Workshop on Theory and Practice of Provenance

Event Report

James Cheney
University of Edinburgh
jcheney@inf.ed.ac.uk

Abstract

Provenance, or metadata about the creation, influences upon, or other history of objects or data, has attracted attention in a wide variety of contexts in computer science over the last few years. This event report describes a recent workshop on "Theory and Practice of Provenance", intended as a forum for presenting novel ideas about provenance and encouraging interaction among provenance researchers.

1. Introduction

Provenance is metadata about the creation, influences upon, and other history of objects or data. In computer systems, such metadata is often used for security and profiling, and is essential for making informed judgments about data quality, integrity, and authenticity. Recent research in a variety of settings (databases and data warehouses, geographic information systems, scientific workflows, grid computing, and the Semantic Web) has begun addressing the problem of recording and exploiting provenance. In addition, forms of provenance are now used in several areas of advanced computer systems research such as probabilistic databases, incremental computation, information-flow security, file synchronization, and annotation management systems. Other topics, such as version control and archiving, may also benefit from better understanding of provenance.

However, provenance research has so far been carried out in relative isolation in a number of disciplines of computer science, such as databases, scientific computation, systems, and security; moreover, topics such as information flow security [5], dependence analysis in programming languages [2], causal models in artificial intelligence [4], and traceability in software engineering appear related, yet there has been little investigation of applications of this work to provenance. We believe that more meaningful interaction between the theory and systems communities, between academia and industry, and between different branches of computer science is needed to make progress on understanding and implementing reliable, efficient and scalable provenance management in computer systems.

The First Workshop on *Theory and Applications of Provenance* (TaPP) was held on February 23, 2009 in San Francisco, California, just before the 2009 USENIX Conference on *File and Storage Technology* (FAST 2009). The goals of the workshop included presenting cross-disciplinary, foundational and highly speculative research, raising the profile of provenance research, and increasing interaction between provenance researchers in several subdisciplines, the broader systems community, and industry. We believe that TaPP made significant contributions to all of these goals.

TaPP had a lightweight, but formal peer-review process with a nine-member program committee representing a variety of backgrounds, including databases, scientific workflow computation, software engineering, security, and systems. TaPP attracted 22 submissions, comprising 13 long papers and 9 short papers. Submissions were reviewed by at least three members of the program committee. Five long papers were selected for long presentations, and another five long papers and three short papers were selected for short presentations. This led to a crowded, but lively schedule.

The workshop attracted 38 participants (including speakers and organizers), representing a healthy mixture of academic and industrial backgrounds. TaPP was fortunate to be able to invite two excellent speakers using support from eSI. They were Joe Halpern (Cornell University) and Margo Seltzer (Harvard University). In the rest of this event report, we list the presentations and abstracts, and where appropriate discuss additional highlights of the workshop, particularly discussions following the invited talks.

TaPP 2009 was supported by the United Kingdom eScience Institute (eSI) through a Theme Program on "Principles of Provenance", and significant organizational support was provided by USENIX. USENIX has also indicated enthusiasm for supporting future iterations of the workshop. The proceedings of TaPP are published online by USENIX [3], and a complementary event report that provides further details about the audience questions and discussion has been published in the USENIX newsletter [1].

2. Contributions

2.1 Session 1: Joe Halpern, Cornell University

Professor Halpern's talk, entitled "Causality, Responsibility, and Blame: A Structural-Model Approach" focused on

mathematical models of concepts such as *causality*. The concept of causality has troubled philosophers and scientists for hundreds of years, and scientists have been wary of the concept because of Hume's critical analysis of causality as a subjective (or psychological) rather than objective phenomenon. Halpern, together with Pearl, Chockler and Kupferman, has made valuable contributions towards mathematically characterizing the concepts of *cause* and *actual cause*. We invited Professor Halpern based on our belief that mathematical models of causality are important for understanding provenance, especially forms of provenance based on intuitions such as "cause", "dependence", or "influence". We believe the lively discussion during and after the talk confirms our belief that the provenance community can benefit from learning more about the models of causality developed by Halpern and Pearl [4].

2.2 Session 2: Provenance and Security

A Formal Model of Provenance in Distributed Systems. Issam Souilah, University of Southampton, UK; *Adrian Francalanza*, University of Malta, Malta; Vladimiro Sassone, University of Southampton, UK

Abstract: We present a formalism for provenance in distributed systems based on the π-calculus. Its main feature is that all data products are annotated with metadata representing their provenance. The calculus is given a provenance tracking semantics, which ensures that data provenance is updated as the computation proceeds. The calculus also enjoys a pattern-restricted input primitive which allows processes to decide what data to receive and what branch of computation to proceed with based on the provenance information of data. We give examples to illustrate the use of the calculus and discuss some of the semantic properties of our provenance notion. We conclude by reviewing related work and discussing directions for future research.

Towards Semantics for Provenance Security. *Stephen Chong*, Harvard University

Abstract: Provenance records the history of data. Careless use of provenance may violate the security policies of data. Moreover, the provenance itself may be sensitive information, necessitating restrictions on the use of both data and provenance to enforce security requirements. This paper proposes extensional semantic definitions for provenance security. The semantic definitions require that provenance information released to the user does not reveal confidential data, and that neither the provenance information given to the user, nor the programs output, reveal sensitive provenance information.

Scalable Access Controls for Lineage. Arnon Rosenthal, Len Seligman, *Adriane Chapman*, and Barbara Blaustein, The MITRE Corporation

Abstract: Lineage stores often contain sensitive information that needs protection from unauthorized access. We build on prior work for security and privacy of lineage information, focusing on complex conditions and scalable administration. We use Attribute-Based Access Control (ABAC) to express conditions based on many attributes, instead of roles. We then make administration and management more scalable, instead of managing large, monolithic access predicates for each object. To do so, we first support modular traceability and maintainability for separate concerns (e.g. security, legally mandated privacy, organizationally mandated privacy). We then provide constructs to manage authority when multiple administrators must collaborate. We show that these security techniques are needed for easy lineage security administration.

2.3 Session 3: Provenance for Web, Grid and Digital Libraries

On Explicit Provenance Management in RDF/S. Graphs P. Pediaditis, *G. Flouris*, I. Fundulaki, and V. Christophides, ICS-FORTH

Abstract: The notion of RDF Named Graphs has been proposed in order to assign provenance information to data described using RDF triples. In this paper, we argue that named graphs alone cannot capture provenance information in the presence of RDFS reasoning and updates. In order to address this problem, we introduce the notion of RDF/S Graphsets: a graphset is associated with a set of RDF named graphs and contain the triples that are jointly owned by the named graphs that constitute the graphset. We formalize the notions of RDF named graphs and RDF/S graphsets and propose query and update languages that can be used to handle provenance information for RDF/S graphs taking into account RDFS semantics.

Application of Named Graphs Towards Custom Provenance Views. *Tara Gibson*, Karen Schuchardt, and Eric Stephan, Pacific Northwest National Laboratory

Abstract: Provenance capture as applied to execution oriented and interactive workflows is designed to record minute detail needed to support a "modify and restart" paradigm as well as re-execution of past workflows. In our experience, provenance also plays an important role in human-centered verification, results tracking, and knowledge sharing. However, the amount of information recorded by provenance capture mechanisms generally obfuscates the conceptual view of events. There is a need for a flexible means to create and dynamically control user oriented views over the detailed provenance record. In this paper, we present a design which leverages named graphs and extensions to the SPARQL query language to create and manage views as a server-side function, simplifying user presentation of provenance data.

Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. *Michael Factor*, Ealan Henis, Dalit Naor, Simona Rabinovici-Cohen, Petra Reshef, and Shahar Ronen, IBM Research Lab in Haifa, Israel; Giovanni Michetti and Maria Guercio, University of Urbino, Italy

Abstract: A growing number of digital objects are desig-

nated for long term preservation - a time scale during which technologies, formats and communities are very likely to change. Specialized approaches, models and technologies are needed to guarantee the long-term understandability of the preserved data. Maintaining the authenticity (trustworthiness) and provenance (history of creation, ownership, accesses and changes) of the preserved objects for the long term is of great importance, since users must be confident that the objects in the changed environment are authentic. We present a novel model for managing authenticity in long term digital preservation systems and a supporting archival storage component. The model and archival storage build on OAIS, the leading standard in the area of long-term digital preservation. The preservation aware storage layer handles provenance data, and documents the relevant events. It collocates provenance data (and other metadata) together with the preserved data in a secure environment, thus enhancing the chances of their co-survival. Handling authenticity and provenance at the storage layer reduces both threats to authenticity and computation times. This work addresses core issues in long-term digital preservation in a novel and practical manner. We present an example of managing authenticity of data objects during data transformation at the storage component.

Steps Toward Managing Lineage Metadata in Grid Clusters. *Ashish Gehani* and Minyoung Kim, SRI International; Jian Zhang, Louisiana State University

Abstract: The lineage of a piece of data is of utility to a wide range of domains. Several application-specific extensions have been built to facilitate tracking the origin of the output that the software produces. In the quest to provide such support to extant programs, efforts have been recently made to develop operating system functionality for auditing filesystem activity to infer lineage relationships. We report on our exploration of mechanisms to manage the lineage metadata in Grid clusters.

2.4 Session 4: Margo Seltzer, Harvard University

Professor Seltzer spoke on the topic "The State of Provenance in 2019". She gave a talk prepared for the 11th Workshop on Theory and Practice of Provenance, which she expects to take place ten years in the future. Her talk envisages a world in which provenance research has borne fruit in many areas of computer science — provenance is part of all storage systems, databases, and applications; provenance is secure and queryable, and search engines such as "Poogle" offer provenance-aware informational retrieval. Professor Seltzer then outlined the development of this state of affairs, from the work that has already (really) been done in 2009, to her predicted view in 2019. She predicts that the main thrusts of this development will require major progress on *storing and querying* provenance, *standards* for transmitting and integrating provenance (encompassing both file systems and network protocols), and *formalism* or theoretical understanding of provenance. Furthermore, security concerns will come to dominate as provenance becomes essential for accountability for sensitive data, and once the technology has been developed, significant effort will also be needed to establish standards and governmental policy that will legislate good provenance practices.

We look forward to hearing an updated version of this talk in 2019.

2.5 Session 5: Provenance Systems I

A Framework for Fine-grained Data Integration and Curation, with Provenance, in a Dataspace. *David W. Archer*, Lois M.L. Delcambre, and David Maier, Portland State University

Abstract: Some tasks in a dataspace (a loose collection of heterogeneous data sources) require integration of fine-grained data from diverse sources. This work is often done by end users knowledgeable about the domain, who copy-and-paste data into a spreadsheet or other existing application. Inspired by this kind of work, in this paper we define a data curation setting characterized by data that are explicitly selected, copied, and then pasted into a target dataset where they can be confirmed or replaced. Rows and columns in the target may also be combined, for example, when redundant. Each of these actions is an integration decision, often of high quality, that when taken together comprise the provenance of a data value in the target. In this paper, we define a conceptual model for data and provenance for these user actions, and we show how questions about data provenance can be answered. We note that our model can be used in automated data curation as well as in a setting with the manual activity we emphasize in our examples.

The Case for Browser Provenance. *Daniel W. Margo* and Margo Seltzer, Harvard University

Abstract: In our increasingly networked world, web browsers are important applications. Originally an interface tool for accessing distributed documents, browsers have become ubiquitous, incorporating a significant portion of user interaction. A modern browser now also reads email, plays media, edits documents, and runs applications. Consequently, browsers process large quantities of data, and must record metadata, such as history, to help users manage their data. Most of the metadata that modern browsers record is actually provenance - metadata that captures the causality and lineage of data obtained via the browser. We demonstrate that characterizing browser metadata as provenance and then applying techniques from the provenance research community enables new browser functionality. For example, provenance can improve both history and web search by indicating contextual and personal relationships between data items. Users can also answer complex questions about the origins of their data by querying provenance. Our initial results suggest these features are feasible to implement and could perform well in modern browsers.

Provenance as Data Mining: Combining File System Metadata with Content Analysis. *Vinay Deolalikar* and

Hernan Laffitte, Hewlett Packard Labs
Abstract: Provenance describes how an object came to be in its present state. Thus, it describes the evolution of the object over time. Prior work on provenance has focused on databases and the file system. The database or file system is enhanced or augmented in order to capture additional information about the historical evolution of document collections, and thus answer the provenance question. We address the question of provenance for unstructured information (i.e., document corpora from file systems) but without any enhancements to the file system. To provide a solution in this setting, we model the provenance problem in such a setting as a problem of data mining. We show that data mining can provide provenance information for repositories of unstructured information, including chains of historical evolution. Thus, we do not require any additions to the file system, and we can operate on legacy documents. Experimental results indicate a strong performance of our approach.

This presentation led to an interesting discussion on how to objectively evaluate such "provenance mining" techniques, given that we cannot easily construct "gold standard" test data in the absence of systems that already record provenance.

2.6 Session 6: Provenance Systems II

Story Book: An Efficient Extensible Provenance Framework. *R. Spillane*, Stony Brook University; R. Sears, University of California, Berkeley; C. Yalamanchili, S. Gaikwad, M. Chinni, and E. Zadok, Stony Brook University
Abstract: Most application provenance systems are hard coded for a particular type of system or data, while current provenance file systems maintain in-memory provenance graphs and reside in kernel space, leading to complex and constrained implementations. Story Book resides in user space, and treats provenance events as a generic event log, leading to a simple, flexible and easily optimized system.

We demonstrate the flexibility of our design by adding provenance to a number of different systems, including a file system, database and a number of file types, and by implementing two separate storage backends. Although Story Book is nearly 2.5 times slower than ext3 under worst case workloads, this is mostly due to FUSE message passing overhead. Our experiments show that coupling our simple design with existing storage optimizations provides higher throughput than existing systems.

Making a Cloud Provenance-Aware. *Kiran-Kumar Muniswamy-Reddy*, Peter Macko, and Margo Seltzer, Harvard University
Abstract: The advent of cloud computing provides a cheap and convenient mechanism for scientists to share data. The utility of such data is obviously enhanced when the provenance of the data is also available. The cloud, while convenient for storing data, is not designed for storing and querying provenance. In this paper, we present desirable properties for distributed provenance storage systems and present design alternatives for storing data and provenance on Amazon's popular Web Services platform (AWS). We evaluate the properties satisfied by each approach and analyze the cost of storing and querying provenance in each approach.

Transparently Gathering Provenance with Provenance Aware Condor. *Christine F. Reilly* and Jeffrey F. Naughton, University of Wisconsin-Madison
Abstract: We observed that the Condor batch execution system exposes a lot of information about the jobs that run in the system. This observation led us to explore whether this system information could be used for provenance. The result of our explorations is Provenance Aware Condor (PAC), a system that transparently gathers provenance while jobs run in Condor. Transparent provenance gathering requires that the application not be altered in order to run in the provenance system. This requirement allows any application that can run in Condor to also run in PAC. Through SQL queries, PAC is able to answer a wide range of questions about the files used by a job and the machines that execute jobs.

3. Conclusions

We believe that there is a clear need for a combination of both new theoretical insights and new systems research in tackling the many challenges of data provenance. The Theory and Practice of Provenance workshop series will be supported by USENIX and will, we hope, develop into a forum for encouraging, recognizing and disseminating great new ideas in this area.

Acknowledgments We would like to gratefully acknowledge the USENIX staff for their support in organizing TaPP 2009, particularly Casey Henderson, Jane-Ellen Long, Devon Shaw, and Ellie Young.

References

[1] Event report: First workshop on theory and practice of provenance (TaPP '09). *;login: The USENIX Magazine*, 34(4):84–90, June 2009. Available online at: `http://www.usenix.org/publications/login/2009-06/`.

[2] Martín Abadi, Anindya Banerjee, Nevin Heintze, and Jon G. Riecke. A core calculus of dependency. In *POPL*, pages 147–160. ACM Press, 1999.

[3] James Cheney, editor. *First Workshop on the Theory and Practice of Provenance, February 23, 2009, San Francisco, CA, USA, Proceedings*. USENIX, 2009. `http://-www.usenix.org/event/tapp09/tech/`.

[4] J.Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach–part I: Causes. *British J. Philos. Sci.*, 56:843–887, 2005.

[5] Andrei Sabelfeld and Andrew Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.