LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-480340

"Problems in Web-Based Open Source Information Processing for IT Early Warning"

K. Grothoff, M. Brunner, H. Hofinger, C. Roblee, C. Eckert

April 13, 2011

Web Intelligence for Information Security Workshop 2011
Lyon, France
August 22, 2011 through August 27, 2011

Disclaimer

Problems in Web-Based Open Source Information Processing for IT Early Warning

Krista Grothoff*, Martin Brunner†, Hans Hofinger†, Christopher Roblee‡ and Claudia Eckert*†
*Faculty of Informatics, IT Security Lab, Technische Universität München
Email: {kgrothoff@sec.in.tum.de, claudia.eckert@sit.fraunhofer.de}
†Fraunhofer Institute for Secure Information Technology SIT, Garching bei München, Germany
Email:{martin.brunner, hans.hofinger, claudia.eckert}@sit.fraunhofer.de
‡Lawrence Livermore National Laboratory – Email: cdr@llnl.gov

***Abstract*—IT early warning systems are a promising paradigm for early detection and mitigation of network attacks. While network-based indicators of emerging attacks are currently being exploited in the development of network early warning systems, other data points are available which could be used to detect, verify and correlate suspected malicious network activity. One source of such data is openly available text from time-sensitive sources on the Internet. In this paper, we discuss the practical possibilities and issues of building a text-based IT early warning sensor based upon our experiences in design and early implementation. In particular, we highlight open research issues which must be addressed before such sensors can be effectively implemented and discuss the nature of solutions which must be developed.**

***Keywords*-early warning; open source information processing; text analysis; semantics, event correlation**

I. Introduction

The concept of early warning (EW) applies to many catastrophe protection domains; natural disasters, epidemics and political unrest all start with seeds of trouble and produce increasing evidence that an event of interest is about to occur (or has already occurred). Early warning systems (EWSs) attempt to detect clues that undesirable events are in progress at the earliest possible stage in order to provide responders with the opportunity to plan, mitigate and respond.

EWSs are not just being developed for physical events, however. Given modern society's reliance on computing and networks, network-based attacks represent an event-based threat with increasingly broad (and costly) impacts. The primary goal of IT early warning systems (IT-EWSs), an emerging area of fundamental research, is to detect attacks in their nascent stages. Based on early detection, IT-EWSs are able to inform operators about security-related incidents which may have negative impact on their assets, enabling them to quickly take appropriate countermeasures.

In this paper, we identify the generic, practical challenges encountered during the design and ongoing implementation of an IT-EW sensor which uses web-based text as a source for identifying emerging and ongoing attacks. We discuss issues with selection and acquisition of usable and relevant source texts, approaches to text analysis and the related goals and issues inherent in text analysis in this domain. Further, we consider hinderances and solutions to the problem of evaluation, and suggestions for research necessary to progress forward in the domain.

II. IT early warning systems

From a high-level perspective, the ability to detect an attack early in its evolution is dependent upon three things: the structure of the attack's evolution, what evidence the attack leaves at a particular stage, and how that evidence is acquired (or "sensed"). Network-based attacks have a specific evolution over time related to vulnerabilities, when exploits are available, when mitigative measures are available, and when such measures are effectively deployed. Sensing in-progress attacks is usually done by means of strategically placed sensors throughout a network. How these sensors work and what analysis is done to determine that an attack is in progress is specific to the particular system in question.

A. Timeline of a network attack

The initial phases of network-based attacks typically follow a timeline similar to the following: first, a vulnerability is discovered by the community, product developers, or malicious entities looking for vulnerabilities to exploit. The vulnerability is then eventually disclosed. After a vulnerability has been disclosed in some venue, a proof-of-concept exploit is released at some point demonstrating its viability — corresponding advisories typically show up during this phase; finally, after a (generally) short period of time, initial attacks using this exploit (or modified ones) are observed in the wild, often instrumented within malware to automate propagation and infection.

This timeline is illustrated in Figure 1; note the initial divergence between explicit vulnerability disclosure (bottom line) and the lack thereof ("0-day disclosure" – upper line).

Systems often remain unpatched for quite some time. Thus, even after patches are released, malware outbreaks are initially largely unaffected by mitigating measures, often leading to the infection of a large number of systems. This lag between patch release and real mitigative impact allows for collection and analysis of data for the generation of intrusion-detection and malware signatures; these can be used to confirm threats, recognize variants released via other vectors, and so on.

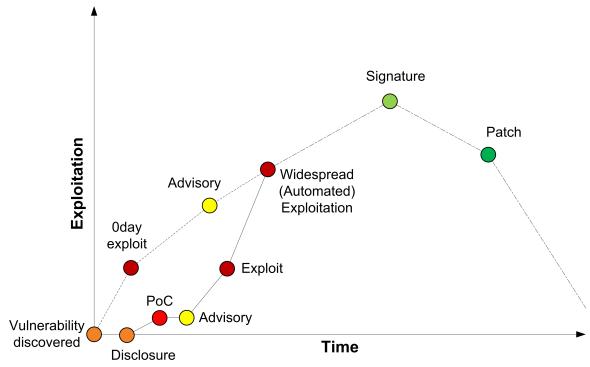

Figure 1. General exploit timeline

With the above-referenced timeline in mind, it is important to note situations where the existence of a vulnerability or exploit could have been detected earlier if initial indicators had been taken into account. For example, based on analysis of the Conficker worm[1], the first incidents indicating an emerging, serious hazard were publicly available weeks before the first Conficker variant was released, and months passed until the Conficker payload was actually activated. While it is very unlikely that highly sophisticated and targeted attacks can be detected by IT-EWSs in the early stages, we believe that their application to widespread attacks such as Conficker is feasible and that varied source and sensor types create real added value for IT early warning.

B. IT early warning

The primary goal of IT-EWSs is to produce early holistic and continously updated situational awareness based on evidence of emerging or ongoing hazards. To that end, intrusion detection systems (IDSs) are often instrumented for IT early warning, acting as one type of possible sensor within the detection component of an EWS. Because traditional IDSs offer only a local situational view, distributed, collaborative IDS approaches were developed in which a number of (different) sensors are placed on the Internet. Detected events, such as anomalies, are then correlated in order to detect emerging attacks in real time. IT-EWSs follow a similar approach in the sense that the detection component is typically based on a network of distributed sensors; however, IT-EWSs consist of many parts fulfilling different purposes, such as detection, aggregation, correlation, analysis and visualization.

Evidence on the wire, however, is not the only possible indicator that an attack is at hand. Humans also produce signs that an event of interest is underway: companies disclose vulnerabilities; hackers brag about exploits; analysts discuss anomalies in public fora. Evidence of emerging network attacks appears not only in network data, but also in language-based evidence created by participants, victims and observers. Text-based attack information is already consumed by security analysts, often at great time cost. These individuals must sort through large quantities of information on a daily basis. Automatic acquisition and processing of relevant information from these sources would be not only of benefit to analysts but could also add to the information collected by other sensors in order to broaden the picture of network threats.

[1]http://www.confickerworkinggroup.org/wiki/uploads/Conficker_Working_Group_Lessons_Learned_17_June_2010_final.pdf

III. Alternate Sensors: Practical Text-based Event Detection and Correlation

Attack detection in early warning systems relies on collection of corroborating data in order to correlate events from a number of sources. While network-based sensors might all collect the same kind of information about a particular kind of measureable network anomaly, humans might describe an ongoing system failure that can be correlated with the anomalous network behavior (or even outline a future attack which displays this behavior). Being able to use text-based sensors to increase the amount and depth of available relevant information would be a worthwhile addition to an IT-EWS's toolbox.

Design and implementation of text-based early warning components is, however, far from trivial. Source identification and acquisition as well as text analysis may sound like solved problems, but both generic and domain-specific concerns exist which make realizing automated text analysis for such systems challenging. Even when appropriate sources are identified, implementation of tools for automated acquisition of relevant core text from online sources is extremely labor-intensive and the resultant programs are often fragile. Data analysis and rigorous evaluation of results are the most challenging fundamental research questions inherent in development of such systems, with defined solutions only partially available; common statistical solutions for text analysis are, in our experience, insufficient, and there is as yet no concrete corpus of attack-based event-related text publicly available against which researchers can run their analysis engines. This creates difficulties in evaluating the efficacy of solutions, making both identification of concrete goals and evaluation of their implementation difficult.

A. Source selection

Good source selection is an obvious prerequisite for an effective sensor. The quality of any analysis is directly impacted by the quality of the input data extracted from its sources. In our case, the effectiveness of the entire system depends on the likelihood that the sources being analyzed contain the desired event data in some form.

This is more complicated than simply looking for web pages mentioning vulnerabilities and exploits. Sources must be evaluated according to timeliness, reliability, structure, and relevance. It is also important to note that highly derivative sources (summary news articles, CVE reports[2] and vendor disclosures) are likely to be much

[2]Common Vulnerabilities and Exposures (CVE) reports are standardized entries in the National Vulnerabilities Database (see http://nvd.nist.gov).

more structured (and thus easily parseable) than earlier information in fora, feeds containing real-time observations of activity. In general, the later in the attack cycle a source is published, the more organized it potentially is. Because this is a sensor for early warning, we cannot simply rely on more "structured" texts; we must also consider sources that have freeform text or use different languages, dialects and registers of speech. CVE or news report text tends toward the very formal and regular end of the language spectrum; informal discussion is much less structured and predictable. We cannot avoid the less structured sources simply because the language seems harder to deal with if we want to use it for early warning.

B. Acquisition

Not only is extracting information in a way that makes sense a challenging problem, but web-based sources present particular problems for which general solutions do not appear to exist.

Web sources are not written in a regular, predictable manner. Content providers often exploit the fact that (i) invalid HTML can be displayed correctly by browsers, and (ii) syntactic tricks[3] can be employed such that content displays "correctly" in a browser but is nearly impossible to extract and reassemble coherently with automated tools. This is a means to deter web scrapers from plagiarizing content which also frustrates data extraction. It is an enormous stumbling block for data acquisition, and that there are not generalized solutions is evidenced by the fact that companies charge exhorbitant fees for tools to assist in the extraction of relevant web-page elements on a per-site basis (without solving the problem of recomposition for text analysis)[4].

Furthermore, per-site solutions are remarkably fragile. Our experience using site-specific parsing modules is that small changes in web layout break a site's solution, necessitating a manual fix. This makes automating data collection high-maintenance and labor-intensive.

Some text-based EWSs such as BioCaster [1] avoid this problem by limiting sources to RSS feeds. This is insufficient for our purposes for two reasons. First of all, while RSS feeds are much more regular, most sites use only partial RSS feeds, restricting entries to shorter stubs (in order to foil scrapers) which then link to the full (HTML) article. Secondly, many important early sources of information are not available via RSS.

This data issue is not part of our main research problem; however, source extraction remains an inherent and difficult part of the text-based sensor problem with no simple solutions and must be addressed in order to build a real, functioning sensor.

C. Analysis

A primary goal for text-based sensors is the automated integration of the attack-related data an analyst would extract from web sources into an IT-EWS's analysis pipeline. This makes it very clear that domain-specific information must be present somewhere within the system.

There is no getting around this point; our experimentation suggests that semantically-unaware statistical analysis (topic modeling, clustering, etc) using various automatic measures (e.g. tf-idf and n-gram frequencies) for identification of related documents does not seem to cluster documents by event; rather, when successful in grouping related documents at all, these analyses cluster by attack type (e.g. buffer overflow), product name, or other non-event-related data. However, as we used purely lexical measures, and different indicators of the same attack may not have common vocabulary, this is not entirely unexpected. While statistical methods or simple keyword-based filtering of documents may may be appropriate within the context of data analysis or a relevance-filtering pass, it is clear that without additional knowledge of the domain and the ability to extract facts from those documents, very little is likely to be gained for automatic analysis.

Since further semantic knowledge is necessary for data extraction and intepretation, the question is where this domain-specific knowledge is located in the system. In other domains and other subfields of information security, this problem is solved by creating domain-specific ontologies and reasoning systems to represent and process the necessary semantic information in the extracted data ([2], [3], [4]). The only other work we are aware of in open source information processing for IT-EWSs avoids automated semantic analysis entirely by explicitly keeping humans in the system loop; Dörges, et al [5], propose an open source ticket-based report ranking system which collects reports automatically but leaves it to human operators to rank and judge the reports' relevance. As our stated goal is to have an automated system, this is not a sufficient solution in the text-based sensor context.

The challenge is to find the right approach for determining what texts are relevant and extracting whatever salient data is present. In health-based early warning, a semantics-based ontological approach is already used to both detect potential disease outbreaks and correlate reports cross-lingually and in different registers [1]. Semantic approaches to other large-scale data problems are being used in other domains as well; the BioInfer group has successfully used dependency-grammars linked to an ontology to capture protein-interaction information from the text in biology research papers [3]. Manual processing of the natural language in these texts to extract such data is a monumental task. Our problem is similar; we have a large body of textual information to process from sources of different quality, language, and size. We anticipate we can specify the needed information within concept structure and use a combination of tools to map language to those structures. Those data structures are then ready to be used for detection, correlation and classification tasks. How this is done depends on the textual reflection of what events look like in text, which is discussed in the next section.

[3] e.g. separating sentences across embedded tables, putting segments of paragraphs within ad content, etc.

[4] Some examples of scraping tools can be found at http://scrapy.org, http://visualwebripper.com, and http://www.automationanywhere.com

D. Event Analysis and Evaluation of Results

To the best of our knowledge, there has as yet been no publically available comprehensive analysis of the anatomy of a network attack as it presents itself in open source texts, and, more specifically, the evolution and textual cues related to such an attack. The fact that such a textual reflection of event-related data exists is evidenced by the daily work of network security analysts.

A concrete understanding of how these data points appear in natural language and form a picture of an event is necessary for progress in text-based automated early warning. The extant work in semantic representation of security topics (e.g. Raskin, et al [2], who have heavily explored the ontological structure of information security topics and applications to semantic inference; Benali, et al's ontology for the translation of alerts from firewalls based upon incompatible system logs from different manufacturers [6]; and CAPEC[5] – a MITRE community resource containing a taxonomy-based dictionary for Common Attack Pattern Enumeration and Classification) should be exploited in developing text-based sensors; however, choosing a specific representation requires a thorough analysis of text-based event evidence to determine adequate specification of concepts for both attacks and the progression inherent in the attack cycle. Inference methodologies can then be introduced based upon textual observations to fill in the picture of an ongoing attack.

Representations aside, a critical issue in domain-specific text analysis is evaluation. This is not a hypothetical problem; our attempts to subject a collected database of roughly 35,000 articles to clustering techniques made clear the fact that we had no rigorous method for showing success or failure, only subjective evaluations of observed phenomena. One standard method for addressing this problem is to generate what is known as a *gold standard corpus*; these corpora are annotated corpora used, among other things, for evaluation in a domain and for statistical training. The development of the publicly available semantically-tagged BioInfer corpus in bioinformatics [3], for example, is significant work toward the goal of evaluating information extraction tasks in biological literature.

Two similar types of corpora would be of intrinsic value within the early warning domain: an event-tagged corpus, allowing researchers to determine if their algorithms successfully identify known events, and a semantically-tagged corpus for domain grammar evaluation and training. Corpora can be manually created from historical web texts based on known past events, and detection and analysis of events in that corpus using given time points can be used both for understanding the evolution of events and for analyzing their detection potential over time.

Development of such corpora has other benefits such that this task, above others mentioned in this paper, has priority for future research. The understanding of event data in text will be facilitated by the real-text analysis required to develop an event corpus; further, without this information, evaluation of results is extremely subjective and the ability to show that real systems identify information correctly is limited or impossible. Our group has begun work on creation of such an event corpus as a necessary first step toward creating text-based early warning sensors, as initial efforts at text analysis showed that evaluation and understanding of events in text was sorely lacking.

[5]http://capec.mitre.org

IV. Conclusion and Future Directions

The development of text-based sensors for integration with traditional network sensor data is a promising direction in IT early warning, but without understanding the binding between the detailed underlying semantics of network attacks and the evolution of attacks over multiple documents in a timeline, further progress cannot be made. Real systems require solid means for evaluating correctness, and in natural language analysis, this implies the development of a corpus against which results from various approaches can be compared. These gaps must be filled for advancement in text-based IT-EWSs, along with development of appropriate analysis methods for both the text itself and the attack data once extracted as well as selection of appropriate and extractable sources. We believe that creating an event corpus is a first step toward opening up the field and understanding attacks, and current efforts are directed toward that end.

Acknowledgment

The authors would like to thank a number of people after the paper has been through the review process.

References

[1] A. Kawazoe, H. Chanlekha, M. Shigematsu, and N. Collier, "Structuring an event ontology for disease outbreak detection," *BMC Bioinformatics*, vol. 9, pp. 1–8, 2008.

[2] V. Raskin, J. M. Taylor, and C. F. Hempelmann, "Ontological semantic technology for detecting insider threat and social engineering," in *Proceedings of the 2010 workshop on New security paradigms*, ser. NSPW '10. New York, NY, USA: ACM, 2010, pp. 115–128.

[3] S. Pyysalo, F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski, "BioInfer: a corpus for information extraction in the biomedical domain," *BMC Bioinformatics*, vol. 8, no. 1, pp. 50+, 2007.

[4] N. Collier, A. Kawazoe, L. Jin, M. Shigematsu, D. Dien, R. Barrero, K. Takeuchi, and A. Kawtrakul, "A multilingual ontology for infectious disease surveillance: rationale, design and challenges," *Language Resources and Evaluation*, vol. 40, no. 3, pp. 405–413, Dec. 2006.

[5] T. Dörges and J. Sander, "Integrating open source information: Rumors & facts in early warning." Presented at the 1st European Workshop on Internet Early Warning and Network Intelligence (EWNI2010), 2010.

[6] F. Benali, S. Ubéda, and V. Legrand, "Collaborative approach to automatic classification of heterogeneous information security," in *Proceedings of the 2008 Second International Conference on Emerging Security Information, Systems and Technologies*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 294–299.