DATA PROVENANCE IN THE INTERNET OF THINGS: VIEWS AND CHALLENGES

Zuhaib Bari Mufti[1] and Mahmoud Elkhodr[2]

[1]School of Computing, Engineering, and Mathematics, Western Sydney University, Sydney, Australia
[2]School of Engineering and Technology, Central Queensland University, Sydney, Australia

ABSTRACT

The IoT is a rich and dynamic network of interconnected networks where various devices share information, create knowledge and perform actuations events. In such an environment, it is important to precisely trace the origin of data and the events that contributed to their changes. This concept has long been known as provenance. This paper attempts to shade some lights on the importance of data provenance in the IoT, its application, and the challenges associated with data provenance in the IoT.

KEYWORDS

Internet of Things, Wireless Network, Data Provenance, Security & Trust

1. INTRODUCTION

Earlier forms of provenance appeared as a method to validate the authenticity of an artefact by examining an object's origin, ownership or any modifications made to the item (Dogan, 2016a). In a world entangled in a mesh of connected networks i.e. the Internet of Things (IoT), provenance becomes even more vital to keep track of events, the source of information, decisions, and origin of data and the metadata. E-Science relies on provenance to measure the quality of the data[1]. Nowadays, data provenance is no longer just concerned with finding the origin of the data, but it extends to include the capacity of tracking any events or modification made to the data. Example includes the followings applications[2]:

- Creating a file and any subsequent modifications to it and defining the ownership and accessibility is a form of File Systems provenance[1].
- Administrative systems and intrusion detection aided by logging system events is a form of Operating systems provenance.
- Similarly, compliers and run time errors can be detected by tagging the source line using compilers.
- Records of any insertion, modification and deletion are an application of provenance in curated databases[2].
- Browsing history is considered a form of web browsing provenance.

Additionally, several financial institutions are required by laws to record the source and origin of each digital transaction. This highlights the importance of provenance in the financial industry where each paper notes and its origin is treated as provenance. Intelligence and hospital systems are some of the prime users of provenance information[2]. A discrete information system having

Dhinaharan Nagamalai et al. (Eds) : CSEN, NCWC - 2018
pp. 01–07, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81201

adequate relevance, capable of undergoing classification into various domains for the purpose of evaluation can be considered as Intelligence. Hospital records and related data protected by the Health Care Portability and Accountability Act (HIPAA) act makes it obligatory to record and store all hospital records and data in addition to managing proper authorised access to the data[3].Information and lineage data used as provenance must possess some inherent technical features in order for it to be reliable. Some of these features are as follows:

- Information about every action performed on data needs to be preserved and stored completely[4].
- Ensuring that no manipulation of the data with a malicious intent takes place (Integrity).
- Provenance data should be available readily without any hassle (Availability).
- By providing authorised access to provenance data, confidentiality of the information can be ensured[5].
- Provenance data in the E-science field must be obtained in an economically feasible manner.
- Provenance data must be stored and available in such a way that the privacy of a person is not compromised, especially in the IoT[6]. Systems involving data provenance data need to deal with diverging aspects of ensuring that no outside entity or system is able to access the data and at the same time data within the system is readily available and shared among authorised entities for transparency[5].

2. APPLICATIONS OF DATA PROVENANCE

Some of the most common applications of provenance have been listed below:

DIAGNOSTICS:

Provenance has been used for debugging and detecting real time anomalies in a distributed system [7] If a monitoring system is based on declarative monitoring, there is a provision to analyse the network traffic which indirectly can be employed for detecting an intrusion[8]. SeNDlog can dynamically trace changes to a routing table and helps in generation of an alarm if the number of changes made are above a certain threshold value. Once an alarm has been generated a distributed recursive query on the network performance can trace the origin of any malicious activity[9].

SECURITY:

Data provenance covers historical data in addition to real time data as well. This helps in finding correlations in the network pattern of an attacker; thus, helping in the security of vital assets. Locating the source or filtering the IP address from the traffic is a typical example[10]. Annotations can be used in data provenance to help identifying potential attacker as well as tracing back information for forensic analysis[8]. Provenance can also be used to identify any malicious packets dropping in a sensor network[11].

ACCOUNTABILITY:

Data Provenance ensures a proper accountability for an action as well as data. In conventional forensic analysis, call-details consisting of information, time and location of the call are a form of data provenance. Network Provenance can be also used to manage trusts in a distributed environment[12].

TRUST:

By enabling a network of information where nodes are capable of tracing the origin of data, effective trust policies can be implemented[8]. Multi-hop networks and Body Sensor Networks rely also on data provenance to ensure trust[13]. Provenance can be used in quantifying trust, which enables sensors to process information from trusted nodes only (Wenchao et al., 2008).

OPTIMIZATION:

Monitoring of a system and tracing important events using data provenance in sensor networks can help in optimization of resources[2]. Resource allocation and finding bad routes or draining nodes are good examples.

DEVELOPMENT PROCESS:

Provenance logs can be used to capture changes in a network before and after an event takes place[14]. Comparing the snapshots before and after to see the changes in energy and other resources and using provenance to gauge the dependencies of a system can help in the development of a smooth process.

RECOVERY:

Provenance is often used to restore a system after a failure and for success validation[2]. In a sensor network, it is vital to not only identify the points of failure but also to avoid those which cause system anomalies. Provenance of graphs plays a key role in scenarios requiring troubleshooting as well.

3. DATA PROVENANCE CHALLENGES IN THE IOT

The IoT proposes various revolutionary concepts by employing millions, even billions, of tiny sensor or actuators nodes collecting and communicating information just about everything[15]. The volume of data collected in such a large network will have a high velocity, volume and divergent variety. This augments the significance of analysing the data for trustworthiness establishment in order to make better decisions. Therefore, it is becoming increasingly important to analyse a distributed network for possible anomalies and to pinpoint any erring node. These capabilities are some of the functional requirements needed to provision for Network Accountability and forensic analysis. Therefore, provenance of information or data plays a critical role in such environments. On the other hand, in an IoT smart based environment, the flow of information is relayed ultimately through the open Internet. It is a well-known security principle that the Internet is insecure. Therefore, it is essential to have reliability, trust, accountability and similar security principles addressed by employing a strong provenance enabled system.

To this end, as new, comples and dynamic data exchanged by IoT devices gets published on the Internet -where platforms accessing, publishing and modifying the data can be also diverse-, it becomes important to address the lineage, trustworthiness, reliability and accuracy of data in the IoT[16]. While papers' provenance has been employed in several systems, the IoT poses some unique challenges to the provisioning of data. Some of the challenges are listed below.

SECURITY

Data transmitted through an IoT system is extremely susceptible to attacks by a third party[17]. If provenance of data is insecure, it can result in a breach of sensitive information. The challenge is to impart enough confidentiality so that provenance can be accessed by only authorised individuals. Under certain circumstances, identity and location of the IoT device needs to be secured above all as the device may be more valuable than the data it sends. A robust security mechanism should incorporate confidentiality, integrity, privacy and availability of the information[18]. However, a high level of heterogeneity coupled with the massive scale in which IoT devices are likely to be deployed complicates the security issue of data provenance in the IoT[18].Moreover, IoT devices lack the computational power and energy requirements to incorporate complex security solutions such as encryption, cryptography, public key and symmetric key infrastructure[19]. Integrity of data provenance to assure a level of trust should be considered as well. This demands the use of cryptographic hashes algorithms which are extremely difficult to implement in the IoT due to the resource constraint feature of IoT devices[17].

BIG DATA

The massive volume of data produced by sensor networks in the IoT can result in the generation of petabytes of data, thus resulting in additional computational burden on the already fragile system[20]. Some researchers point out to the fact that Big Data and IoT need to be treated in tandem rather than as separate entities[21]. Querying and tracing Provenance information in such a system to point out the anomalies and other faults in the system is extremely difficult. Data Provenance may consume a lot of network resources, which in turn may hamper the operational efficiency of the system[17]. To ensure that Metadata is readily available upon request, there is a need to design systems which have a very low computational overhead to ensure smooth performance[22].

INDEXING:

A complete list of provenance in an IoT environment is practically impossible owing to the large nature of information. Hence, an indexing scheme is normally used[22]. However, it is likely that information can't be queried in a conventional manner wherein looking-up an attribute to retrieve the data is common. Users often have to query the dataset, which is essentially a subset of an attribute. Even in XML-based schema used for mapping names and values may prove not to be sufficient without the help of additional structures.

MULTIPLE CONSUMERS:

IoT data can have potentially vast and diverse range of consumers, with clients possessing divergent requirements. Some clients may need data on a real time bases, whilst others may just need to archive the provenance data. For example, while managing a smart city environment, provenance data may be required dynamically to make better decisions and rectify any anomalies in a system. Therefore, adequate flexibility is required for the provenance of data in the IoT.

TRANSFORMATION OF DATA:

Sensors in an IoT network collect data and pass or route them to other sensors, which may modify the information before passing it on to a more computationally powerful device. In other cases, actuators may receive data modified by various sensors during the transfer phase and thus, it becomes necessary to overcome the challenges encountered in representing such a complex provenance of information[17].

QUERYING INFORMATION:

Just tracing the lineage of data and its object may not be adequate for future systems and powerful querying tools need to be deployed to meet the cybersecurity challenges of next generation [23]. Records may need to be queried based on the context and requirements while maintaining the confidentiality at the same time may be essential [17].

INTEROPERABILITY:

IoT devices need to work in an extremely interoperable environment to ensure that the data collected by the sensors is successfully delivered to the target location. Also, various intermediate nodes or platforms are capable of reading or modifying the data. In such as case, Data Provenance demands that all the devices present in a system to be interoperable by having sufficient features to use each other's data. Keeping in view the limited computational power and resources of IoT devices and ensuring security of the system, achieving efficient interoperability in the IoT is still not an easy task[24]. IoT devices are manufactured by different vendors and may use different networking and routing protocols and often there is no standard or regulation yet in place to ensure uniformity and interoperability of devices.

DATABASE MANAGEMENT:

IoT data can be discrete, continuous, and dynamic. Certain data can be descriptive or based on environmental factors. Other can be in the form of addresses such as RFID tag format[25]. As the number of IoT devices may run into Billion coupled with limited computational capability of devices, it is almost impossible to adhere to IPv4 protocol for IoT Devices. Thus Internet Engineering Task Force (IETF) has introduced various protocols for IoT based in IPv6 addressing format[26]. But in doing so the header size has been increased from 32 bit to 128 bit addressing scheme, thus making it extremely difficult for resource constrained IoT devices to implement the system[25]. Thus, traditional databases may not provide a complete solution for such a complex system and it becomes imperative to deploy innovative and non-traditional databases.

An innovative approach is needed to cope up with the challenges associated with data provenance in IoT. In this case numerous protocols have been put forward such as the 6LoWPAN (IPv6 over low power wireless personal area network) protocol which is specially designed for resource constrained devices. The protocol is based on IPv6 and ensures universality, stability and additional features for IoT devices[26]. 6LoWPAN protocol suite specifically targets the integration of IPv6 and MAC (Media Access Control) and physical layers used in IEEE 802.15.4 standard. It is pertinent to mention that the maximum frame size of 127 bytes supported by IEEE 802.15.4 standard hinders the use of IPv6 and MAC header. By incorporating such a technology, it is possible to address various security and provenance issues using symmetric key and public key cryptography solutions.

One must also considers that not all IoT devices can transmit data. Hence, IoT gateways are used in some cases to bridge between the IoT devices with the Internet. Therefore, helping in harnessing the full potential of the technology[27]. The gateways provide a mechanism to ensure the computational power of IoT devices does not need to be high enough to increase the overall cost of the system, but at the same time they are able to smoothly operate in tandem with external applications and computational devices without compromising the efficiency and effectiveness of the system. Constricted application Protocol (CoAP) for device to device communication is employed to enables IoT devices to use the Representational state transfer (REST) mechanism which is similar to HTTP. This enables data provenance to be written using standard HTTP queries, which helps in mitigating the complexities of collecting provenance of data in IoT

applications. The use of several NoSQL like CouchDB, MongoDB etc. databases to store provenance data is recommended as they enable extensive flexibility during storing and retrieving of information.

4. CONCLUSIONS

The IoT with its diverse and heterogeneous nature of communications requires the provisioning of data provenance. Undoubtedly challenges associated with data provenance, especially in the IoT are enormous owing to the constrained resources available to IoT devices. Certain areas such as in the health and security domains demand elaborated provenance mechanisms whereas such intricacies may not be desired in simple IoT application such as controlling lighting in a smart building. Our future work will look into solutions that employs a middleware to leverage the overhead associated with the provision of data provenance in the IoT.

REFERENCES

[1] J. Cheney, S. Chong, N. Foster, M. Seltzer, and S. Vansummeren, "Provenance: a future history," in Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications, 2009, pp. 957-964: ACM.

[2] G. Dogan, "A Survey of Provenance in Wireless Sensor Networks," Adhoc & Sensor Wireless Networks, vol. 31, 2016.

[3] R. Lange, "Provenance aware sensor networks for real-time data analysis," University of Twente, 2010.

[4] R. Hasan, R. Sion, and M. Winslett, "The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance," in FAST, 2009, vol. 9, pp. 1-14.

[5] S. Bauer and D. Schreckling, "Data provenance in the internet of things," in EU Project COMPOSE, Conference Seminar, 2013.

[6] D. Gollmann, "Computer security," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 5, pp. 544-554, 2010.

[7] A. Singh, P. Maniatis, T. Roscoe, and P. Druschel, "Distributed monitoring and forensics in overlay networks," 2006: EuroSys.

[8] W. Zhu, E. Cronin, and B. T. Loo, "Provenance-aware secure networks," Departmental Papers (CIS), p. 387, 2008.

[9] W. Zhou, E. Cronin, and B. T. Loo, "Provenance-aware declarative secure networks," 2007.

[10] S. Savage, D. Wetherall, A. Karlin, and T. Anderson, "Network support for IP traceback," IEEE/ACM transactions on networking, vol. 9, no. 3, pp. 226-237, 2001.

[11] S. Sultana, E. Bertino, and M. Shehab, "A provenance based mechanism to identify malicious packet dropping adversaries in sensor networks," in Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on, 2011, pp. 332-338: IEEE.

[12] M. Blaze, J. Feigenbaum, J. Ioannidis, and A. D. Keromytis, "The role of trust management in distributed systems security," in Secure Internet Programming: Springer, 1999, pp. 185-210.

[13] K. Govindan et al., "Pronet: Network trust assessment based on incomplete provenance," in MILITARY COMMUNICATIONS CONFERENCE, 2011-MILCOM 2011, 2011, pp. 1213-1218: IEEE.

[14] K.-K. Muniswamy-Reddy, Foundations for provenance-aware systems. Harvard University Cambridge, 2010.

[15] H.-S. Lim, Y.-S. Moon, and E. Bertino, "Provenance-based trustworthiness assessment in sensor networks," in Proceedings of the Seventh International Workshop on Data Management for Sensor Networks, 2010, pp. 2-7: ACM.

[16] O. Hartig, "Provenance Information in the Web of Data," LDOW, vol. 538, 2009.

[17] A. Alkhalil and R. A. Ramadan, "IoT Data Provenance Implementation Challenges," Procedia Computer Science, vol. 109, pp. 1134-1139, 2017.

[18] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," Computer networks, vol. 76, pp. 146-164, 2015.

[19] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," Business Horizons, vol. 58, no. 4, pp. 431-440, 2015.

[20] C. Perera, R. Ranjan, L. Wang, S. U. Khan, and A. Y. Zomaya, "Big data privacy in the internet of things era," IT Professional, vol. 17, no. 3, pp. 32-39, 2015.

[21] S. Sahu and Y. Dhote, "A Study on Big Data: Issues, Challenges and Applications," International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), vol. 4, no. 6, pp. 10611-10616, 2016.

[22] A. Chebotko, J. Abraham, P. Brazier, A. Piazza, A. Kashlev, and S. Lu, "Storing, indexing and querying large provenance data sets as RDF graphs in apache HBase," in Services (SERVICES), 2013 IEEE Ninth World Congress on, 2013, pp. 1-8: IEEE.

[23] V. Gazis et al., "Short paper: IoT: Challenges, projects, architectures," in Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on, 2015, pp. 145-147: IEEE.

[24] P. Grace, J. Barbosa, B. Pickering, and M. Surridge, "Taming the interoperability challenges of complex iot systems," in Proceedings of the 1st ACM Workshop on Middleware for Context-Aware Applications in the IoT, 2014, pp. 1-6: ACM.

[25] J. Cooper and A. James, "Challenges for Database Management in the Internet of Things," IETE Technical Review, vol. 26, no. 5, pp. 320-329, 2009/09/01 2009.

[26] Z. Sheng, S. Yang, Y. Yu, A. Vasilakos, J. Mccann, and K. Leung, "A survey on the ietf protocol suite for the internet of things: Standards, challenges, and opportunities," IEEE Wireless Communications, vol. 20, no. 6, pp. 91-98, 2013.

[27] B. Kang and H. Choo, "An experimental study of a reliable IoT gateway," ICT Express, 2017/04/27/ 2017.