

Research Statement

Web Version

Christine F. Reilly

<http://pages.cs.wisc.edu/~chrisr>

Data management systems have transformed our society's relationship with information. The basic function of these systems is to allow us to quickly store and retrieve data. On a day-to-day basis, a person encounters data management systems when they do their banking, modify their personal information at work or school, and search the internet. Data management systems have also played a critical role in managing vast amounts of data for scientific research.

Computer scientists have done an excellent job at developing systems that manage data. Database management systems are able to efficiently store and quickly retrieve structured data, such as student records and bank data. Other systems manage unstructured data, such as the text contained in web pages. Web search engines demonstrate how quickly information can be retrieved from unstructured data. One limitation of most data management systems is that they have no record of the origin and history, also known as provenance, of the data. Uses of provenance include: finding the source of errors in the output of a program, determining if the input files or program have been updated since the program was last run, and identifying data that was created using a machine that is later found to be producing erroneous output.

My dissertation research focuses on how to collect provenance. Many provenance collection systems are either designed to be used within a database management system, for a single application, or with programs that are designed in a specific way. While these systems are able to collect very useful provenance, their use is restricted to specialized conditions. The main contribution of my research is demonstrating that it is possible to design a system that transparently collects provenance from applications that store data in files. A transparent provenance system can be used by programs from a wide range of disciplines, and does not require that the user change her programs in order to obtain provenance.

Dissertation Work

The overall goal of my dissertation work is to determine if it is possible to construct a system that transparently collects a useful amount of provenance. With an ideal provenance system, a user could retrieve the provenance of a single piece of data. One issue I face is that the desired granularity of data differs between various programs. For some programs it is adequate to know that a specified output file was created using a certain input file. Other applications require finer grained provenance. For example, if the input and output files contain lines of data, the desired provenance of one line in the output file is the line or lines in the input file that contributed to the line of output data. My dissertation first examines file-level coarse grained provenance, then turns to finer grained provenance.

Coarse Grained Provenance: I first addressed coarse grained provenance in the context of Condor, a distributed job execution system. Condor is used on over 100,000 CPUs in universities, government labs, and private companies worldwide. A wide variety of applications are run in Condor, representing a diverse set of fields that includes computer engineering, biology, chemistry, physics, finance, and insurance. My work began as part of a larger project, Quill, that aimed to collect the operational data that is exposed while Condor runs [1]. This operational data includes information about when and on what machine a job ran, what files were used by a job, the status of machines in the cluster, and the activities of users in the system. I recognized that portions of this operational data were useful for provenance [4]. In order to improve the provenance gathered from jobs that run in a distributed system, I gave Quill the ability to gather additional information about files, and called the resulting system Provenance Aware Condor (PAC) [3]. This system provides useful provenance in the case where the desired data granularity is a file. I chose to operate at file-level granularity because Condor handles data at the level of files.

As an example of how PAC is used, consider an output file that was created by a simulation that reads two input files. By querying PAC, a user can find the location and version of the program and input files that created the output file. If the result is unexpected or seems odd, the user can examine the input files and program for suspected errors. The data from PAC can also be used to confirm that the output file was created using the current versions of the input files and program. One weakness of this approach is that the provenance includes both of the input files even if no data was used from one of the files.

The main contribution of my work on coarse grained provenance is demonstrating that it is possible to alter an existing computing system to collect provenance while not requiring users to change how they interact with the system. As part of a case study, I demonstrated that coarse grained provenance is not adequate for some applications. This finding leads to the next part of my work, examining whether it is possible to transparently gather fine grained provenance.

Fine Grained Provenance: My current work looks for a way to obtain finer granularity provenance, using the same case study as I used for coarse grained provenance. This involves rewriting the case study program in a declarative programming language called Xlog, a variation of datalog. By adding provenance gathering capabilities to Xlog, I expect to be able to gather fine grained provenance that allows a user to determine what items in the input files contributed to a specific item in the output file. While I am currently focusing the case study program, Xlog could be used to gather provenance from many applications.

Although requiring that the application be written in Xlog contradicts part of my transparency requirement, this approach can still be used by a wide range of programs. In many cases, I expect the benefit of obtaining fine grained provenance to outweigh the drawback of writing the application in a specialized language.

Future Work

My plans for future work entail continuing my current research, expanding my research into the broader area of scientific data management, and exploring methods for increasing the diversity of students who study computer science. As part of conducting this research, I am interested in pursuing opportunities for interdisciplinary collaboration and including undergraduate students on my research team.

Continuing current work: The next step in my research is to combine the systems I used for coarse grained and fine grained provenance collection. This combination will result in a system that has the fine granularity provenance provided by Xlog, and obtains operational data about each run of the program from PAC. Another near-term goal is to add a provenance-enabled user interface to the web page that is produced by the case study program. This interface will allow a user to click on a piece of information on the web page and obtain the provenance of that information. Additionally, I plan to examine the overhead of storing provenance and to explore methods for managing this constantly growing data set.

Scientific data management: My overall research interests are learning how scientists use data and discovering ways to improve how scientists manage data. I first became interested in this topic during my graduate studies in environmental engineering. My environmental engineering master's thesis included a simulation that consumed and produced a lot of data [2]. I now know that better data management practices would have allowed me to write more efficient programs and to reduce the need to repeat runs of the program. There are two major aspects to my long term research on data management: improving how scientists manage experiments and data, and gaining an understanding of how scientists relate to their data.

I plan to use my knowledge of computer science to improve the methods that scientists use to manage experiments and data. Part of this will involve educating scientists about certain aspects of computer science. For example, the combined knowledge of data structures and database management systems allows a scientist to write programs that efficiently manipulate data and quickly store and query the data. I would like to continue to explore data management strategies that can be applied to many disciplines. Additionally,

I am interested in devising creative solutions for data management problems that are specific to one discipline, or possibly even specific to one experiment.

The other focus of my long term research is gaining an understanding of how scientists relate to their data. Through my Ph.D. minor in Science and Technology Studies I became interested in how the work of science is done. By understanding how scientists do their work I expect to be able to better help them manage their data and other computing activities.

Increasing Diversity: The third aspect of my future work is examining strategies for increasing the diversity of students who study computer science. One idea that I plan to examine is whether introducing students to computer science in the early part of their science and engineering studies will encourage more students from underrepresented groups to major in computer science. This idea is rooted in my own experience. My first exposure to computer science occurred during my junior year of college. After taking a programming course that was required for my environmental engineering major I discovered that I was very interested in computer science and eventually switched my field of study. I think I would have majored in computer science if I had been exposed to it in high school or during the first year of college.

Plans for Interdisciplinary Collaboration: In order to implement these long term research goals, I plan to collaborate with scientists in other fields. Developing methods for managing scientific data will require that I collaborate with scientists who have data-intensive experiments. I am most familiar with simulations of environmental, geological, and ecological processes, and I expect that my broad education in science and engineering has prepared me to collaborate on a wide range of topics including chemistry and genetics. Given that examining how scientists relate to their data and developing methods for increasing the diversity of computer science students both require an understanding of what people think and do, these aspects of my research would benefit from collaboration with social scientists.

Undergraduate Research: Working on research projects is a very useful experience for undergraduate students. As an undergraduate, I found that research projects provided an opportunity for creative thought, and were valuable experiences when I was making the decision to attend graduate school. Undergraduates bring fresh and unique ideas to a research project and I look forward to including them on my research team.

Because the amount of data our society needs to manage is continuously growing, I expect my interest in developing strategies for managing large amounts of data to remain relevant for a long time. I look forward to continuing my research, and I am excited for the opportunity to share my passion for computer science research with students.

References

- [1] Huang, Jiansheng, Ameet Kini, Erik Paulson, Christine Reilly, Eric Robinson, Srinath Shankar, Lakshmikanth Shrinivas, David DeWitt, Jeffrey Naughton. An overview of Quill: A passive operational data logging system for Condor. <https://www.cs.wisc.edu/condordb>. 2007.
- [2] Reilly, C.F., and C.N. Kroll. Estimation of low streamflow statistics using baseflow correlation. *Water Resources Research*, 39(9), September 2003.
- [3] Reilly, C.F. and J.F. Naughton. Transparently gathering provenance with Provenance Aware Condor. First Workshop on the Theory and Practice of Provenance, San Francisco, California, February 2009.
- [4] Reilly, Christine F. and Jeffrey F. Naughton. Exploring provenance in a distributed execution system. In *Proceedings of the International Provenance and Annotation Workshop*, Chicago, Illinois, May 2006.