On Provenance and Privacy

Susan B. Davidson

University of Pennsylvania

Provenance is a double-edged sword. On the one hand, it enables transparency, understanding the ”why” and ”where” of data, and reproducibility of results. On the other hand, it potentially exposes intermediate data and the functionality of modules within the workflow. However, a scientific workflow often deals with proprietary modules as well as private or confidential data, such as genomic or medical information. Hence providing exact answers to provenance queries over all executions of the workflow may reveal private information. In this talk we discuss potential privacy issues in a scientific workflow - module privacy, data privacy, and provenance privacy - and frame several natural questions: (i) Can we formally analyze module, data or provenance privacy giving provable privacy guarantees for an unlimited/bounded number of provenance queries? (ii) How can we answer provenance queries, providing as much information as possible to the user while still guaranteeing the required privacy? Then we look at module privacy in detail and propose a formal model. Finally we point to several directions for future work.

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, p. 1, 2010.