CU-led consortium will test new confidential methods to mine census data

By Bill Steele

With modern computing power, data from the U.S. Census Bureau, the Internal Revenue Service, law enforcement agencies and other branches of government can be combined to answer important public policy questions. The trick is to do this without violating people's privacy.

Abowd

The National Science Foundation (NSF) has provided funding for Cornell economist John Abowd and colleagues to develop techniques that enable social scientists to use government-obtained data while maintaining the confidentiality that both law and ethics demand.

This program has now been enlarged with a $2.9 million NSF Information Technology Research grant to expand the types of data to be made available and to ensure that the reprocessed data are valid. The grant is to a consortium headed by Cornell and including Carnegie Mellon University, Duke University, the University of Michigan, Argonne National Laboratory and the Census Bureau.

Abowd is the Edmund Ezra Day Professor of Industrial and Labor Relations at Cornell and director of the Cornell Institute for Social and Economic Research (CISER). He also is a distinguished senior research fellow of the Census Bureau.

These days, "anonymizing" the data -- taking the name and address off a census form -- isn't nearly enough. Widely available public databases make it possible to identify individuals based on combinations of, for example, income level, occupation, geographic area and age. Geospatial databases can associate a street address with its exact latitude and longitude, and probably tell you the size of that household's electric bill and which cable services it subscribes to. Data on businesses can be even more transparent. How many sheet metal fabrication shops are there, for example, in Ithaca?

One of Abowd's solutions is to create synthetic data: a data set of "virtual households" that, taken together, produce the same overall statistical result as the original set of census forms or other records. Another is "coarsening," in which small groups of households or businesses are blended into single records.

One of the early products of this work is Quarterly Workforce Indicators Online at http://lehd.dsd.census.gov/led/01/004/, which allows businesses and local governments to see where the jobs are, for what kind of workers, how much workers can expect to earn and employers to pay, drilling down to individual counties or workforce investment areas.

But social scientists are naturally suspicious of synthetic data. Does it really produce the same statistical results as the actual census microdata from which it's derived? Abowd plans to test this by running actual research projects on both real and synthetic data.

Scientists currently are allowed access to census microdata through nine Research Data Centers (RDCs) -- one located at CISER -- that provide encrypted links to confidential databases in physically secured settings. To use the RDCs, researchers must meet confidentiality requirements and demonstrate that their research will

One of the goals of the project is to create "virtual RDCs," which would provide access to synthetic data over the Internet.

A key component of the project, so far only partially funded by the grant, will be the creation at the Census Bureau of a new 256-processor parallel computer, supported by Intel, Unisys and SAS Institute, to be used for the creation of synthetic data and evaluation of the proposed products.

October 7, 2004

| Cornell Chronicle Front Page | | Table of Contents | | Cornell News Service Home Page |