asks the caller to pay a bill fraudulently.
- *Possible fraud*: suspected fraudulent support telephone numbers, where the business description provided during the telephone call does not agree with the Web page that contains the number. For example, the business description on the Web page claims to be the manufacturer's maintenance number. However, during the phone call, the business is described as a third-party maintenance business.

- *Non-fraudulent*: normal telephone numbers.

The percentage of the three labels applied using the two methods are shown in Table 1.

Table 1: Labeling results

	Propagation	Propagation + VIPS
Fraudulent	74%	78%
Possibly fraudulent	11%	12%
Non-fraudulent	15%	10%

As shown in the table, while VIPS decreases the recall of the detection of telephone numbers from 10.0% to 9.5%, it increases the precision of the numbers in the first bucket.

Comparison with Web spam detection methods

We compare the performance of our approach with the traditional Web spam detection methods TrustRank and Anti-TrustRank. TrustRank and Anti-TrustRank share the same idea of trust/anti-trust propagation. TrustRank starts from the seed set of reputable pages, whereas Anti-TrustRank starts from the seed set of spam pages. Both algorithms compute the Trust/Anti-Trust scores recursively using the following equation:

$$\mathbf{t}^* = \alpha_B \cdot \mathbf{T} \cdot \mathbf{t}^* + (1 - \alpha_B) \cdot \mathbf{d}$$

where $\mathbf{t}^*$ is the vector of TrustRank/Anti-TrustRank scores, α_B is the decay factor for biased PageRank, $\mathbf{T}$ is the transition matrix and $\mathbf{d}$ is the normalized static score distribution vector whereby the scores of seeds are 1 and those of others are 0.

We obtain the seed sets of official support telephone numbers and fraudulent numbers and construct the co-occurrence graph of numbers to take advantage of the main idea of TrustRank/Anti-TrustRank and detect the fraudulent numbers. However, a problem that occurs is that the link relations between Web pages are directional, whereas the co-occurrence relations between telephone numbers on Web pages are not. Therefore, we regard each edge on the co-occurrence graph of telephone numbers as bidirectional. We also compare the performance of the Good-Bad Rank algorithm proposed in (Liu et al. 2013) by subtracting the BadRank score from the GoodRank score. We rank the results of each algorithm in descending order of degree of fraudulence. To compare the performance of our approaches with the Web spam detection methods, we randomly sample 200 telephone numbers from the results of each algorithm and apply manual labeling. The label strategy is as described previously: we regard the first two labels as indicating fraudulent numbers and the last label as indicating a normal number. We compute the true positive rate and the false positive rate and compare the ROC curves and the AUC (Area Under Curve) values. The results are shown in Figure 3 and Table 2.

From the results, we can see that propagation from both sides performs better than propagation from only one side since Good-Bad Rank algorithm achieves a higher AUC value than TrustRank and Anti-TrustRank. Our approach outperforms the traditional Web spam detection methods.

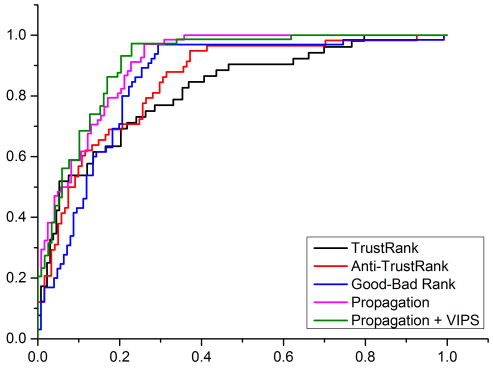

Figure 3: Comparison of ROC curves of fraudulent support telephone number identification algorithms

Table 2: Comparison of AUC values of fraudulent support telephone number identification algorithms

Method	AUC value	Improvement compared with TrustRank
TrustRank	0.8200	–
Anti-TrustRank	0.8485	3.5%
Good-Bad Rank	0.8508	3.8%
Propagation	0.9067	10.6%
Propagation + VIPS	**0.9111**	**11.1%**

Both algorithms with and without VIPS improve the AUC value by more than 10% compared with TrustRank. Additionally, after segmenting Web pages into blocks using VIPS, the propagation algorithm obtains a higher AUC value because VIPS separates unrelated telephone numbers that appear on the same Web page, and the detection of fraudulent numbers gains precision.

Conclusion

Previous studies demonstrate that fraudulent support telephone numbers on the Web not only degrade search engine user experience but also cause substantial harm to society. In this paper, we propose an approach to automatically identify fraudulent support telephone numbers. We start by extracting telephone-number-related queries and selecting a small set of seed official support telephone numbers and seed fraudulent numbers. Then, we construct a co-occurrence graph according to the co-occurrence relations between the telephone numbers on search engine result pages or page blocks. Finally, we diffuse the trust/anti-trust scores of the seed numbers on a co-occurrence graph to detect additional fraudulent telephone numbers. Experimental results demonstrate that our approach can detect 9.5% of all of the telephone numbers on the co-occurrence graph as fraudulent or possibly fraudulent with 90% precision. Our propagation algorithms outperform the traditional Web spam detection methods with an AUC value of up to 0.91. In addition, the segmentation of Web pages into blocks contributes to improved precision in the identification of fraudulent support telephone numbers.

Acknowledgments

This work was supported by Natural Science Foundation (61073071). Part of the work has been done at the Tsinghua-NUS NExT Search Centre, which is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490). We also benefit a lot from discussion with Rongbo Luan, Jun Ma, Tangdong Li, Miaomiao Li and Yanjun He from Baidu inc. at the early stage of work.

References

Abernethy, J.; Chapelle, O.; and Castillo, C. 2008. Web spam identification through content and hyperlinks. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, 41–44. ACM.

Becchetti, L.; Castillo, C.; Donato, D.; Leonardi, S.; and Baeza-Yates, R. A. 2006. Link-based characterization and detection of web spam. In *AIRWeb*, 1–8.

Benczúr, A. A.; Csalogány, K.; and Sarlós, T. 2006. Link-based similarity search to fight web spam. In *In AIRWEB*. Citeseer.

Cai, D.; Yu, S.; Wen, J.-R.; and Ma, W.-Y. 2003. Extracting content structure for web pages based on visual representation. In *Web Technologies and Applications*. Springer. 406–417.

Cai, D.; He, X.; Wen, J.-R.; and Ma, W.-Y. 2004a. Block-level link analysis. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 440–447. ACM.

Cai, D.; Yu, S.; Wen, J.-R.; and Ma, W.-Y. 2004b. Block-based web search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 456–463. ACM.

Cao, H.; Jiang, D.; Pei, J.; He, Q.; Liao, Z.; Chen, E.; and Li, H. 2008. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 875–883. ACM.

Castillo, C.; Donato, D.; Gionis, A.; Murdock, V.; and Silvestri, F. 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*, 423–430. ACM.

Chang, J.-H., and Lee, K.-H. 2010. Voice phishing detection technique based on minimum classification error method incorporating codec parameters. *Signal Processing, IET* 4(5):502–509.

Fetterly, D.; Manasse, M.; and Najork, M. 2004. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*, 1–6. ACM.

Gyongyi, Z., and Garcia-Molina, H. 2005. Web spam taxonomy. In *First international workshop on adversarial information retrieval on the web (AIRWeb 2005)*.

Gyöngyi, Z.; Garcia-Molina, H.; and Pedersen, J. 2004. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 576–587. VLDB Endowment.

Krishnan, V., and Raj, R. 2006. Web spam detection with anti-trust rank. In *AIRWeb*, volume 6, 37–40.

Liu, Y.; Cen, R.; Zhang, M.; Ma, S.; and Ru, L. 2008a. Identifying web spam with user behavior analysis. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, 9–16. ACM.

Liu, Y.; Zhang, M.; Ma, S.; and Ru, L. 2008b. User behavior oriented web spam detection. In *Proceedings of the 17th international conference on World Wide Web*, 1039–1040. ACM.

Liu, X.; Wang, Y.; Zhu, S.; and Lin, H. 2013. Combating web spam through trust-distrust propagation with confidence. *Pattern Recognition Letters*.

Maggi, F.; Sisto, A.; and Zanero, S. 2011. A social-engineering-centric data collection initiative to study phishing. In *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, 107–108. ACM.

Maggi, F. 2010. Are the con artists back? a preliminary analysis of modern phone frauds. In *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, 824–831. IEEE.

Ntoulas, A.; Najork, M.; Manasse, M.; and Fetterly, D. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, 83–92. ACM.

Salem, O.; Hossain, A.; and Kamala, M. 2010. Awareness program and ai based tool to reduce risk of phishing attacks. In *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, 1418–1423. IEEE.

Wang, X.; Zhang, R.; Yang, X.; Jiang, X.; and Wijesekera, D. 2008. Voice pharming attack and the trust of voip. In *Proceedings of the 4th international conference on Security and privacy in communication netowrks*, 24. ACM.

Yu, S.; Cai, D.; Wen, J.-R.; and Ma, W.-Y. 2003. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the 12th international conference on World Wide Web*, 11–18. ACM.