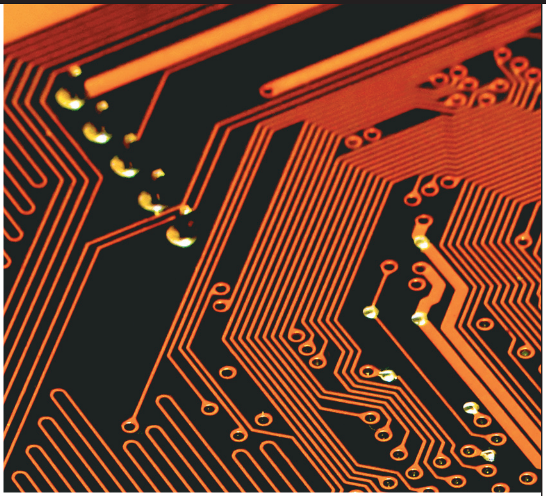
Network Analysis of Scientific Workflows: A Gateway to Reuse



Wei Tan, *University of Chicago and Argonne National Laboratory*

Jia Zhang, *Northern Illinois University*

Ian Foster, *University of Chicago and Argonne National Laboratory*

Online workflow repositories let scientists share successful experimental routines and compose new workflows from best practices and existing service components. The authors share the results of a social-network analysis of the myExperiment workflow repository to assess the state of scientific workflow reuse and propose the CASE framework to facilitate such reuse.

To accelerate data-intensive scientific exploration, many disciplines including biology and biomedicine have adopted workflows as data-pipeline orchestrators and Web services as computational components. A scientific workflow precisely describes a multistep procedure to streamline a composition of tasks and the dataflow among them.[1] Services-computing technology enables scientists to expose data and computational resources as Web services so that they become publicly available to other researchers. A scientific workflow thus may utilize published Web services as tasks to speed up workflow composition. The "Scientific Workflows" and "Web Services" sidebars provide more details.

Business workflows are generally organization specific and rarely shared across company boundaries. In contrast, the scientific world is more open, and researchers often publish workflows to share experimental routines with colleagues, who can either use those workflows unchanged or compose new ones from best practices and existing service components. Several domain-specific online workflow repositories have evolved in recent years, including the UK-based myExperiment project (www.myexperiment.org), which has collected more than 1,000 life-science workflows.[2] The advent of these online repositories makes it possible to assess the state of scientific workflow reuse.

Software engineers commonly reuse components to attain higher quality and productivity.[3] Many scientific workflow development tools such as Taverna[4] similarly allow scientists to design a workflow using available Web services, and dedicated repositories help scientists find these services. For example, BioCatalogue,[5] a sister project to myExperiment, has catalogued more than 1,600 life-science Web services. Such online repositories have opened a gateway to scientific workflow reuse.

To advance the state of the art in service-oriented science,[6] we analyzed the workflows stored at myExperiment. Applying social-network analysis techniques,[7] we aimed to answer two questions: What is the current usage pattern of services in scientific workflows, and how can this knowledge be extracted to facilitate reuse? Based on our study's results, we propose a new framework named CASE—Collection, Annotation, Search, and rEcommendation—to support scientific workflow reuse.

Published by the IEEE Computer Society

0018-9162/10/$26.00 © 2010 IEEE

NETWORK METRICS AT A GLANCE

We downloaded myExperiment workflows via its REST (Representational State Transfer) API[2] on 20 March 2010. We were interested only in the repository's 599 Taverna formatted workflows; the other workflows are less structured and some are completely freestyle.

We analyzed the structure of each Taverna workflow serialized in an XML-based language. We found that 280 of the workflows contained at least one Web service and that altogether there were 118 unique services. Because our goal was to identify the current usage pattern of services in workflows, we focused on these 280 workflows and 118 services.

We abstracted these workflows and services into a *workflow-service network*, an undirected graph in which nodes represent workflows or services and edges represent the inclusive relations between them—that is, a workflow is connected to a service if it calls the service. From this network, we derived two additional networks: a *workflow-workflow network* in which two workflows are connected if they comprise services in common, and a *service-service network* in which two services are connected if they appear in some workflow together. We used Pajek,[7] a widely used social-network analysis tool, to produce all three graphs.

Table 1 summarizes the myExperiment dataset used in our study, including some metrics of the original and derived networks.

WORKFLOW-SERVICE RELATION AND DERIVATIONS

We parsed the myExperiment workflows to create the workflow-service relation Q, formalized as an $m \times n$ matrix, where m is the number of workflows (280) and n is the number of services (118):

$$Q = [q_{ij}],\ 0 \le i \le m,\ 0 \le j \le n,$$

where $q_{ij} = 1$ if workflow i contains service j.

We derived two more relations, W and S, from Q as follows:

$$W = Q \cdot Q^{\mathrm{T}} = [w_{ij}],\ 0 \le i, j \le m,$$

where w_{ij} = number of services shared by workflows i and j, and w_{ii} = number of services in workflow i; and

$$S = Q^{\mathrm{T}} \cdot Q = [s_{ij}],\ 0 \le i, j \le n,$$

where s_{ij} = number of workflows where both

SCIENTIFIC WORKFLOWS

A scientific workflow precisely defines a multistep procedure to seamlessly integrate and streamline local and remote heterogeneous computational and data resources to perform in silico scientific exploration.[1] Scientific and business workflows overlap in some requirements and features, and some tools, such as Sedna,[2] adopt the industry-standard Business Process Execution Language (BPEL) for scientific workflows. However, fundamental differences exist between scientific and business workflows.[3] One is that the execution model of scientific workflows is dataflow oriented, while that of their business counterpart focuses on control-flow patterns and events.[4] This difference is somewhat analogous to that between procedural and functional programming.[5] Consequently, many research groups have explored various dataflow-based models and languages, resulting in several scientific workflow authoring and management tools including Taverna (www.taverna.org.uk), Kepler (https://kepler-project.org), and Triana (www.trianacode.org).

References

1. C. Goble and D. De Roure, "The Impact of Workflow Tools on Data-centric Research," T. Hey, S. Tansley, and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009, pp. 137-145.
2. B. Wassermann et al., "Sedna: a BPEL Based Environment for Visual Scientific Workflow Modelling," I.J. Taylor et al., eds., *Workflows for e-Science: Scientific Workflows for Grids*, Springer, 2007, pp. 428-449.
3. B. Ludäscher et al., "Scientific Workflow Management and the Kepler System," *Concurrency and Computation: Practice & Experience*, Aug. 2006, pp. 1039-1065.
4. U. Yildiz, A. Guabtni, and A.H.H. Ngu, "Business versus Scientific Workflows: A Comparative Study," *Proc. 2009 Congress on Services* (SERVICES 09), IEEE CS Press, 2009, pp. 340-343.
5. V. Curcin and M. Ghanem, "Scientific Workflow Systems—Can One Size Fit All?" *Proc. 4th Cairo Int'l Biomedical Eng. Conf.* (CIBEC 08), IEEE Press, 2008, pp. 1-9.

WEB SERVICES

A Web service is a programmable Web application component that has a standard interface and is universally accessible through standard network protocols.[1] In the current service-oriented science paradigm, technologies, components, and experimental routines are increasingly wrapped in various services. Scientists can leverage such published services to quickly compose new scientific workflows.[2] A Web service is typically accessed via the Simple Object Access Protocol (www.w3.org/TR/soap) or Representational State Transfer.[3] SOAP is more heavyweight but able to perform rigid type checking, while REST is more lightweight without extra XML markup. MyExperiment provides a REST API for users to fetch stored workflows.

References

1. L.-J. Zhang, J. Zhang, and H. Cai, *Services Computing*, Springer, 2007.
2. C. Goble and D. De Roure, "Curating Scientific Web Services and Workflows," *EDUCAUSE Rev.*, Sep./Oct. 2008, pp. 10-11.
3. R.T. Fielding, "Architectural Styles and the Design of Network-Based Software Architectures," doctoral dissertation, Donald Bren School of Information and Computer Science, Univ. of California, Irvine, 2000.

Table 1. Overview of myExperiment dataset.

Data	Value
Taverna workflows with at least one Web service	280
Unique services	118
Operations	179
Average services per workflow	1.36
Average workflows per service	3.22
Average collaborators per service	1.44
Largest component of service-service network	31 percent

services *i* and *j* are invoked and s_{ii} = number of workflows where service *i* is invoked.

Relation *Q*

Q represents the 280 workflows that contain services. In the visualization of *Q* shown in Figure 1a, yellow diamonds represent workflows, green circles represent services, and an edge between a diamond and a circle indicates that the workflow calls the service. We performed a statistical analysis of *Q*, with the results summarized in Figure 1b. Most workflows contained few services (76 percent of workflows invoked only one service); no workflow contained more than four. On average, each workflow that we considered consumed 1.36 services. Meanwhile, most services participated in only a few workflows (50 percent of services participated in a single workflow). Thirty-one services were called by two workflows and only four utility services by more than 20 workflows.

Degree centrality. From the dataset, we sought to identify the highly used services and workflows that invoked more services. We therefore configured Pajek such that node size represents its degree centrality or popularity—that is, the larger a node is, the more nodes

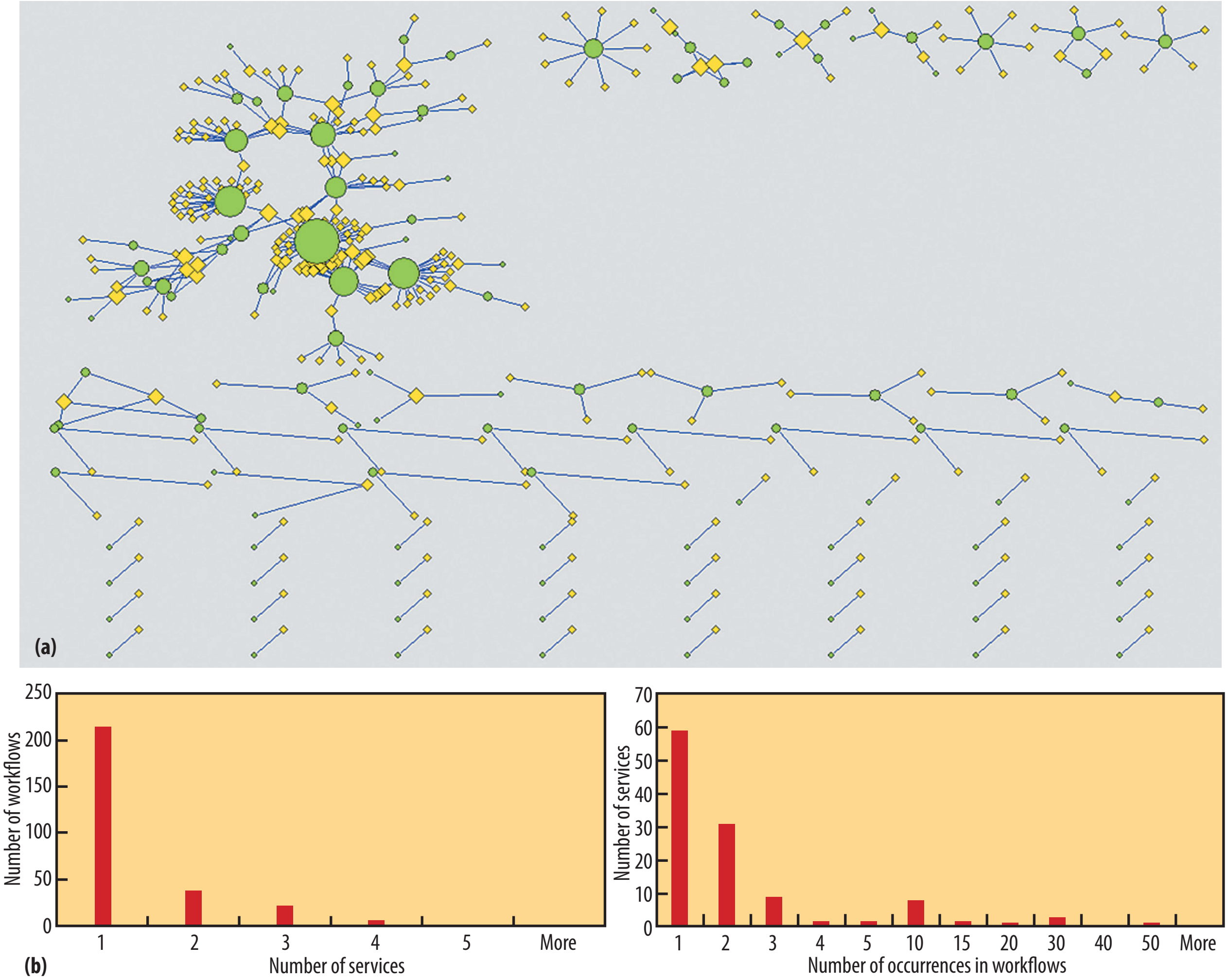


Figure 1. Workflow-service relation *Q*. (a) Visualization of *Q* with degrees. (b) Histogram of the number of services per workflow and the number of occurrences in workflows per service.

Table 2. Top six myExperiment services in degree centrality.

Service URL	Number of workflows appeared	myExperiment ranking	BioCatalogue ranking
http://soap.genome.jp/KEGG.wsdl	50	1	4
http://xml.nig.ac.jp/wsdl/Blast.wsdl	26	2	1
http://xml.nig.ac.jp/wsdl/Ensembl.wsdl	24	3	11
http://phoebus.cs.man.ac.uk:8081/axis/EnsemblListner.jws?wsdl	21	4	N/A
http://www.ebi.ac.uk/Tools/webservices/wsdl/WSDbfetch.wsdl	16	5	2
http://www.ebi.ac.uk/Tools/webservices/wsdl/WSInterProScan.wsdl	14	6	12

it connects to. The larger green circles in Figure 1a imply that more workflows use the services; the larger yellow diamonds imply that the workflows use more services as components.

We found that the highly reused services are a small set of utility services widely employed by bioinformaticians. Table 2 lists the top six services ranked by their degree centrality in descending order. For example, the top-ranked service is the Kyoto Encyclopedia of Genes and Genomes (KEGG), which appears in 50 workflows. To evaluate our findings, we also examined (on 2 June 2010) the six services' BioCatalogue (www.biocatalogue.org) popularity rankings. The BioCatalogue dataset consists of 1,630 registered biology services, each with metadata, including popularity as measured by the number of times viewed. As the table shows, five of the services also have high BioCatalogue rankings (the sixth is currently inactive and thus not listed). This analysis confirmed that services frequently reused in myExperiment workflows also attract more interest in BioCatalogue.

Betweenness centrality. In addition to popularity, we examined how information flows through different services and workflows, aiming to identify the hinge services or workflows in myExperiment. In social-network analysis, *betweenness*[8] is a node's centrality measure: it evaluates the connectivity of a node in its context, which is the number of shortest paths in the network that pass through a given node. Nodes that occur on many shortest paths between other nodes have higher betweenness than those that do not.

Figure 2 shows the largest strongly connected components in *Q* with betweenness centrality. For example, workflow w148 connects service s287 (Blast) with service s293 (KEGG); workflow w43 connects service s286 (EBI InterProScan) with s287. Comparing Figures 1 and 2, we find that w148 and w43 both have high betweenness values but low degree values. This indicates that although they aren't directly connected to (invoke) many services, they are on many geodesics between other pairs—that is, they're hinge nodes in terms of information flow in the network.

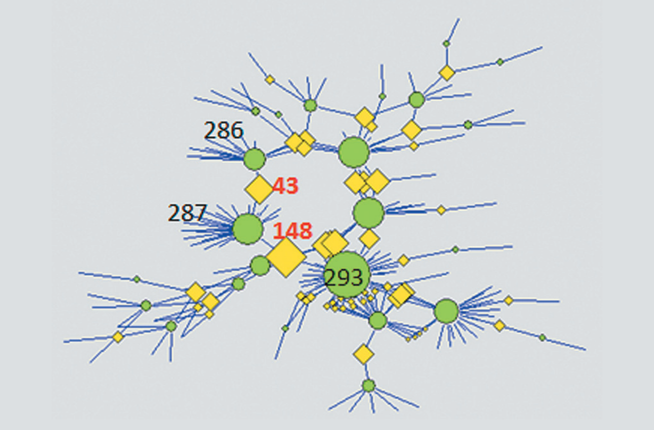


Figure 2. Visualization of the largest strongly connected components in *Q* with betweenness centrality.

Relation *W*

In social-network analysis, a *clique*[9] is a maximal complete subgraph of three or more nodes, all of which are directly connected to one another. It usually represents an interest group whose members tend to have more homogeneous opinions and share more common traits. In workflow-workflow relation *W*, two workflows (nodes) are connected if they both invoke common services. Therefore, a clique in *W* refers to a group of workflows that invoke common services. In other words, the group of workflows comprising a clique may share some common goals or requirements.

Figure 3 is a visualization of *W*, wherein each node represents a workflow, the node's size connotes the number of services used, and the thickness of an edge indicates the number of services shared by the two workflows at both ends. The dense areas are cliques of workflows sharing common utility services. Overlapped cliques may also imply some common interests or goals.

Relation *S*

A workflow may be viewed as a recipe documenting how services collaborate to fulfill a scientific experiment's requirement. Therefore, service-service relation *S* can be

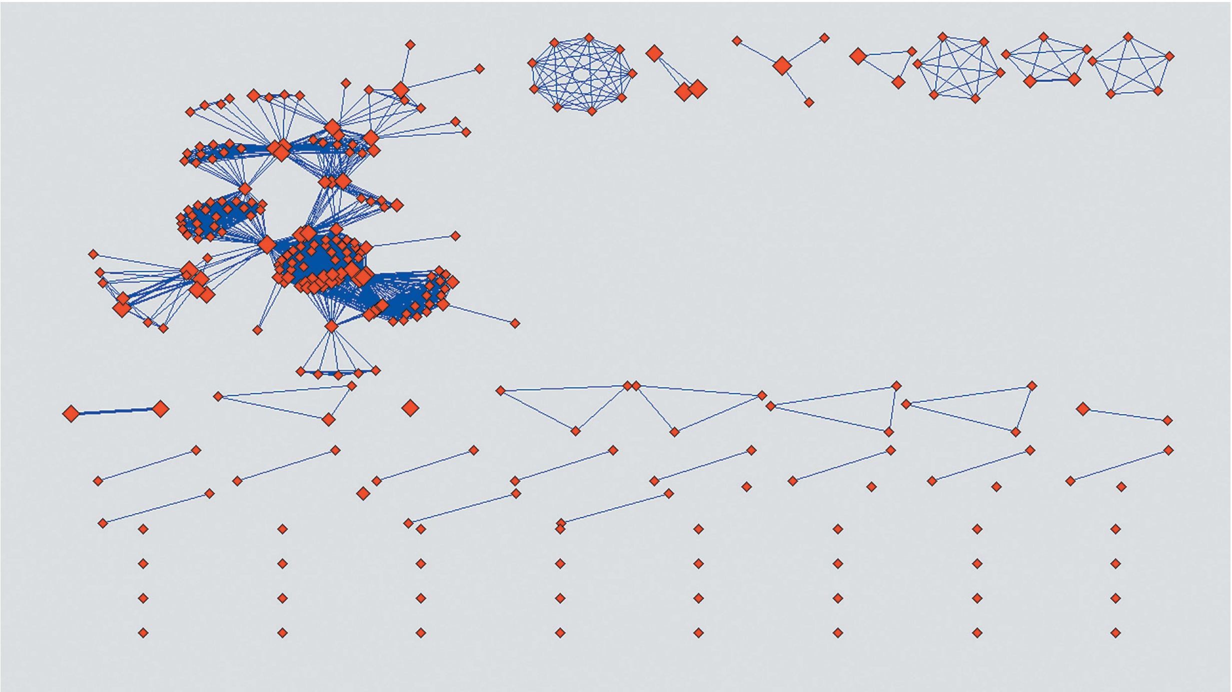


Figure 3. Visualization of workflow-workflow relation *W*. The dense areas indicate cliques.

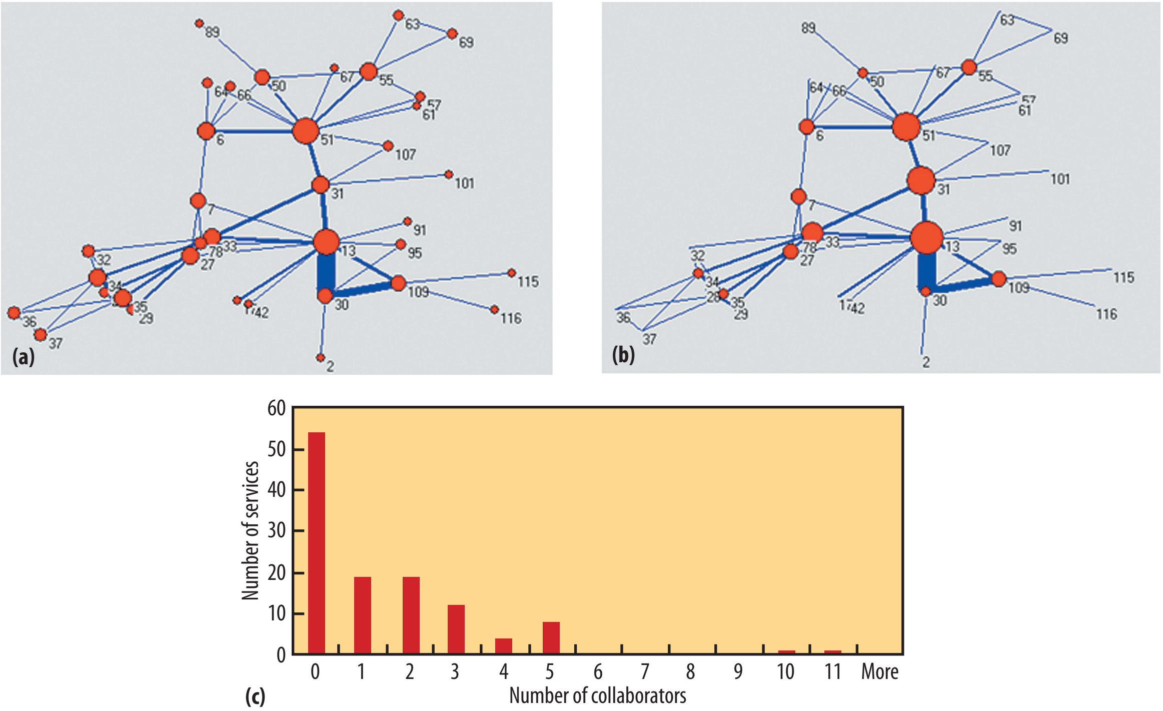


Figure 4. Service-service relation *S*. (a) Visualization of a portion of *S* with degree centrality. (b) Visualization of a portion of *S* with betweenness centrality. (c) Histogram of the number of collaborators for services.

seen as a collaboration network among services—that is, services appearing in the same workflow collaborate with one another.

Figures 4a and 4b illustrate the degree and betweenness of *S*, respectively. The size of a node is proportional to its degree and its betweenness, respectively. An edge's thick-

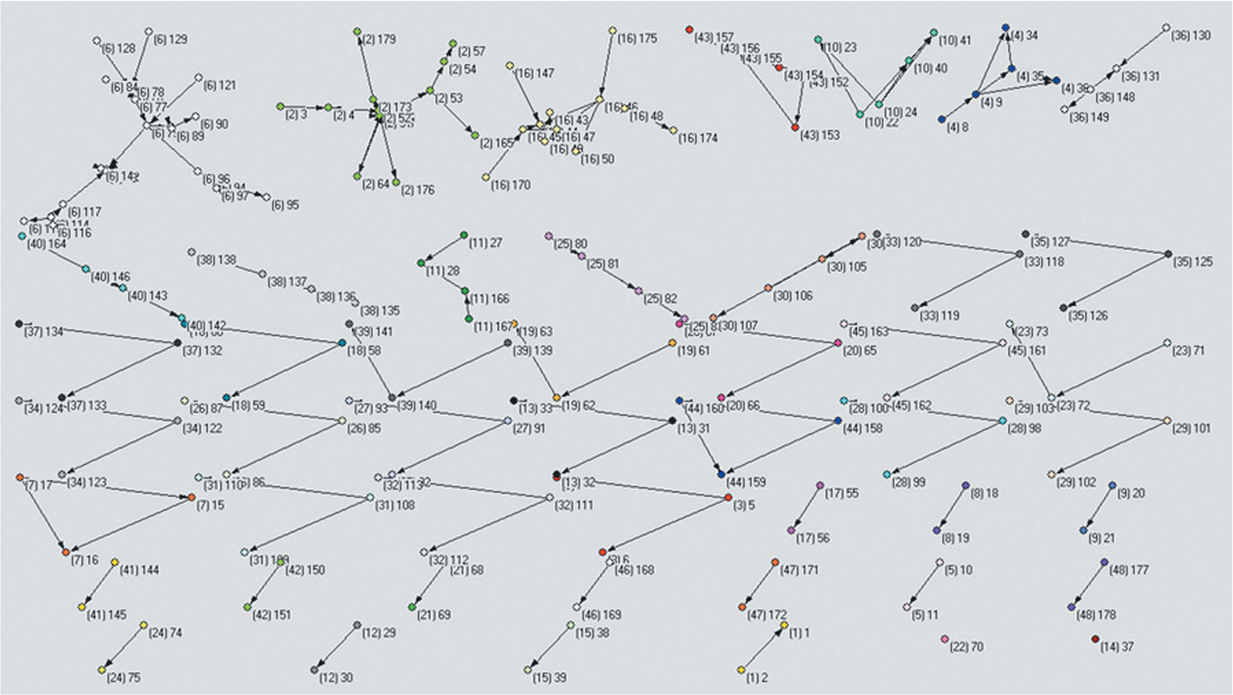


Figure 5. Visualization of operations invocation network *S'*. Operations are grouped into weakly connected components called clusters, with each assigned a unique number.

ness is proportional to the number of workflows that share the two services on its ends. The figures highlight the most connected portion of S and neglect some isolated services.

A service with high degree centrality collaborates with more peer services in all workflows; a service with high betweenness centrality means that service collaborations are more likely to go through it—that is, it has more control over the information flow between services. In Figure 4, for example, service s31 has relatively higher betweenness compared to its degree, while service s6 has a higher degree compared to its betweenness. (A given node's absolute values of degree and betweenness centrality aren't comparable, only its relative values, in a network. For example, a node may have higher degree centrality and lower betweenness centrality than other nodes in a network, and vice versa.)

To highlight the collaborative relationship among services, Figure 4c shows the distribution of the number of collaborators for services. On average, a service has only 1.44 collaborators; 54 services (46 percent) have no collaborator at all. Most services have connections with only a couple of others.

In social-network analysis, a *component* is a set of network nodes connected by some relationship such that they are strongly connected. In this case, the largest component in relation *S* covers only 31 percent of it. This indicates that services in myExperiment largely function individually rather than work together to form a linked research community.

Operations invocation network

To study finer-grained collaboration among services in our dataset, we zoomed in on *S* to the operation level. We derived a directed relation *S*', shown in Figure 5, by examining the invocation relations among service operations. Nodes are operations in services, and a directed edge represents a data link between two operations in some workflow.

Operations in *S*' are grouped into weakly connected components called *clusters*, with each cluster assigned a unique parenthesized number. For example, at the top right corner of Figure 5, cluster 4 comprises five operations (8, 9, 34, 35, and 36). Altogether, there are 179 operations grouped in 48 clusters.

Based on *S* and *S*', we can define two levels of collaboration between services. The collaboration relationship among services invoked in the same workflow in *S* is *weak*. Compared to *S*, *S*' contains operation-level information of both intra- and interworkflow invocation sequences. Thus, a *strong* collaboration relationship between two services implies a direct operation invocation between them in some workflow. For example, operation o35 calls o34 in one workflow and o36 in another.

ANALYSIS OF FINDINGS

Network analysis of scientific workflows in the myExperiment repository produced answers to both of the questions guiding our study.

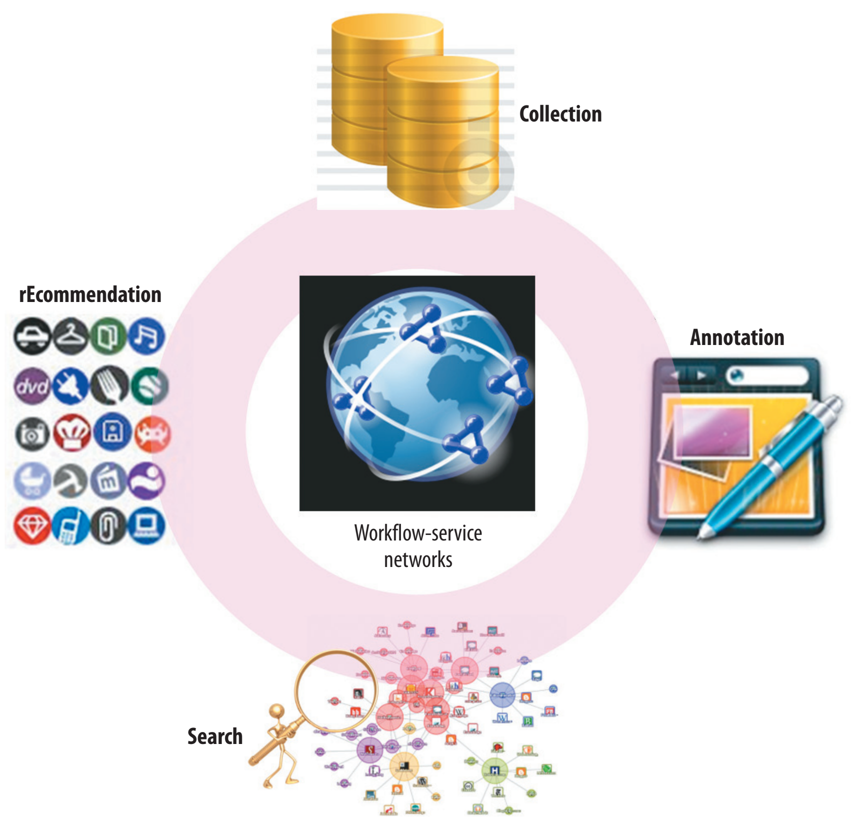


Figure 6. The CASE framework for workflow reuse is centered on workflow-service networks and their associated knowledge.

What is the current usage pattern of services in scientific workflows? Three findings are significant:

- The use of life science services is low in myExperiment workflows, and only a couple of utility services are frequently used.
- Frequently used services in myExperiment workflows are also popular in BioCatalogue.
- Services used in myExperiment workflows largely function individually without collaborating with each other.

In summary, current service reuse in scientific workflows is unsatisfactory.

How can this knowledge be extracted to facilitate workflow reuse? Our work demonstrates the effectiveness of constructing a workflow-service network and its derived networks. The usage pattern embedded in these networks provides quantitative answers to the following four relationships:

- workflow-service: how workflows use services;
- service-workflow: how services are used in different workflows;
- workflow-workflow: how multiple workflows use common services; and
- service-service: how services collaborate with one another.

Such knowledge embeds the best practice of using services in workflows, and therefore is well suited to feed into a recommendation system to facilitate services-oriented workflow reuse.

CASE FRAMEWORK FOR WORKFLOW REUSE

Advances in social-network analysis and recommendation systems, which accumulate the wisdom of crowds, can help scientists discover relevant workflows and services and adapt them to their own explorations, much as biomedical researchers use publication repositories such as PubMed (www.ncbi.nlm.nih.gov/pubmed) to discover relevant findings. Toward this end, we propose the CASE framework for services-oriented scientific workflow reuse. CASE is an ongoing effort and, as Figure 6 shows, is centered on workflow-service networks and their associated knowledge.

Collection

Workflows and services are incrementally collected from centralized repositories such as myExperiment and BioCatalogue as primary data sources. Additional information may be collected from Web servers hosting individual services, publication libraries like PubMed, websites of participating research institutions, and so on. Workflow-service networks are built and stored in CASE and serve as the information collection index that binds the four CASE components.

Annotation

Annotation (www.w3.org/2001/Annotea) is widely used to facilitate knowledge sharing. For example, Taverna lets authors annotate workflows and BioCatalogue lets users annotate services. Such volunteer-based human actions, however, may lead to fragmentary and inconsistent annotations scattered in disjointed resources. CASE integrates annotations generated from various heterogeneous data sources such as author annotations at different levels (for example, workflow, service, or data channels), user comments at runtime, best practices, and statistical data of existing scientific workflows and services, including popularity and usage patterns. To ensure performance, such annotations are stored independently of corresponding workflows and services. Automatic annotation elicitation, generation, and analysis instruments support services-oriented scientific workflow discovery, composition, and adaptation.

Search

CASE uses Apache Lucene (http://lucene.apache.org), an open source search engine, to index the information collection and associated annotations. Users can carry out full-text search to find artifacts of interest. In addition, the workflow-service networks support relation-aware search. For example, relation *W* can be used to locate workflows providing similar functions, and relation *S*' to predict a given operation's most likely next step. An interactive GUI lets users visualize the search interface and results to navigate through the artifacts, zoom into details, or zoom out to global connections.

Recommendation

CASE's ultimate goal is to provide recommendation support in workflow composition. When a scientist is building a workflow in some integrated development environment, the CASE recommendation plug-in for this IDE observes the context, such as the user's profile and the incomplete workflow. The plug-in then communicates with the recommendation component in CASE and offers relevant suggestions. Examples include a collection of related services (referring to relation Q), a sequence of operations in a newly added service (referring to relation *S*'), a workflow snippet to produce a data object given the existing data objects in the incomplete workflow, and so on. Recommendation can be either passive (requested explicitly by users) or proactive (automatically delivered when CASE perceives such a need).

Scientific workflow repositories open a door to workflow reuse. Our study applied social-network analysis to mine and analyze the myExperiment workflow repository, focusing on service usage patterns. The results indicate that services are currently reused in an ad hoc style instead of a federated manner. This observation suggests a need for techniques that help domain scientists dynamically locate related services and workflows and reuse successful processes to attain their research purposes. Our proposed CASE framework addresses this challenge. In the future, we plan to enrich its recommendation services and study its impact on helping workflow reuse and composition in real scientific exploration.

Acknowledgments

We thank the caBIG (Cancer Biomedical Informatics Grid) community, sponsored by the National Cancer Institute, for various application scenarios motivating our network analysis and CASE framework. We thank Carole Goble and David De Roure for supporting our assessment of Taverna, myExperiment, and BioCatalogue. We also thank Ravi Madduri for constructive discussions on workflow reuse in caBIG. The work described in this article is partially supported by the National Cancer Institute, the National Institutes of Health under contract N01-CO-12400, and the National Science Foundation under grant IIS-0959215.

References

1. I.J. Taylor et al., eds., *Workflows for e-Science: Scientific Workflows for Grids*, Springer, 2007.
2. C.A. Goble et al., "myExperiment: A Repository and Social Network for the Sharing of Bioinformatics Workflows," *Nucleic Acids Research*, 25 May 2010, pp. W677-W682.
3. T. Xie et al., "Data Mining for Software Engineering," *Computer*, Aug. 2009, pp. 55-62.
4. T. Oinn et al., "Taverna: Lessons in Creating a Workflow Environment for the Life Sciences," *Concurrency and Computation: Practice & Experience*, Aug. 2006, pp. 1067-1100.
5. J. Bhagat et al., "BioCatalogue: A Universal Catalogue of Web Services for the Life Sciences," *Nucleic Acids Research*, 19 May 2010, pp. W689-W694.
6. I. Foster, "Service-Oriented Science," *Science*, 6 May 2005, pp. 814-817.
7. W. de Nooy, A. Mrvar, and V. Batagelj, *Exploratory Social Network Analysis with Pajek*, Cambridge Univ. Press, 2005.
8. Y.-L. Luo and C.-H. Hsu, "An Empirical Study of Research Collaboration Using Social Network Analysis," *Proc. 2009 Int'l Conf. Computational Science and Eng.* (CSE 09), vol. 4, IEEE CS Press, 2009, pp. 921-926.
9. G. Groh and V. Rappel, "Towards Demarcation and Modeling of Small Sub-Communities/Groups in P2P Social Networks," *Proc. 2009 Int'l Conf. Computational Science and Eng.* (CSE 09), vol. 4, IEEE CS Press, 2009, pp. 304-311.

***Wei Tan** is a research professional associate at the Computation Institute, a joint institute of the University of Chicago and Argonne National Laboratory. Contact him at wtan@mcs.anl.gov.*

***Jia Zhang** is an associate professor in the Computer Science Department at Northern Illinois University. Contact her at jiazhang@cs.niu.edu.*

***Ian Foster** is director of the Computation Institute. He is also a senior scientist and Distinguished Fellow at Argonne National Laboratory as well as Chan Soon-Shiong Scholar and Arthur Holly Compton Distinguished Service Professor in the Department of Computer Science at the University of Chicago. Contact him at foster@mcs.anl.gov.*