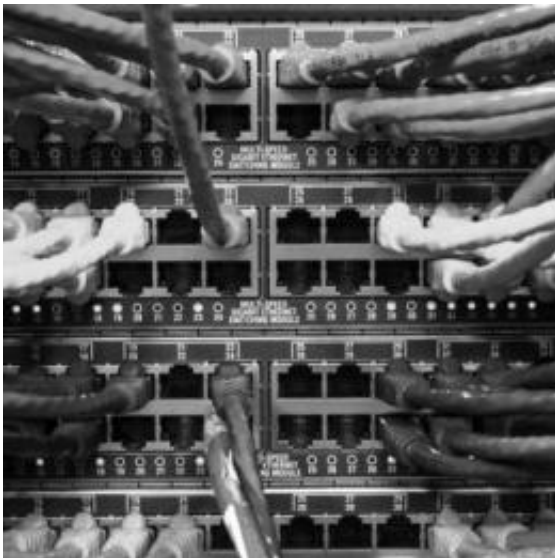# New tools will make sharing research data safer in cyberspace

September 25 2012

Harvard researchers will receive a four-year NSF grant totaling nearly $5 million to study and enhance the privacy of research data. Credit: Adapted from a photo by Ken Fager / Flickr, under a Creative Commons license (BY-NC-SA 2.0).

The real-time data of cyberspace, detailing every like, dislike, spur of the moment thought—and more—provide unprecedented opportunities for research by scientists from all areas.

No longer limited to narrow focus groups, painstaking in-person surveys, or artificially controlled studies, researchers today have a far easier time compiling and manipulating large data sets. At the same time, however,

sharing such data can be fraught with risks.

Now, researchers at Harvard University will receive a four-year grant totaling nearly $5 million from the National Science Foundation's Secure and Trustworthy Cyberspace (SaTC) program to study and enhance the privacy of research data. The "Privacy Tools for Sharing Research Data" project will develop methods, tools and policies to further the tremendous value that can come from collecting, analyzing, and sharing data while more fully protecting individual privacy.

Salil Vadhan, Vicky Joseph Professor of Computer Science and Applied Mathematics at the Harvard School of Engineering and Applied Sciences (SEAS), will serve as the lead investigator of the multi-school, cross-departmental effort that draws upon Harvard's renowned expertise in the social sciences, law, government, statistics, and computer science.

"The Internet and, in particular, social networking sites, provide an amazingly powerful platform for researchers to gather, mine, and share data on human behavior and interactions," explains Vadhan, who conducts research in theoretical computer science. "Even with the best intentions and safeguards in place, however, the risk of personal information leaking out remains high."

While the academic community is eager to share data in an open-access manner, researchers face the risk that by sharing data they may be putting their subjects at risk and, even worse, potentially violating the privacy of individuals who may not even know their data was being used.

Given the complexities involved in ensuring privacy for social science research, Vadhan will be joined in the endeavor by Gary King, Albert J. Weatherhead III University Professor at Harvard University and Director of the Institute for Quantitative Social Science (IQSS); Latanya Sweeney, Professor of Government and Technology in Residence at

Harvard University and Director of the Data Privacy Lab; and Phil Malone, Clinical Professor of Law at Harvard Law School (HLS) and Director of the HLS Cyberlaw Clinic at the Berkman Center for Internet & Society at Harvard.

Additional participants in the grant include Edo Airoldi, Assistant Professor of Statistics at Harvard; Stephen Chong, an Assistant Professor of Computer Science at SEAS; Merce Crosas, Director of Product Development for IQSS; Micah Altman, Director of Research for MIT Libraries and Non-Resident Senior Fellow at the Brookings Institution; and Cynthia Dwork, Distinguished Scientist at Microsoft Research Silicon Valley.

The project was incubated by SEAS' Center for Research on Computation and Society (CRCS), in collaboration with IQSS and the Berkman Center, and with the support of a gift from Google, Inc.

Academics are often prevented from collaborating and tapping into what could be a gold mine for the study of social interactions and human nature, due to legitimate concerns about personal privacy. Likewise, useful data from commercial sites like Netflix or Facebook often remains locked up due to ethical concerns and past cases where supposedly anonymous data has been re-identified.

"Only a few pieces of information can often uniquely identify a person in data," says Sweeney, an expert on data privacy. "Today's data-rich networked society makes de-identifying data increasingly difficult as so much data can be brought to bear. As datasets grow to include millions of people and hundreds of details about each person, harms from accidental releases become significant. Yet we cannot risk leaving data in isolated silos. Enormous benefits are possible to society from sharing data widely with researchers and to individuals from having copies of their own data. It is important to develop ways to share data widely while

providing privacy protection."

The ethical questions raised by projects where data proved to be re-identifiable were, in fact, what inspired the team to propose their research project, which will make it safer to share and study personal data on the web.

"In recent years, the computer science research community has developed a rich mathematical theory for how to protect privacy while analyzing and sharing data," says Vadhan. "We are looking to advance and refine this theory so as to meet the particular needs of social science researchers, as well as develop policy and legal instruments that will work together with the computational tools to protect privacy while enabling data sharing."

The explosion of personal data and the desire to share it digitally have far outpaced the original mandates of the Institutional Review Boards (IRBs) that were established in the 1950s to protect research subjects. To be effective in the new online arena, the IRB protocols and tools must extend beyond the lab notebook and into the virtual world.

"The problems of sharing clean data among trusted researchers have always been there, but on a much smaller scale," says King, one of the leading experts on quantitative social science. "Our project will help formulate standards and expand the pool of those we can share data with. "We hope to take research collaboration to entirely new levels—protecting the public and at the same time, helping to further research that could have profound social benefits."

With corporations and scholars eager to study aggregate data from online health assessments, genomic testing websites, and even online learning platforms (to understand how students learn)—and amid a proliferation of privacy lawsuits—the effort is as timely as it is critical to ensure

intellectual progress.

The new tools will be tested and deployed at the IQSS Dataverse Network, an open-source digital repository that offers the largest catalogue of [social science](#) datasets in the world.

In addition to bolstering the research infrastructure for social scientists, the ideas developed in this project have the potential to benefit society more broadly, offering solutions that may help with the thorny data privacy issues in many other domains, including public health and electronic commerce.

Provided by Harvard University