Security, Big Data & Algorithmic Accountability

The NSF Secure & Trustworthy Computing Program

Jeremy Epstein, Deputy Division Director
Computer and Network Systems
NSF research programs in Big Data . . .

Encompass Research, Cyberinfrastructure, Education and Training, and Community Building

Cover algorithmic, statistical, and mathematical foundations of data science; new techniques, technologies, and methodologies, including hardware and software approaches; and innovative uses of data for scientific discovery and action

Similar to SaTC, Big Data projects cut across divisions and directorates at NSF
Big Data & SaTC

Big Data is transforming research in all areas of science and engineering, cybersecurity included.

The accumulation of large amounts of personal data by government, companies, and other organizations has important implication for the privacy and security of personal data.
Recent projects & workshops at the intersection of SaTC and Big Data

- Secure Data-Intensive Computing on Hybrid Clouds
- Privacy Preserving Computation in Big Data Clouds
- Workshop: Advancing ethics for trustworthy cyberspace and data analytics
Secure Data-Intensive Computing on Hybrid Clouds

Data-intensive computations traditionally has been done on individual organizations’ internal systems due to concerns with low-cost public clouds adequately protecting sensitive user data.

For cloud-based solutions to be practical, privacy concerns must be addressed.

**Challenge:** existing cryptographic techniques tend to be too heavy-weight to manage large amounts of data.

**Solution:** develop privacy-aware MapReduce system that partitions components across public/private clouds according to security levels required by data, in way that is efficient and secure.

Project involves industry collaborators, and advances may be useful for wide range of computing jobs (e.g., commercial data analysis; DNA analysis; intrusion detection).
Privacy Preserving Computation in Big Data Clouds

Privacy is vital to freedom of creativity and innovation, and must be protected if we are to achieve maximum benefits from harnessing big data.

For cloud-based solutions to be practical, privacy concerns must be addressed.

**Challenge:** the ability to perform efficient big data computations in the cloud has great potential for data analytics related to health, advertising, and other domains, but there are many concerns with user privacy.

**Solution:** The PrivacyGuard project is a practical framework that seeks to enhance privacy-preserving distributed computation by creating algorithms, systems, and tools that guarantee end-to-end privacy throughout a data analytic job.

- Designing formal mechanisms for privacy requirements for data release (e.g., associating data release with usage framework to restrict the analyses that may operate)
- Developing set of guards intended to audit and enforce compliance during analysis
- Devising proactive strategy to prevent information leakages associated with mining output

Integration of research with curriculum development of Georgia Institute of Technology contributes to ensuring future data scientists are aware of privacy.
Big data analytics centers have the potential to address a broad set of problems (e.g., health disparities, natural disasters; social stability in urban settings)
At the heart of many of these problems are central ethical questions related to privacy, inequalities, validity, and use

**Challenge:** Collaboration between users and developers of new analytic tools is important for creating meaningful ethical practices; however, a disconnect exists between developers and users

**Solution:** Hold a workshop that brings together a diverse set of stakeholders with different perspectives, (e.g., developers; ethnographers of scientific practices; users), with goal of building framework for big data creation that emphasizes consideration of ethical issues
USACM Seven Principles for Algorithmic Transparency and Accountability

1. Awareness
2. Access and redress
3. Accountability
4. Explanation
5. Data Provenance
6. Auditability
7. Validation and Testing
Jeremy Epstein
Deputy Division Director, Computer & Network Systems
Directorate of Computer & Information Science & Engineering
jepstein@nsf.gov
703-292-8338