Malicious Web Page Detection Based on Anomaly Behavior

Chia-Mei Chen
Department of Information Management, National Sun Yat-Sen University, Taiwan
Outline

1. Introductions
2. The Proposed Approach
3. System Implementation and Experiment
4. Conclusions
With the rapid development of the computer networks, people nowadays are dependent on the Internet increasingly.

Browsing webpage is insecure due to the vulnerabilities of browsers and web applications.
The common vulnerability of web applications

Information Leakage: 81%
Cross-site scripting: 91%
SQL Injection: 36%
Broken Access Controls: 78%
Broken Authentication: 67%

(Stuttard & Pinto, 2007)
The evading mechanisms used by hackers somehow make the behavior of malicious web pages different from normal web pages.

We find out some special and interesting characters of malicious web pages through three aspects:

- injection media
- obfuscation
- and redirection

We present a new malicious web page detection algorithm based on anomaly behavior detection.
The Proposed Approach

The architecture of proposed system, WPC (Web Page Checker)
Web Page URL Extraction Module
The Proposed Approach

- Web page URL extraction module:
  - Tracing and recording suspicious HTTP request URLs.
  - Providing a connection topology about the target web page.

- Web page crawler module:
  - Crawling back resources requested by invisible JavaScript or iframe tags.
Behavior Extraction Module: 
- Webpage encoding detection.
- Sensitive keywords splitting detection.
- Sensitive keywords encoding detection.
- Redirection detection.
- Unreasonable coding styles detection.
MoBR Module
The Proposed Approach

MoBR module:
- Using templates to address common malicious web page species or family based on semantic and signature.
Based on our observation, we identify the most important characters of malicious web pages.

A formula is used for behavior scoring to detect anomaly behavior of malicious web pages based on expert knowledge.
WPC (Web Page Checker) alarms the web page with scores above threshold.

Behavior Scoring Formula:

\[
\text{SCORE}_{\text{anomaly-behavior}} = (RR + SKSR + SKER + SKSER) \times 100 + \\
(\text{Depth} + \text{UCSR-eval} + \text{UCSR-document.write}) \times 50 + \\
(\text{AlgoExMD Rate} + \text{MET}) \times 20
\]  
(2)
<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>Brief Description</th>
<th>Symbol</th>
<th>Importance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redirection Rate</td>
<td>Redirection Rate is defined as the number of pages which are identified as having redirection behavior.</td>
<td>RR</td>
<td>Level 1</td>
</tr>
<tr>
<td>Sensitive Keywords Splitting Rate</td>
<td>SKSR is defined as the number of pages which are identified as having sensitive keywords splitting behavior.</td>
<td>SKSR</td>
<td>Level 1</td>
</tr>
<tr>
<td>Sensitive Keywords Encoding Rate</td>
<td>SKER is defined as the number of pages which are identified as having sensitive keywords encoding behavior.</td>
<td>SKER</td>
<td>Level 1</td>
</tr>
<tr>
<td>Sensitive Keywords Splitting Encoding Rate</td>
<td>SKSER is defined as the number of pages which are identified as not only having sensitive keywords splitting behavior, but also sensitive keywords encoding behavior.</td>
<td>SKSER</td>
<td>Level 1</td>
</tr>
<tr>
<td>Depth</td>
<td>In our definition, the depth is defined as the height of a tree. In tree data structure, the height of a node is the length of the longest downward path to a leaf from that node. And the height of the root is the height of the tree. (Tree (data structure).)</td>
<td>Depth</td>
<td>Level 2</td>
</tr>
<tr>
<td>Unreasonable Coding Styles Rate - using eval() method</td>
<td>UCSR-eval is defined as the number of pages which are identified as having unreasonable coding styles using eval() method.</td>
<td>UCSR-eval</td>
<td>Level 2</td>
</tr>
<tr>
<td>Unreasonable Coding Styles Rate - using document.write() method</td>
<td>UCSR-document.write is defined as the number of pages which are identified as having unreasonable coding styles using document.write() method.</td>
<td>UCSR-document.write</td>
<td>Level 2</td>
</tr>
<tr>
<td>AlgoExMD Rate</td>
<td>AlgoExMD Rate is defined as the number of pages which are identified as malicious web pages by AlgoExMD algorithm in MoBR module.</td>
<td>AlgoExMD Rate</td>
<td>Level 3</td>
</tr>
<tr>
<td>Max Encoded Times 2009/7/28</td>
<td>Encoded Times is defined as the number of times a web page is encoded. In our observation, malicious web pages may encode themselves recursively. And MET is defined as the max number of times a web page is encoded of total tested web pages.</td>
<td>MET</td>
<td>Level 3</td>
</tr>
</tbody>
</table>
Our implementation of WPC:

- A plug-in for Internet Explorer 6.
  - Developing a DLL for IE 6.
Experiments
System Implementation and Experiment design

- Comprehensive comparison.

![Graph showing comprehensive comparison of different detection software]

<table>
<thead>
<tr>
<th>detection software</th>
<th>Avira AntiVir</th>
<th>Kaspersky</th>
<th>McAfee</th>
<th>ESET NOD32</th>
<th>TrendMicro</th>
<th>AlgoExMD</th>
<th>WPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>false positive rate</td>
<td>0.40%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2.01%</td>
<td>2.81%</td>
</tr>
<tr>
<td>false negative rate</td>
<td>11.92%</td>
<td>36.42%</td>
<td>42.38%</td>
<td>90.73%</td>
<td>34.44%</td>
<td>16.56%</td>
<td>0%</td>
</tr>
<tr>
<td>accuracy</td>
<td>95.25%</td>
<td>86.25%</td>
<td>84.00%</td>
<td>65.75%</td>
<td>87.00%</td>
<td>92.50%</td>
<td>98.25%</td>
</tr>
<tr>
<td>error rate</td>
<td>4.75%</td>
<td>13.75%</td>
<td>16.00%</td>
<td>34.25%</td>
<td>13.00%</td>
<td>7.50%</td>
<td>1.75%</td>
</tr>
</tbody>
</table>
The contributions of WPC:

- A new anomaly behavior aspect for malicious web page detection.
- Client-side solution for detecting malicious web pages.
  - the system implementation and deployment are not difficult.
- Real-time protection for Internet browsers.
Thank you!
Q&A.