D. Hühnlein, H. Roßnagel, C. Schunck, M. Talamo (Eds.): Open Identity Summit 2016,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 175

Identity Mining Vs Identity Discovering: a new approach based on data mining in the context of Big Data

Constantina Caruso[1], Andrea Dimitri[2], and Massimo Mecella[3]

Abstract: The economy of an advanced country is, every day more, based on complex information systems and interconnected networks that made its cyberspace. Security in this cyberspace is an essential requirement. In Italy a national lab for Italian government has been constituted. In this framework identity and identity management systems has been studied. The depicted scenario defines new open questions and new challenges. In this paper we propose to deal with identity management in complex systems using analytical tools coming from anomaly detection for big data.

Keywords: identity management, big data, data mining, analytics, identity management system.

1 The context

The economy of an advanced country is, every day more, based on complex information systems and interconnected networks that made its cyberspace. Security in this cyberspace is an essential requirement.

To effectively face this new challenge, professors and researchers from 34 Italian universities have joined their efforts to constitute the Italian Cyber Security National Lab. In October 2015 a white book ("Il futuro della cyber security in Italia", [CD15]) established targets and future goals of this institution.

In this research container, the Universities of Roma Tor Vergata and La Sapienza, with the University of Bari Aldo Moro have been nominated responsible for the area Intelligence – Big Data Analytics.

In this organizational framework the data mining activity has been oriented to the particular task of identity discovering in complex systems for managing and adding security in Identity Management Systems; because of the new huge contexts (mobile, IoT, cloud computing), which stress the traditional activities of Identity Management Systems, experience and tools coming from Big Data Analytics have been considered.

Identity discovering in complex systems can enhance their security because, analyzing

[1] University of Bari,Aldo Moro, Piazza Umberto I, 1, 70121 Bari, Italy, costantina.caruso@uniba.it
[2] University of Rome Tor Vergata, Via Orazio Raimondo 18, 00173 Rome, Italy, andrea.dimitri@uniroma2.it
[3] University of Rome La Sapienzia, Piazzale Aldo Moro, 5, 00185 Rome, Italy, mecella@dis.uniroma1.it

the logs associated to the identity activities, we can compare discovered identities and previous known, formal identities of a complex service system thus obtaining factual information about novel or anomalous behaviour [SK15].Threats description

The progresses of technology in media and end devices and the necessity of complex resources sharing modified the architecture of the Organizations that cooperate to provide services. A lot of organizations, either public or private, without common hierarchic relationships, cooperate to administer an area and/or to provide services [SI01]. We refer to these organizations with the expression “Complex Systems”.

One of the emerging tendencies is to build and develop distributed systems which have mobile and dynamic service endpoints (smart phones are typical examples) and, at the same time, computational, storage and networking resources are centralized (cloud storage, cloud networking, cloud computing) [DA05][DT07].

In this complex and explicitly not hierarchical structure, every elementary organization possesses its security system but this does not guarantee that the resulting complex system is secured [RA08].

2 Digital identity management in complex systems

A digital identity is a set of attributes owned by an entity used by computer systems to represent an agent (person, organization, application, or device) [RA08] [BG02]. Its management is typically delegated to Identity and Access Management (IAM) which enables the right individuals to access the right resources at the right times and for the right reasons.

Classical IAM methods are generally designed to use tables-based architectures for storing entities attributes and involve four basic functions: creation, management and deletion of identities, the access function i.e. the control that data used by an entity to access to services are right; the service function which delivers personalized, role-based, online, on-demand, multimedia content to entity w.r.t. its authorizations; the “identity federation” function: the system that relies on federated identity must authenticate an entity without knowing its password.

Many large organizations have IAM infrastructure for user provisioning, Single Sign On (SSO) and identity governance but to track all the aspects of user activity is very difficult. In a recent research survey, security professionals identified “user behaviour activity monitoring” as the weakest area of security monitoring.

Classical approaches are inadequate in a business and technical landscape dramatically modified by described complex systems: what do you do when you are asked to build an identity and access management system that can handle up to billions of identities that have to be stored, managed and controlled in real-time?

Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety) [SY14][IE14].

In the majority of cases, Big Data security analytics is applied to security data such as network packets, meta data, emails, transaction systems to help teams to detect malware, phishing sites and online frauds because it can help enterprises to address unprecedented information risk arising from two conditions:

1. dissolving organizations network boundaries; corporate applications and data are increasingly accessed through cloud services and mobile device, introducing new threat vectors and making the systems become more vulnerable to data misuse and theft [CH13];

2. more sophisticated cyber attackers which bypass traditional defences, static threat detection measures and signature-based tools.

Security data encompasses any type of information that could contribute for a complete control of the organization and its risks.

A more agile approach based on dynamic risk assessments, the analysis of vast volumes of data and real-time security operations is essential to provide meaningful security. The Security for Business Innovation Council [SB16] advises organizations to move to an intelligence-driven security model.

We must face the differentiation of login and user authentication functions and the exponential growth of logging data ought to billions of digital identities (machine-to-human and machine-to-machine).

An authentication process verifies the identity claimed by or for a system entity. User authentication methods use three approaches: “something you know”, “something you have” and “something you are” [RA08].

The “something you know” verification is typically the use of a password to access on-line services. This method is the weakest because people often use weak passwords which are simple to find. Despite its weakness, the one-factor authentication is still used but it is not sufficient for organizations which need to differentiate their services and, as a consequence, the typologies of access points and login functions.

“Something-you-have” methods identify the entity by means of an object, physical or virtual, it owns: bank token, Kerberos ticket, ATM card, credit card, smart card are use cases of these methods. “Something-you-are” methods use user's biometric measures and are really strong authentication methods but they are very expensive.

A valid alternative to the three more classical methods seems to be the entity authentication based on “something you do”. There are very recent but promising studies which uses Big Data in identification methods for network traffic [SY14], for

recognizing fake identities in social media networks [SK15] [WE15] to identify social predators, all by using behaviour characteristics of users. To identify right or fake behaviours implies recognizing anomalies of the analyzing contest; in [IE14] a call detail record based anomaly detection method is presented; it analyzes the users' calling activities and detects the abnormal behaviour of user movements in a real cellular network. A major problem that many network/IT system administrators face is to detect a defect in user activity from a pool with many users and millions of transactions; the proposed approach reduces the cost of data processing compared to traditional data warehouse technologies. Event correlation is important when identifying anomalies and workflows: in [KM15] and in [RB15] and in Hadoop[AH16], cluster is used to study a scalable security event aggregator for situational awareness. The project Eagle [GS15] proposes a highly scalable system, based on Hadoop[AH16], capable of monitoring multiple eBay clusters in real-time. When a user performs any operation in the cluster, Eagle matches current user action against his prior activity pattern and raises an alarm if it suspects anomalous action.

Big Data technologies can reduce overall attack surface of IAMs addressing these issues:
1.clean up the access list by quickly identifying rogue accounts or users who haven't accessed applications for a prolonged period;
2. manage correctly separation of duties (SODs); IAM security analytics can clearly show business process relationships and find conflicts related to compliance and risk; this can help business managers to establish and manage correctly SODs;
3. to manage privileged users; analytics can be used to identify the privileged accounts and create a security system which alerts when anomalous behaviour is identified;
4. to secure IAM by discovering anomalous workflows activated by malicious entities (human or software);
5. to secure honest users by quickly identify fake identities.

3 The Big Data Analytics approach to anomaly detection in IAM: the Lab approach

Many large Organizations are already using security analytics tools to streamline processes and improve IAM oversight efficiency. From a security perspective, this can help enterprises reduce their overall attack surface and lower IT risk. Big data means the collection of hard amount of data to analyze them in a unique framework. Big data means, also, correlation of data coming from heterogeneous sources, meaning structured data versus unstructured data. This approach requires new technologies and new mathematical tools.

Many challenges are to be considered to plan a big data strategy [BC12][LA14]:
- interdisciplinary: manipulation of big data have to put together data mining, machine learning, information retrieval, natural language processing , statistics, applied mathematics, other than a strict interlacement with domain experts;

- quantity vs quality of data: the common vision deriving from big data privileges the quantitative. The step over is to develop statistical algorithms based on quality-aware methodologies;
- just in time predictions: a security system is to be able to notice a bug contextually and before it becomes a real threat. To prevent a bug means to evaluate stability and robustness of the workflow associated to that service;
- low user/organization impact: when importing a new technology in a context when an old one is already running, the impact has to be evaluated carefully; if the new technology has a strong impact in the old one, the switch-off could be not easy (often impossible);
- high security: classical security is highly related to cryptography; in this context the SSL (secure socket layer) protocol is a tunnel where clear data are redirected without considering their structure. This will be only partially feasible in the context of big data.

Atomic activities of the previous goals, planned in the context of the lab, are:

- analysis of algorithms of data analysis, statistical data analysis and learning and development of new algorithms and mathematical tools;
- entropy analysis and analysis of transaction entropy; entropy and conditioned entropy are the basis for the building of an index of forecast for the future state of a variable;
- anomaly detection systems and anomaly management;
- new algorithms for the manipulation of huge quantities of data;
- cryptographic systems for big data; in particular, we concentrate our research activities on the development of algorithms and protocols based on homomorphic encryption.

References

[SI01] Simon, H.A. (2001). Complex systems: The interplay of organizations and markets in contemporary society. Comp. & Math. Org. Th.y, Aug. 7(2), 79-85.

[SK15] K.D.B.H. Subasinghe, S.R.Kodithuwakku. “A Big Data Analytic Identity Management Expert System for Social Media Networks”, 2015 IEEE Int. WIE Conference on Electrical and Computer Engineering (WIECON-ECE).

[CH13] Chibber, A (2013)."Security analysis of cloud computing” Int. J.of Advanced Research in Engineering and Applied Sciences 2(3): 2278-6252.

[RB15] H. Reguieg, B. Benatallah, Hamid R., Motahari Nezhad, F. Toumani. “Event Correlation Analytics: Scaling Process Mining Using MapReduce-Aware Event Correlation Discovery Techniques”. IEEE Trans. on Services Computing. 2015.

[DA05] Dimitri A., Arcieri F. A prevention strategy for security: a bayesian approach to anomaly detection. IFIP Int. Fed. for Inf. Proc.. Sprinter Boston 2005.

[DT07] Dimitri A., Talamo M. A Meta Analysis Framework based on Hierarchical Mixture Model. Adv. and App. in Statistics. Volume 7, Issue 3, (Dec. 2007).

[BC12] R. Baldoni, G. Chockler: Collaborative Financial Infrastructure Protection - Tools, Abstractions, and Middleware. Springer 2012.

[LA14] G. Lodi, L. Aniello, G. A. Di Luna, R. Baldoni: An event-based platform for collaborative threats detection and monitoring. Inf. Syst. 39: 175-195,2014.

[MT15] A. Moroni, M. Talamo, and A. Dimitri. 2015. Adoption factors of NFC Mobile Proximity Payments in Italy. In Proc. of the 17th Int. C. on H-C Inter. with Mobile Devices and Services (MobileHCI '15). ACM, NY, USA, 393-399.

[CD15] C. Caruso, A. Dimitri, M. Mecella, M. Talamo Intelligence e Big Data Analytics, Il Futuro della Cyber Security in Italia, 2015, pagg. 50-53. CINI-Laboratorio Nazionale di Cyber Security.

[RA08] Ross J.Anderson, 2008. Security Engineering: a guide to building dependable distributed systems (2 edition), Wiley Publishing.

[BG02] Francesco Bergadano, Daniele Gunetti, and Claudia Picardi.“User authentication through keystroke dynamics”. ACM Transactions on Information and Systems Security. 5, 4 (November 2002), 367-397.

[SY14] Sung-Ho Yoon, Jun-Sang Park and Myung-Sup Kim. “Behaviour Signature for Big Data Traffic Identification”. 2014 Int. Conf. on Big Data and Smart Comp.

[IE14] Ilyas Alper Karatepe, Engin Zeydan. “Anomaly detection in cellular network datausing Big Data analytics”. European Wireless 2014.

[KM15] Jinoh Kim, Ilhwan Moon, Kyungil Lee, Sang C. Suh, Ikkyun Kim. “A scalable security event aggregator for situational awareness”. 2015 IEEE First Int. Conf. on Big Data Computing Service and Applications.

[GS15] Chaitali Gupta, Ranjan Sinha, Yong Zhang. “Eagle: User Profile-based Anomaly Detection for Securing Hadoop Clusters”. 2015 IEEE Int. Conf. on Big Data.

[WE15] Estée van der Walt, J.H.P. Eloff. “A Big Data Science experiment – Identity Deception Detection”. 2015 Int. Conf. on Computational Science and Computational Intelligence.

[AS16] Apache Spark, a general engine for large-scale data processing, http://spark.apache.org, 2016.

[AH16] Apache Hadoop, an open-source software project for reliable, scalable, distributed computing; http://hadoop.apache.org/, 2016.