



AFRL-RI-RS-TR-2012-273

FINE-GRAIN DOCUMENT ACCESS MODELING AND MONITORING

GEORGIA TECH RESEARCH CORP.

NOVEMBER 2012

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the DARPA Public Release Center and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RI-RS-TR-2012-273 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:



AMANDA P. OZANAM
Work Unit Manager



WARREN H. DEBANY JR.
Technical Advisor, Information
Exploitation & Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.					
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) NOVEMBER 2012		2. REPORT TYPE FINAL TECHNICAL REPORT		3. DATES COVERED (From - To) MAR 2011 – JUN 2012	
4. TITLE AND SUBTITLE FINE GRAIN DOCUMENT ACCESS MODELING AND MONITORING				5a. CONTRACT NUMBER FA8750-11-C-0136	
				5b. GRANT NUMBER N/A	
				5c. PROGRAM ELEMENT NUMBER 62303E	
6. AUTHOR(S) Calton Pu and Christopher Grayson				5d. PROJECT NUMBER CIND	
				5e. TASK NUMBER GA	
				5f. WORK UNIT NUMBER IT	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Georgia Tech Research Corporation 505 10 th St. NW Atlanta, GA 30332-0001				8. PERFORMING ORGANIZATION REPORT NUMBER N/A	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIGA 525 Brooks Road Rome NY 13441-4505				10. SPONSOR/MONITOR'S ACRONYM(S) N/A	
				11. SPONSORING/MONITORING AGENCY REPORT NUMBER AFRL-RI-RS-TR-2012-273	
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited. DARPA DISTAR Case No: 20117 Dated 6 Nov 2012					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT This project builds software tools to store documents in the PPS (Partial Persistent Sequence) system, monitor accesses to PPS documents, and analyze the access patterns to find cyber insider attacks. PPS uses a simple API access authorization scheme to thwart direct data transfer cyber insider attacks. A combination of fine-grained access monitoring data and statistical analysis on combined sub-sessions supports the effective detection (and prevention) of next generation cyber insider attacks that only steal minimized and targeted data transfers through legitimate APIs.					
15. SUBJECT TERMS Cyber insider threat; collusion in insider threats; partial persistent sequences; statistical detection of insider					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT SAR	18. NUMBER OF PAGES 22	19a. NAME OF RESPONSIBLE PERSON AMANDA P. OZANAM
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (Include area code) (315) 330-4517

TABLE OF CONTENTS

1	SUMMARY	1
2	Introduction	1
2.1	Purpose of Report and Project	1
2.2	Mission Description	2
2.3	Project Members	2
3	Methods, Assumptions, and Procedures	3
3.1	Technical Objectives and Methods	3
3.2	Classification of Cyber Insider Threats (Assumptions).....	3
3.3	Summary of Project Achievements	4
4	Results and Discussion	4
4.1	PPS Infrastructure	5
4.2	Monitoring and Detection Tools.....	5
4.3	Malware Test Platform	7
4.4	Demonstration and Evaluation	7
4.5	Concrete Observables	8
4.6	Concrete Metrics for Evaluation	9
5	Conclusion	12
6	References	13
6.1	Software tools and documents	13
	Appendix A: MTP Manual.....	14
	List of Acronyms.....	18
	Glossary of Technical Terms.....	18

LIST OF TABLES

Table 1	List of Project Members.....	2
Table 2	List of Concrete Observables	9
Table 3	Concrete Metrics for Evaluation	9

1 SUMMARY

The main achievement of the project has been obtained by a set of software tools that demonstrate the successful detection of colluding team cyber insider attacks in various dimensions (time, space, value) through fine-grained monitoring of document access sessions and statistical analysis of combinations of apparently innocuous access sessions. The demonstration has been achieved through the observables and evaluation metrics detailed in the following subsections.

The software tools include: a document storage and access metadata storage system (the PPS), a document access monitoring tool (a Firefox plug-in), document translation tools (e.g., PDF to HTML), and an attack generator (Malware Test Platform) that can generate various forms of cyber insider attack, including the various forms of collusion. Using these software tools, we demonstrated the effective detection of stealthy cyber insider attacks through statistical analysis tools such as various correlation analyses methods.

2 Introduction

2.1 Purpose of Report and Project

This final report is a deliverable summarizing the effort expended by the Georgia Institute of Technology (GT) team in support of AFRL and DARPA on Contract FA8750-11-C-0136. Part of the Cyber Insider Threat (CINDER) program. The project Principal Investigator is Prof. Calton Pu.

The objective of this project is to develop a prototype system for detecting document access activities that indicate cyber insider attacks, including collusion among multiple cyber attackers. The attacks covered range from indiscriminate bulk read/bulk copy to the detection of evasive techniques such as low intensity, fine granularity, and high value/low volume access of sensitive document content by cyber attacker software. At the basic level (called Generation 1 in this project), indiscriminate accesses by cyber attackers is through controlled access APIs by the customized PPS document storage system. This is analogous to some existing tools (e.g., DLP software) that work on network traffic and standard file systems. More sophisticated cyber insider attacks (called Generation 2 in this project) that combine apparently normal access sessions are detected through statistical analyses of correlation to valuable document content. This is achieved by our software tools to monitor all access sessions (from PPS) and

correlation analysis of document access sessions by potentially colluding cyber insider attackers.

2.2 Mission Description

Evasive Cyber Insider Mission - Our mission of interest lies beyond the Generation 1 cyber insider malwares that employ bulk access tactics. These can already be detected through existing intrusion detection techniques that rely on threshold-based filtering and analysis of outbound information-flows, e.g., Data Loss Prevention (DLP) tools such as OpenDLP.

Beyond bulk access, we anticipate *evasive* cyber insider attacks (Generation 2) that will use several strategies to escape generic thresholds on coarse-grain metrics. We will investigate primarily three evasion strategies (and their combinations) that can be used by individual cyber insiders or cyber insider teams working in collusion: time division, space division, and value-driven. First, time division reduces the intensity of detectable activity by separating periods of activity in time. Second, space division reduces the granularity of detectable activity by dividing the work among various members of a colluding team. Third, the value-driven approach takes advantage of insider attack semantics (stealing of information) by reducing the total amount of detectable activity and accessing only high value parts of a document.

2.3 Project Members

Table 1 List of Project Members

Member Name	Position	Duration of Work
Pu, Calton	Principal Investigator	03/11 - 06/12
Liu, Ling	Technical Lead	04/11 - 09/11
Grayson, Christopher	Research Scientist	04/11 - 06/12
Suh, James	Application Developer	08/11 - 04/12
Malkowski, Simon	Graduate Research Assistant and Project Manager	04/11 - 04/12
Doo, Myungcheol	Graduate Research Assistant	08/11 - 04/12
Kisung, Lee	Graduate Research Assistant	08/11 - 04/12
Wang, De	Graduate Research Assistant	08/11 - 04/12
Yang, Yi	Graduate Research Assistant	08/11 - 04/12
Park, Junhee	Graduate Research Assistant	05/11 - 04/12

Wu, Qinyi	Graduate Research Assistant	05/11 - 07/11
Mohan, Paritosh	Graduate Research Assistant	08/11 - 04/12

3 Methods, Assumptions, and Procedures

3.1 Technical Objectives and Methods

The technical deliverables of the project are divided into three main components:

1. The document access infrastructure, including the PPS (partial persistent sequence) API and a customized relational database implementation of PPS for this project. This is discussed in Section 4.1.
2. The insider access monitoring and detection mechanisms for HTML and TXT documents stored in PPS and accessed through a web browser, specifically, the Firefox, and translators for documents stored in other formats such as PDF. This is discussed in Section 4.2.
3. The MTP (Malware Test Platform), a powerful document access generator that simulates a variety of insider accesses to documents. This is discussed in Section 4.3.

The progress made on these deliverables is measured by the concrete observables listed in Section 4.5 and evaluated according to the concrete metrics listed in Section 4.6. These concrete observables and metrics are a subset of and derived from the original observables and evaluation metrics developed during the initial period of performance, in collaboration with the CINDER program independent evaluators at MIT Lincoln Labs. As the program progressed and focused specifically on cyber insider activities, the project focused on the concrete observables and metrics listed in Sections 4.5 and 4.6.

3.2 Classification of Cyber Insider Threats (Assumptions)

We use an adversarial learning-based classification of cyber insider threats to put our project in perspective. The simplest category of cyber insider attacks, called Generation 1, consists of bulk transfer of data that can be detected using current DLP tools. Generation 1 cyber insider attacks are oblivious to defense tools.

The next category, called Generation 2, consists of cyber insider attacks that are aware of threshold-based defense tools such as DLP and intrusion detection software. They maintain a low profile by staying below the threshold set by DLP and intrusion detection tools. There are various techniques for staying below the threshold, which are modeled as *collusion* in this project. A Generation 2 cyber insider attack may use time division (accessing a document over a period of time through several short sub-sessions by the same insider), space division (accessing a document through several sub-sessions by different insiders), and value-based (reading the document in a set of sub-sessions by one insider, and copying only the valuable parts of the document in a short sub-session by another insider). This project is focused on Generation 2 cyber insider attacks.

Further evolution of cyber insider attacks include the emulation of human reading patterns to escape detection mechanisms that attempt to distinguish human access from cyber tool access. Due to time constraints this project did not explore the cyber insider attacks that emulate human access patterns (Generation 3).

3.3 Summary of Project Achievements

The main achievement of the project has been obtained by a set of software tools that demonstrate the successful detection of colluding team cyber insider attacks in various dimensions (time, space, value) through fine-grained monitoring of document access sessions and statistical analysis of combinations of apparently innocuous access sessions. The demonstration has been achieved through the concrete observables and evaluation metrics detailed in Sections 4.5 and 4.6.

The software tools include: a document storage and access metadata storage system (the PPS), a document access monitoring tool (a Firefox plug-in), document translation tools (e.g., PDF to HTML), and an attack generator (Malware Test Platform) that can generate various forms of cyber insider attack, including the various forms of collusion. Using these software tools, we demonstrated the effective detection of stealthy cyber insider attacks through statistical analysis tools such as various correlation analyses methods.

4 Results and Discussion

The project's main results are software tools described in Sections 4.1, 4.2, and 4.3. The results are evaluated and discussed in Sections 4.4, 4.5, and 4.6.

4.1 PPS Infrastructure

We have designed and implemented a customized version of the PPS (Partial Persistent Sequences) representation of documents. This version of PPS can be described as a customized file system with significant metadata collection and querying capabilities that are absent in typical file systems. At an abstract level, PPS collects all data access information (for both reads and writes) at fine granularity (currently at cursor and mouse pointer resolution level) and store it with the relevant document component.

From a data structure point of view, PPS is implemented as a tree, where top level nodes contain the actual words of a document, and the children nodes of each word store its access information. Concretely, we store both the data and metadata in a relational DBMS (currently MySQL).

The PPS provides four major functions. First, it supports read/write of document data. Second, it monitors and records all accesses to document data. Third, the access monitor can detect anomalous (bot program) access and block them immediately. Fourth, PPS provides a simple query facility for the metadata, so we can extract appropriate information for further analysis.

4.2 Monitoring and Detection Tools

While the PPS infrastructure benefited significantly from previous work (it builds on Qingyi Wu's dissertation, which was completed in summer 2011), the access monitoring tools were built from ground up specifically for this project. These tools consist of a GUI, a logging facility to capture document access metadata, and various statistical analysis tools.

For the GUI, we have chosen the web browser interface, since it is the most popular GUI for many applications, including document access. Our implementation is a Firefox plug-in (Javascript implementation) for HTML documents, plus translators for TXT (easily translated into HTML) and PDF (somewhat more involved translation). A parallel project (funded by other sources at George Mason University) built similar monitoring tools for Microsoft Word.

The Javascript implementation of Firefox plug-in has been used by more than 50 students in Prof. Pu's classes in Fall 2011 and Spring 2012 semesters for their paper reading assignments. (This work was carried out after suitable IRB approval. We did not collect any participant identification data, but assume that each student used his/her own account for access.) Their reading access patterns have been recorded and they

are being analyzed for a paper on the experimental analysis of real world document access patterns.

The initial implementation of the Firefox plug-in (Javascript implementation) was able to capture the time spent on a document at the granularity of approximately a paragraph. This metadata is collected into the PPS backend database for subsequent statistical analysis. The granularity of capture was refined to a line and further refinements are feasible, with some additional overhead.

The Firefox GUI read access monitor collects access statistics and sends data to the PPS database at a tunable time interval (usually set at 20 seconds). At this interval, our access monitor is virtually undetectable by human users and has very low overhead (a few percent CPU usage on a PC). If additional overhead is acceptable, fine-grained real-time data collection may enable the real-time detection of insider activity.

The GUI displays the analysis results both as a graph and a heatmap for the visualization of document reading behavior. The display part of the statistical analysis tool reproduces the graph and heatmap visualization of the plug-in. The GUI includes the collusion analysis that combine several sub-sessions for visualization and subsequent statistical analysis to detect collusion among various sub-sessions.

The development of statistical analysis tools contributed to finding collusion in stealthy cyber insider attacks. This is facilitated by displaying the data in visually meaningful ways (e.g., graphs and heatmaps) and correlate reading access data with functional models of fine-grain read access for each document that may indicate insider activity.

Specific objectives achieved:

1. Sensors: the Firefox plug-in (and the Javascript implementation) as the first reader/finder recorders, to record the time a reader (including cyber insider attackers) spends at each point of the document.
2. Test data set: a selection of open documents from various sources (e.g., Wikipedia) for testing of the finder recorder, annotated with document value curve.
3. Statistical analysis and visualization tools: the visualization component, functional models of reading behavior, API for representing such functional models, statistical analysis algorithms that correlate reading access data (the experimentally recorded reading time) with functional models (document value curve) to find the cyber insiders.

4.3 Malware Test Platform

One of the main challenges in the evaluation of our software tools is the generation of relevant stealthy cyber insider attacks. This is a challenge since no details of such attacks have been published in the open literature. Our approach to address the challenge is the development of a test tool, the Malware Test Platform (MTP), which is capable of generating the various combinations of colluding attacks, including the time division, space division, and value division, as well as their combinations.

In the time division collusion attack, portions of a document are read and copied by a cyber insider attacker over several sub-sessions. The time division attack is designed to be effective against typical DLP tools since it only takes a small amount of data each time, to stay below the threshold of DLP detection. The MTP generates read/copy sessions (by the same insider) at increasingly finer granularities to simulate time division attack.

In the space division collusion attack, portions of a document are read and copied by several cyber insider attackers in different sub-sessions. The space division attack is also designed to be effective against threshold-based DLP tools for the same reasons of time-division collusion attack. The MTP generates read/copy sessions (by different insiders) at increasingly finer granularities to simulate space division attack.

In the value division collusion attack, only the valuable portions of a document are taken by one or several cyber insider attackers. We assume the attacker does not have the capability to guess the valuable parts of the document without reading it first. Consequently, value division collusion attack includes sub-sessions that read and copy portions of the document. The MTP generates read/copy sessions of various portions (by the same or different insiders) to simulate value division attack and the combination of various collusion attacks.

The MTP is designed modularly to support other Generation 2 cyber insider attack modes if conceived.

4.4 Demonstration and Evaluation

Using the MTP, we are able to demonstrate the effectiveness of the software tools in monitoring document access and then combining the various sub-sessions for statistical analyses that detect collusions of various forms. The initial statistical analysis tools simply combine the sub-sessions for analysis, without concern about the particular collusion approach. This exhaustive search worked well up to about 7 sub-sessions,

when the combinatorial explosion caused the analysis to take longer than an hour on a PC.

To address the scalability challenge, we introduced several optimization techniques, e.g., eliminating the analysis of “unnecessary” combinations that have been covered in other combinations. These optimization techniques, plus simple heuristics such as flagging of very short sub-sessions (e.g., lasting only a few seconds) as immediately “non-human” access, make our analysis tools very close to covering all combinations.

For Generation 1 cyber insider attacks, the PPS system allows access only through authorized APIs, which consists of a single API: the Firefox plug-in. All other accesses are considered unauthorized and rejected. Thus our system is able to prevent Generation 1 cyber insider attacks, independent of any threshold.

For Generation 2 cyber insider attacks that use the Firefox plug-in API, we monitor the document access patterns at fine granularity (cursor and mouse pointer positions at intervals of a fraction of a second). The access patterns are analyzed for collusion. Our evaluation shows that our monitoring and analytical tools are completely effective in detecting Generation 2 cyber insider attacks generated by MTP. Details of evaluation are discussed in Section 4.6 below.

4.5 Concrete Observables

The project investigated primarily the three concrete observables listed in the table below, derived from the original observables and metrics. The first concrete observable is the document access through the PPS API. Assuming that PPS API has not been compromised, its use outside of our browsers immediately indicates cyber insider activity. The second concrete observable consists of document access patterns recorded by our browser plug-ins, e.g., the Firefox Javascript plug-in. This observable includes primarily read operations with associated metadata (time, location). The reading pattern indicates the reader’s ability to find high value information in the document. The third concrete observable includes the amount of data copied by the reader, which indicates the actual reader interest.

Table 2 List of Concrete Observables

Observable	Applicable Dimension(s)	Delivered Sensors	Other Possible Sensors
Document Access API	PPS document access API detection	PPS tools and access monitors	Other storage systems with access monitors
Document Access Patterns (Accumulated amount of time spent on each part of the document)	Evasive cyber insider detection (time division, space division) and collusion analysis	PPS document access pattern recorder for various file types and their readers	Click level recorders, which would require significant post-processing
Copied Content / information	Clipboard operations that copy document content for potential exfiltration	OpenDLP and PPS content copy recorder for editors of PPS documents	Other DLP tools, if applicable

4.6 Concrete Metrics for Evaluation

Table 3 Concrete Metrics for Evaluation

Metric Area	Metric	How Measured	Success Criteria	Current State of the Art
Data Access Detection (effective)	D1. Machine PPS interface access detection	PPS data structure accessed through the native (machine) interface instead of a GUI.	Immediate detection of unauthorized document access by MTP.	N/A currently
Data Access Detection (effective)	D2. Access volume (bulk) detection	A large number of documents are accessed within a short period of time.	MTP accesses 20 documents within a 1-hour window. (Tunable threshold)	Available in DLP tools
Data Access Detection (effective)	D3. Access speed detection	Rapid scan of documents (too fast for human access).	MTP accesses 20 documents within a 5-minute window. (Tunable threshold)	N/A currently
Data	D4. High value	Scan of documents	MTP accesses K	N/A

Access Detection (effective)	access detection	that find high value parts.	valuable parts of a document. (Tunable threshold)	currently
Data Access Detection (effic/scal a)	D5. Access detection overhead	Measurement of response time for PPS data access operations	MTP access to PPS documents with response time under 100ms	Similar to file system, but N/A for PPS
Collusion Detection (effective)	C1. Time division combinatorial detection	Combinatorial analysis of multiple sessions by single insider accessing parts of a document sequentially.	Detection of multiple MTP sessions that cover the entire document (up to 20 sessions)	N/A currently
Collusion Detection (effective)	C2. Space division collusion detection	Combinatorial analysis of multiple sessions by multiple insiders accessing parts of a document possibly in parallel.	Detection of multiple MTP sessions that cover the entire document (up to 20 sessions)	N/A currently
Collusion Detection (effective)	C3. Value-driven collusion detection	Combinatorial analysis of multiple sessions by multiple insiders accessing parts of a document in sequence and/or in parallel.	Detection of multiple MTP sessions that cover the most valuable parts of a document (up to 20 sessions)	
Collusion Detection (effective)	C4. Time/Space combination detection	Combinatorial analysis of multiple sessions by multiple insiders accessing parts of a document in sequence and/or in parallel.	Detection of multiple MTP sessions that cover the entire document (up to 20 sessions)	N/A currently
Collusion Detection (effective)	C5. Time, Space, Value combination collusion detection (T/V, S/V, T/S/V)	Combinatorial analysis of multiple sessions by multiple insiders accessing parts of a document in sequence and/or in parallel.	Detection of multiple MTP sessions that cover the most valuable parts of a document (up to 20 sessions)	N/A currently
Collusion Detection (effic/scal a)	C6. Scalability of naïve combinatorial analysis	Run the analysis for all combinations of a set of sessions.	Find maximum number of sessions (K) that can be combined in exhaustive analysis	N/A currently

Collusion Detection (effic/scalability)	C7. Heuristics to improve scalability of collusion analysis	Use heuristics to improve detection efficiency (e.g., most likely collusions first)	Increase the number of sessions that can be combined in heuristic analysis to 5K	N/A currently
Collusion Detection (effic/scalability)	C8. A variety of statistical correlation metrics and aggregation methods for collusion analysis	A number of different metrics that can be used in evaluating the likelihood of collusion to increase robustness against adversarial attack tuning and specification	At least five separate statistical correlation algorithms are integrated in collusion analysis and combination analysis	
PPS Storage System (effective)	P1. Storage and retrieval of various document formats	PPS support for HTML, PDF, Word, and text formats	Demo of document access for the target formats; realistic evaluation tests using MTP	N/A currently
PPS Storage System (efficiency)	P2. PPS data access indistinguishable by human reading	PPS access (read / write) unnoticeable at human reading speed	A PPS document accessed and read via the GUI should take no longer than twice load time from normal systems	N/A currently
PPS Storage System (qualitative scalability)	P3. Extensible and flexible PPS metadata support	PPS metadata storage and query supports the analyses in Generation 2 and beyond	PPS metadata extensions will support all the analyses outlined in this metrics table	N/A currently

We acknowledge that these concrete metrics are a subset of and derived from the original observables and metrics developed in collaboration with the MIT Lincoln Labs independent verification and validation (IV&V) team. Furthermore, we acknowledge that the actual metrics used in the evaluation of delivered software tools (the Fine Grain Document Modeling system) and MTP are a subset of, and derived from the concrete observables and metrics listed above. This is due to the instantiation of evaluation metrics to the specific and actual implementation of software tools and MTP, which support a substantial subset of these metrics and in some cases more detailed metrics derived from these listed. Using the derived actual metrics, we have demonstrated the effectiveness and scalability of our software tools, with positive conclusions on the derived metrics.

The IV&V team conducted independent tests on the delivered software tools using the derived actual metrics that form a subset of concrete metrics listed, as described in their report. Examples of metrics that were not directly supported by MTP and so could not be tested include C6, C7, and C8 (and their scalability). Within the context of the actual subset of metrics tested and verified, we quote the first paragraph of their Result Summary (page 2):

Overall, the tests showed the MTP successfully produced the anomalous activity and the reading patterns were captured by the Fine Grain Document Modeling system. For the remainder of the paper, both the Fine Grain Document Modeling system and the MTP will be called the System Under Test (SUT). The SUT was able to simulate and detect single and multiple agents accessing single and multiple documents in short bursts or over extended periods of time. The system also detected access to high value portions of documents by single and multiple agents during short or prolonged access, as well as access to different portions of the same document over time.

5 Conclusion

This project built software tools to store documents in the PPS system, monitor accesses to PPS documents, and analyze the access patterns to find cyber insider attacks. We use a simple API access authorization scheme to thwart Generation 1 cyber insider attacks (direct data transfers). A combination of fine-grained access monitoring data and statistical analysis on combined sub-sessions supports the effective detection of Generation 2 cyber insider attacks (minimized/targeted data transfers through legitimate APIs).

6 References

6.1 Software tools and documents

VMware Workstation server image (all software):

Available by request to authors and upon approval by sponsor.

Malware Test Platform (MTP):

Available by request to authors and upon approval by sponsor.

Malware Test Platform Documentation:

Available by request to authors and upon approval by sponsor.

Test Bed Manual:

Available by request to authors and upon approval by sponsor.

Appendix A: MTP Manual

Getting started

In order to begin using the software that we have developed, boot the VMware image in VMware Workstation. Upon booting, the accounts that can be used to log in are as follows

- Chris Grayson // Delicious_ham_sandwich
- root // Delicious_ham_sandwich

Upon logging in, both the MySQL and Apache HTTP daemons will need to be started. In order to do so, open up terminal and do the following.

- su root; cd /etc/init.d; ./mysqld start; ./httpd start; (if logged in as Chris Grayson)
- cd /etc/init.d; ./mysqld start; ./httpd start; (if logged in as root)

Once all of this is done our software should be ready to go.

Creating an account for our PPS system

Next, in order to use the malware test platform or any of the other tools we have developed, you will need to create an account through the web-based GUI. To do so, simply do the following:

***NOTE** When using the web interface for any activity, it is best to zoom it out to the proper zoom level. This is typically done by holding CTRL and scrolling the mouse wheel. The proper zoom level is shown below.*



The screenshot shows a web interface titled "Welcome to the Infosphere Research Project!". It contains two side-by-side forms. The left form is titled "Log in" and has two input fields labeled "E-mail" and "Password", followed by a "Submit" button. The right form is titled "Register" and has four input fields labeled "E-mail", "Password", "Re-enter Password", and "Registration Code", followed by a "Submit" button.

Figure 1 Screenshot of Login Page

1. Open up a web browser
2. Navigate to <http://localhost/info2>
3. Enter in all necessary credentials in the "Register" box. The registration code to use is "Delicious_ham_sandwich"
4. Hit the "Submit" button
5. On the subsequent page, scroll to the bottom and click on the "I Agree" button
6. Upon hitting the "I Agree" button your account creation will be confirmed and you will be redirected to the log in page.

You now have an account with our PPS system. In order to access any other tools through the web GUI, navigate to <http://localhost/info2/login.php>, enter your credentials, and log in.

Uploading a document

To upload a document, do the following:

1. Log in to the web GUI
2. Click on the "Upload document" button in the "Upload a new document" box.
3. On the subsequent page, click on the "Browse..." button to find the document you would like to upload on your local.
4. Once you have found and selected the appropriate document, you can enter a description of the document in the allotted space.
5. Once you are satisfied with your description (not having a description is fine) click on the "Upload" button
6. After an amount of time proportional to the size of the document you should receive a message stating that your document was successfully uploaded.

Viewing a document

To view a document, do the following:

1. Log in to the web GUI
2. In the "Read a document" box, select the document you would like to read from the drop down
3. Once you have the document selected, click on the "Start reading" button
4. Allow the document to fully load before attempting to scroll / read
5. Once you are done reading, click on the "Done reading" link in the bottom right-hand corner of the page

Annotating a document

To annotate a document, do the following:

1. Log in to the web GUI
2. In the "Annotate and existing document" box, select the document you would like to annotate from the drop down
3. Once you have the document selected, click on the "Start annotating" button
4. Allow the document to fully load before attempting to scroll / read / annotate
5. To annotate a given region, hold down the CTRL key and click and drag the resulting box over the area of the document you would like to annotate.
6. Once you release the mouse key, the box will become dark blue. At this point you can assign a value to the annotated area using the arrow keys in the top right-hand corner of the blue box. Should you want to delete the box, click on the 'X' in the top right-hand corner of the blue box.

7. Once you have annotated all sections of the document that you would like, click on the "Save annotation" link in the bottom right-hand corner of the page

Viewing your reading patterns

To view the reading patterns that you have created using the PPS GUI system, do the following:

1. Log in to the web GUI
2. In the "View document reading patterns" box, select "View own reading patterns" from the drop down
3. Click on the "View access patterns" button
4. Once on the reviewing tool page, you can select how you would like to view your own reading patterns from the drop down ("View individual patterns" or "View patterns joined by document")

Logging in to the Malware Test Platform

To log in to the malware test platform, do the following:

1. Create an account with the PPS system using the web GUI
2. Open the malware test platform by double-clicking the icon on the desktop
3. Log in to the malware test platform using the credentials you provided upon registering via the web GUI

Deploying malware to the PPS data structure

To deploy malware to the PPS data structure, do the following:

1. Log in to the malware test platform
2. At the main menu for the MTP, click on "Deploy malware"
3. On the deploy malware frame, choose a malware type that you would like to deploy from the drop down list. It should be noted that some of the types of malware depend on there being multiple users in the DB and/or annotations for documents.
4. Fill out the GUI components for the type of malware you would like to deploy
5. Click on the "Deploy" button a single time
6. Watch the log to make sure everything was deployed successfully

Removing malware from the PPS data structure

To remove malware from the PPS data structure, do the following:

1. Log in to the malware test platform
2. At the main menu for the MTP, click on "View deployed malware"
3. On the monitor deployed malware frame, select the row in the deployed malware table that corresponds to the malware you would like to remove
4. Click on the "Delete malware" button

5. Watch the log to make sure everything was removed successfully

Analyze PPS structure for malware activity

To analyze the PPS structure for malware activity, do the following:

1. Open a browser and navigate to <http://localhost/info2/libraries/analysisPage.php>
2. Choose an analysis type corresponding to the mission and metrics documentation from the drop down
3. Choose a sub-type from the second drop down
4. Fill out necessary criteria depending on the analysis you plan on running (some analyses require no additional input). Please note that input is not validated and as such input should be properly formed by the user (this is straight-forward, as IDs are integers and valid ranges are specified).
5. Click on the "Scan" button

Please note that some of the analyses will take longer than others, and some will run in to scalability issues. For testing purposes, please evaluate functionality and hold off on scalability until we provide an updated VM.

Database structure

To access the database, use the user name "root" with the password "new-password".

The database schema is as follows:

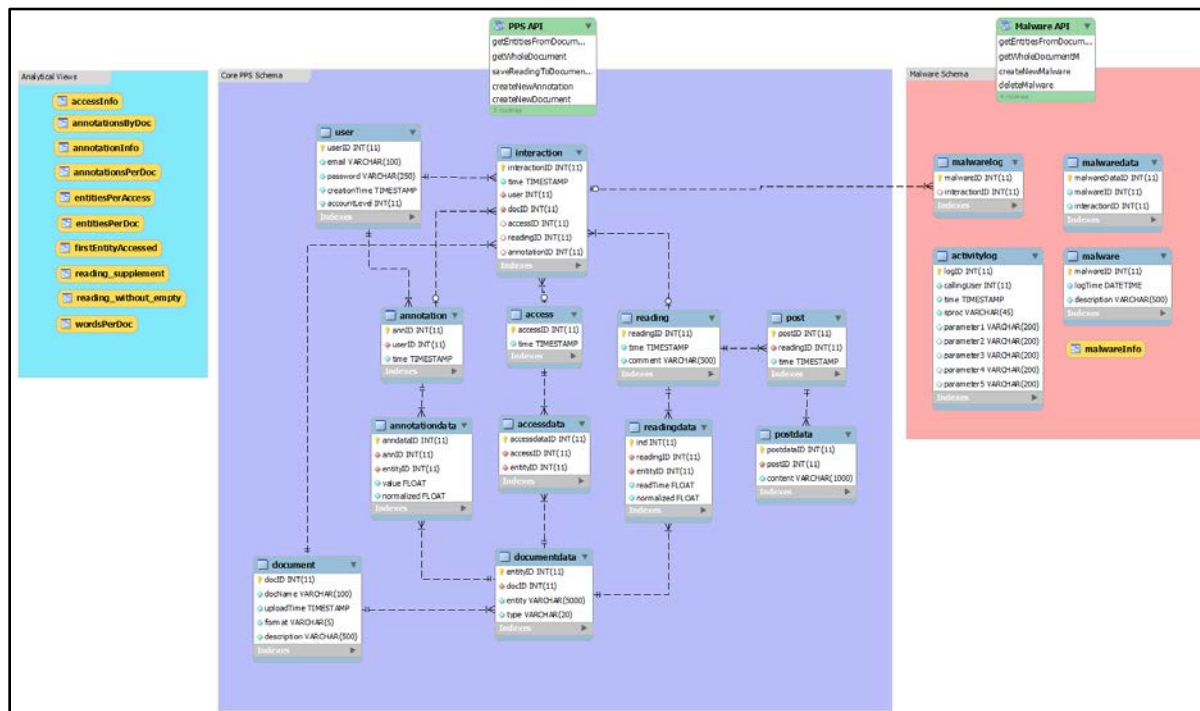


Figure 2 PPS System Architecture

List of Acronyms

- API: Application Program Interface, software module access interface definition for other programs to use this module.
- DLP: Data Loss Prevention, techniques and software tools to detect unauthorized data sharing, typically through messages sent over networks.
- GUI: Graphical User Interface, a program used by human users to access software tool functionality.
- HTML: hypertext markup language, widely used web document format.
- HTTP: a network protocol for transmitting messages over the World Wide Web.
- IPR: Interim Program Review, project review during the project execution.
- IRB: Institutional Review Board, a board that monitors and approves human-related research.
- MTP: Malware Test Platform, a software tool that generates cyber insider attacks for many documents.
- PDF: Portable Data Format, a standard document representation format.
- PPS: Partial Persistent Sequence, a customized data structure to store data and metadata of documents; the current implementation uses a relational database management system (MySQL).
- T/V, S/V, T/S/V: combination of cyber insider collusion attack modes: T=time division, S=space division, V=value based. T/V means time and value combination. S/V means space and value combination. T/S/V means time, space, and value combination.
- TXT: plain text file type for storing documents in common operating systems such as Windows.

Glossary of Technical Terms

- Apache: an open source web server, used in the project to manage GUI and web accesses.
- Generation 1: cyber insider attacks that use direct/bulk data transfers.
- Generation 2: cyber insider attacks that stay below certain thresholds by using collusion methods.
- MySQL: an open source relational data management system, used in the project to implement PPS.
- OpenDLP: open source software tool that implements many DLP functions.
- VMware: a generic name for the virtual machine environment provided by a family of commercial virtual machine monitors produced by the company of same name.