

Handling provenance, including mathematical proofs, in
cyberinfrastructure-oriented data processing

Ann Q. Gates, Olga Kosheleva, Vladik Kreinovich, Sa-aat
Niwitpong, Paulo Pinheiro da Silva, Leonardo Salyandia

Traditionally, computations used to be several orders of
magnitude faster than communications. As a result, to avoid the
drastic slowdown caused by communications, researchers tended
to bring all the data into a central location and process this
data there. For example, NASA maintained a central depository
of satellite images, US Geological Survey tried to collect and
store all geophysics-related data, etc.

Centralized data storage also enabled researchers to take care
of the fact that different instruments at different locations
produce data in different formats; the centralized storage
enabled the researchers to pre-transform all the data into a
standard format. The downside of this pre-transformation is
that it is very time consuming; as a result, by the time the
existing data is transformed, a large amount of new data
appears.

Lately, web communications have become much faster -- to the
extent that it is often much faster to get the results from a

nearby computer than from your own disk space. Thus, it makes sense to avoid time-consuming data transfer and data transformation, to keep all the data in their original storage place and in their original format -- and to design cyberinfrastructure enabling the automatic on-demand transfer and format transformation. Such cyberinfrastructure has been successfully developed for astronomy, bioinformatics, geology, and many other applications. Our NSF-supported CyberShare center aims at expanding this successful technology to other application areas.

Cyberinfrastructure brings new aspects to the important questions of accuracy and reliability of the results of data processing. To answer these questions, we must keep track of the origin (provenance) of the data and of the algorithms that are used for pre-processing the data.

For gauging accuracy and reliability of data, we must modify the existing statistical techniques so that they can easily handle decentralized data and decentralized algorithms.

For handling algorithms provenance, we need to handle different types of such provenance ranging from expert opinion on heuristic techniques to experimental confirmation of semi-heuristic numerical methods to formal proofs of algorithm

correctness. In line with the main ideas behind cyberinfrastructure, it is desirable to combine and process these provenances without moving them to a central location. This necessitates, e.g., a need to keep the proof of correctness of the combined algorithm de-centralized, with proofs of component correctness stored at local machines.

In this talk, we describe techniques and algorithm for handling such provenance issues in cyberinfrastructure.

Keywords: cyberinfrastructure; provenance; statistical analysis of decentralized data; decentralized proofs