2014 IEEE International Congress on Big Data

Knowledge Discovery from Big Data for Intrusion Detection Using LDA

Jingwei Huang, Zbigniew Kalbarczyk, and David M. Nicol
Information Trust Institute, University of Illinois at Urbana-Champaign
Email: {jingwei,kalbarcz,dmnicol}@illinois.edu

***Abstract*—This paper explores a hybrid approach of intrusion detection through knowledge discovery from big data using Latent Dirichlet Allocation (LDA). We identify the "hidden" patterns of operations conducted by both normal users and malicious users from a large volume of network/systems logs, by mapping this problem to the topic modeling problem and leveraging the well established LDA models and learning algorithms. This new approach potentially completes the strength of signature-based and anomaly-based methods.**

I. Introduction

National Center for Supercomputing Applications (NCSA) collects a large volume of forensic data from security monitoring and system logs everyday; those data are also in many different formats, because they are generated from a variety of monitoring systems. Those data are a rich source for security knowledge discovery and intrusion detection; at the same time, those big data and the complexity of the problems also present a great challenge.

In NCSA, security is monitored at both network and host layers. At the network layer Bro IDS is used to perform deep packet inspection of network traffic going through the border router for detection of anomalous activity. 4.5GB of logs are collected daily. The network flow (netflow) collectors, such as *Argus* and *nfdump*, are distributed such that flows are monitored and logged from the border router and from the internal routers. The measurements give visibility to traffic within the internal subnets as well as to the traffic going in and out of the network. 2GB of netflow logs are collected daily. In host layer, operations are logged by *syslog* and file change logs are generated by the *File Integrity Monitor*. When moving to new generation of supercomputing, the amount of collected data grows rapidly, e.g. BlueWaters (petascale facility in NCSA) collects 3.7TB of syslogs over its first 336 days of services. Based on those logs, also blacklists and other external information from partners' sites, as well as security policies, security analysis tools generate various alerts for anomalies and signature matches [1], [2].

Many intrusion detection technologies have been developed [3]; signature-based technologies catch the known attacks effectively, but are not effective to the rapid growing new types of attacks; anomaly-based technologies detect "abnormal" behaviors, therefore this approach may catch the unknown attacks but typically has high false positive rate.

This paper aims to explore a new approach to knowledge discovery from big data for intrusion detection by using Latent Dirichlet Allocation (LDA) [4], [5]. LDA is a powerful topic modeling technique developed in Natural Language Processing and Machine Learning. LDA is able to identify the latent semantics of raw data, e.g. topics of text, and has been successfully used in large scale information retrieval and document classification. We envision that intrusion detection can be mapped into a topic modeling problem and be solved by a LDA approach. This proposed LDA approach is a hybrid approach, which may complement the features of signature-based and anomaly-based approaches.

II. Approach

In order to use LDA for intrusion detection, we need to map the intrusion detection problem into the topic modeling problem. To this end, the list of all monitored events is regarded as "*vocabulary*"; a logged event is regarded as a "*word*" in that *vocabulary*; each type of security incidents (such as intrusion or anomalies caused by insiders) is regarded as a "*topic*", which is represented by a probability distribution over the *words* (in the *vocabulary*); typical types of normal operations are also regarded as *topics*; a collection of events done by a user in a period of time is regarded as a "*document*"; a *document* is a mixture of *topics*, i.e. a *document* is modelled as a probability distribution over *topics*; a collection of all *documents*, i.e. all sets of logged events, is called a *corpus*, from which *topics* are identified through LDA learning. In runtime, the *topic distribution* of each new "*document*" (a set of new logged events) is calculated; if a new *document* has a high probability distribution over some security breach *topics*, then this set of logged events represented by that *document* may be a security breach, and a security warning should be issued.

III. A Use Case

To illustrate the above approach, consider the following scenario. Let us assume that the following list of possible types of events are monitored and collected by security monitoring systems. w1: login from an unknown IP address, according to the user profile; w2: multiple logins from the same IP address (multiple users use the same external IP address to sign in); w3: SSH connection; w4: suspicious IRC (Internet Relay Chat) connection (e.g. IRC traffic on a non-standard port); w5: internal scan; w6: external scan; w7: SSH scan; w8: downloading from publicly available

978-1-4799-5057-7/14 $31.00 © 2014 IEEE
DOI 10.1109/BigData.Congress.2014.111

IEEE computer society

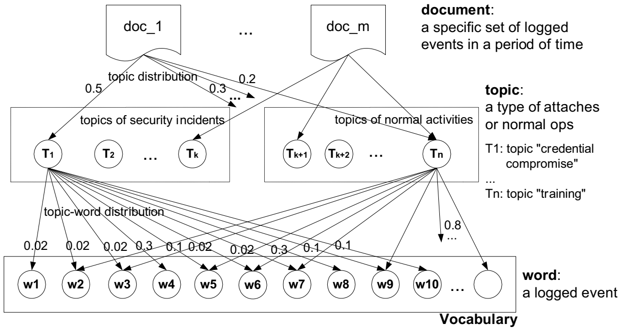

Figure 1. Example of LDA modeling for intrusion detection

exploit repositories or watchlist; w9: rename file types; w10: connections to an unused IP block within the network; ...

The collection of those event types forms the *vocabulary*, as shown in the bottom level in figure 1. In the figure, each "*document*" represents a set of the logged operations conducted by a user (a legitimate user or an intruder/malicious insiders/malware) in a specific period of time; each "*word*" in the document represents a logged event.

What a user is trying to do is unknown to the security monitoring system; however, *although a single event does not reveal the user's intent, a collection of logged events may reveal the intent*. This hidden intent, or latent semantics in the terminology of topic modeling, is modeled as a "*topic*", which is defined as a probability distribution over the *vocabulary* (the set of monitored events), called *topic-word distribution*. In this way, each *topic* represents a class of security incidents, e.g. *credential compromise*; or represents a class of normal activities conducted by users, e.g. "IT training". The *topics* are illustrated in the middle level in figure 1. The "*topics*" of a "*document*" (a set of logged events) do not appear explicitly as *words* do, but can be identified by using LDA modeling.

Assume that based on an in-depth analysis of data logs on security incidents, we established that topic *credential compromise* has a probability distribution over a set of logged events ("*words*"), as shown in fig. 1. Similarly, there exist *topics* corresponding to the normal behavior of a legitimate user, e.g. *training*. Those topics (defined by topic-word distributions), illustrated in the middle layer in fig. 1, can be identified or learnt from a set of *documents* (a collection of all security logs and alerts) by using a LDA algorithm.

At runtime, a new "*document*", i.e. a set of new logged events, associated with observable system activities over a predefined period of time, is collected. Based on the logged events (*words*) contained in the *document*, its topics distribution is calculated as follows: "credential compromise" (0.5), "Training" (0.2), and other *topics* of legitimate activities (0.3). The obtained probability distribution indicates that the observed behavior corresponds in a high probability to a security incident credential compromise (rather than legitimate activity); hence an alert should be raised.

Generally, the security incident *topics* represent malicious behavior; the normal activity *topics* represent normal behavior; so, in the topic distribution of a *document*, if a probability on a security incident *topic* is higher than a predefined threshold, an alert should be issued; if the probabilities on normal activity *topics* are fairly high and the probabilities on all security incident *topics* are lower than their thresholds, the behavior reflected by the *document* can be regarded as normal; if all probabilities on all *topics* are low and there exist a big error between the *document* (the real set of logged events) and the generated *document* from LDA topic model, then it is the sign of that the behavior reflected by the *document* is beyond the current *topics*, and further investigation needs to be conducted.

IV. Summary

We proposed a hybrid approach to knowledge discovery from big data for intrusion detection using LDA, in which the *topics* capture the"patterns" of both security incidents and normal activities, by using "bag of words" and probability distribution, so that this new pattern representation has flexibility to capture attacks in a wider range. More interestingly, the topics identified by LDA catch the latent semantics of a collection of the logged events, which gives a higher level description about what a user is doing or intends to do by that specific set (or sequence) of operations. This latent semantics is critical to intrusion detection, and can be identified by LDA modeling. We will use MapReduce to handle a large volume of logs and to conduct LDA learning in a timely manner.

Acknowledgment

This material is based on research sponsored by AFRL and AFOSR, under agreement number FA8750-11-2-0084.

References

[1] A. Sharma, Z. Kalbarczyk, J. Barlow, and R. Iyer, "Analysis of security data from a large computing organization," in *Proc. DSN'2011*, pp. 506–517.

[2] A. Pecchia, A. Sharma, Z. Kalbarczyk, D. Cotroneo, and R. K. Iyer, "Identifying compromised users in shared computing infrastructures: A data-driven bayesian network approach," in *Proc. SRDS'2011*, pp. 127–136.

[3] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, pp. 18 – 28, 2009.

[4] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.