Guest Editorial: The Provenance of Online Data

Guest Editorial: The Provenance of Online Data

ADRIANE CHAPMAN, University of Southampton
JAMES CHENEY, University of Edinburgh
SIMON MILES, King's College London

ACM Reference format:
Adriane Chapman, James Cheney, and Simon Miles. 2017. Guest Editorial: The Provenance of Online Data. *ACM Trans. Internet Technol.* 17, 4, Article 33 (August 2017), 3 pages.
https://doi.org/10.1145/3108938

1 INTRODUCTION

Across many domains, there is a need to trace how data has been created, manipulated, and disseminated. This has led to strong recent interest in technology for modelling and reasoning about *provenance*. Provenance is information about the entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability, or trustworthiness. It is itself data, commonly represented as a directed acyclic graph linking these elements (entities, activities, and agents) to the earlier elements that influenced them. Provenance is becoming a key Internet technology, and the World Wide Web Consortium has standardised PROV as a representation for exchanging provenance on the (Semantic) Web. It is also important in a number of other settings to address the problems that arise in a distributed, internetworked world: for example, the need to document the sources of information in order to establish their trustworthiness, and the need to secure critical systems from attackers who can, thanks to the Internet, be located anywhere in the world.

Access to provenance information underlies our ability to interpret and to judge the reliability of data, whether on the Web, in databases, or within and between applications. Despite much recent progress, it is still uncommon for people or software to have access to the provenance of online data. The formal requirements for provenance, including issues such as correctness, completeness, and security of provenance, are not yet fully understood. The question of what is semantically useful provenance and how to capture it is still open, as are benchmarks that could be used to measure the performance of proposed systems. Moreover, as the patterns of use on the Internet change, with greater prevalence of crowdsourcing of information and services, virtualisation of applications in clouds, location-aware streaming, and so on, both the technological and social requirements on provenance are evolving.

Authors' addresses: A. Chapman, Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK; email: Adriane.Chapman@soton.ac.uk; J. Cheney, Laboratory for Foundations of Computer Science, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK; email: jcheney@inf.ed.ac.uk; S. Miles, Department of Informatics, King's College London, Strand, London WC2R 2LS, UK; email: simon.miles@kcl.ac.uk.

1533-5399/2017/08-ART33
https://doi.org/10.1145/3108938

2 THIS SPECIAL ISSUE

In general, handling of provenance offers open problems in several broad categories: capture, storage, querying/usage, and administration including security. Provenance relates to a number of other communities and topics, but over the past 10 years we have seen provenance research begin to take shape as a distinctive field, with its own events such as the International Provenance and Annotation Workshop (IPAW, biennial since 2006) and the workshop on Theory and Practice of Provenance (TaPP, annual since 2009), as well as a number of other events. This special issue on the Provenance of Online Data recognises the increasing maturity and coherence of the provenance community, and collects four contributions that address some of the challenges listed above.

In the first article, "Taming the Costs of Trustworthy Provenance through Policy Reduction" by Bates et al., provenance is collected to help secure web servers or other Internet-accessible systems. Provenance tracking at the operating system level has been explored over the past decade, often motivated by security goals. Recording fine-grained provenance, a typical system under load potentially imposes substantial overheads due to extra time spent logging actions and extra storage space needed to retain them for later inspection. However, typically the vast majority of the activity is benign, repetitive, or security-irrelevant. Bates et al. seek to decrease the burden of recording provenance in security scenarios by leveraging the information-flow policies enforced in security-enhanced operating systems such as SELinux. Their system, ProvWalls, decreases the storage required for some typical applications by up to 88%, without significant recording slowdown. This is an important step towards making provenance-based security practical for production systems, because it not only improves performance but also may remove security-irrelevant information, easing later analysis.

In "A Canonical Form for PROV Documents and its Application to Equality, Signature, and Validation" by Moreau, the focus is on preventing provenance information from being tampered with. This work provides a means to allow a party to check that provenance data has not been tampered with since creation. Digital signatures provide a well-understood means of verifying that data has not been modified, but relies on the format of the data to remain exactly the same, which is problematic when the same information can be expressed in multiple ways and may change during communication. This issue is solved in various serialization formats, such as XML and RDF, by providing a single canonical form into which data in that format can be transformed prior to signing and validation. However, this only tackles the potential differences at the level of serialization, whereas provenance data in a model such as PROV can be expressed in differing ways even if the serialization remains canonical, due to the flexible semantics it allows. To address this problem in PROV, and to help in general to provide a means for testing equality of PROV documents, Moreau specifies a canonical form for PROV, and discusses and evaluates its uses and limitations.

Provenance has been touted as the mechanism to understand where data came from and what happened to it. "Managing Provenance of Implicit Data Flows in Scientific Experiments" by Neves et al. adds an interesting twist to the usage problem. Traditionally, incomplete provenance is augmented by enabling capture points from different tools and perspectives and merging the provenance that is generated from several viewpoints. In addition to furthering an understanding of the provenance capture problem, the techniques developed within this work allow scientists to understand data evolution throughout the lifecycle of the project in addition to what happens within the context of the workflow tools used by the scientist.

Finally, in "PROV2R: Unstructured Processes Provenance Analysis" by Stamatogiannakis et al., the problem of capturing unstructured and unanticipated events and application usage is tackled. Taint analysis and record and replay technologies are adapted and analysed for use as a method

of capturing provenance. This method would change the capture problem from application-by-application modification to a general purpose tool. The requirements and performance of using established taint tracking systems for provenance capture is studied. In addition, the problem of what information is appropriate provenance is reviewed.

ACKNOWLEDGMENTS

We would like to thank the authors of all submissions, the referees, and the editorial staff of Transactions on Internet Technology, all of whom have worked hard to make this special issue possible and keep it on schedule.