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

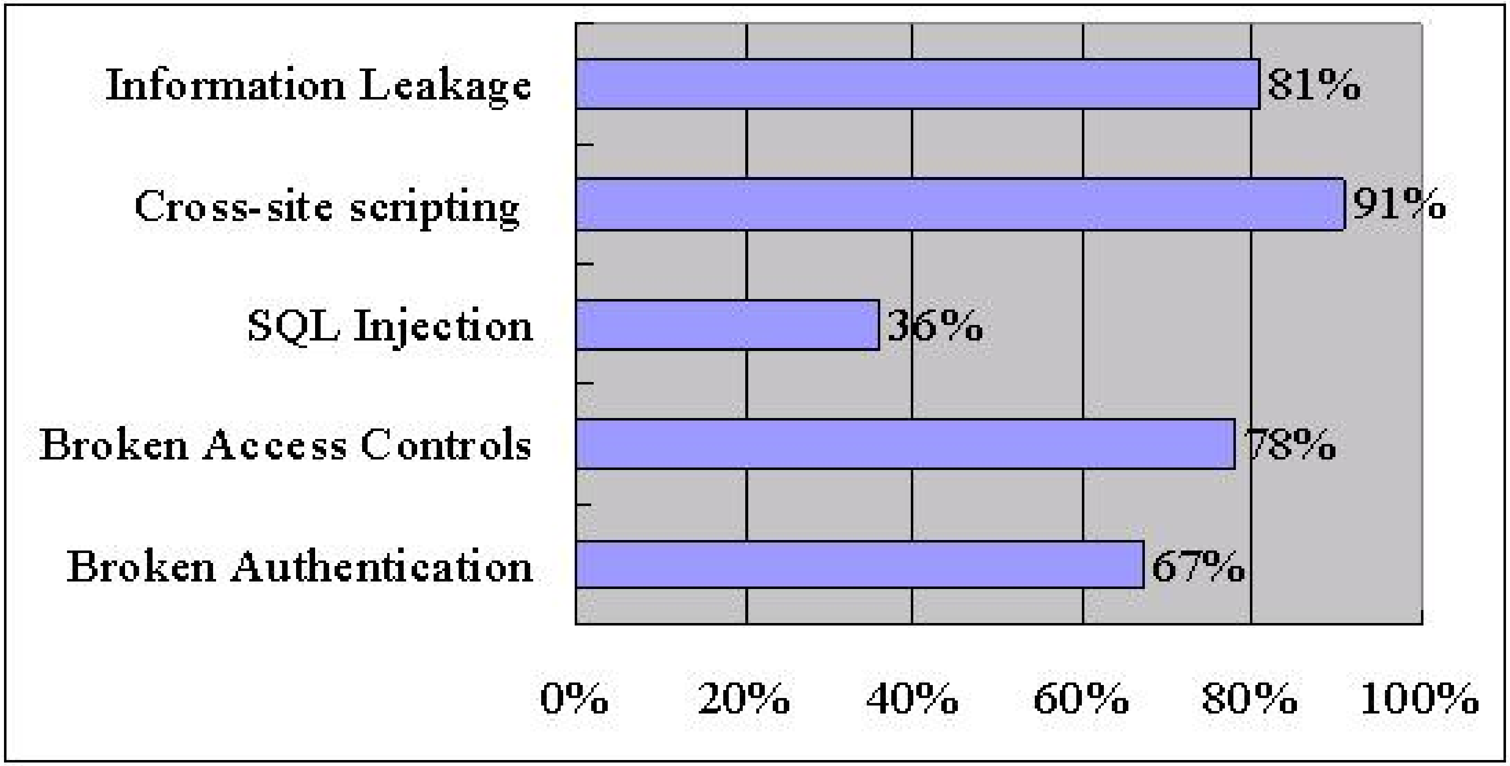
4 Conclusions

Background Introductions

- 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.

Background Introductions

The common vulnerability of web applications

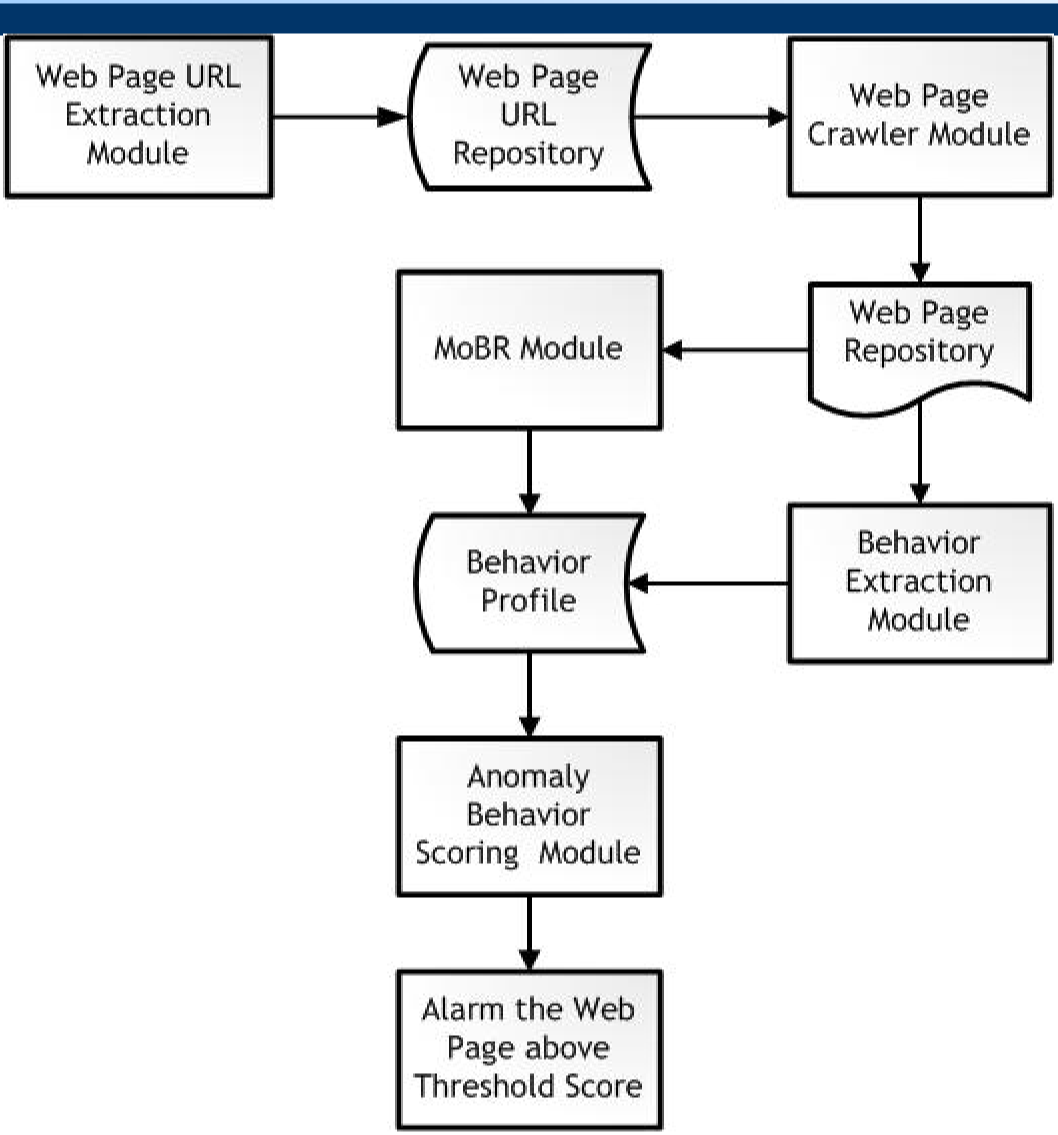


(Stuttard & Pinto, 2007)

Motivation Introductions

- 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 Proposed Approach


Web Page URL Extraction Module
Web Page URL Repository
Web Page Crawler Module
MoBR Module
Web Page Repository
Behavior Profile
Behavior Extraction Module
Anomaly Behavior Scoring Module
Alarm the Web Page above Threshold Score

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

The Proposed Approach

- 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.

Anomaly Behavior Scoring Module

The Proposed Approach

- Based on our observation, we identify the most important characters of malicious web pages.

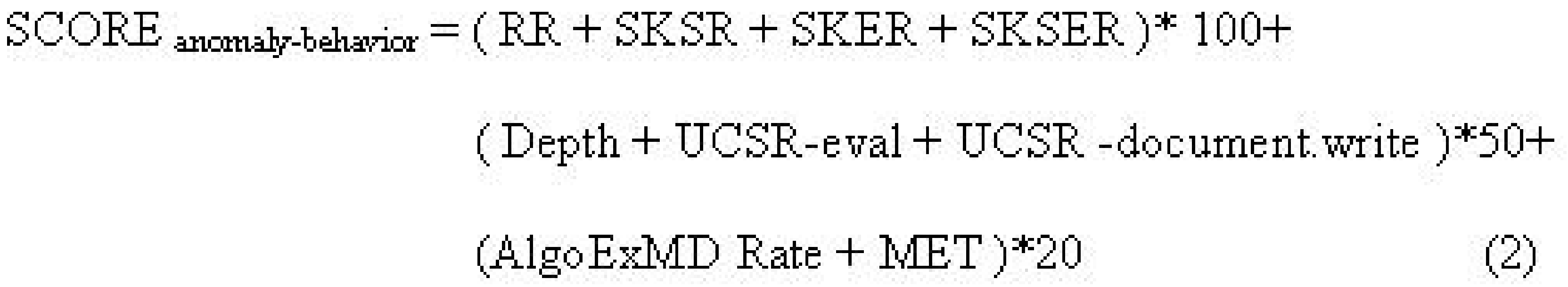
- A formula is used for behavior scoring to detect anomaly behavior of malicious web pages based on expert knowledge.

Anomaly Behavior Scoring Module

The Proposed Approach

- **WPC (Web Page Checker) alarms the web page with scores above threshold.**

- **Behavior Scoring Formula:**



$$\begin{aligned} \text{SCORE}_{\text{anomaly-behavior}} = {} & (\text{RR} + \text{SKSR} + \text{SKER} + \text{SKSER})*100+ \\ & (\text{Depth} + \text{UCSR-eval} + \text{UCSR -document.write})*50+ \\ & (\text{AlgoExMD Rate} + \text{MET})*20 \qquad (2) \end{aligned}$$

Predictor Variables	**Brief Description**	Symbol	Importance Level
Redirection Rate	Redirection Rate is defined as the number of pages which are identified as having redirection behavior.	RR	Level 1
Sensitive Keywords Splitting Rate	SKSR is defined as the number of pages which are identified as having sensitive keywords splitting behavior.	SKSR	Level 1
Sensitive Keywords Encoding Rate	SKER is defined as the number of pages which are identified as having sensitive keywords encoding behavior.	SKER	Level 1
Sensitive Keywords Splitting Encoding Rate	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.	SKSER	Level 1
Depth	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).*)	Depth	Level 2
Unreasonable Coding Styles Rate - using eval() method	UCSR-eval is defined as the number of pages which are identified as having unreasonable coding styles using eval() method.	UCSR-eval	Level 2
Unreasonable Coding Styles Rate - using document.write() method	UCSR-document.write is defined as the number of pages which are identified as having unreasonable coding styles using document.write() method.	UCSR-document.write	Level 2
AlgoExMD Rate	AlgoExMD Rate is defined as the number of pages which are identified as malicious web pages by AlgoExMD algorithm in MoBR module.	AlgoExMD Rate	Level 3
Max Encoded Times	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.	MET	Level 3

System Implementation

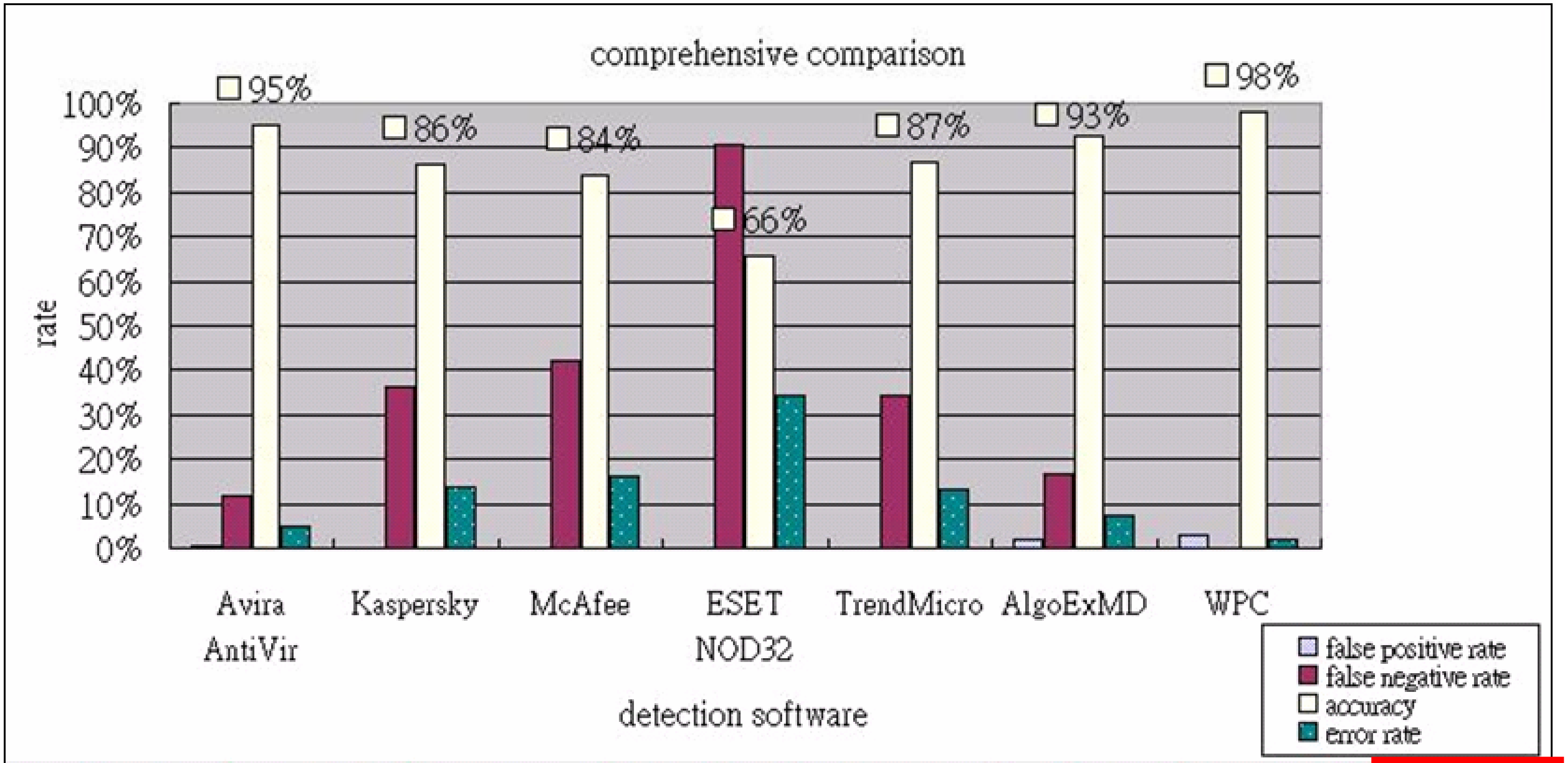
System Implementation and Experiment design

- **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.



detection software	Avira AntiVir	Kaspersky	McAfee	ESET NOD32	TrendMicro	AlgoExMD	WPC
false positive rate	0.40%	0%	0%	0%	0%	2.01%	2.81%
false negative rate	11.92%	36.42%	42.38%	90.73%	34.44%	16.56%	0%
accuracy	95.25%	86.25%	84.00%	65.75%	87.00%	92.50%	98.25%
error rate	4.75%	13.75%	16.00%	34.25%	13.00%	7.50%	1.75%

Conclusions

- 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.