Data Provenance and Security

Patrick McDaniel
Pennsylvania State University

An unanticipated consequence of the Internet age is a pervasive loss of context. Users and organizations receive information in incredible volumes from both internal processes and far-flung, untrusted, and sometimes unknown sources.

Information is often filtered, sampled, repackaged, condensed, or altered to suit any number of purposes. Over time, the entropy of these processes causes information to lose its essential validity. Far too often, we're left with data without knowledge.

Context is necessary for good decision-making. Whether attempting to determine if a website is legitimate or diagnosing a patient ailment, understanding the origins and processes that formed the data is essential to a successful outcome. However, this understanding is frequently cobbled together from vague notions of physical origins and assumed data-handling practices. Intelligence failures, poor medical diagnoses, and monetary losses resulting from poor source data and attribution are far too common in both physical processes and computer systems.

Data provenance documents data's genesis and subsequent modification as it's processed in and across systems. Manifest as annotations, logs, or workflow histories, provenance indicates the data's origins and pedigree. Data provenance and security are symbiotic. Good security leads to accurate, timely, and detailed provenance, and good provenance lets systems and users make good security decisions. Unfortunately, support for provenance in information systems is largely experimental at this time.

Definition and History of Provenance

In its strongest form, data provenance supports information and process integrity by documenting the entities, systems, and processes that operate on and contribute to data of interest. This serves as an unalterable historical record of the data's lifetime and its sources.[1] We can think of provenance as a collection of annotations bound to the tracked data,[2] analogous to that required for physical evidence in modern legal systems, for instance, crime scene inventories and chains of custody. The scope of provenance is a reflection of its desired use. Provenance can be measured on high-level objects, such as websites or datasets; low-level objects, such as database records or files; and everything in between.

Separating *why* from *where* provenance is useful. *Why provenance* tracks the reasons that particular data is in the state it's in. Often, this is a record of the processes and inputs that created or transformed the data. *Where provenance* enumerates the entities that contributed to the data. Consider a patient x-ray image document used in a healthcare system. The provenance record for that document would identify the x-ray inputs; any postprocessing operations, such as image enhancements; the annotations created by specialists (why provenance) as well as technicians, nurses, and doctors who created and modified the record; the hosts on which the software executed; and the times and dates the processes were formed (where provenance).

Data provenance has had its widest adoption in the scientific computing community.[3] Provenance is often recorded in these environments by capturing data from instrumented workflows and systems that process large datasets. The database community has recently explored how to support provenance collection in database records and streams by retaining query structures and input. Others have examined the use of these features to measure provenance in social networks and general information retrieval processes.

A more recent area of focus is provenance in operating and storage systems, which have the advantage of being difficult to circumvent, but are expensive and hard to deploy. The Provenance-Aware Storage System uses a modified Linux kernel, collecting elaborate information flow and workflow descriptions at the OS level.[4] Further efforts defined

1540-7993/11/$26.00 © 2011 IEEE

a query language and attempted to embed provenance information directly in applications and the OS. Others have begun to look at whole-system provenance that tracks all inputs and process interactions on a given system, which provides a detailed forensic record for failure analysis and intrusion detection in high-assurance applications.

Other provenance efforts have focused on provenance information collection, semantic analysis, and dissemination, but little has been done to ensure the data's security and privacy. Whether tracing sensor data from a pipeline or tracing dependencies between clinical data in a drug trial, it's essential that the provenance be secure against manipulation. Failure to provide such protection leaves the supported system open to misuse. For example, sensor readings could be manipulated to induce or ignore catastrophic failures or mislead drug developers, researchers, and agencies governing drug experiments (for example, the US Food and Drug Administration).

Applications of Provenance

There have been longstanding calls for provenance in computer systems. A 2009 report prepared for the chairman and ranking member of the US Senate Committee on Homeland Security and Governmental Affairs highlighted provenance as one of three key future technologies for securing our national critical infrastructure.[5] The report cited a need to ascertain sensor data provenance as it's recorded and aggregated in cyberphysical systems, such as the electrical grid and SCADA environments.

Consider a hypothetical initiative to collect electronic voting data provenance. In such a system, one could trace voter registration, ballot definitions, polling place records, vote tallies, and outcomes to the software, devices, and actors creating and manipulating the election data. Such a system would vastly improve the transparency and ability to validate elections. Election officials and concerned citizens could simply review a public provenance record to identify and explain any problems resulting from elections. Possibly more important, these records could disprove or confirm claims of malfeasance and potentially recover from failed processes. Faulty machines, incorrect processes, and bad actors could be unambiguously identified.

Here, the advantages of a provenance record clearly cut across security and performance goals. When correctly used, provenance enhances essential functions such as system calibration, auditing, and quota and billing management. Provenance records are further instrumental in assessing and documenting compliance with regulatory policy, ensuring privacy, identifying and repairing data inconsistencies, understanding process flows and efficiencies, enabling cost efficiencies by removing redundancies, and identifying and recovering from malicious behavior.

Industrial efforts in provenance are increasing. The healthcare industry has been aggressively investing in provenance technology. Such investment is primarily motivated by cost efficiencies and requirements for documented compliance with Health Insurance Portability and Accountability Act regulations. Similar efforts are ongoing in the legal, biotech, and manufacturing sectors.

The Challenges of Provenance

Supporting provenance in current information systems presents several technical challenges. Where do you store the provenance record? Past provenance systems have augmented storage systems (file systems, databases) with means to annotate data per some system policy. However, most of these systems are tailored to a specific environment and purpose, making their generalization difficult. Other options include keeping centralized or federated repositories for collecting and warehousing provenance data, potentially at the cost of reliability and availability. Extending such records to span multiple systems is largely unexplored.

Provenance records can grow significantly over time, adding data storage, transmission, and processing costs, so provenance systems must be vigilant in managing data. Pruning irrelevant or outdated records can mitigate provenance size costs. However, such filtering assumes you know what's interesting before you need to know it, which isn't always the case.

Understanding and harmonizing the semantics of data and its provenance log is also often problematic. Standardizing ontologies and using globally unique identities can help this process, but providing detailed and universally meaningful provenance in open environments such as the Internet is problematic at best.

Finally, there's an enormous security challenge in providing provenance. When presented with a provenance record for a photograph received over the Internet, how do you know that record is legitimate? How do you ensure a secure binding between a record and the data it describes? What structures are necessary to make such a record usable and reliable in practice? Even for closed, highly structured, and well-managed environments such as healthcare systems, these challenges are daunting.

How we will find answers to these challenges remains unclear. However, increasing support from funding agencies and industrial sectors is providing opportunities to make progress. The research has just begun, but many systems and studies are helping to illuminate problems and solutions.

Our increasing reliance on cybersystems mandates well-informed decision-making. This is particularly true when such decisions influence its users' health, safety, and security. The next generation of computing systems must be able to base such decisions on not only the manifest data but also an understanding of where that data came from and how it evolved. In the absence of such information, we're destined to relive the history of poor choices made on poorly understood data. Future provenance systems will strive to provide that essential context and thereby provide the means to make our systems safer and more secure. □

References

1. P. Buneman, S. Khanna, and W.C. Tan, "Data Provenance: Some Basic Issues," *Proc. 20th Conf. Foundations of Software Technology and Theoretical Computer Science* (FST TCS 00), Springer-Verlag, 2000, pp. 87–93.
2. J. Bentley, "Programming Pearls: Self-Describing Data," *Comm. ACM*, vol. 30, no. 6, 1987; http://portal.acm.org/citation.cfm?id=315733.
3. R. Bose and J. Frew, "Lineage Retrieval for Scientific Data Processing: A Survey," *J. ACM Computing Surveys*, vol. 37, no. 1, 2005, pp. 1–28.
4. K.-K. Muniswamy-Reddy et al., "Provenance-Aware Storage Systems," *Proc. Usenix Ann. Technical Conf.*, Usenix Assoc., 2006.
5. M.N. Wybourne, M.F. Austin, and C.C. Palmer, "National Cyber Security Research and Development Challenges," Inst. for Information Infrastructure Protection, 2009; www.thei3p.org/docs/publications/i3pnationalcybersecurity.pdf.

Patrick McDaniel *is an associate professor at Pennsylvania State University's Computer Science and Engineering Department. Contact him at mcdaniel@cse.psu.edu.*