Multi-Blogs Steganographic Algorithm Based on Directed Hamiltonian Path Selection

Gang Luo, Xingming Sun and Lingyun Xiang
School of Computer and Communication, Hunan University, Hunan, 410082, China

Abstract: In this research, a steganographic algorithm based on the directed Hamiltonian path selection in the complete digraph mapped from multi-blogs with same article has been proposed. Firstly, we can regard n different blogs referring to a same article as n different virtual vertices and then connect them virtually to construct a complete digraph. As there exist n! different directed Hamiltonian paths at most in a complete digraph with n vertices, after numbering all directed Hamiltonian paths, a large number converted from arbitrary secret information smaller than n! can be expressed as a certain Hamiltonian path. In the process of the actual realization, the cited links of the same article are used to indicate the location of the former vertex of the current in a selected Hamiltonian path. In the information-extracting process, we can recover the whole Hamiltonian path by tracing the cited links and then decode this Hamiltonian path to retrieve the hidden information. Based on the theoretical analysis and the experimental proof, it demonstrates that the proposed steganography has good imperceptibility and security.

Key words: Steganography, directed Hamiltonian path, complete digraph, blog

INTRODUCTION

Steganography (Rabah, 2004) is the art and science of hiding information by embedding information within other, seemingly innocuous cover messages (Cachin, 1998; Wang and Wang, 2004). It has the advantage of avoiding the attention of a third party when transmitting secret information. Therefore, it is more effective than the traditional encryption technology in hiding and transmitting secret information. In recent years, steganography has been greatly improved and many superior algorithms have been proposed so far. According to the type of cover messages (Cox et al., 2005; Castiglionea et al., 2007; Socek et al., 2007), steganography can be categorized into image steganography, video steganography, audio steganography, text steganography and so on. At present, the image steganography is the research hotspot and is most mature technology (Kharrazi et al., 2006; Noda et al., 2007; Chen, 2008). With the Internet arguably becoming most successful medium for open-information distribution, some researches on steganography related with the Internet have been done (Fisk et al., 2002; Goel et al., 2007), which are, however, usually limited to taking the Network protocols (such as the TCP/IP family of Internet protocols) as the carrier to hide information, or direct application of image, text and other steganography under Internet background. Because the image, text and other cover messages are the information forms of distributing abroad through the Internet, the stego-messages generated by the corresponding steganography can be directly transmitted through the Internet.

However, new technologies based on Internet have been developed and widely used, each always having its unique property. If the existing steganographic methods are directly applied to the cover messages existing on the Internet, the optimal effect of hiding information cannot be achieved. However, little work is done on steganographic method based on unique properties of such technologies, which can achieve better performance, in the literature currently.

Now we take the blog (Lindahl and Blount, 2003) technology as an example to illustrate its unique properties. Blog is the new form of information distribution and communication on Internet developed in the past several years. The current blogs are generally based on WEB2.0 technique, which makes people collaborate and share information online. Since blog has many advantages (Nardi et al., 2004; Secco, 2005) compared with the traditional interactive network technologies, it has been rapidly developed. According to the report published by Technorati (2007), which is the world’s largest blog search engine, on April 5, 2007, the global number of bloggers is greater than 70 million and the average of daily increment of blogs is 120,000, namely increasing 1.4 new bloggers per second; it generates
new blog articles at a speed of 1.5 million per day, or
17 sec⁻¹; among the 100 most popular Web site globally,
22 of them are blog sites. Considering the simplicity of
analysis, the term blog in this paper denotes the blog and
the similar technology (space technology) with the same
characteristics. In other words, the number of users of the
current blog as well as space using the similar technique
already reaches a very huge number. Therefore, research
on steganography based on unique property of blog is
valuable in terms of both its theory and application.
However, little related work can be found both on the
Internet and in the literature.

Since Blog has the characteristic of being easily
applied for and a large number of articles are being shared
by different blogs, this research proposes a steganographic
algorithm based on these characteristics. First, choose n
different blogs simultaneously referring to an identical
article and then take the location of the article as n
different virtual vertices and connect them to construct a
complete digraph. There are n! different directed
Hamiltonian paths in a complete digraph with n vertices.
Therefore, after numbering all directed Hamiltonian paths,
a large number less than n! can be expressed as a certain
Hamiltonian path. In other words, the secret information
after being converted to a large number can be hidden
by selecting a specific Hamiltonian path. Both the
theoretical analysis and experimental proof demonstrate
that the proposed steganography has good
imperceptibility and security.

**CHARACTERISTIC ANALYSIS OF BLOG**

Blog usually has the following advantaged
characteristics for hiding information:

- The simplicity of applying for blogs. Now there are
many providers offering free blog. Anyone can apply
for a large number of free blogs without any cost. For
steganography, the simplicity of applying for blogs
always means that the continual replacement of the
blogs containing hidden information can avoid the
secret information being found. For this study, the
ability of using amount of blogs at the same time is
the necessary condition of the proposed
steganographic algorithm.

- It is easy to publish an article as it has relatively
fewer constraints than publishing articles on BBS
(Bulletin Board Service) and so on. As the blog
content is basically determined by the applicant,
blogger can publish any articles of arbitrary thematic
without breaking the law on his own blog, which is
impossible for publishing article on other sites such
as BBS, because the article may be deleted by the
administrators for reasons such as improper thematic
and so on. For steganography, the ease with
publishing article means that the type of cover
messages for hiding information has tremendous
diversity and the transfer of secret information is
much more difficult to be interrupted.

- The form of the published blog content is
unrestrained. Compared with formal web sites, the
form of the blog articles is relatively arbitrary. There
is no strict restriction to ensure that the published
content has to be in accordance with some standard
form. Upon this point, BBS also has the same
characteristic. For steganography, the requirements
for the form of the published content is fewer, the
redundant space used to hide secret information is
greater. This characteristic is not utilized in this
research, but it is used in some research on the
steganography based on BBS and so on (Shirali-
Shahreza and Shirali-Shahreza, 2007; Topkara et al.,
2007).

- It always happens that the same article is referred in
different blogs and normally the site of the quoted
article is given when the article is cited from other
site. Despite the large amount of original articles, the
number of republished articles is much greater and
articles about some hotspots will be republished
abroad and frequently. This characteristic is the
foundation of the proposed algorithm.

**MULTI-BLOGS STEGANOGRAPHIC ALGORITHM
BASED ON DIRECTED HAMILTONIAN PATH
SELECTION**

Because of the first characteristic of blog, we can
easily obtain many blogs and republish the same article
on multi-blogs that we own based on the fourth
characteristic mentioned above. The blog sites referring
to the same article can be regarded as a vertex. After
virtually connecting these vertices, a complete digraph
can be constructed. A larger number can be hidden by
selecting different directed Hamiltonian path in the
complete digraph. In the concrete design of algorithm, the
cited links of the same article are used to indicate the
former vertex of the current in the directed Hamiltonian
path. And after converting the secret information into a
large number, secret information can be hidden in the
certain directed Hamiltonian path. In the following
algorithm, if no special instruction is mentioned, the
subscript of sequence begins with 1 rather than 0.
The approach of mapping the same article in different blogs into a directed Hamiltonian path: Because when republished in various blogs, the republished article M often contains a cited link to indicate the location of the source. Then we can take each article M in different blogs as a vertex and connect every two vertices to construct a complete digraph D. The cited link in the article is regard as one property of the vertex which is used to point to other vertex in D. Generally speaking, most of the cited links are in confusion when the articles are cited arbitrarily. In such case, a directed Hamiltonian path does not exist. Figure 1a shown a complete digraph mapped from six blogs. In Fig. 1a, the dashed arrow denotes the virtual edges of the complete digraph. These edges do not actually exist. They are just used to express that the arbitrary vertices can point to each other. The start of solid arrow denotes the vertex mapped from the article actually existing in the blog. The vertex of solid arrow pointing to denote the location of the cited link in the arrow' start pointing to. For example, arrow from the vertex 6 pointing to vertex 1 denotes that the article mapping to the vertex 6 has a cited link, which points to the location of article mapping to vertex 1. The black vertex in the figure indicates that it has the property of the cited link, which is not meaningful for us, such as the vertex mapped from the original location of M, which does not include cited link, or the vertex mapped from M which points to an invalid location. Thus, there does not exist not any cited arrow starting from the black vertices.

But if we adopt an appropriate approach to modify the cited links in M's in order to make the links point to the different vertices in D, thus a directed Hamiltonian path can be constructed. If the original location is taken as a starting vertex, it will find a directed Hamiltonian path in the opposite direction compared with the direction which the cited links point to Fig. 1b shows the result of modifying the cited links from the Fig. 1a. The modifications are: modifying the cited link of vertex 2 to point to vertex 5 instead of vertex 1, modifying the cited link of vertex 3 to point to vertex 2 instead of an invalid link. Then, a directed Hamiltonian path (shown by the solid directed edges in Fig. 1b) is determined, which is in the opposite direction compared with the direction of the cited links point to. The start vertex of the Hamiltonian path is vertex 1, the end is vertex 4 and the direction is the opposite of which the cited links point to.

For any n selected Blogs with the same article M included, the vertex that includes a cited link pointing to the location out of these n Blogs can be regarded as the start of the selected Hamiltonian path, i.e., we can easily map the position of article M published in n different blogs into a virtual complete digraph and then construct a Hamiltonian path by modifying the cited links in the republished M.

In the same way, a reverse mapping can be constructed to map the selected directed Hamiltonian path in the complete digraph into the blogs pointed out by the cited links of M in n blogs.

The selection algorithm of directed Hamiltonian path: According to the graph theory, any directed Hamiltonian path in a complete digraph can be expressed as an integer less than n'. That is to say a given number can be converted to a corresponding special directed Hamiltonian path. Therefore, we can select a special directed Hamiltonian path to hide the number m, which is converted from secret message.

Algorithm 1: The algorithm of selecting directed Hamiltonian path.

452
**Input:** A complete digraph D with N vertices, a number m less than n!

**Output:** The vertex sequence of the selected directed Hamiltonian path

- Orderly add each vertex in the digraph D into a temporary list TempList.
- Set i = 1 and m' = m.
- \( k_i = \frac{m'}{(n-i)!} + 1 \)
- \( m' = (m' \mod (n-i)) \)
- Append TempList[k_i] to the tail of output; Delete TempList[k_i] from TempList;
- i = i + 1
- If the number of vertices in TempList is larger than 1, then return to step 3.
- Append TempList[1] to the tail of output and return the output.

Here, a simple example will be used to describe the result of the above algorithm. Suppose we have 10 vertices numbered from 1 to 10 orderly and an integer \( m = 1234567 \) (much less than 10!), a Hamiltonian path can be determined according to m. The vertex sequence of the selected directed Hamiltonian path calculated by Algorithm 1 is (4, 5, 7, 10, 8, 1, 3, 2, 9, 6). When \( m = 2345678 \), the corresponding sequence is (7, 5, 2, 4, 10, 3, 8, 6, 1, 9).

**The extraction algorithm:** above analysis, For a given M, we can easily obtain the complete digraph D to which the n blogs map and the n blogs are selected by determining whether it contains the article M or not. The serial number of the former vertex can be obtained by the cited link of current vertex. Here the former vertex of one vertex is called the father vertex for short. With the complete digraph D and the father vertices of all vertices in D, the selected Hamiltonian path in D can be reconstructed.

**Algorithm 3:** The reconstruction algorithm of the selected directed Hamiltonian path

**Input:** The complete digraph D with n vertices, the serial number of father vertex obtained by the cited links for every vertex

**Output:** The vertex sequence of the selected directed Hamiltonian path

- Construct a doubly linked list, each node of the list is designed to store 4 values: 
  - node. PointId stores the serial number i of the corresponding vertex in D,
  - node. PointParentId stores the serial number p of the father vertex of vertex i,
  - node. Parent stores the pointer that points to the father node of current node,
  - node. Child stores the pointer that points to the child node of current node.

- Initialize the node number of the doubly linked list as n+1, set node. Parent = nil, node. Child = nil for all nodes; As the first node is virtual, node. PointId and node. PointParentId is set to NULL;

For other nodes, set the value of node. PointId as the serial number i of the corresponding vertex in D, set the value of node. PointParentId as p, of the father vertex of vertex i;

Set i = 1.
• Make node.Parent of the i-th node point to the p-th node, make node.Child of the p-th node point to the i-th node, i = i + 1, repeat step 3 until i=n.
• Initialize node = the virtual node.
• If node.Child = nil, then return the sequence of result and exit.
• Append node.Child.PointId to the tail of the result sequence and set node = node.Child, return to step 5.

After obtaining the selected Hamiltonian path in the complete digraph D by the Algorithm 3, a large number denoted by the Hamiltonian path can be extracted. Algorithm 3 is the reverse process of Algorithm 1.

**Algorithm 4:** The extraction algorithm of the number denoted by the certain directed Hamiltonian path

**Input:** The complete digraph D with n vertices, the vertex sequence List of the selected directed Hamiltonian path

**Output:** The number m

• Orderly add the serial number of the every vertex in digraph D into a temporary list TempList.
• Initialize m = 0 and i = 1.
• k_i = TempList. index of (List [i]), namely k_i expresses the position of List [i] in TempList.
• m = m + (k_i-1)*(n-i)!!
• Delete TempList[k_i] from TempList and i = i + 1.
• If the number of vertices in TempList is larger than 1, then return to step 3;
• Return m;

Assume we have 10 vertices numbered from 1 to 10, respectively in ascending order, if the vertex sequence of the selected Hamiltonian path is (4, 5, 7, 10, 8, 1, 3, 2, 9, 6), then the corresponding code m should be 1234567 calculated by Algorithm 4; if the vertex sequence is (7, 5, 2, 4, 10, 3, 8, 6, 1, 9), then the corresponding code m should be 2345678. Figure 1b has shown a Hamiltonian path, whose vertex sequence is (1, 6, 5, 2, 3, 4). As the complete digraph has six vertices, thus the corresponding code m should be 114, which can be calculated by Algorithm 4.

Based on the analysis above, we can achieve the extraction of secret information.

**Algorithm 5:** The extraction algorithm of multi-blogs based on directed Hamiltonian path selection

**Input:** The specified article M in n blogs

**Output:** Secret message m

• Construct a complete digraph D using the n blogs, make the cited link L_i of M in an arbitrary i-th blog point to the blog that corresponds to the father vertex p_i of i-th vertex in D.
• Using Algorithm 3, obtain the list List that stores the vertex sequence of the Hamiltonian path reconstructed by using the cited links;
• Using Algorithm 4, calculate the corresponding code m extracted from the special Hamiltonian path.
• Return m

**ANALYSIS OF THE PROPOSED ALGORITHM AND EXPERIMENT**

The analysis of imperceptibility: The traditional steganographic algorithm makes use of the redundant data to hide information. But the proposed steganographic algorithm adopts a different strategy. As we just modify the cited link of the article instead of modifying the content of cover messages such as cover image, cover video and cover text. Besides, after the cited links are modified, they will still point to a real locations and the article where the cited link locate at is the same as the article that the cited link points to. Therefore, the algorithm does not bring abnormality on content of a single article or make any statistic changes, that is to say, under the condition of the existing general steganalysis algorithms (Chandramouli et al., 2003; McBride et al., 2005; Lysyanskaya and Meyerovich, 2006), the proposed steganographic algorithm in this research has better imperceptibility than the existing steganographic algorithms.

If one steganalyst knows the principle of this proposed algorithm, he/she could design a steganalysis algorithm by making use of some specific characteristics owned only by this algorithm. We call this kind of steganalysis algorithm particular steganalysis algorithm. Normally, the steganographic methods usually do not have the ability of resisting such attacks, which come from the future unforeseen. Therefore, we only analyze how difficult to implement particular attacks. In order to successfully detect the existence of the secret information embedded by the proposed algorithm, the steganalysis algorithm must accurately locate the n blogs used. As if the n blogs are not determined, the further analysis would be of no significance. To accurately locate these blogs, the steganalysis algorithm must filter all the articles in the existing blogs to find out all the identical articles and record the cited links one by one. Through such mass statistics, it is possible to find out the blog groups that
transmit secret information many times; and further statistical analysis would detect the existence of the secret information transmitted. Thus it is possible to accurately detecting the used n blogs. But the computational cost will be very expensive, as the steganalysis must have the highest information processing capability to deal with more than 70 million existing blogs and hundreds of millions web pages of similar techniques and this number is continually expanding each second.

The analysis of security: The security of the steganographic algorithm mainly refers to the difficulty of decoding in case of the secret information being discovered. Normally, the security of the steganographic algorithm depends on the encryption algorithm used to encrypt the secret information before embedding. This is the last defensive line of steganographic algorithms. The encryption algorithm is not mentioned in this research, as the secret information is encrypted by default. In addition to the security provided by the encryption algorithm, a further security will be provided by this research before the steganalysis directly attacking the encryption algorithm, as it is difficult to obtain the number sequence of the selected blogs. In this research, only if the correct serial numbers of all the n blogs and the order is known, the algorithm will output the exact result and it is possible for steganalysis to obtain the correct secret information encrypted. As we know, the number of possible permutations of n blogs is $P_n^m = n!$, the complexity of exhaustive search would be very high, so the suitable permutations will be effective in improving the difficulty for steganalysis to correctly extract secret information. The commonly used methods of permutation according to their security level (from strong to weak), are listed as follows: random mapping table, the sort after encryption, the sort after hash, direct sort and so on.

The analysis of capacity: Based on earlier analysis, the capacity of the steganographic algorithm proposed in this research at a single time depends on the number of used blogs. Suppose the capacity is $C$ bits, then $C = \log_2(n!)$. The relationship between the capacity of hiding information and the number of used blogs is shown in Fig. 2.

When compared with the steganographic method based on image or others, the proposed method has relatively small communication capacity at a single time. Therefore, if a large amount of secret information needs to be transmitted, more times of communication are needed.

Experiments: First, we register 11 blogs on Internet, then use MD5 to hash the links of blogs and number them orderly (Table 1).

Then, we successfully embed the secret information string HNU using the proposed steganographic algorithm. For convenience of demonstration, we directly encoded the secret information without encrypting it in advance. The ASCII code of string HNU in hexadecimal is 484555, which is 7373645 in decimal. Thus, we need to hide a large number $m = 4738645$.

According to Algorithm 1, the calculated vertex sequence of the selected Hamiltonian path is (2, 5, 1, 8, 4, 7, 10, 9, 3, 11, 6), Making use of the sequence, we can obtain the republished links and the corresponding cited links. The article we used to republish for hiding
Table 2: Transformation between the vertex sequence of the Hamiltonian path and the cited links

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>The published link</th>
<th>The cited link</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://blog.sina.com.cn/e/blog_5088469010@68.html">http://blog.sina.com.cn/e/blog_5088469010@68.html</a></td>
<td><a href="http://netcote998.blogspot.com/diary_13500671.shtml">http://netcote998.blogspot.com/diary_13500671.shtml</a></td>
</tr>
<tr>
<td>8</td>
<td><a href="http://live.bkoke.com/viewdiary_20971968.html">http://live.bkoke.com/viewdiary_20971968.html</a></td>
<td><a href="http://blog.sina.com.cn/e/blog_5088469010@68.html">http://blog.sina.com.cn/e/blog_5088469010@68.html</a></td>
</tr>
<tr>
<td>9</td>
<td><a href="http://hi.baidu.com/lypiglet/blog/item/b688856d111a1df683169481.html">http://hi.baidu.com/lypiglet/blog/item/b688856d111a1df683169481.html</a></td>
<td><a href="http://blog.bchinese.net/netcore/archive/2008/01/29/129567.aspx">http://blog.bchinese.net/netcore/archive/2008/01/29/129567.aspx</a></td>
</tr>
<tr>
<td>3</td>
<td><a href="http://windlegend.blog.hexcan.com/16742860_d.html">http://windlegend.blog.hexcan.com/16742860_d.html</a></td>
<td><a href="http://hi.baidu.com/lypiglet/blog/item/b688856d111a1">http://hi.baidu.com/lypiglet/blog/item/b688856d111a1</a></td>
</tr>
</tbody>
</table>

information is the famous poetry <Clouds and Waves> written by Rabindranath Tagore. Transformation between the vertex sequence of the Hamiltonian path and the cited links is shown in Table 2. Only the first vertex of the sequence has a cited link which points to a location out of the used 11 blogs and others point to the published location of poetry <Clouds and Waves> in blog which is the former vertex compared with the current vertex in sequence.

Thus, the information has been successfully embedded. Extraction process includes firstly acquiring all the cited links and then using the extraction algorithm to extract the embedded original information.

CONCLUSION

The blog technology has some advantages, for example, easily applying for and large amount of articles republished in different blogs with the cited link given. Through the analysis on these characteristics, this research makes multi-blogs refer the same article and then maps the locations of the articles with same content to a complete digraph and finally selects a certain directed Hamiltonian path from the complete digraph to embed secret information. Based on the theoretical analysis and experimental proof, the proposed method has been proven to be imperceptible and secure.

ACKNOWLEDGMENTS

This research is supported by National Natural Science Foundation of China (No. 60573045), National Basic Research Foundation Program of China (No. 2006CB303000) and Key Program of National Natural Science Foundation of China (No. 60736016).

REFERENCES


