Privacy-Preserving Workflow Repositories

Susan B. Davidson, University of Pennsylvania

Need for Workflow repositories: Data intensive science requires computational analysis to enable discoveries. This computational analysis is frequently multi-step, combining a variety of different analysis and formatting steps within a workflow. Creating a workflow for a new type of analysis can be difficult, but may be facilitated by reusing components of existing workflow or by modifying portions of an existing workflow. For this, it is necessary to be able to access searchable/queryable repositories of workflows.

Some repositories of workflows are currently being created, many that are local to a particular workflow system and some that are globally accessible, e.g. myExperiment.org. Some of these repositories enable search by asking users to associate keywords with the submitted workflows. However, this is a primitive form of search, and can be improved by taking advantage of the hierarchical structure of workflows, i.e. the fact that some of the steps are themselves subworkflows. In this case, there are opportunities to return as the result of a search a smaller result than the entire workflow (see, e.g. [7]).

Need for Provenance information: Once a workflow has been designed it will be executed many times, generating a large amount of ``in-silico" experimental data which must also be managed by the system. To help users understand how this data was generated, tools for capturing provenance are being developed in systems such as myGrid/Taverna, Kepler and VisTrails. By maintaining *provenance* information about the sequence of processing steps (module executions) used to produce a data item, as well as the parameter settings and intermediate data items passed between module executions, the validity and reliability of data can be better understood and results can be made reproducible. In particular, users can ask queries over this provenance information such as: ``What were the input data and parameter settings for BLAST (a particular processing step) in this workflow execution?" (direct provenance information), ``What downstream data was affected by this particular data?" (direct and transitive provenance information), ``Does this data depend directly or indirectly on that data?" (a reachability query), or ``In how many executions of this workflow was the alternative BLAST chosen to implement the alignment search?" (an aggregate query over all executions of a particular workflow).

Need for Privacy in Provenance: However, with the availability of provenance information, privacy becomes an issue, as authors/owners of workflows may wish to keep some of this provenance information private. For example,intermediate *data* within an execution may contain sensitive information, such as the social security number, a medical record, or financial information about an individual. Although users with the appropriate level of access may be allowed to see such confidential data, making it available to all users through a workflow repository, even for scientific purposes, is an unacceptable breach of privacy. Beyond data privacy, a *module* itself may be proprietary, and hiding its description may not be enough: users without the appropriate level of access should not be able to infer its behavior if they are allowed to see the inputs and outputs of the module. Finally, details of how certain modules in the workflow are connected may be proprietary, and therefore showing how data is passed between modules may reveal too much of the *structure* of the workflow. There is thus an inherent tradeoff between the utility of the information shown inn response to a search/query and the privacy guarantees that authors/owners desire.

References and some relevant papers

[1] U. Braun, A. Shinnar, and M. Seltzer. Securing provenance. In *USENIX HotSec, The 3rd USENIX Workshop on Hot Topics in Security*, USENIX HotSec, pages 1–5, Berkeley, CA, USA, July 2008. USENIX Association.

[2] A. Chebotko, S. Chang, S. Lu, F. Fotouhi, and P. Yang. Scientific workflow provenance querying with security views. *WAIM*, pages 349–356, July 2008.

[3] S. B. Davidson, S. Khanna, S. Roy, J. Stoyanovich, V. Tannen, and Y. Chen. On provenance and privacy. In ICDT, pages 3{10, 2011.

[4] S. B. Davidson, S. Khanna, V. Tannen, S. Roy, Y. Chen, T. Milo, and J. Stoyanovich. Enabling privacy in provenance-aware workflow systems. In CIDR2011, pages 215-218.

[5] Y. Gil, W. K. Cheung, V. Ratnakar, and K. kin Chan. Privacy enforcement in data analysis workflows. In *PEAS*, 2007.

[6] Y. Gil and C. Fritz. Reasoning about the appropriate use of private data through computational workflows. In *Intelligent Information Privacy Management, Papers fromthe AAAI Spring Symposium*, pages 69–74, March 2010.

[7] Z. Liu, Q. Shao, and Y. Chen. Searching workflows with hierarchical views. *PVLDB*, 3(1), 2010.