Big Data for Security: Challenges, Opportunities, and Examples

Pratyusa K. Manadhata

HP Laboratories
HPL-2012-216

Keyword(s):

Big data; big data analytics; data mining; inference; malware detection; malicious domain detection

Abstract:

This is the age of big data. Enterprises collect large amounts of data about their operations and analyze the data to improve all aspects of their businesses. Big data for security, i.e., the analysis of very large enterprise data sets to identify actionable security information and hence to improve enterprise security, however, is a relatively unexplored area. Enterprises routinely collect terabytes of security relevant data, e.g., network logs and application logs, for several reasons such as availability of cheap storage and need for regulatory compliance and post hoc forensic analysis. But we face a situation where more is less; the more data we collect, the less is our ability to derive actionable information from the data. Our research group is trying to move toward a scenario where more is more; we aim to design and implement algorithms and systems to identify security relevant information from large enterprise datasets. The more data we collect, the more value we derive from the data. Our approach opens up new opportunities by combining data from multiple sources in an enterprise and from multiple enterprises. We, however, face many challenges, e.g., legal, privacy, and technical issues regarding scalable data collection and storage and scalable analytics platforms for security. Our group is currently focusing on several big data problems. In this talk, we will briefly describe the problems and then focus on one example - scalable and reliable identification of infected hosts in an enterprise network and of malicious domains visited by the enterprise's hosts. We model the identification problem as an inference problem over very large graphs derived from enterprise datasets. We will describe our experience of applying the inference approach to datasets collected from multiple enterprises worldwide. Joint work with Marc Eisenbarth, Stuart Haber, William Horne, Prasad Rao, and Sandeep Yadav.

External Posting Date: October 6, 2012 [Fulltext] Approved for External Publication
Internal Posting Date: October 6, 2012 [Fulltext]

Big Data for Security: Challenges, Opportunities, and Examples

Pratyusa K. Manadhata
HP Labs
5 Vaughn Drive, Suite 301
Princeton, NJ 08540
manadhata@hp.com

ABSTRACT

This is the age of big data. Enterprises collect large amounts of data about their operations and analyze the data to improve all aspects of their businesses. Big data for security, i.e., the analysis of very large enterprise data sets to identify actionable security information and hence to improve enterprise security, however, is a relatively unexplored area. Enterprises routinely collect terabytes of security relevant data, e.g., network logs and application logs, for several reasons such as availability of cheap storage and need for regulatory compliance and post hoc forensic analysis. But we face a situation where more is less; the more data we collect, the less is our ability to derive actionable information from the data.

Our research group is trying to move toward a scenario where more is more; we aim to design and implement algorithms and systems to identify security relevant information from large enterprise datasets. The more data we collect, the more value we derive from the data. Our approach opens up new opportunities by combining data from multiple sources in an enterprise and from multiple enterprises. We, however, face many challenges, e.g., legal, privacy, and technical issues regarding scalable data collection and storage and scalable analytics platforms for security.

Our group is currently focusing on several big data problems. In this talk, we will briefly describe the problems and then focus on one example– scalable and reliable identification of infected hosts in an enterprise network and of malicious domains visited by the enterprise's hosts. We model the identification problem as an inference problem over very large graphs derived from enterprise datasets. We will describe our experience of applying the inference approach to datasets collected from multiple enterprises worldwide.

Joint work with Marc Eisenbarth, Stuart Haber, William Horne, Prasad Rao, and Sandeep Yadav.

Bio: Pratyusa K. Manadhata is a researcher at HP Labs with a current research focus on big data analytics for security. He obtained his Ph.D. degree in computer science from CMU and worked on very large scale machine learning techniques for malware detection at Symantec Research Labs.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection (e.g., firewalls)*; K.6.5 [**Management Of Computing and Information Systems**]: Security and Protection

General Terms

Security

Keywords

Big data, big data analytics, data mining, inference, malware detection, malicious domain detection

BADGERS'12, October 15, 2012, Raleigh, North Carolina, USA.
ACM 978-1-4503-1661-3/12/10.