A Taxonomy of Hyperlink Hiding Techniques

Guang-Gang Geng1, Yan-Ming Zhang2, Xiao-Dong Lee1, Wei Wang1, An-Lei Hu1
1China Internet Network Information Center
2Computer Network Information Center, Chinese Academy of Sciences
3National Laboratory of Pattern Recognition
4Institute of Automation, Chinese Academy of Sciences
Beijing 100190, P. R. China
{gengguanggang,lee,wangwei,huanlei}@cnnic.cn;ymzhang@nlpr.ia.ac.cn

ABSTRACT

Hidden links are designed solely for search engines rather than visitors. To get high search engine rankings, link hiding techniques are usually used for the profitability of black industries, such as illicit game servers, false medical services, illegal gambling, and less attractive high-profit industry, etc. This paper investigates hyperlink hiding techniques on the Web, and gives a detailed taxonomy. We believe the taxonomy can help develop appropriate countermeasures. Study on 5,583,451 Chinese sites’ home pages indicate that link hidden techniques are very prevalent on the Web. We also tried to explore the attitude of Google towards link hiding spam by analyzing the PageRank values of relative links. The results show that more should be done to punish the hidden link spam.

Categories and Subject Descriptors
I.2.6 [Artificial Intelligence]: Learning; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Measurement, Experimentation, Algorithms

Keywords
Web Spam, Link Hiding, Hidden Spam, Spam Detection

1. INTRODUCTION

Most Web surfers depend on search engines to locate information on the Web. Link analysis algorithms [12], such as PageRank [13] and HITS [10], are usually used for search engines ranking. Link analysis algorithms assume that every link represents a vote of support, in the sense that if there is a link from page x to page y and these two pages are authored by different people, then the author of page x is recommending page y. In particular, PageRank is the basis of Google’s search technology [4].

Web spammers try to mislead search engines to make a high rank in search results [8]. In this context, hyperlink hiding techniques are often used to deceive search engines. Spammers hope that many small endorsements from these pages with hidden links result in a sizable PageRank for the target page. Several questions naturally arise: what link hidden techniques are the spammers using; how prevalent are hidden spam links on the Web; and are the hidden pages punished by the search engines? This paper attempts to answer those questions.

The rest of sections are organized as follows. Section 2 presents a literature review. Section 3 gives a comprehensive taxonomy of current hidden link spam techniques. Section 4 describes the experimental analysis on 5,583,451 Chinese Web sites. At last, section 5 draws the conclusion.

2. RELATED WORK

Hidden links are designed to increase link popularity, which are invisible for visitors [15]. Google considers hyperlinks hidden by small characters as deception [7]. Z. Gyongyi et al. point out that hidden links are often used in honey pot to boost the ranking of the spam pages [9]. They further present a comprehensive taxonomy of current spamming techniques and survey content hiding techniques, where spam links hidden by avoiding anchor texts or tiny anchor images are mentioned [8].

To the best of our knowledge, there is no previously published literature that directly studied how prevalent, successful, or varied hidden link spam techniques are on the Web. This paper attempts to study hidden link spam in detail. It is hoped that the findings can help in developing appropriate countermeasures.

3. HYPERLINK HIDING TECHNIQUES

There are many different ways to hide links from visitors while leaving it perfectly viewable to search engines. In this section, we will examine current hyperlink hiding techniques used by spammers and attempt to categorize them based on their features. Just as the work on JavaScript redirection spam [5], we present short examples to show the hidden techniques really used by spammers. Simple techniques are presented first and are followed by more advanced ones.

3.1 A: Making Anchor Text Font Color the Same as Background Color

The simplest and oldest method that spammers use to create hidden links is to make the font of anchor text the same color as the background. Here is one example.

```html
&lt;span style="background:white;" &gt;
  &lt;a href="target.html" style="color:white">anchor text&lt;/a&gt;
&lt;/span&gt;
```

In this example, the color scheme is defined in the HTML
document. Color schemes can also be defined in an attached cascading style sheet file (CSS). Sometimes, spammers also consider background images. They set the image color to be the same as the font color, which is relatively harder to detect.

### 3.2 B: Making Anchor Text Font Color Almost Match Background Color or Background Image

Instead of setting the font color to entirely match the background color, some spammers and web masters set their font colors to almost match the background color. The idea behind this method is that they believe that they are thwarting the search engines’ software detection systems by slightly changing the color of the text.

```html
<div style="background-color:white;">
  <a href="target.html" style="color:#FEEFEE">anchor text with similar color with background</a>
</div>
```

### 3.3 C: Setting Tiny Anchor Text or Placing the hyperlinks in a Tiny Block

Making tiny anchor text is another hyperlink hiding method. This way, the hyperlink can be set small enough, such as 1 pixel high, even 0 pixel. Here’s a simple example of that.

```html
<a href="target.html" style="font-size:0px">keyword</a>
```

In HTML, the div element is often used for generic organizational or stylistic applications. Spammers can also use div to set the link size. The following is another example.

```html
<div style="font-size:0px;">
  <a href="target.html">invisible anchor text</a>
</div>
```

Perhaps the most common use of div element is to carry class or id attributes in conjunction with CSS to apply layout, typographic, color, and other presentation attributes to parts of the content. In the previous example, the `font-size:0px` can also be defined in a CSS file. Besides, div block size can be set via width and height attributes. For example, `div style="width:1px;height:1px;"`, where the div size is 1 pixel.

Another example of hiding a hyperlink via tiny scrolling block is presented below.

```html
<marquee scrollAmount=1 width=1 height=1>
  <a href="target.html">keywords</a>
</marquee>
```

In this example, `target.html` is put in a scrolling block with area $1 \times 1$ pixel, which is invisible to Web users.

### 3.4 D: Disguising Anchor Text as Plain Text

Sometimes, spammers insert hyperlinks into a paragraph, where the anchor text looks like plain text. Here’s a paragraph of text on a site:

```
The SEO company follows strict rules to insure the clients website reach the top of search engines and stay there.
```

A user wouldn’t see any hyperlinks, even if they moused over every word in the paragraph. But if you happened to click on just the right word, you’d get whisked away to a SEO site. Actually, there is a hidden link under the anchor text “SEO company”. If you view the source of the page, here’s what you’ll see:

```
The <a href="http://www.seomarketleaders.com" onMouseOver="window.status='';return true;" style="cursor:text;color:black;text-decoration:none;">SEO company</a> follows strict rules to insure the clients website reach the top of search engines and stay there.
```

Using a similar method, a link can be hidden in a small character – for example, a hyphen in the middle of a paragraph.

### 3.5 E: Placing Hyperlinks in High-Speed Scrolling Blocks

The `<marquee>` tag is a non-standard HTML element which causes text to scroll up, down, left or right automatically [14]. Although the W3C advises against its use in HTML documents, it’s still widely used. `SCROLLAMOUNT` attribute sets the speed of the scrolling. A bigger value for `SCROLLAMOUNT` makes the marquee scroll faster. If the `SCROLLAMOUNT` value is big enough, the scrolling block will be invisible to the naked eye. Here is a simple example.

```
<marquee height=1 width=8 scrollamount=3000>
  <a href="target.html">keywords</a>
</marquee>
```

The default `SCROLLAMOUNT` value is 6. The value in the example is 3000, which is too fast to see. Similar effects can also be achieved through the use of JavaScript or HTML `blink` element [14] [1].

### 3.6 F: Putting Links outside the Screen

Using cascading style sheets, you have the option to absolutely or relatively position any division. Using absolute position, you can simply position the text you wish to hide any number of pixels off the screen to the left of the window. Here are some example codes:

If you put that in your style sheet and then assign the class “hiddenclass” to your div, then the div will display 977 pixels to the left of the visible screen - i.e., it will not appear on the screen. Here is an example:

The absolute position can also be set in the div directly as follows:
The example is easy to understand, which is a packaging of the method described in section 3.7. In the above codes, the class "hiddenclass" to your div, the hyperlinks in the div block will not appear in the browser window.

3.8 H: Hiding Hyperlinks via JavaScript

JavaScript is an open source programming language commonly implemented as part of a web browser in order to create enhanced user interfaces and dynamic websites [6]. Google claims that search engines have difficulty accessing JavaScript [7]. In 2011, labnol.org reported that Google indexes JavaScript based Facebook comments, but there is no clear report that Google parsers JavaScript codes on the whole Web. This fact encourages spammers to hide hyperlinks by the aid of JavaScript. Here is a simple example:

```
<script language="javascript" type="text/javascript">
    var r='';
    for(var i=0;i<s.length;i+=2){
        var sxx=parseInt(s.substring(i,i+2),16);
        r+=String.fromCharCode(sxx);
    }
    return r;
</script>
```

The above JavaScript codes are designed in rather vague terms. The elements of array `r` are written with ASCII characters. The last line of the above JavaScript codes is `document.getElementById('q1000').style.display="none"`, which makes all the content, including hyperlinks, in the div named q1000 invisible. In order to avoid presenting the whole style assignment directly, a script can build up the style assignment via string concatenation. One very straightforward example is presented below.

```
<script type="text/javascript">
    document.getElementById('q1000').style.display="n" + "o" + "ne";
</script>
```

What is worse, JavaScript as a programming language, has many functions and operators, which throw off a human readers. The following codes show the flexibility of JavaScript.

```
<script language="javascript">function HexTostring(s){
    var r='';
    for(var i=0;i<s.length;i+=2){
        var sxx=parseInt(s.substring(i,i+2),16);
        r+=String.fromCharCode(sxx);
    }
return r;
}
</script>
```

These codes are essentially equivalent to the previous example, yet look completely different.
3.9 I: Hiding Hyperlinks via Cloaking or Redirection Techniques

Cloaking is a Web spam technique in which the page presented to the search engine spider is different from that presented to the user’s browser [16]. Some spammers hide target hyperlinks using cloaking technique. Similarly, spammers also use redirection techniques to hide targeting hyperlinks. Among the redirection spam techniques, JavaScript based redirection is the most notorious and difficult to catch [5]. Wu et al. [16] and Chellapilla et al. [5] have conducted comprehensive studies of cloaking and redirection techniques respectively, so the techniques will not be repeated here. However, it’s important to point out that we do not consider the redirected target URL, but the hyperlinks in the redirection page as hidden links. For example, A redirects to B, and C is a hyperlink in page A. In this paper, C is a hidden link, but B is not seen as a hidden link.

3.10 J: Hiding Hyperlinks in Pull-Down Menu

Pull-down menu is also called a drop-down menu, which is a menu of commands or options that appears when you select an item with a mouse. A drop down menu can make it easier to display a large list of choices - since only one choice is displayed initially, the remaining choices can be displayed when the user activates the dropdown. Some spammers insert target hyperlinks into a long pull-down list, which are hard to find.

3.11 K: Inserting Links into Long Title or Meta Tags

Generally, web browsers show the preceding part of a long title. Thus, some spammers insert urls into long title. Similarly, meta tags provide structured meta data about a Web page and they are used for search engines. Although they have been the targets of spammers for a long time and search engines consider these data less and less, there are pages still using them.

3.12 L: Hiding Div “Below” the Visible Layer

Another sneaky way to hide a hyperlink from Web users while keeping it available to the search engines is to put the hyperlinks in a layer that is “behind” the visible layer. The CSS z-index property specifies the stack order of an element, which is supported in all major browsers. An element with a greater stack order is always in front of an element with a lower stack order. One example hiding hyperlinks via z-index is presented below.

```html
<div id="front" style="position:absolute; z-index:1">  
  <img src="image.gif" />
</div>

<div id="back" style="position:absolute; z-index:-1">  
  <a href="target.html" target="_blank">keywords</a>
</div>
```

The codes show that the second div has a negative stack order, which determines the target.html is behind the image.gif. Besides z-index, “overflow:hidden” can also hide the hyperlinks below the visible layer. Here is a simple example.

```html
<style type="text/css">
#spam{width:99px;height:20px;overflow:hidden;position:absolute;}
#spam a{display:block;line-height:20px;text-decoration:none;}
</style>

<div id="spam">
  <a href="/">&#160;</a>
  <a href="target.html" title="keywords">keywords</a>
</div>
```

In the example above, target.html is covered by a non-breaking space.

4. PREVALENCE OF LINK HIDING TECHNIQUES

In this section, we study the prevalent of hidden spam links, and how prevalence of the variety of techniques described in Section 3. We further study whether the hidden pages are punished by Google via PageRank analysis.

We carried out the analysis on 5,583,451 Chinese homepages (http://www. + domain name), including .com, .net and .cn domain names. To detect the Web pages with hiding links, we first train a cost sensitive naive bayes classifier on 63 pages with hidden links and 181 normal pages. The cost sensitive model ensures a high recall of pages with hidden links. Then, we filtered the 5,583,451 pages with the trained model. The detection results contain quite a few false alarms, but it’s enough for us to analyze the prevalence of hidden links. By random sampling from the suspicious set and carrying out manual verification, we approximately determined the number of pages with hidden links. Table 1 tabulates the statistics in detail.

| Table 1: Percentage occurrence of hidden link spam among Chinese Web pages |
|------------------------|--------|------------------|
| URL Type               | Count / Total | Percentage |
| government.cn          | 994/41405 = 1/42 | 2.4%          |
| .com/.net/.cn          | 81765/(5583451 – 41405) = 1/68 | 1.48%         |

It is noticed that a number of Chinese pages use hyperlink hiding techniques. In comparison with ordinary pages, government pages are more likely to contain hidden links. The underlying reason is that the government sites usually have better credibility. Spammers consider that the recommendations from these sites will help to boost the ranking of target sites. Is it really the case? We will try to answer it by analyzing the PageRank values of the hidden links.

To analyze the prevalence of the variety of techniques described in Section 3, we randomly sampled 160 pages with hidden links from government.cn set, and randomly sampled 118 pages with hidden links from .com/.net/.cn set. Each sampled hidden link spam page was manually analyzed. All the 278 samples were labeled with the types of techniques they used. Besides, all the hidden links are extracted for further analysis. In total, 9864 target hyperlinks are hidden in the
278 pages. 

Table 2 describes the prevalence of hidden link techniques in detail.

Table 2: Prevalence of different link hiding techniques

| techniques | number(percentage) => number of hidden links |
|------------|---------------------------------------------|
| A          | 6 (2.2%) => 39                              |
| B          | 3 (1.1%) => 29                              |
| C          | 8 (2.9%) => 33                              |
| D          | 19 (6.8%) => 21                             |
| E          | 8 (2.9%) => 117                             |
| F          | 68 (24.5%) => 4333                          |
| G          | 30 (10.8%) => 1876                          |
| H          | 111 (39.9%) => 3210                         |
| I          | 5 (1.8%) => 122                             |
| J          | 9 (3.2%) => 31                              |
| K          | 15 (5.4%) => 6                              |
| L          | 3 (1.1%) => 47                              |
| All        | 285 (285/278=102.5%) => 9864                |

The table shows that the 278 Web pages contain 9864 hidden links. F, G and H are the most popular link hidden techniques, which account for 75.3% of that total. These three techniques can be easily used to hide multiple hyperlinks. It can be observed that some of the 278 web pages contain more than one link hidden technique.

Are the 278 source pages and the 9864 target pages punished by the search engines? We do not know the detailed ranking strategy of commercial search engines, but we can explore this problem from a side by analyzing the PageRank values of the source and target hyperlinks.

Google provides a public interface, toolbarqueries.google.com, for querying the PageRank values. Given the 278 source urls and the overwhelming target hyperlinks do not have hierarchical path, in the form of \( http://++hostname \). We compared the PageRank values of the hostnames in our analysis. A hostname is a domain name assigned to a host computer, which is usually a combination of the host’s local name with its parent domain’s name. For example, hostname(\(http://en.wikipedia.org/wiki/Hostname\)) = en.wikipedia.org. In the rest of the section, the PageRank of a hyperlink or url refers to the PageRank of it’s corresponding hostname.

Table 3 shows the average PageRank values of source links, target hidden links and randomly selected 29994 hostnames from DNS resolution logs. Table 3 shows that the average PageRank values of source urls is 2.216, which means that spammers usually select the reputable pages to hide target hyperlinks. And the 9864 hidden links have an average PageRank value 1.32, which is higher than that of the randomly selected hostnames. The result means that Google does not establish the effective punitive mechanism for the hidden links. There is certainly another possibility: Google has punished these hidden links, which should have higher PageRank values if they do not cheat.

We further analyzed whether the links hidden in government sites have higher PageRank values. Our study doesn’t support this conclusion. The average PageRank value of links hidden in gov.cn sites is 1.319; however, that of links hidden in other sites is 1.342.

Figure 1 describes the distribution of PageRank values of 278 source links.

It can be observed that spammers tend to insert the hidden hyperlinks into pages with high PageRank values. And in this way, spammers hope to boost the PageRank values of the target links.

Figure 2 describes the distribution of PageRank values of 9864 hidden links. From figure 2, we can see that 53.9% hidden hyperlinks have 0 PageRank values. However, quite a number of spam links have high PageRank values. Figure 2 shows that more than 5.4% hidden links PageRank values are greater than or equal to 5.

Table 3: Comparison of average PageRank values

| Number | source Urls | hidden links | randomly selected hosts |
|--------|-------------|--------------|-------------------------|
| Average PageRanks | 2.216 | 1.320 | 1.139 |

1The data is freely available on http://gengguanggang.wix.com/gggeng/#/hidden-link-spam/cqz1.
Next, we analyzed the average PageRank values of target hyperlinks using different hidden techniques, which are described in detail in figure 3.

![Figure 3: The average PageRank values of target links hidden with different techniques.](image)

Figure 3 shows that the hyperlinks hidden via H, I or J techniques have greater PageRank values. Links hidden with H and I techniques have big PageRank values is pretty easy to understand for Google ignoring the JavaScript codes. However, why hyperlinks hidden in pull down menu have big PageRank values is puzzling. Maybe the spam links scatter in the drop down list, which is relatively hard for search engines to recognize.

Finally, we analyze the high-frequency words in the anchor texts of the 9864 target hyperlinks. The top 10 high-frequency keywords and the corresponding types are tabulated as follows.

| Keywords | Term Frequency | Type            |
|----------|----------------|-----------------|
| 博彩     | 328            | Gambling sites  |
| 百家乐   | 258            | Gambling sites  |
| 私服     | 239            | Illicit game servers |
| 全讯网   | 221            | Gambling sites  |
| 魅动乾坤 | 172            | Online novels   |
| 医院     | 167            | Medical services |
| 太阳城   | 104            | Gambling sites  |
| 六合彩   | 95             | Gambling sites  |
| 代孕     | 73             | Medical services |
| 网址之家 | 65             | Navigation sites |

Figure 4: The high-frequency words in the anchor texts of the target hyperlinks.

The statistics show that gambling sites, personal game servers and medical services are the main types of the hidden links. Most of the sites belong to shady or illegal industries.

5. CONCLUSION AND FUTURE WORK

In this paper we presented a variety of commonly used link hiding techniques, and organized them into a taxonomy. We analyzed the prevalence of common link hiding techniques on the web and discussed whether search engines punish the hidden links via PageRank analysis. Just as the previous work on Web spam [8] [5], we argue that such a structured discussion of the subject is important to raise the awareness of the research community. Given that most of the sites using link hidden techniques are shady or illegal industries, more should be done to punish the hidden link spam.

In the future, we should pay more attention to two things. The first is studying link hidden spam on a bigger data set, which includes multilingual samples. The second is developing a proper countermeasure to address the problem as a whole, despite the variety of different link hidden techniques. One possible solution draws support from maturing optical character recognition techniques (OCR) [11]. The motivation is that as a computer vision technique, OCR can only read the visible content on the Web page like humans. The snapshot of a Web page can be easily taken via some softwares, such as wkhtmltopdf [3] and snapshotter [2]. All the visual text on the snapshot image can be recognized via OCR techniques as textVector. If an anchor text does not exist in the textVector, the corresponding hyperlink is identified as hidden link. And, of course, the relative position of anchor text should also be taken into account.

Acknowledgment

We would like to thank Xiao-Tong Yuan for his help in our labeling efforts. This paper is supported by grants National Natural Science Foundation of China (Nos. 61005029 and 61103138).

6. REFERENCES

[1] Blink element — Wikipedia, the free encyclopedia, 2013. [Online; accessed 20-January-2013].
[2] Snapshotter, 2013. [Online; accessed 20-February-2013].
[3] wkhtmltopdf, 2013. [Online; accessed 20-February-2013].
[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1):107–117, 1998.
[5] K. Chellapilla and A. Maykov. A taxonomy of javascript redirection spam. In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, pages 81–88. ACM, 2007.
[6] D. Flanagan. JavaScript: the definitive guide. O’Reilly Media, Incorporated, 2006.
[7] Google. Webmaster guidelines - webmaster tools help, 2013. [Online; accessed 17-January-2013].
[8] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In First international workshop on adversarial information retrieval on the web (AIRWeb 2005), 2005.
[9] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, pages 576–587. VLDB Endowment, 2004.
[10] J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632, 1999.
[11] S. Mori, H. Nishida, and H. Yamada. Optical character recognition. John Wiley & Sons, Inc., 1999.
[12] A. Ng, A. Zheng, and M. Jordan. Stable algorithms for link analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and
[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[14] Wikipedia. Marquee element — Wikipedia, the free encyclopedia, 2013. [Online; accessed 19-January-2013].

[15] Wikipedia. Spamdexing — Wikipedia, the free encyclopedia, 2013. [Online; accessed 17-January-2013].

[16] B. Wu and B. Davison. Cloaking and redirection: A preliminary study. In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'05), 2005.