NSF-funded Workshop to Develop a Research Agenda for Service Innovation

Position Statement by: Laurie Garrow

There is a growing need for business analytics skills. Ninety-seven percent of companies with revenues exceeding $100 million are using or plan to use some form of business analytics [1]. The U.S. faces a shortage of 140,000 to 190,000 people with deep analytical skills, as well as a shortfall of 1.5 million data-savvy managers with the know-how to analyze big data to make effective decisions [2]. Companies and universities venturing into the era of "big data analytics" face several new research problems.

One common problem faced by many industries pursing big data applications is the need to combine detailed information about consumers with other data while protecting the security and privacy of individual consumers. For example, economists who want to model consumer welfare impacts of mergers have access to public datasets that contain detailed supply information (in the form of published flight schedules), aggregate demand and fare information (in terms of the number of customers who flew from market A to B through connecting city C in a particular quarter and the average quarterly fare paid by these consumers), and little to no customer demographic information (e.g., gender, age, income). As a second example, there has been increasing interest among the transportation planning community in using passively collected datasets to model travel behavior. This movement, which represents a significant change in modeling philosophy, is important as all metropolitan planning organizations above a certain population size are required to maintain a travel demand model to assess economic, environmental and social impacts associated with major transportation improvements. Developing methods to fuse data from social media sites and location-based devices (such as cell phones) with household demographic data offers the potential to revolutionize how cities plan their infrastructure needs. However, similar to the airline application, the underlying datasets are at different levels of aggregation and partially complete.

In both of these examples, it is theoretically possible to "fill in the gaps" of our knowledge by developing estimation methods that combine these data sources at various levels of aggregation with other data, such as Census data and/or targeted marketing data. Targeted marketing data is sold by firms that compile detailed attitude and lifestyle behavior information about "all" adults in the U.S.; this information is sold to companies to help them target products to specific market groups. Conceptually, detailed information about individuals can be fused across datasets: for a certain subset of individuals, we "know" who made an airline purchase, where the individual lives, and by linking it to targeted marketing data, what their age, gender, income, and household composition is. Similarly, we "know" that a cell phone belongs to a certain individual, where that individual lives and, by linking it to targeted marketing data, their demographic information. Theoretically, while it is statistically advantageous to be able to link data at the individual consumer level, practically, this raises intense privacy concerns. Consequently, there is a research need to: (1) develop methods to fuse multiple individual-level consumer data that ensures privacy concerns are addressed; (2) develop methods to fuse individual-level consumer data with more aggregate data, and enhance theories that can help researchers identify and quantify statistical biases introduced in using more aggregate data to represent individual characteristics.

[1] See http://www.sas.com/resources/asset/busanalyticsstudy_wp_08232011.pdf, page 2

[2] Manyika, et al. (2011).Big Data: The Next Frontier for Innovation, Competition, and Productivity. Report by the McKinsey Global Institute.