PLUS: Provenance for Life, the Universe and Stuff*

Adriane Chapman, M. David Allen, Barbara Blaustein, Len Seligman, Chris Wolf, Michael Morse, Arnon Rosenthal

The MITRE Corporation

{achapman, dmallen, bblaustein, seligman, cwolf, arnie}@mitre.org, michael.morse04@gmail.com

ABSTRACT

In this demonstration, we exhibit a new type of provenance system, one that is not tied to any particular domain, closed-world system or use. The PLUS provenance system was inspired by government requirements to enable provenance capture, storage and use across multi-organizational systems. PLUS is general enough to interact across open-world distributed systems, often without administrative access to those underlying distributed systems. It captures and stores provenance, permits user annotations, and provides tools for analyzing the provenance on the basis of those annotations. Due to the need to share provenance across many organizations, much attention has been paid to provenance access and security. We highlight all of these features via a demonstration using an Emergency Preparedness and Response (EP&R) scenario.

1. INTRODUCTION

PLUS [5] is a provenance manager developed by The MITRE Corporation that addresses requirements of our U.S. government customers not addressed by previous work, including:

- "Open world" collection in distributed, heterogeneous environments;
- Attribute-based access control that support flexible sharing of provenance across different organizations, classes of users, and privilege levels;
- Techniques to provide more informative provenance when the sensitivity of certain nodes or edges precludes sharing the entire graph, and;
- Flexible annotation management over provenance, which enables a number of important analysis applications.

In addition, we have done a number of experiments with both synthetic data and a large-scale simulation environment that have shown PLUS to scale effectively, as discussed below. We have demonstrated PLUS using a number of government-inspired scenarios spanning defense, intelligence analysis, homeland security, and emergency preparedness and response (EP&R).

In this demonstration, we will use an emergency preparedness and response (EP&R) scenario that highlights the importance of open world collection, the ability to protect sensitive information by substituting less sensitive information in provenance graphs, and an application using user annotations to understanding the downstream impacts of a data modification attack.

1.1. Current Provenance Systems

Provenance, or the history of information, has garnered interest in government, commercial and scientific circles. Topics of provenance study include capture [6, 11], storage [7], reasoning [3, 9], security [12, 19], usability [8, 14, 17], etc. However, all provenance systems to this point have been applied to "closed world" systems. A closed world system contains at least one of the following properties:

- The underlying application or systems are known in advance and provenance enabled.
- A provenance administrator has administrative privileges for the systems and applications in use.
- Full knowledge of either the data or processes is known in advance.

These assumptions work very well for scientific applications [2, 8, 10, 11, 13, 14, 17, 18], within relational databases[3, 9], and for specific applications [6, 15]. However, the real world is much messier. Below we describe current provenance systems, and then highlight the area in which their use is infeasible.

Workflow-Based. In workflow-based systems, such as [2, 14, 17], the user defines a series of processes and data. Workflows in these systems are defined explicitly; the user declares exactly the series of steps that will be performed. Upon execution of the workflow, provenance is captured by observing the system as it executes. Complete provenance capture of the workflow run is possible and is used, essentially, to recreate a scientific lab notebook. However, only provenance for events that occur within the workflow system can be captured.

Application-Based. In application-based provenance systems, particular applications are provenance capture enabled. For instance, in ES3 [10], the applications used by scientists for data analysis are modified to capture provenance of their use. Other applications that wish to be provenance-aware build this capability directly into the application [6]. This method allows complete capture of provenance information and detailed knowledge of the application execution. Unfortunately, every time

* While the answer to Ultimate Question of Life the Universe and Everything is 42 (according to Douglas Adams), a lot of provenance is required to understand where that data came from.

VLDB '10, September 13-17, 2010, Singapore.

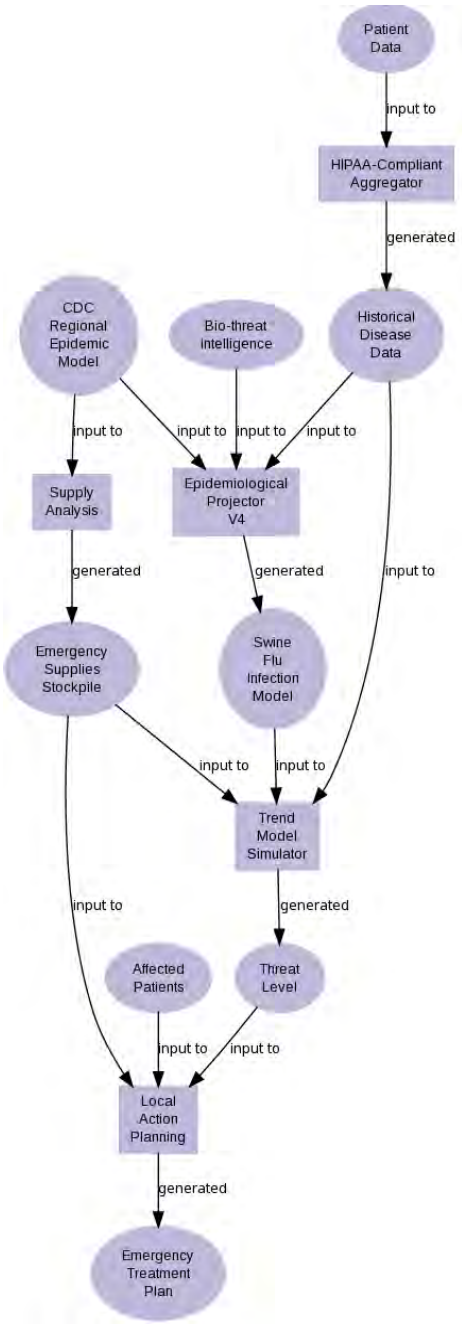
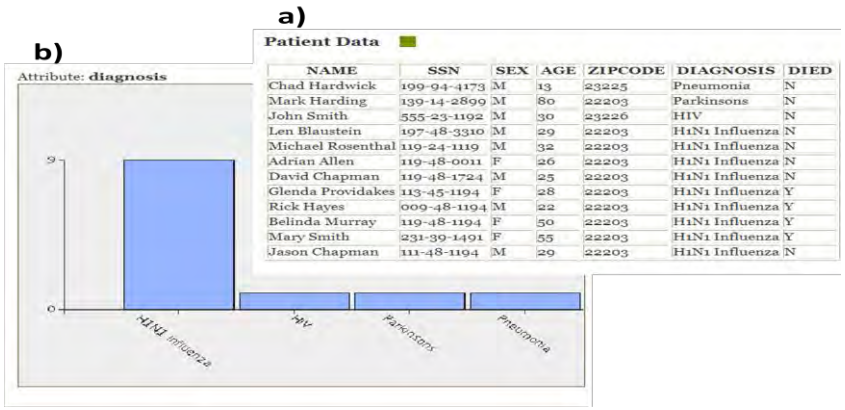


Figure 1: The provenance gathered from an Emergency Preparedness and Response Multi-Organizational scenario. Circles represent data; boxes processes.



NAME	SSN	SEX	AGE	ZIPCODE	DIAGNOSIS	DIED
Chad Hardwick	199-94-4173	M	13	23225	Pneumonia	N
Mark Harding	139-14-2899	M	80	22203	Parkinsons	N
John Smith	555-23-1192	M	30	23226	HIV	N
Len Blaustein	197-48-3310	M	29	22203	H1N1 Influenza	N
Michael Rosenthal	119-24-1119	M	32	22203	H1N1 Influenza	N
Adrian Allen	119-48-0011	F	26	22203	H1N1 Influenza	N
David Chapman	119-48-1724	M	25	22203	H1N1 Influenza	N
Glenda Providakes	113-45-1194	F	28	22203	H1N1 Influenza	Y
Rick Hayes	009-48-1194	M	22	22203	H1N1 Influenza	Y
Belinda Murray	119-48-1194	F	50	22203	H1N1 Influenza	Y
Mary Smith	231-39-1491	F	55	22203	H1N1 Influenza	Y
Jason Chapman	111-48-1194	M	29	22203	H1N1 Influenza	N

Figure 2: (a) Underlying (fake) patient data used to ultimately produce the Emergency Treatment Plan in Figure 1. (b) The surrogate version of the data for an uncleared Emergency Responder.

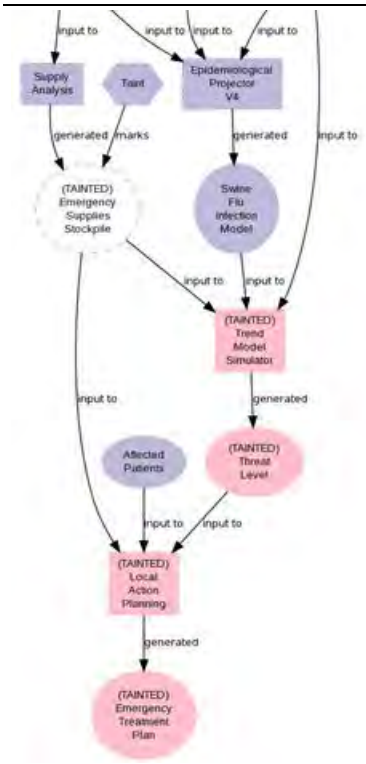


Figure 3: An example of a tainted Emergency Supplies Stockpile and the notification of data users that rely upon it.

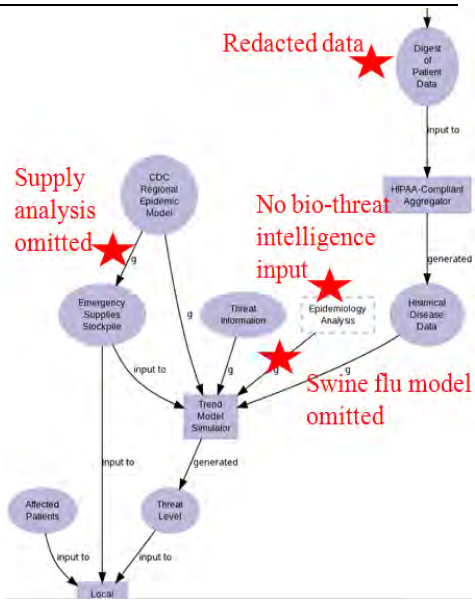


Figure 4: The provenance information from Figure 1 protected with surrogated nodes and edges.

a scientist finds a new application to assist with her job, it too must be provenance-enabled (entailing additional software development) for continued complete capture, administrator-like powers, or at least the ability to modify or "wrap" applications with a separate provenance capture program.

System-Based. The third type of provenance system is a system-based approach, as embodied by PASS [15], Karma [18] and PreServ [11]. Of the techniques described so far, this technique can capture provenance from the most diverse set of applications, and does not require pre-planning on the part of the user. PASS [15] observes the operating system and records all system calls. PreServ [11] and Karma [18] provide the ability to capture provenance across The Grid, which is a system of distributed machines, but is still a closed system in that it is logically centralized. However, the provenance calls must be inserted manually, raising the burden on the system engineer, and again

assuming administrative rights for all applications targeted by the user.

Relational. There are several systems that explore provenance in the relational world [3, 9]. The very closed nature of the relational database – well defined relational algebra, single system, etc – allow a more in-depth probe of provenance techniques. The provenance within these systems is complete, but only for actions that occur within the given relational system.

1.2. Our Contributions

Among our U.S. government customers, it is common for data to flow across organizational boundaries and for each autonomous stakeholder to use and transform the data using their own applications and processes, subject to their additional security concerns. Because data sharing partners are constantly evolving their agreements and exchanges, provenance capture must cross system and organization boundaries, there is usually no one who can pre-plan all data manipulation, or use only provenance-enabled applications, or piece together provenance gathered only on specific systems. The PLUS system [5] was developed with these distributed capture and usage needs in mind.. To be useful in an "open universe", provenance capture must:

- Capture provenance across multiple systems with no assumption of control over those systems, and
- Capture provenance from legacy systems that are not provenance aware.

In addition, to be useful to users within this distributed environment, the provenance system must:

- Enable annotation, for domain-dependent analysis;
- Enable many diverse players to specify access controls to specific provenance information; and
- Enable those access controls to be honored, while providing provenance that is as informative as possible.

1.3. Demonstration Scenario

In order to highlight these contributions, we demonstrate an EP&R multi-organizational scenario. Figure 1 shows the provenance information for this demonstration. Provenance reflects processing in multiple organizations including: hospitals, the Centers for Disease Control (CDC), local government, etc. Moreover, data in the scenario has different sensitivities and must be protected appropriately; e.g. patient data is sensitive while historical disease data is public.

Users of our system will be able to explore the data at various security levels, assess the impact of known suspect data items, trace data derivation across organizational boundaries, and discover new collaborators with similar data usage patterns.

2. SYSTEM DESIGN

2.1. Distributed Capture Methods

The first step for a provenance-enabled distributed system is the ability to capture the provenance information. Similar to [11, 18], we supply an API that any legacy system can call to log provenance information. However, in addition to this basic service, we have focused the system on "coordination" points that are often used in cross-organizational data sharing. For example, an Enterprise Service Bus (ESB) is often used to coordinate applications comprised of data and components from many different organizations. To this end, we have modified MULE, a popular open source ESB, to automatically capture and report provenance for all messages passed [1]. Our MULE-based provenance collector is the first provenance capture facility of which we are aware to collect provenance in heterogeneous multi-organizational environments. This capture technique scales effectively and does not noticeably impact the underlying systems.

2.2. PLUS Provenance Storage

Once provenance information is captured, it must be stored for later use. PLUS utilizes a MySQL database for provenance storage, and models provenance as a directed acyclic graph (DAG), $G = (N, E)$, containing a set of nodes, N, and a set of edges, E. Each node has a set of features describing the process or data it represents, e.g., timestamp, description, etc. Edges in the graph denote relationships, such as usedBy, generated, inputTo, etc., between nodes in the graph. A provenance graph may include disconnected subgraphs.

A data node can represent any object the user wishes to register, for example, strings, files, XML messages, relational data items of arbitrary granularity, etc. The data itself is not stored in the PLUS system for security and archiving reasons. However, enough "breadcrumbs" are maintained to point from the provenance information back to the original data as stored by the owning organization. For instance, the patient data shown in Figure 2a is stored in the hospital system, not within PLUS, but can be reached based on pointers such as connection information and the SQL query used to fetch the relation.

Several underlying methods are implemented to allow easy provenance traversal, such as "Bling" (for "backwards lineage") and "Fling" (for "forward lineage"). In other words, PLUS provides the capability for a user interested in a particular data item to trace exactly the data and processes that were used to create it (Bling) and how it was subsequently used (Fling).

2.3. Data Taint: An Application of Provenance Annotations

As in most prior work, the provenance that PLUS captures is immutable. However, in exploring our customers' desired uses of provenance, it quickly became clear that many of them required a flexible facility for adding annotations to provenance information. For example, a user may want to enter an opinion about his confidence in a particular piece of information or to note special circumstances that surrounded a particular process execution. These additional annotations (not provenance information) are mutable, since any of these assessments might change. Such annotations are essential in cross-organizational information sharing, in which a user from Organization2 who uses data from Organization1 may have no knowledge of how that data was generated, and whether it can truly be used for her purposes. In such cases, provenance together with user assessments of confidence and social networking friend-of-a-friend relationships can do a great deal to increase trust in information.

For instance, using these annotations, has the ability to help our customers understand the consequences of a data modification cyber attack. In our running example, suppose that analysts discover that an attacker has altered information about the Emergency Supplies Stockpile. In the past, an organization would correct the problem locally, but any other users of that data would

be blind to any actions they had already taken on the basis of bad information. PLUS provides the ability for a user to annotate the suspect data as being "tainted." This taint marking is then propagated to all data and processes that rely upon it, and can be seen by the data owners in a different organization (as in Figure 3), thus alerting them to a potentially serious issue.

2.4. Access Control and Surrogates

A fundamental requirement of any system that shares provenance across many organizations is the ability to protect those organizations' data sensitivity interests. Organizations may be comfortable with very different levels of sharing. For example, a hospital may be comfortable stating that it gave Patient Data to the CDC, but sharing the actual Patient Data with other provenance users would violate HIPAA. Meanwhile, an intelligence agency may not want the existence of BioThreat Intelligence to be divulged at all in the provenance store to users below a particular clearance.

PLUS provides two capabilities to assist with these security issues. First, it allows organizations to specify how the information they submit to the provenance store should be released [16]. This is accomplished through privilege predicates assigned to the nodes and the ability to restrict access to edges. These privilege predicates are compared to a user's authorization and determine whether a given node or edge should be displayed in the provenance information for that user.

In addition, PLUS provides the ability to provide alternative information to the users not authorized to see the underlying provenance information [4]. If the owning organization provides a surrogate node (via a Surrogate Generating Function[1]) and the user's authorization is high enough to see the surrogate, then the surrogate information will be shown. Figure 2b shows an example of a surrogate data node; the patient data has been redacted to show general trends in disease but no identifying information. This has the effect of maintaining better graph connectivity to take advantage of the Bling, Fling and Taint features discussed earlier. Compare Figure 4 to Figure 1 to see how sensitive information may be protected through surrogates.

3. CONCLUSION

In this work, we demonstrate the first provenance system designed to capture and use provenance in open world systems. Previous approaches all exist within a closed world assumption: the underlying system is under the control (or accessible by) the provenance collector and all provenance should be shared with the user. While this assumption holds for workflow systems, for provenance collected via applications on TheGrid, or for OS provenance loggers, many systems do not follow this paradigm. There are a plethora of systems that are distributed across organizations, and that are created by stringing applications together that have no common administrator. In this distributed, open world environment, it is currently impossible to capture or share provenance information.

Thus, it is now possible to capture provenance information outside of the closed world assumed by provenance collection methods up to this point. Moreover, the special use requirements of distributed environments (in terms of data tracing and security) are also addressed. We see this work as an initial step toward enterprise-scale (and even multi-organizational) provenance systems.

[1] An SGF is an arbitrary computation that the node owner provides, which when invoked, will return a different account of the data suitable for release to the querying user

4. REFERECES

[1] M. D. Allen, A. Chapman, B. Blaustein, and L. Seligman, "Provenance Capture in the Wild," in submission IPAW, 2010.

[2] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance Collection Support in the Kepler Scientific Workflow System," IPAW, pp. 118-132, 2006.

[3] O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom, "ULDBs: Databases with Uncertainty and Lineage," VLDB Seoul, Korea, pp. 953-964, 2006.

[4] B. Blaustein, A. Chapman, L. Seligman, M. D. Allen, and A. Rosenthal, "Surrogate Parenthood: Protected and Informative Provenance Graphs," in submission PVLDB, 2010.

[5] B. T. Blaustein, L. Seligman, M. Morse, M. D. Allen, and A. Rosenthal, "PLUS: Synthesizing privacy, lineage, uncertainty and security," ICDE Workshops, pp. 242-245, 2008.

[6] P. Buneman, A. Chapman, and J. Cheney, "Provenance Management in Curated Databases," ACM SIGMOD, pp. 539-550, 2006.

[7] A. Chapman, H. V. Jagadish, and P. Ramanan, "Efficient Provenance Storage," SIGMOD, pp. 993-1006, 2008.

[8] S. Cohen-Boulakia, O. Biton, S. Cohen, and S. Davidson, "Addressing the provenance challenge using ZOOM," Concurrency and Computation: Practice and Experience, vol. 20, pp. 497-506, 2008.

[9] J. N. Foster, T. J. Green, and V. Tannen, "Annotated XML: Queries and Provenance," PODS, pp. 271-280, 2008.

[10] J. Frew, D. Metzger, and P. Slaughter, "Automatic capture and reconstruction of computational provenance," Concurr. Comput. : Pract. Exper., vol. 20, pp. 485-496, 2008.

[11] P. Groth, S. Miles, and L. Moreau, "PReServ: Provenance Recording for Services," UK OST e-Science second AHM, 2005.

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

[13] P. Missier, K. Belhajjame, J. Zhao, and C. Goble, "Data lineage model for Taverna workflows with lightweight anotation requirements," in IPAW, 2008.

[14] P. Missier, S. M. Embury, M. Greenwood, A. Preece, and B. Jin, "Managing information quality in e-science: the qurator workbench," SIGMOD, pp. 1150-1152, 2007.

[15] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. I. Seltzer, "Provenance-Aware Storage Systems," USENIX, pp. 43-56, 2006.

[16] A. Rosenthal, L. Seligman, A. Chapman, and B. Blaustein, "Scalable Access Controls for Lineage," in Theory and Practice of Provenance, 2008.

[17] C. E. Scheidegger, H. T. Vo, D. Koop, J. Freire, and C. Silva, "Querying and Re-Using Workflows with VisTrails," SIGMOD, 2008.

[18] Y. Simmhan, B. Plale, and D. Gannon, "Karma2: Provenance Management for Data Driven Workflows," Journal of Web Services Research, vol. 5, 2008.

[19] J. Zhang, A. Chapman, and K. LeFevre, "Fine-Grained Tamper-Evident Data Pedigree," in Secure Data Management. Lyon, France, 2009.