A Secure Provenance-Aware Model for Internet of Things
Sabah Suhail and Choong Seon Hong
Department of Computer Engineering, Kyung Hee University, Yongin, 446-701 Korea
{Sabah, shong}@khu.ac.kr

Abstract
In current cyber-physical space, IoT represents interconnection of highly heterogeneous and differentiated networked entities, collecting data and delivering services to multifarious domains including environment monitoring, energy management, health-care system, and industrial automation. However, despite the perceived comforts of global connectivity, accessibility and ubiquity, the Internet of Things (IoT) encounter various challenges ranging from tracing and storing veracious sensory data, utilizing collected sensory data for decision-making and activity reconstruction. In this paper, we have discussed requirements and challenges grounded on IoT technical infrastructure. We have proposed a provenance-aware IoT model capable of identifying compromised nodes by monitoring the abnormal behavior of nodes via provenance-based chained-events.

1. Introduction:
Provenance is a meta-data describing the complete lineage of data and processes chain. Provenance-aware system has the ability to resolve issues including tracking ownership of data, transformation (input parameters) applied on data and environment settings in which data is processed or evolved at granularity level. Provenance being beneficial in auditing, result reproducibility, forensics investigation and quality assessment have been extensively focused in various domains including databases, scientific workflows, distributed systems, and networks[7].

Internet of Things—a paradigm connecting heterogeneous networked things to facilitate goods and services in global supply chain, e-health, and various other domains has a high risk of security and privacy breach of the involved stakeholders. Furthermore, the trustworthiness of data being generated is highly questionable. Therefore, provenance can be integrated with IoT to track inconsistencies in data and to guarantee the integrity and authenticity of data. For instance, in the health-care system, drug tracking events can determine medicine in-take pattern of patients. Moreover, the life-log provenance data may aid the medical consultants to monitor the causes of severity of any particular chronic disease; which drug should be substituted to achieve optimum sudden recovery. Similarly in smart agriculture, to analyze the crop yielding keeping in view the fertilizer application, nutrients, soil mapping, weather data, machinery, livestock etc. The farming association can track yearly crop production and can be able to identify the factors which may enhance the quality of agriculture products based on the information obtained from data provenance. Such meta-data may help in making agronomic decision.

The main contribution of this paper is to devise a provenance-aware model that is capable of monitoring sensory devices in order to construct chained-events based on provenance meta-data. The paper is organized as follows. Section 2 discusses the security challenges, Section 3 provides an overview of the proposed methodology. Section 4 shows simulation results. Finally, we conclude the paper with future work.

2. Security Challenges:
The multifaceted requirements and challenges are based on the IoT technical architecture (for instance, RFID-based IoT, WSN-based IoT and IP-based IoT). Such requirements need to be evaluated in terms of provenance to enforce authenticity and integrity of data streams generated by IoT.

For instance, traceability issues in RFID-based IoT [5]. Consider a supply chain scenario in which a composite object O having integrated parts p1, p2 and p3 attributed with tags t1,t2 and t3. A potential solution to resolve traceability problem is to associate provenance data with each object's tag which can facilitate the tracker to identify the status of any individual part.

Gathering provenance is another key feature which raises the question, “provided the constraints, at which level provenance should be gathered to attain optimal information about secure nodes”. In this regard, provenance collection can be classified as level-based provenance collection and phase-based provenance collection. Since each of the levels is mapped to a phase thereby provenance data may duplicate itself which may facilitate confidence of correctness in sensors provenance metadata. Furthermore, the collection of provenance at any level or phase also effect the level of accuracy of provenance data. For instance, provenance collected at node level can depict more specialized details as compared to the data that may arrive at application with the high probability of suffering from tampering.

The provenance collection in IP-based IoT[6] can be classified as:

I: Level-based provenance collection:
a) Collection of provenance at node level: The resource-constrained things/nodes in IoT vary in terms of energy supply, for instance, battery-oriented and battery-less nodes focusing on low energy consumption or lowest power mode/sleep mode. Under such circumstances, provenance collection may suffer because of the unavailability of dormant mode of sensor at a particular instant or due to energy supply constraints.
b) Collection of provenance at network level: In IP-based IoT [6], provenance collection at network level may lead to other issues including integration of provenance capture mechanisms at either network layer or data link layer.
c) Collection of provenance at process/application level: Provenance can be collected at application level but it may not be trustworthy as the attacker may already have forged the sensor data while in transit.

II: Phase-based provenance collection:

a) Collection of provenance at bootstrapping phase: The bootstrapping process denotes the joining thing in the IoT at a particular time and location. Since device identity and security parameters are provided to the device during this phase hence, trust bootstrapping between the nodes of different vendors is of significant importance. Provenance can store this information to efficiently track the associated metadata with each node.
b) Collection of provenance at operational phase: The information
including decommissioning, re-ownership of devices and maintenance needs to be gathered during the operational phase.

To ensure the reliability of provenance data itself, provenance data must be stored in encrypted format to a to enforce confidentiality. Similarly to enforce access rights, the users may provide credential (user name/ password or secret key) in order to access the data legitimately. In that case, only the authorized users, for example, medical consultants will be allowed to query data in order to determine the patient's medical history including drug prescription, resulting symptoms of any particular medicine and other health related issues. To ensure trustworthiness of data, the message digest of each provenance chain can be computed to identify tamper-evident data [4].

3. Proposed Methodology:
The problem of integrating secure provenance into IoT can be classified into various sub-problems including a) extracting sensor data; b) inferring contextual information and other relevant details from streaming data; c) collecting and maintaining provenance data; and d) securing provenance data. In addition, challenges associated with secure provenance for instance confidentiality, integrity and access rights are required to be addressed in the context of IoT are among the primary objectives of this proposed approach. However, we are primarily concerned with collection, storage and querying of provenance where as other problems can be targeted with available techniques in the literature.

The sensor network $N$ can be modeled as graph $G(N, L)$ where $N$ represents the set of nodes and $L$ represents the set of links such that each $l_i$ correspond to a pair of node $n_i$ and $n_j$. For collecting node-level provenance, we will maintain a provenance graph ($P_{node}$) for each $N_i$.

![Fig. 1 Provenance Graph](image1)

A high-level architecture (Fig. 2) of proposed approach is adapted from [1]. However, to meet the additional requirements, the proposed approach has added other modules.

The streaming unstructured data acquired from smart devices including sensors and actuators will be gathered at a Data Semantics Evaluator (DSE). DSE plays a vital role in structuring the data and forward it to sensor database and Provenance Handler (PH). DSE will classify the data as contextual information (higher-level details) and related details (low-level details). For example, high-level detail may include available parking lot while low-level detail may include occupancy detection sensor number. The DSE will pass on related details ($P_{node}$) to PH which is responsible for managing provenance information. PH may ask for contextual information from DSE, if required.

DSE will forward selective contextual information ($C_{node}$) to the sensor database because data users are only interested in abstraction-level data. The stakeholders of such high-level contextual information are usually doctors, medical experts or insurance companies. To facilitate the user for querying, they may be provided with any interface (web browsers or application).

![Fig. 2 Secure Provenance-aware model for IoT](image2)

PH will be responsible for maintaining sensors meta-data or provenance. PH will maintain a separate secure provenance database $S_{node}$. This database will consist of all structured chained information necessary to reconstruct any event or activity to figure out any issue. For instance, querying secure provenance database can let the medical expert know what type of medicine $m$ he has advised to patient $p$ at given time $t$.

Consider a scenario in which a wellness recommendation system (for example, Mining Mind Platform [3]) suggest the user to perform exercise or inform about any other food consumption tips related to balanced diet. If such suggestions are kept on recorded by some mechanism then they will help users or even medical experts to track the food consumption pattern against any user for any given period of time.

4. Simulation Results:
We have performed experiment in Contiki’s network simulator Cooja. Cooja runs deployable Contiki code. In this simulations, we use emulated Tmote Sky nodes for performing IoT-Wormhole-IDS.

![Fig. 3 Log-listener](image3)

Consider an IoT networked scenario $N_i$ having a designated receiver mote $R_i$ and 4 sender motes ($S_1$, $S_2$, $S_3$ and $S_4$). Suppose a malicious node $M_1$ managed to enter into the $N_i$ and execute wormhole attack. A wormhole attack is an attack in which the attacker records the packets and tunnel or retransmit them to another network. Through wormhole attack, a malicious node $M$ is capable of tunneling the information to other malicious motes ($m_\text{a}$, $m_2$, $m_3$, ..., $m_n$) in network $N_i$. To identify the malicious node under

1 https://github.com/parvanpomgle/IoT-Wormhole-IDS
such scenario, we can keep track of log information with respect to timed events in provenance database. Provided that under no failure of nodes the sensor nodes are stationary after deployment; by observing patterns of each authenticated mote from log-listeners (Fig. 3), we can identify the abnormal behavior of malicious mote.

**Algorithm 1:**
1) For each node n:
   a) Collect \( P_{\text{for}} \) for each \( n \) where \( n \subseteq N_s \) from \( P_{\text{graph}} \)
   b) Update \( P_{\text{for}} \) for any failure node \( n_f \)
2) Pass \( P_{\text{for}} \) containing \( f_{\text{for}} \) to PH
3) PH stores the \( P_{\text{for}} \) in \( S_{\text{db}} \)
4) Pass \( C_{\text{for}} \) to \( S_{\text{db}} \)

![Fig. 4 Wormhole Attack](image)

Provenance meta-data can be collected from log-listeners (Fig. 3) and organized based on Algorithm 1. Therefore, \( P_{\text{for}} = N_{\text{foruptime}} + N_{\text{forsequence}} + N_{\text{forstatus}} \)
where \( P_{\text{for}} \) represents the provenance data for \( i^{th} \) network, \( N_{\text{foruptime}} \) represents attributes including IP, data value, packet sequence number etc., \( N_{\text{forsequence}} \) represents time-stamp and \( N_{\text{forstatus}} \) represents the status of \( i^{th} \) node. Hence querying from provenance data can help to reconstruct any event or to figure out abnormal activity going on in \( N_s \).

5. **Conclusion:**
In this paper, we have proposed a provenance-aware IoT model that is capable maintaining a log of streaming data for inferencing. However, due to the constrained-based IoT devices, the process of collecting and storing provenance is not quite easy and therefore not addressed yet. Our solutions to these challenging problems and to incorporate provenance in other IoT architectures (RFID and WSN) will be the subject of the forthcoming publications.

**Acknowledgement:**
This work was supported by the Industrial Core Technology Development Program(10049079 , Development of Mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) *Dr. CS Hong is the corresponding author.

**References**


