

USING PROVENANCE IN SENSOR NETWORK APPLICATIONS FOR FAULT-TOLERANCE AND TROUBLESHOOTING

Position Paper

Gulustan Dogan and Theodore Brown

Graduate Center, City University of New York, New York City, U.S.A.

Keywords: Provenance, Fault-tolerance, Troubleshooting, Sensor Networks.

Abstract: Provenance is a rapidly progressing new field with many open research problems. Being related to data and processes, provenance research is at the cross-roads of research from several research communities. With the huge amount of information and processes available in sensor networks, provenance becomes crucial for understanding the creation, manipulation and quality of data and processes in this domain too. Sensors collaboratively carry out sensing tasks and forward their data to the closest data processing center, which may further forward it. Provenance provides the means to record the data flow and manipulate snapshots of the network. Consequently given enough data, provenance can be used in sensor network applications to find out causes of faulty behavior, to figure out the circumstances that will affect the performance of the sensor network, to produce trustworthy data after elimination of the causes, etc. In this position paper, we describe provenance work in the sensor network community to sketch a panoramic view of the recent research and give a provenance model of a binary target localization sensor network as a real life example to show how provenance can be used in sensor network applications for fault-tolerance and troubleshooting.

1 INTRODUCTION

Wireless sensor networks are used in many applications such as battlefield surveillance, air pollution monitoring, forest fires detection, biological, chemical attack detection. Due to their nature, wireless sensor networks are more error-prone than traditional networks. However most of the sensor network applications are real-time and mission critical. Therefore fault-tolerance and troubleshooting become very crucial in order to sustain the network. In this position paper, we argue that sensor networks should have provenance support for many benefits including fault-tolerance and troubleshooting.

Although sensor networks is an area of research for years, provenance management should be a concern too in order to have an understanding of how the results are obtained for fault tolerance and troubleshooting purposes. In some sensor networks such as ad hoc sensor networks, in which data is copied, moved, created, updated and deleted in an uncontrollable way, provenance can play an important role in deciding about data qualities such as trustworthiness, accuracy, verifiability.

Maintaining provenance information makes it po-

ssible to have a clearer picture of the movement of the data and its manipulation in a sensor network by tracking the evolution of the data systematically. There can be many reasons why the data values generated are not accurately received at the sink. For instance one reason could be the sensors themselves may be sending faulty data. Provenance can be used to keep track of the state of the sensors and more generally to find out the causes of faulty behavior, to figure out the circumstances that will determine the connectivity of the network, and to produce trustworthy data after elimination of the causes.

We present a dataflow-oriented provenance model for sensor networks. Although our dataflow-oriented provenance model is generic, we use a particular scenario to support our argument modeling it on a proximity binary target localization sensor network. This model works best with networks sensing dynamic objects, and we introduce a provenance directed network-level fault-tolerance mechanism by using the cognitive strength of provenance models. Our provenance model reduces the limitations of faulty data by decreasing the possibility of errors in wireless sensor networks. By determining faulty data early in the stream, our model also makes it advanta-

geous to have a self-adjusting sensor network so that the sensor data that is produced results in more accurate results at the receiving end.

This paper is organized as follows. Related work is presented in Section 2. In Section 3 we give some background information on provenance. In Section 4 we describe how provenance information can be used in sensor network applications for fault-tolerance and troubleshooting. In Section 5 we describe the provenance model for a target localization sensor network. Section 6 concludes the paper.

2 RELATED WORK

Provenance in the sensor network community is an area of research that is new and open to many directions. Provenance management in sensor networks should be considered in order to have an understanding of how the results are obtained. Having this motivation, although not as extensive as provenance research in database (Cui et al., 2000; Buneman and Tan, 2007; Moreau et al., 2008; Cheney et al., 2009) and eScience community (Davidson and Freire, 2008a; Barseghian et al., 2010; Crawl and Altintas, 2008; Feng and Lee, 2008; Freire et al., 2008), there has been research on provenance in sensor networks community.

To our knowledge, there is not any work on using provenance specifically for fault tolerance in sensor networks. Some research has leveraged provenance in extracting metadata from weather sensors (Stephan et al., 2010), in answering domain specific complex queries (Patni et al., 2010; Park and Heidemann, 2008a), in building provenance aware sensor data storage (Ledlie et al., 2005). In our previous work we used provenance for network restructuring (Dogan et al., 2011) and assessing trust (Govindan et al., 2011).

The nature of provenance in sensor networks is different from eScience and database community in several ways (Park and Heidemann, 2008a), this is why research on provenance in sensor network community should be done more extensively to make robust provenance systems for sensor networks.

3 BACKGROUND

Provenance has been defined broadly as the origin, history, chain of custody, derivation or process of an object. In disciplines such as art, archeology, provenance is crucial to value an artifact as being authentic and original (Cheney, 2010). However prove-

nance has also become a crucial component in fields that rely on digital information. For instance Homeland Security and Governmental Affairs highlighted provenance as one of three key future technologies for securing their critical infrastructure (Wynbourne et al., 2009). Provenance has grown in importance in its use in helping to understand how the digitally captured data is manipulated at the source and used at the destination.

The literature often divides provenance into data and workflow provenance (Moreau and Ludascher, 2007). Data provenance gives a detailed record of the derivation of a piece of data that is the result of a transformation step (Tan, 2007) whereas workflow provenance is the information or metadata that characterizes the processing of information from input to output (Davidson and Freire, 2008b).

4 PROVENANCE FOR FAULT-TOLERANCE AND TROUBLESHOOTING

Provenance is needed to be transmitted to where a description of the data namely metadata is needed. Basically, if there is any input data (data provenance), any process chain (workflow provenance) or information flow (dataflow provenance), provenance can be included. There are many areas in sensor networks where provenance can be used. In this paper, we concentrate on showing how provenance can improve fault-tolerance and troubleshooting capability of sensor networks.

In a sensor network, there can be many unpredictable events such as broken sensors, unstable connections, lack of energy which will cause corruption in system. For example a sensor can get many streams at the same time such as temperature, audio and video. When it transmits the stream, the metadata will specify what kind of stream the sent data is, when it was transmitted, the id of the node transmitting the stream, whether the data was modified or not. If later this data is found to be faulty, by using dataflow provenance graph the responsible node can be determined and marked as an untrusted node. Another example will be if a video sensor changes its angle, zoom, or resolution, provenance can record the path the image travels (Tilak et al., 2005) which can be later used to track the source if the user is not satisfied with the image.

It is beneficial to store provenance information for historical value for the sensor data (its metadata). Mostly sensor data is treated as data with a real time

value but Ledlie et al point out that it has historical value too if thought in a broader sense. To heal, adjust and manage sensor networks, historical sensor data is required. For instance, for finding the patterns related to snowfall and traffic, sensor data monitoring traffic and snowfall and their provenance data can be useful (Ledlie et al., 2005). In our model, dataflow provenance graphs are stored in a central storage which is further described in Section 4. The stored historical metadata of paths is a good source in determining the behavior patterns of the network. We can make use of the historical metadata to extract statistical information that is related to fault-tolerance such as misreport frequency of a node, unsuccessful transmissions of a link, inconsistent data retrievals, etc.

Furthermore, provenance ensures authenticity of data which is helpful in creating a safe platform for fault-tolerance and troubleshooting. In sensor networks, data can be modified along its path to its destination and can be used for maliciously. If there is provenance tracking on the node, the data can be verified as authentic or not. Although this verification is not foolproof, it adds a layer of protection. For instance the system can be configured so that before a change, the identity of the node making the change would be verified. If the node's identity cannot be verified, this will be the sign of an attack and the modifications will not be allowed, making the network more resistant and fault-tolerant.

By making use of logs of provenance information, tracing and monitoring can be done efficiently and changes between two time slots can be found out if the later information is considered faulty. The logs can be used in troubleshooting and fault-tolerance making network more fault-tolerant as mistakes will be more quickly detected at a closer destination before data travels long distances.

4.1 Where Provenance Comes In

As stated above provenance answers questions such as "how was data created?", "on what other data does computed data value depend upon?", "how do the ancestries of these two data differ?"(Muniswamy-Reddy, 2010). Therefore, in sensory data systems, sensors can also have the provenance information of data dependency. For instance there will be a data dependency representing the data binding relation between the outputs of sensor nodes A, B, C, D, E and input of fusion node F shown in Figure 1. Operation of fusion node F depends on availability of the data provided by nodes A, B, C, D, E. This dependency provenance will be helpful in analyzing data flow graphs of sensor networks.

Dataflow provenance is the path that data is transmitted over until it comes to the fusing node. Then if there is a misreport, using dataflow provenance we can find out the sensor creating the false information and it can be replaced by reconfiguring the network or waking up an appropriate sensor before a system malfunction.

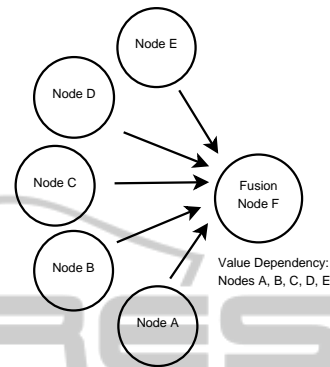


Figure 1: Value dependency between nodes.

4.2 How to Model Provenance

For standardization purposes, a provenance model called Open Provenance Model (OPM) has being crafted (Moreau and Ludascher, 2007). An alternative model is the W3C SSN-XG. We will use OPM for modeling provenance and for lack of space we will not further discuss this latter model. However the OPM model does not support some requirements that are specific to sensor networks such as recording provenance of streaming data, capturing times between sensing. It is an ongoing research issue to completely adapt this model to sensor networks (Park and Heidemann, 2008b) and it is in our future agenda.

In OPM, provenance is modeled as a directed acyclic graph (DAG). The nodes in the DAG represent objects whose provenance the system describes. The edges between objects indicate relationships between them. Both the nodes and edges can have attributes. For nodes, the attributes consist of information such as the name of the object it represents and the object type. For edges, the attributes indicate the type of relationships between objects. In our system, as energy is an important consideration, we keep the forwarded provenance data as small as possible and we do not label attributes of nodes in our provenance graphs. We only transmit id of the node, edge label and data at each transmission.

Apart from being modeled as a DAG, our model specifically is an Information Flow Model (Sabelfeld and Myers, 2003). Information flows between a source object and its destination object. The

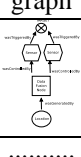
graph	timestamps	localization
	2011-01-19 03:14:07	λ
.....

Figure 2: Central storage scheme.

source object is the ancestor of the destination object (Muniswamy-Reddy, 2010). This model will be more useful as we are interested in finding out the path that created the target localization decision. In our system, dataflow provenance is kept as directed graphs in Central Provenance Repository as illustrated in Figure 2.

5 AN EXAMPLE: TARGET LOCALIZATION SENSOR NETWORK

To better illustrate our concepts, we will examine a system that makes use of proximity binary sensors. We assume that environment that sensors are in is not under attack so that we can assume that the provenance information is assured to be accurate. In proximity-based wireless sensor networks, the likelihood of the target position is calculated using the binary values reported by proximity binary sensors. A proximity sensor acts as a tripwire i.e. it reports a detection when a target close by triggers it. Examples of these sensors are seismic, acoustic, passive, infrared; they can be deployed in large numbers because of their low cost. The binary proximity behavior in sensors is achieved by implementing simple energy detection algorithms where the signal is compared to a threshold. If the signal exceeds the threshold, the sensor node reports a “1” meaning a detection, otherwise a “0” is reported for no detection. A network of such sensors is used to localize and track targets (Le and Kaplan, 2010). Provenance data is captured in our model for this network as a support for fault-tolerance and troubleshooting of the target localization. A detailed picture of open provenance model of our system is illustrated in Figure 3. Basically network snapshots are taken at time intervals when a target is detected and they are stored in the Central Provenance Repository. Later these network snapshots are used in order to get a better understanding of the behavior of the network for fault-tolerance and troubleshooting. Traditional sensor network systems deal with real time data and they lack the ability to remember past detections and which nodes were actively participat-

ing in the target localization. As clearly can be seen in Figure 3 below, our system records the dataflow pictures.

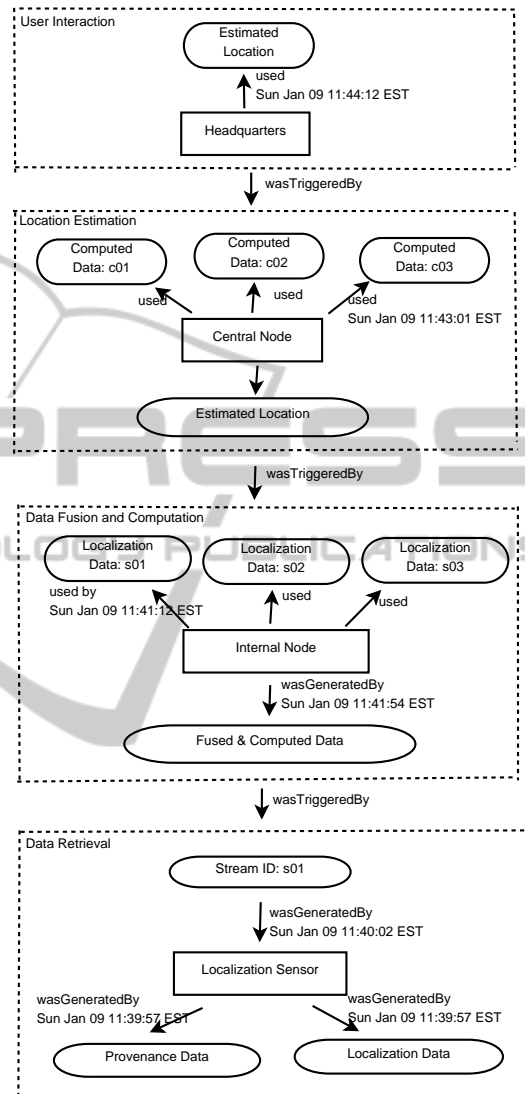


Figure 3: OPM graph of binary localization.

5.1 Fault Tolerance and Troubleshooting

Consider the following scenario: A target steps into the field and trips one or more sensors. Target localization is done at the base node within headquarters. Administrators may see that the target is localized wrongly. Provenance will be helpful at this point to understand the reasons behind this. In our model, provenance graphs of localizations (data flow from one node to another) can be examined and causes for

any change or fault can be detected. After finding the root cause, the network can be reconfigured eliminating the faulty nodes.

There are many possible exceptions in a sensor network which some of them can be listed as follows, sensor node failure, deadline expiry, resource unavailability, path loss and unpredictable multipath (Bal et al., 2010). On the other hand Zahedi et al characterize a sensor measurement to be in one of several states including normal, noisy, spike, frozen, saturation, bias, spike, oscillation (Zahedi et al., 2008). Our dataflow-oriented sensor network model that is illustrated in Figure 3 will capture and tolerate failure patterns due to faulty data. As our flow model has value dependencies stored, fault-tolerance will be doable. For example when a sensor is failing, it can be replaced and retransmission can be done. Another example is, since provenance data flows to the central provenance repository, the problem can be found out earlier that the node is out of energy and be replaced before it causes system to provide faulty or incomplete data.

At the headquarters, administrators can find out wrong localizations. In traditional sensor networks, it is hard to find faulty nodes because there may not be historical data available. Network constantly detects targets and reports their locations but the network snapshot at the time of the detection is not accessible. However in our model, every time a target detection is sensed, the result and corresponding provenance DAG is sent to the Central Storage. The scheme of the Central Storage is given in Figure 2. As the system has access to past right and wrong detections and the network snapshots at these times, it can find out the common patterns in false detections. For instance, comparing past 50 wrong detections and the provenance graph, a graph traversal algorithm can figure out the common pattern in the graphs such as “every time the target is nearby this group of sensors, the localization is done wrongly”. Network can be healed by replacing these nodes or omitting them. Although building a graph mining algorithm is not in the scope of this paper, the graph mining algorithm for bug finding in graphs developed by Abdelzaher et al is suitable to our model (Khan et al., 2009).

On the other hand, central repository of provenance information makes restructuring for the purpose of healing the network easier. For instance, if a faulty node has to be removed, it can be replaced by a close accurate node. For determining the closest accurate node, we can search the central provenance repository. Central repository also makes possible examining redundant sources that may come from different parts of the network. For instance a mobile de-

vice on a vehicle and a fixed sensor can sense the same data but the data from the vehicle may be captured in the network later after it moves far from where it collected the data. By using the provenance repository, we can trace back to the ancestors of the data and see that both data are captured in the same location and they are redundant.

6 CONCLUSIONS AND FUTURE WORK

In this position paper, we have built a provenance model for sensor networks. We illustrated our model on a binary target localization network. Our model makes it possible to build fault-tolerant sensor networks leveraging historical dataflow information. There are many opportunities for further research of using provenance in sensor networks. Open Provenance Model (OPM) was not designed primarily for sensor networks, it is lacking some important features such as handling real-time provenance. A model of provenance collection and dissemination for sensor networks requires further work. There are many research problems to consider. For instance provenance information may overwhelm the amount of actual data being collected. By storing the provenance data in a central store, errors can be found more efficiently than if the provenance data is stored within the network. However the amount of additional communications should be weighed with this. The best way to model the provenance, what to keep as provenance and how to make use of it in an efficient way can depend on the application, but general guidelines are still not well understood. As a future direction, after adapting OPM to sensor networks, we would like to do simulations of our revised model on real-world data sets.

ACKNOWLEDGEMENTS

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy right notation here on.

REFERENCES

- Bal, M., Shen, W., and Ghenniwa, H. (2010). Collaborative signal and information processing in wireless sensor networks: a review. In *2009 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3240–3245.
- Barseghian, D., Altintas, I., Jones, M., Crawl, D., Potter, N., Gallagher, J., Cornillon, P., Schildhauer, M., Borer, E., and Seabloom, E. (2010). Workflows and extensions to the kepler scientific workflow system to support environmental sensor data access and analysis. *Ecological Informatics*, 5(1):42–50.
- Buneman, P. and Tan, W. (2007). Provenance in databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1171–1173. ACM.
- Cheney, J. (2010). Causality and the semantics of provenance. *Arxiv preprint arXiv:1004.3241*.
- Cheney, J., Chiticariu, L., and Tan, W. (2009). Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474.
- Crawl, D. and Altintas, I. (2008). A provenance-based fault tolerance mechanism for scientific workflows. *International Provenance and Annotation Workshop (IPAW)*.
- Cui, Y., Widom, J., and Wiener, J. (2000). Tracing the lineage of view data in a warehousing environment. *ACM Transactions on Database Systems (TODS)*, 25(2):179–227.
- Davidson, S. and Freire, J. (2008a). Provenance and scientific workflows: challenges and opportunities. In *SIGMOD Conference*, pages 1345–1350. Citeseer.
- Davidson, S. and Freire, J. (2008b). Provenance and scientific workflows: challenges and opportunities. In *SIGMOD Conference*, pages 1345–1350. Citeseer.
- Dogan, G., Brown, T., Govindan, K., Khan, M., Abdelzaker, T., Mohapatra, P., and Cho, J. (2011). Evaluation of network trust using provenance based on distributed local intelligence. *MILCOM*.
- Feng, T. and Lee, E. (2008). Real-time distributed discrete-event execution with fault tolerance. In *IEEE Real-Time and Embedded Technology and Applications Symposium*, pages 205–214. IEEE.
- Freire, J., Koop, D., Santos, E., and Silva, C. (2008). Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21.
- Govindan, K., X., W., Khan, M., Dogan, G., Zeng, K., Powell, G., Brown, T., Abdelzaker, T., and Mohapatra, P. (2011). Pronet: Network trust assessment based on incomplete provenance. *MILCOM*.
- Khan, M., Abdelzaker, T., Han, J., and Ahmadi, H. (2009). Finding symbolic bug patterns in sensor networks. *Distributed Computing in Sensor Systems*, pages 131–144.
- Le, Q. and Kaplan, L. (2010). Target localization using proximity binary sensors. In *Aerospace Conference, 2010 IEEE*, pages 1–8. IEEE.
- Ledlie, J., Ng, C., and Holland, D. (2005). Provenance-aware sensor data storage. *IEEE Computer Society*.
- Moreau, L. and Ludascher, B. (2007). The first provenance challenge. *Concurrency and Computation: Practice and Experience*.
- Moreau, L., Ludascher, B., Altintas, I., Barga, R., Bowers, S., Callahan, S., Chin, J., Clifford, B., Cohen, S., Cohen-Boulakia, S., et al. (2008). Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418.
- Muniswamy-Reddy, K.-K. (2010). *Foundations for Provenance-Aware Systems*. PhD thesis, Harvard University, Massachusetts.
- Park, U. and Heidemann, J. (2008a). Provenance in sensor-net republishing. *Provenance and Annotation of Data and Processes*, pages 280–292.
- Park, U. and Heidemann, J. (2008b). Provenance in sensor-net republishing. In Freire, J., Koop, D., and Moreau, L., editors, *Provenance and Annotation of Data and Processes*, volume 5272 of *Lecture Notes in Computer Science*, pages 280–292. Springer Berlin / Heidelberg.
- Patni, H., Sahoo, S., Henson, C., and Sheth, A. (2010). Provenance Aware Linked Sensor Data. In *2nd Workshop on Trust and Privacy on the Social and Semantic Web, Co-located with ESWC2010, Heraklion Greece*.
- Sabelfeld, A. and Myers, A. C. (2003). Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21.
- Stephan, E., Halter, T., and Ermold, B. (2010). Leveraging The Open Provenance Model as a Multi-Tier Model for Global Climate Research. In *Proc. of 3rd International Provenance and Annotation Workshop (IPAW10), Troy, NY*.
- Tan, W.-C. (2007). Provenance in databases : Past, current, and future. *IEEE Data Engineering Bulletin*, 30:”3–12”.
- Tilak, S., Chiu, K., Abu-Ghazaleh, N., and Fountain, T. (2005). Dynamic resource discovery for sensor networks. *Embedded and Ubiquitous Computing*, pages 785–796.
- Wynbourne, M., Austin, M., Palmer, C., on Homeland Security, U. S. C. S. C., and Affairs, G. (2009). *National Cyber Security Research and Development Challenges Related to Economics, Physical Infrastructure and Human Behavior: An Industry, Academic and Government Perspective*. Institute for Information Infrastructure Protection.
- Zahedi, S., Szczodrak, M., Ji, P., Mylaraswamy, D., Srivastava, M., and Young, R. (2008). Tiered architecture for on-line detection, isolation and repair of faults in wireless sensor networks. In *Military Communications Conference, 2008. MILCOM 2008. IEEE*, pages 1–7. IEEE.