University of Amsterdam

System and Network Engineering, MSc

Research Project 1

Security Intelligence Data Mining

Research Proposal

Diana Rusu
Diana.Rusu@os3.nl

Nikolaos Petros Triantafyllidis
Nikolaos.Triantafyllidis@os3.nl

January 18, 2015

Abstract

Acknowledgement

We would like to express our gratitude and appreciation for all support, expert knowledge and guidelines during this research project. It has been a great opportunity for us to explore the real feel of working with a successful company.

Who should we mention?

- dhr. prof. dr. ir. C.T.A.M. (Cees) de Laat - for the project proposals
- Delloite NL specifically to our first supervisor Henri Hambartsumyan - for giving us access inside their company and offices, for all the experts and employees that we had the chance to meet
  - Henri Hambartsumyan - who proposed Security Intelligence Data Mining
  - Kremers, Joost - supervisor
  - Niels - Cyber Threat Intelligence project
  - Ari Davies - proposed making the plugins also more depth in CTI
  - Gijs Hollestelle - presented the actual CTI platform and the portal to it, proposed phishtank and cleared out the purpose for this project(making some agents for IR system)
- dhr. dr. C.P.J. (Karst) Koymans

Contents

Introduction

With the increasing number of cyber-attacks and the growth of computer crime worldwide, it becomes apparent that IT security is a major concern and crucial survival factor for large companies, organisations and institutions of any sort. Security Operations departments working to ensure confidentiality, integrity and availability for the system infrastructure of their organisation, invest huge parts [1] of their time and effort in detecting threats in real time. A very valuable source of security intelligence, vital to cyber-risk assessment, is information mined from data posted on public sites such as "pastebins" or social networks. However, this is a very cumbersome task due to the lack of Natural Language Processing capabilities in most of the existing tools. Moreover, as recent events have showcased [2], several threats arise from governments and criminal associations originating from countries whose languages use non-latin scripts (Chinese, Russian, Korean, etc.). It is, hence, important to have data mining tools that provide support for such alphabets and languages, since a lot of a security intelligence can be discovered in such texts. The main goals of this research project will be to explore the various public data and detect the most appropriate among them. Moreover, numerous current data analytics techniques as well as their application on security related issues will be assessed. Lastly the above knowledge will be applied on the implementation of a simple system that will work as a proof-of-concept and help determine the technical feasibility, storage requirements and operational cost of such a system. This project was proposed by and will carried out in co-operation with Deloitte NL.

1 Research Questions

This topic is admittedly very open but it can be narrowed down to several specific research questions some of which we will try to answer to some extent. The main question that we will be trying to answer is the following:

How can we effectively use public sources to obtain real time information about security incidents?

This question can be analysed into more specific parts that cover the topic to some extent, as follows:

1. How can the raw data be effectively collected from the public sources?
    - How can we effectively detect the reliable sources?
    - What search terms can we deploy during the retrieval phase?
    - How can the unstructured data be pre-processed?
2. How can the data be analysed in respect to security operations?
    - How can we apply current Data Mining and Analytics techniques on Security issues?
    - How can we derive the risk assessment model from the above?
    - How can we apply the model on new data?
3. How can the collected knowledge be applied on a system implementation?
    - What is a reliable and extensible System Architecture that can be designed?
    - What are the computational and storage requirements of such a system?
    - What extensions can be proposed for that system?

The proposed system extensions can spawn further research questions, namely on the topics of presenting the analysed data, reacting to the real time events and finally assessing the situations that arise and providing feedback to the system.

2 Related work

There is a lot of literature around the field of Data Mining and more recently Web Mining. The most prominent and recent case is the book 'Mining The Social Web' by M.Russel [3] that deals with exploring and mining information from social websites (e.g., Facebook, Twitter, LinkedIn, Google+, GitHub, etc.). There are also several academic papers and books that deal with applying Data Mining to System Security. One example is a system proposed by the university of Minnesota, called MINDS, that employs various Data Mining in Intrusion Detection. The system is described in their paper 'Data Mining for Cyber Security' [4]. Another example is a system proposed by the Dutch company Sentient in co-operation with the Amsterdam Police Force [5] aiming to provide Data Analytics operations automation while on the same time minimising the technical expertise needed by the system user.

3 Experimental installation requirements

Most of the work is going to be carried out on end workstations (desktops, laptops, etc.). Depending on the amount of data collected, additional computational or storage resources might be required. In that case our assigned OS3 servers can be used. As for software requirements, there are several open software tools (Database Systems, IR Systems, Web Crawlers, etc.) that can help us carry out the work. For each component of the system that has to be implemented manually the appropriate programming languages as well as libraries are going to be selected. One example is the very strong Python NLP-toolkit.[6]

4 Information Retrieval System

This chapter introduces a system for collecting information from public sources. First subsection will define what websites are providing interest for detecting threats, malwares and phishing sites. For this specific sources next subsections come as an extension for further investigation and implementation.

4.1 Defining reliable sources

For many companies finding in real time possible attacks became a crucial point. Discovering this type of dangerous threats with a plausible margin of error, it is a challenging task as Internet provides an enormous database with "random messages/information". Hence it is important to gather all data needed from reliable sources. Special attention has to be provided in establishing which are the optimum sources that will provide the seeking information.

First, what is meant by reliable source? Should any type of source or social network be considered as reliable? What about forums and IRC channels? Should we draw our attention over this once as well?

How many real threats or warnings have been encountered till now on the websites that we want to consider as reliable?

Phishtank

4.1.1 API Investigation/Inspection

4.1.2 Collecting data

4.1.3 Implementation

Pastebin

Pastebin [7]it is a webiste where everyone with or without registration can share real time text or code snippets. It attracted many users during past years including malware writers. The enormous flow of information include from database dumps containing e-mails and passwords to harmful backdoor programs. At a deeper examination of the public messages pasted some possible future attacks can be determined.

4.1.4 API Investigation

As for each API, Pastebin offers few options for developers once they obtain the unique Developer API Key and this include creating new pastes, listing trending pastes and pastes from a particular user. Unfortunately, Pastebin API it is not completely promising when it comes to information retrieval from database. Search method using keywords it is provided by "Google Custom Search". Therefore, if one desires to obtain full database it would have to use Google's API. It comes with a price those, if one wants to use it as a free user will be limited to 1000 pastes per day.

4.1.5 Collecting data

4.1.6 Implementation

Twitter

4.1.7 API Investigation

4.1.8 Collecting data

4.1.9 Implementation

Reddit

4.1.10 API Investigation

4.1.11 Collecting data

4.1.12 Implementation

5 Data Analyse

5.1 Threat Types/What are we looking for? Security related data

5.2 Defining data mining model

5.3 Implementation

Conclusions

Implications

Further Research

Ethical implications

The main part of this research comprises of exploring current techniques and their application on IT security, as well as the specification of a system that employs Data Mining techniques to collect security intelligence. In order for the models to be defined some amount of information will have to be gathered. This information will originate solely from public sources and will be mostly historical data. In the unlikely case that any previously unnoticed security issues are encountered they will be handled with discretion and communicated only towards the appropriate targets and only with the approval of the OS3 core team and Deloitte Digital. The collection or storage of personal data is not intended and any collected information will be discarded after the end of this project. The usage of shared computational and network infrastructure will only be used for the needs of this project and within the legal limits.

Appendices

References

[1] Anon, (2015). [online] US cybercrime: Rising risks, reduced readiness, PWC, Available at: http://www.pwc.com/en_US/us/increasing-it-effectiveness/publications/assets/2014-us-state-of-cybercrime.pdf

[2] D. Sanger, N. Perlroth. U.S. Said to Find North Korea Ordered Cyberattack on Sony. [online] Nytimes.com. Available at: http://www.nytimes.com/2014/12/18/world/asia/us-links-north-korea-to-sony-hacking.html?_r=0

[3] Russell, MA (2014). Mining the Social Web, O'Reily Media, USA

[4] V. Chandola et al. Data Mining for Cyber Security, Department of Computer Science, University of Minnesota, Springer, 2006

[5] RCP van der Veer, H.T. Roo, A. van der Zanden, Data mining for intelligence led policing, Sentient, Amsterdam Police Force, Amsterdam, The Netherlands, 2009

[6] Nltk.org, (2015). Natural Language Toolkit - NLTK 3.0 documentation. [online] Available at: http://www.nltk.org/

[7] Pastebin, (2015). Pastebin.com - #1 paste tool since 2002!. [online] Available at:http://pastebin.com/