IPRO 311 Final Report

IPRO 311

1 Planning

1.1 Problem

The main problem the IPRO addressed was that of insider misuse. Contrary to popular belief, the largest threat facing companies on a technology level is not intrusion, but rather insider misuse. Insider misuse is a relatively new area of research in the field of Computer Science, which is one of the reasons it's a relatively unheard of problem. In addition to that, due to intense media bias towards intrusion, people are blissfully unaware of the problem of insider misuse.

Ethically, it's an interesting problem. Companies already closely monitor user activity, but any infringement on perceived privacy in the workplace generally sends everyone flustering left and right. The IPRO had to personally evaluate if we felt the necessary invasion of privacy was worth the overall result which is a more secure and profitable business. Companies can decide which approach they would like to take, but it's hard to believe that a company would choose one that wouldn't help to prevent a loss of money.

1.2 Objectives

The Spring 07 semester of IPRO 311 will focus on the task of implementing a system for capturing and comparing user habits in order to create a misuse detection system that determines if a user is using a computer for nefarious purposes. But what is computer misuse? Generally, when people think about computer crime, they think of the typical 14-20 year old hacker, sitting in his parents cold, dark basement breaking into a CIA mainframe. And while this is a serious problem, it isn't the most common computer crime. Exceedingly more common is the computer misuser, who can fall into two categories. The first is your typical "slacker" who plays flash games all day instead of doing work. The other, more dangerous computer misuser, is the one who abuses their privileges, accesses generally restricted information, and uses it for nefarious purposes. For example, a higher-up exec in a weapons facility is allowed to look at the research for the newest weapon system. Him taking that information and selling it to competitors would be an example of misuse.

This is the first semester of this IPRO, so there is no background information to refer to.

1.3 Task Definition and Durations

NOTE: please see attached Microsoft Project document for the updated WBS. Included in the document are updated summary tasks and hourly durations. For individual tasks, please see §2.5.2

2 Organizing

2.1 Accountability

Team Leader – Yacin Nadji
Software Development Team Leader – Jason Soo
Design Team Leader – Justin Choriki
Query Processing Team Leader – Jong Min Lim

2.2 Software Development Team

The software development sub-team is responsible for two major things. First, they need to develop the automated query logging system to record the following:

Items to record at the Operating System level

* Snapshot of files+binary diff of daily changes
* All keystrokes

  Saved to a file(keylog.txt) with a timestamp every minute

  ```
  Example:
  200701051231:this is test text, this is test text.....
  200701051232:
  200701051233:some more test text after 2 minutes
  ```

* All newly saved files(full listing and actual file)

  Files will be saved in folder saved_files/ with format filename.[usb|floppy|cdr|hd].version, the files will retain the original permissions and ownership as original file

  ```
  Example:
  ls -lh
  -rw-r--r--   1 jwilberd jwilberd   18 Sep  8  2006 testfile.hd.1
  -rw-r--r--   1 jwilberd jwilberd   18 Sep  8  2006 testfile.hd.2
  -rw-r--r--   1 jwilberd jwilberd   18 Sep  8  2006 testfile.hd.3
  -rwxr--r--   1 jwilberd jwilberd   18 Sep  8  2006 testfile2.usb.1
  ```

* All newly opened files(full listing and actual file)
  * Format will be the same as saved files only folder will be called opened_files/
* All newly created files(full listing and actual file)
  * Format will be the same as saved files only folder will be called created_files/
* Running processes
  * Stored in process_list.txt, with timestamps

        * Initial process list
        * New processes
        * Ended processes
    * Memory usage
        * Stored in memory_usage.txt, with timestamps
        * 1 sec intervals
    * CPU usage
        * Stored in cpu_usage.txt, with timestamps
        * 1 sec intervals
    * Disk usage
        * Stored in disk_usage.txt, with timestamps
        * 1 sec intervals
    * Mouse movments

      Record mouse position x,y every second to mouse.txt with timestamp

      ```
      Example:
      20070105123101:100,201
      20070105123102:567,231
      20070105123103:120,350
      ```

    * Raw network data
        * Save to network.txt using pcap's internal format

Items to record at the Information Retrieval level

* The queries
    * Order of queries
    * Time between queries
    * End of session indicator
* Top retrieved documents(20-100)
    * HTML
    * URL
    * Document Ranking
* Click through data
    * Length each link was visited
    * Amount of text was read(scrolled)
    * Order documents were clicked

* External links followed(capture same data as initial link)
* User assisted input
    * User's subjective ranking of documents
    * Record the nature of each task
    * Record purpose of each query
* Each system event will be recorded in a database
    * date
    * time
    * user
    * event type
    * time since last system event
    * time since the last system event of this type
    * query id
* Each system level event can be then mapped back to the query whose results the user was viewing when the event occurred based on the user, date and time the query was issued

After these necessary tools have been implemented, the sub-team will need to code a series of scripts to present the information from several different angles, most notably one allowing the Query Processing Team to make use of the data to run a series of ROC curves to analyze progress. It's also necessary to present the information logged in such a way it will be presentable to anyone at any education level.

Members:
Jason Soo
Yacin Nadji
William Alton
Matt Holmes

2.3 Design Team

The Design Team's responsibility is to design and implement the website, brochure, poster, and final presentation. The individuals were selected based on their technical merit in the design field, and ability to succinctly convey information. The sub-team tasks are not as time consuming as the other sub-team's, so they will help the other sub-teams as necessary throughout the course of the project.

Members:
Justin Choriki
Mark Malanowski
Daniel Hyc
Hee Yeol Jeong

2.4 Query Processing Team

The Query Processing Team is essential to taking all the collected data, throughout the semester, and accomplishing 3 main tasks:

1. Make sure the logging system is functional
2. Analyze the data, searching for patterns between proper use and misuse
3. Present the findings weekly to the team

Members:
Jong Min Lim
Young Cho
Peter Niedzinski
Gerardo Sanchez

2.5 Role and Resource Allocation

2.5.1 Budget

No budget is necessary for the completion of this project. All code not directly developed by the teams have been complimented by various Free Software solutions.

2.5.2 Roles Allocated to Individual Members

Name	Major	Skills & Strengths	Role & Tasks
Yacin Nadji	Computer Science	Programming, Operating Systems, Public Speaking, Management	IPRO Team Leader, Responsible for implementing OS level logging and maintaining the Knoppix bootable CD
William Alton	Computer Science	Programming, Data Mining	Develop IR Level logging and maintain the code-base
Young Cho	Applied Mathematics	Statistical Analysis, Databases	Providing up to date analysis of current query records
Justin Choriki	Mechanical Engineering	Art, Design, Webdesign, Layout	Design Sub-Team Leader, oversees completion of the web site, brochure, poster and final powerpoint presentation
Matt Holmes	Computer Information Systems	Python scripting, text parsing	Parsing log files into presentable format for the Query Processing sub-team
Hee Yeol Jeong	Electrical Engineering	Mathematical and statistical analysis, design, physics	Handling the presentation of the statistical information for the design team. Acts as a liaison between the Design and Query Processing team
Jong Min Lim	Electrical Engineering	Leadership, statistical analysis, circuit design, applied mathematics	Leader for Query Processing sub-team, oversee query processing and participate in statistical analysis
Mark Malanowski	Computer Science	PHP/MySQL, dynamic web programming and design, scripting languages	Back-end web design (PHP/MySQL) and scripting support for Software Development team
Peter Niedzinski	Biology	Public speaking, writing, applied mathematics	Aid in statistical processing and present the information to the team on a regular basis. Provide textual material for IPRO Day. Taking the minutes.
Gerardo Sanchez	Civil Engineering	Applied mathematics and statistical analysis	Aid in statistical analysis of the results
Jason Soo	Computer Science	Scripting, data processing, information retrieval	Software development subteam leader, oversee and develop scripting for data collection
Daniel Hyc	Computer Science	Web design, scripting	Head web developer

2.6 Methodology

Based on the team breakdowns, determining where individuals needed to apply their work was relatively simple. Below is a short breakdown of the major accomplishments handled by the individuals:

Individual	Specific Tasks Completed
Yacin Nadji	OS level logging, Knoppix disc remastering, server Maintenance, querying
William Alton	IR level logging, dataset testing scripts, graph scripts, querying
Young Cho	queries, slides, statistical analysis
Justin Choriki	poster, brochure, queries
Matt Holmes	dataset testing scripts, querying
Hee Yeol Jeong	poster, brochure, queries
Jong Min Lim	query processing, statistical processing, queries
Mark Malanowski	website-related search, website back-end functionality, presentation, queries
Peter Niedzinski	presentation, queries
Gerardo Sanchez	work-breakdown schedules
Jason Soo	dataset testing scripts, logging scripts, presentation, queries
Daniel Hyc	main website design, brochure, poster, final presentation

2.7 Acknowledgments

One of the main references was the paper "*Misuse Detection for Information Retrieval Systems*" by "R. Cathey, L. Ma, N. Goharian, D. Grossman". At the beginning of the semester, it gave us a nice idea of what the problem was, and laid out a few potential solutions that have been tried in the past to let us build up on our own solution. One of the nicest parts was our faculty instructors helped co-author the paper, so for any questions we might've had, we could get the answer directly from the source.

2.8 Table Of Contents

Required Deliverables

1. Abstract → ipro311_abstract.pdf

   A brief introduction to the plan of action for this IPRO. We addressed the problem and gave a potential solution.

2. Project Plan → ipro311_project_plan.pdf

   The Project Plan was an extension of the Abstract, outlining directly how we wanted to approach the problem. It was obviously referenced multiple times throughout the course of the project, and changed as well.

3. Midterm Report → ipro311_midterm_report.pdf

   The Midterm Report provided an estimation of where we are in the course of the project thus far, any changes we have made to our solution, and any potential problems that may arise.

4. Meeting Minutes → ipro311_minutes.pdf

   The Meeting Minutes describe what happened during each meeting, and were used as a reference point for those not present, and those who had deliverables to address for the remainder of the week.

5. Final Presentation → ipro311_presentation.ppt

   The final presentation is a representation of everything the IPRO has done. We started working on the presentation as early as the second month of the IPRO, and slowly but surely, began to fill it with more information.

6. Final Report → ipro311_final_report.pdf

Supplemental Information

7. Team Information → ipro311_team_info.pdf

   IPRO 311 Team Info breaks down each member of the team, the responsibilities they had to undertake, and sub-team responsibilities.

8. Website (not included) → http://www.iit.edu/~ipro311s07

   The website provides very concise information for what the IPRO accomplished, and easy ways to check the results we determined.

9. Poster Image → ipro311_poster.pdf

10. Query Reports → ipro311_query_reports.pdf

    The Query Reports are one of the most influential pieces of the project. Each member was required to issue queries on a certain topic, and based on the information they gathered, they completed a report outlining the information they gathered. This allowed us to collect very realistic queries to attempt to detect misuse.

3 Concluding

3.1 Obstacles

Even a "perfect" project encounters problems along the way, and IPRO 311 was no exception. One of the problems we had that just about everyone IPRO experiences, is one of a select few team members not pulling their weight. It's a natural occurrence in most projects, so fixing this problem was relatively easy, compared to the others. We simply had to keep tabs on our teammates. If I knew the poster is going to be due in a couple of days, I'll ask the Design Team if things are going alright, and if they need any help. It's a very polite way to check in on teammates, and offer support if it's necessary as well. In addition to this, members needed a minimum of 4 hours completed by the Thursday weekly meeting (unless the time issue was previously addressed to the faculty). By Sunday, each member needed to submit a status report detailing all the work completed over the week.

Aside from this, we ran into a few technical problems, namely server trouble. First, the server we were logging information too was kicked off the network and unjustly blocked by OTS. We had to

quickly scramble to write the scripts to support multiple remote hosts, and simultaneously updated to a different machine, while transferring the original server to a new location. After things settled down a few weeks into the semester, data collection continued without a hitch, and allowed us to amass the large data set that we did.

3.2 Results

While nothing has been officially concluded from our results, they certainly show a positive trend which can definitely elevate the foundation we have set up to perpetuate into success. It's lofty thinking to believe a first semester IPRO will accomplish all its goals, let alone any project, but the framework the IPRO has laid down made significant strides in the field of automated misuse detection.

The dataset we generated, amounting to a total of 13 gigabytes in size, is the *only* naturally created misuse dataset currently existing. This, in and of itself, is a sufficient accomplishment, and with the proper data mining techniques, could be utilized into developing automated misuse detection systems. To determine the validity of the dataset, we wrote very simplistic information retrieval algorithms to see if misuse could be detected. On the Operating System side, we were able to detect most instances of misuse using a very simple algorithm (checks to see if statistics lie outside of 3 standard deviations of the mean) and it only returned 1 false positive. On the Information Retrieval side, our results were less positive. This isn't terribly surprising as naive algorithms generally perform *very* poorly. Based on the OS level results, it's safe to say with more developed algorithms, the results will be much more promising.

3.3 Conclusion

IPRO 311 has laid down some very positive groundwork to continue on in later semesters. The main goal of the IPRO has been completed successfully, which was created a non-synthetic misuse dataset to be used for future research and development endeavors.

3.4 Recommendations

Based on the collected dataset, it is imperative research is continued. Due to the field of misuse detection being quite new, it's downright stupid to ignore the potential industrial applications of such a detection system. As the first group to develop such a dataset, it's imperative we also be the first group to generate methods to actively scan for potential misuse. The next steps are to both study the dataset and help it continue to gain a larger amount of user profiles. Intense study of the dataset will yield potential areas of improvement, and also determine possible approaches to developing the software to detect misuse.

3.5 Teamwork

In the beginning, our team had a difficult time "gelling". With a large group comprised mostly of strangers, it's difficult to jump into group mode to get things done effectively. Naturally, with time, things started looking up. The primary way we began truly functioning as a team was to spend

time outside of the IPRO just hanging out, and learning more about each other. The interesting thing you learn about people is as different as we are, we're all human and are generally interested in similar topics. After a few out-of-work sessions, and we were working together like a well oiled machine. The team feels confident in its members, and many of us have become friends outside of the context of the IPRO program.

3.6 Communications

The team corresponded for the most part over email and AIM chats in the beginning, but we slowly gravitated towards more friendly forms of communication such as the phone and face-to-face communication. As mentioned above, the time spent together outside of the IPRO really allowed us to be more comfortable with each other, and as such, more direct and personal means of communication quickly became more prevalent. As soon as the team started working like a team, we also communicated like a team; quickly and effectively.

3.7 Ethical Behavior

For more insight into the ethical issues surrounding the team, please see §(3.1)