Carnegie Mellon

institute for SOFTWARE RESEARCH

Thesis Defense

Institute for Software Research
Societal Computing

Bias and Beyond in Digital Trace Data

Momin M. Malik

Thursday, 9 August 2018, 9am – 12pm
Wean Hall 7500

Social scientific findings, business decisions, and now even public policies are increasingly being made on the basis of digital trace data from sources such as social media platforms, purchase records, emails, and mobile phone sensors. What could go wrong?

In this thesis, I look at ways in which findings made from such digital trace data may be misleading, and how we might make use of such data in light of these limitations. As an example of demographic bias, I take the largest source of combined geographic, temporal, linguistic, and network data, geotagged tweets, and show that such data exhibits heavy geographic and demographic biases. In an empirical demonstration of algorithmic user manipulation, I use a natural experiment to show how design decisions of social media engineers have a causal impact on user behavior and observed network structure. And I argue that sensors in mobile phones measure proximity and co-location but not necessarily interaction as has been claimed, suggesting a different set of theoretically appropriate research questions and study designs than what have been pursued so far.

I then give two examples of study designs with scopes that avoid problems of bias. The first is a partnership with public health researchers, where we demonstrate approaching Twitter not as a means for public health monitoring but as a medium for public engagement. In the second, I design a study using mobile phone sensors in which I use sensor data and survey data to investigate the relationship between co-location and friendship.

This thesis also demonstrates how to operationalize critiques from critical algorithm/data studies, Science and Technology Studies (STS), and sociology into empirical research questions within computer science. By unifying these areas, I demonstrate a model of computational social science that is rigorous both sociologically and in terms of its statistical modeling, and that can responsibly support decision-making, research, and policy.

Document: **http://mominmalik.com/thesis.pdf**

Jürgen Pfeffer *(co-chair)*	Institute for Software Research
Anind K. Dey *(co-chair)*	Human-Computer Interaction Institute
Cosma Rohilla Shalizi	Department of Statistics & Data Science
David Lazer	Northeastern University