On Provenance and Reuse of scientific workflows to analyze big biological data

Sarah Cohen-Boulakia
Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623
Université Paris-Saclay, Bâtiment Ada Lovelace, 91405 Orsay Cedex
cohen@lri.fr

1 Introduction

Typical analysis processes in the Life Sciences are complex, multi-staged, and large. One of the most important challenges is to properly represent, manage, and execute such in-silico experiments. As a response to these needs, scientific workflow management systems have been introduced. They provide an environment to guide a scientific analysis process from design to execution. This area is largely driven by the bioinformatics community and also attracts attention in fields like geophysics or climate research. In a scientific workflow, the analysis processes are represented at a high level of abstraction which enhances flexibility, reuse, and modularity while allowing for optimization, parallelization, provenance tracking, debugging etc. Differences with business and ETL workflows have been studied extensively [7]: scientific workflows have building blocks which are complex user-defined functions rather than relational operators and they are focused on data transformations.

These developments, accompanied by the growing availability of analytical tools wrapped as (web) services, were driven by a series of expectancies [4]: End users of scientific workflow systems, without any programming skills, are empowered to develop their own pipelines; reuse of services is enhanced by easier integration into custom workflows; time necessary for developing analysis pipelines decrease etc. However, despite all efforts, scientific workflows have not yet found widespread acceptance in their intended audience.

In the meantime, it becomes possible to share, search, and compare scientific workflows, opening the door to the exchange of mature and specialized data integration solutions. For example, myExperiment[1] is a portal that hosts more than two thousands scientific workflows while BioCatalogue[2] is a repository of more than one thousand web services to be combined in workflows.

We argue that a wider adoption of scientific workflow systems would be highly beneficial for users but can only be achieved if at least the following two points are considered.

First, provenance in scientific workflows [5] is a key concept and should be considered as a first citizen in scientific workflow systems. The importance of replication and reproducibility has been critically exemplified through studies showing that scientific papers commonly leave out experimental details essential for reproduction, studies showing difficulties with replicating published experimental results, an increase in retracted papers, and through a high number of failing clinical trials. Provenance supports reproducibility and allows assessing the quality of results. Research questions for workflow provenance include comparing workflow runs based on their provenance data and querying provenance information which can be, in turn, used to asses the similarity of workflows.

Second, since the targeted users are mainly non programmers, they may not want to design workflows from scratch. The focus of research should thus be placed on searching, adapting, and reusing existing workflows. Only by this shift can scientific workflow systems outreach to the mass of domain scientists actually performing scientific analysis - and with little interest in developing them themselves. To this end, scientific workflow systems need to be combined with community-wide workflow repositories allowing users to find solutions for their scientific needs (coded as workflows). As a need, to be reused by others, workflows should remain simple to use: a complex workflow composed of dozens of intertwined tasks, in general, is not much easier to understand than a well structured program performing the same analysis.

Our talk will outline the contributions we made to these research questions and draw research opportunities for the database community.

2 Workflow Provenance

In contrast with database provenance [6], transformations occurring in workflows are usually external processes (black boxes), and the log files typically provide mainly object ids. Provenance is more coarse-grained, and the structure of data cannot be reasoned about. Based on such salient differences, provenance in scientific workflows has been gaining interest since the early 2000s followed by the development of a series of International *Provenance Challenges* [8].

In this talk, we will focus on two of our contributions in this domain, performed in collaboration with the University of Pennsylvania. First, we tackle the problem of reducing the large amount of provenance information produced by workflow runs (as an example, French plant phenotyping plateforms may each produce about 5To of raw data per week and 250 To

[1] myexperiment.org
[2] biocatalogue.org

per year, to be possibly considered as input of analysis workflows). ZOOM*userview [2] provides abstraction mechanisms to focus on the most relevant information. Since bioinformatics tasks may themselves be complex sub-workflows, a user view determines what level of sub-workflow the user can see, and thus what data and tasks are visible in provenance queries. More specifically, we formalize the notion of user views, demonstrate how they can be used in provenance queries, provide and implement an algorithm for generating a user view based on the tasks considered as relevant for the user, concretely used to participate to the first provenance challenge.

Second, one major provenance query in scientific workflows (listed in the first provenance challenge) is related to the comparison between two executions of the same workflow. Here, we consider the problem of differing the provenance of two data products produced by executions of the same specification in the PDiffView [1]. Although this problem is NP-hard for general workflow specifications, an analysis of real scientific workflows shows that in majority their specifications can be captured as series-parallel graphs overlaid with well-nested forking and looping. For this restriction, we introduce and implement efficient, polynomial-time algorithms for differencing executions of the same specification and thereby understanding the difference in the provenance of their data products.

3 Scientific workflows Reuse

Our work in workflow reuse performed in collaboration with the University of Berlin (Humboldt) and Manchester is based on a workflow reuse study performed on the users of the myExperiment repository [11].

We provide contributions allowing users to query workflow repositories and find workflows similar to their input workflow. We performed a deep and large comparative review of workflow similarity approaches [9] to compare in isolation different approaches taken at each step of scientific workflow comparison, reporting on an number of unexpected findings. We investigate how these can best be combined into aggregated measures and we make available a gold standard of over 2,000 similarity ratings contributed by 15 workflow experts on a corpus of 1,500 workflows and re-implementations of all methods we evaluated. We then introduce a novel and intuitive workflow similarity measure that is based on layer decomposition [10] which accounts for the directed dataflow underlying scientific workflows, a feature which has not been adequately considered in previous methods.

As another attempt to make scientific workflows easier to (re)use, we introduce techniques to reduce the workflow structural complexity. DistillFlow [3] aims to remove the structural redundancy in workflows designed with Taverna, one of the major scientific workflow systems. More precisely, we identify a set of *anti-patterns* that contribute to the structural workflow complexity and we design a series of refactoring transformations to replace each anti-pattern by a new semantically-equivalent pattern with less redundancy and simplified structure. We then introduce and implement a distilling algorithm that takes in a Taverna workflow and produces a distilled semantically-equivalent Taverna workflow testing our implementation on both the major public repository of Taverna workflows (myexperiment) and on a carefully designed private collection of workflows from the BioVel project.

4 Conclusion

In this talk, we present several collaborative projects to manage and query scientific workflows and their executions. While contributions are mainly related to databases and bioinformatics, they also consider other research domains including graphs, algorithmics and software engineering. More generally, scientific workflows have now reached a level of maturity making them able to deal with large-scale amounts of complex data in production, opening the door to several open research questions directly related to the big data paradigm: how to store, index, query and efficiently analyze the huge and highly distributed amounts of data concretely produced by in-silico experiments. Advances in managing scientific workflows depend – and may have impact on – progress made in other communities such as social networks or system biology.

References

[1] Z. Bao, S. Cohen-Boulakia, S. Davidson, A. Eyal, and S. Khanna. Differencing provenance in scientific workflows. In *Proc. of ICDE, IEEE*, pages 808–819, 2009.

[2] O. Biton, S. Cohen-Boulakia, S. Davidson, and C. Hara. Querying and managing provenance through user views in scientific workflows. In *Proc. of ICDE, IEEE*, pages 1072–1081, 2008.

[3] S. Cohen-Boulakia, J. Chen, C. Goble, P. Missier, A. Williams, and C. Froidevaux. Distilling structure in taverna scientific workflows: A refactoring approach. *BMC Bioinformatics*, 15(1):S12, 2014.

[4] S. Cohen-Boulakia and U. Leser. Search, adapt, and reuse: the future of scientific workflows. *SIGMOD Record*, 40(2):6–16, 2011.

[5] S. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Proc of SIGMOD*, pages 1345–1350. ACM, 2008.

[6] T. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *Proc. of PODS*, pages 31–40, 2007.

[7] B. Ludäscher, M. Weske, T. McPhillips, and S. Bowers. Scientific workflows: Business as usual? In *Proc. of BPM*, pages 31–47. Springer, 2009.

[8] L. Moreau and et al. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, 2008.

[9] J. Starlinger, B. Brancotte, S. Cohen-Boulakia, and U. Leser. Similarity search for scientific workflows. *Proc. of VLDB*, 7(12), 2014.

[10] J. Starlinger, S. Cohen-Boulakia, S. Khanna, S. Davidson, and U. Leser. Effective and Efficient Similarity Search in Scientific Workflow Repositories . *Future Generation Computer Systems*, page 79, Sept. 2015.

[11] J. Starlinger, S. Cohen-Boulakia, and U. Leser. (re)use in public scientific workflow repositories. *Proc. of SSDBM*, 2012.