



## **Final Report**

### **NSF Workshop on Confidential Data Collection for Innovation Analysis in Organizations**

Christian D. Schunn (*University of Pittsburgh*)

Isabel Cruz (*University of Illinois at Chicago*)

Nick Bloom (*Stanford University*)

Evelyne Viegas (*Microsoft Research*)

March 4, 2010

*Acknowledgement.* We would like to thank Julia Lane, Karl Levitt, Sylvia Spengler, and Lenore Zuck for helping to coordinate the work, and all the workshop participants for excellent contributions. This workshop was supported by NSF Award SBE 0943337. The views expressed in this report are the authors' and do not necessarily represent those of the National Science Foundation.

## Table of Contents

<b>TABLE OF CONTENTS</b> .....	<b>2</b>
<b>EXECUTIVE SUMMARY</b> .....	<b>3</b>
RESEARCH ON INNOVATION IN ORGANIZATIONS .....	3
COMPUTER SCIENCE CONFIDENTIALITY RESEARCH .....	3
PROMOTING INTERDISCIPLINARY RESEARCH .....	3
<b>CONTEXT</b> .....	<b>4</b>
DATA FOR A SCIENCE OF INNOVATION .....	4
THE DATA COLLECTION AND AGGREGATION CHALLENGES TO MEASURING INNOVATION.....	5
<b>WORKSHOP OVERVIEW</b> .....	<b>6</b>
<b>DEFINITIONS AND COORDINATION ACROSS DISCIPLINES</b> .....	<b>7</b>
WHAT IS THE DATA PROBLEM ACROSS SOCIAL SCIENCE AND COMPUTER SCIENCE?.....	7
BARRIERS TO COLLABORATION ON CONFIDENTIAL DATA COLLECTION ACROSS COMPUTER SCIENCE AND SOCIAL SCIENCE.....	8
<b>STATE OF THE ART FROM THE DISCIPLINES</b> .....	<b>8</b>
APPROACHES TO PRIVACY IN SOCIAL SCIENCE RESEARCH ON INNOVATION IN ORGANIZATIONS .....	8
COMPUTER SCIENCE RESEARCH ON PRIVACY .....	10
<b>RECOMMENDATIONS FOR THE FUTURE</b> .....	<b>12</b>
RESEARCH ON INNOVATION IN ORGANIZATIONS .....	12
COMPUTER SCIENCE CONFIDENTIALITY RESEARCH .....	13
PROMOTING INTERDISCIPLINARY RESEARCH .....	13
<b>REFERENCES</b> .....	<b>14</b>

## Executive Summary

US economic growth has long been based on cutting edge innovation and organizations. But recently there have been concerns around US productivity growth falling behind that of many Asian countries. On September 9<sup>th</sup> and 10<sup>th</sup>, 2009, a workshop was held at Microsoft, in Seattle, bringing together key individuals from social science and computer science to investigate ways to measure innovation in organizations. The workshop focused on both collecting better data on innovation, organizations and management, and also on addressing the central privacy, security and confidentiality concerns that have historically impeded this data collection. The researchers presented their recent contributions in this area around three themes: 1) third-party data such as US Census data, patent databases, or citation databases; 2) detailed insider data such as internal communications, team video, or team documentation; and 3) broader insider data such as cross-firm surveys. The main recommendations were:

### Research on innovation in organizations

- *Build fewer higher quality public databases.* Historically there have been numerous small organization and innovation databases collected by individuals with limited link-ups. We believe resources could be better spent on a few higher quality public databases.
- *Collect non-standard data on innovation in organizations.* Given the increasing economic importance of intangibles like design practices, organizations, and management, there should be more focus on these alongside traditional R&D measures.
- *Assemble multi-year data.* Given the slow change in innovation practices over time and the interest in examining the impact of policy, it is important to collect time-series panel data on innovation. This should be built into original research proposals.

### Computer science confidentiality research

- *Focus confidentiality research on higher quality databases for public use.* Computer science could develop new tools applied to high value datasets that contribute to the feedback loop in data creation by informing the data creation task.
- *Develop confidentiality approaches that include social scientists as a key user group.* Computer science could provide new ways of processing the data in a secure and private way while allowing for the data to be shared by social scientists.
- *Study innovation activities across information technology modalities.* As the world has become digital, consumed via different modalities like emails and blogs, we believe this program should focus on information exchange as part of innovation in organizations

### Promoting interdisciplinary research

- *Develop sharable problems.* Interdisciplinary collaborations should focus on problem-based issues that are intellectually interesting to both disciplines.
- *Leverage organization input to create sharable testbeds.* Organizations should to be involved in the selection of particular problems to be selected, perhaps creating test-beds that are useful for computer science work and interesting to social science researchers.
- *Invest progressively in interdisciplinary research.* Funding should support multiple PIs, beginning with moderate-size short-term seed funding and continuing with larger funding to groups that demonstrate early success.

## Context

### Data for a Science of Innovation

Even before the recent economic downturn, there have been long-run concerns over falling levels of innovation in the US, especially relative to the apparently increasing rates of innovation in Asia. A National Science Foundation-commissioned report from the National Academy of Engineering warned: “Leadership in innovation is essential to U.S. prosperity and security. In a global, knowledge-driven economy, technological innovation, the transformation of new knowledge into products, processes, and services, is critical to competitiveness, long-term productivity growth, and the generation of wealth. U.S. leadership in technological innovation seems certain to be seriously eroded unless current trends are reversed.”[NAE] It is difficult to overemphasize the economic importance of innovation. Sixty-five percent of total revenues for technology-based companies have come from products that are less than five years old. [K-12] Cross-national studies show a high correlation between patents per million and a nation’s standard of living. [Fagerberg] The Design Council (U.K.) found that companies known for innovative design outperformed the average Financial Times Stock Exchange Index company by 200 percent from 1994 to 2003.

It is hard in the short run to dramatically increase the supply of scientist and engineers, given the long time delays in the education system, and the constraints starting with a decreasing interest in students in science, technology, and mathematics careers at the K-12 level. Further, it is not clear that the market demands large increases in the supply of technical innovators. However, increases in the innovation output of the existing workforce are highly desirable.

For these purposes, building a science of innovation is very useful. NSF has a long-standing interest in studying science and innovation (e.g., with its long-standing Science Technology and Society program as well as division of Science Resources Statistics). Further, NSF is increasingly investing in this area (e.g., with the creation of the Innovation and Organizational Sciences program and the Science of Science and Innovation Policy program).

Many different kinds of social scientists have been working in several different areas to provide new understandings of how organizations innovate, including economists, sociologists, organizational scientists, anthropologists, and psychologists. The range of social scientists reflects the diversity of methodologies required to studying something as complex and inherently multi-level a phenomenon as innovation in organizations.

At the same time Governments and Federal Agencies are struggling to expand research access to data—one clear way to improve the productivity of currently active scientists—because of breaches of cyber infrastructure security, especially those involving unauthorized record linkage. A common challenge is how to make such real-world large-scale data available to researchers to nurture innovation and perform valid experimentation, while maintaining the protection of privacy associated with electronic databases involving individuals, groups, and organizations [Fienberg]. Computer scientists have developed a variety of techniques and built new tools that manage large datasets in ways that could potentially help in supporting and measuring innovation activities. But a better understanding of what data needs to be collected to study

innovation is necessary to successfully apply security and privacy computer scientist techniques or possibly develop new techniques.

### **The Data Collection and Aggregation Challenges to Measuring Innovation**

Although these social scientists have been using many different kinds of data, one fundamental bottleneck to improving a science of innovation is access to high quality data on innovation, including data on innovation inputs, innovation processes, and innovation outputs. From the scientist's perspective, the challenges are many. One immediate obstacle is the collection of new innovation data, given that much of the innovation process is currently unmeasured. The statistical agencies and firms collect public and private research and development data, but this represents just a fraction of the innovation activity. Innovations in intangibles like designs, organizational and management practices have not been systematically studied because of the absence of data on these. As a result a large swathe of innovative activity is unmeasured and therefore extremely hard to develop effective policy for. Second, even when the data exists it needs to be integrated, but this is typically very difficult because these data are not easily linked logically or geographically. For example, privacy concerns frequently make access to much of the data difficult; some data is not easily captured (e.g., data on the thinking processes of innovators) or not easily analyzed given its complex and large size (e.g., the communication processes of innovators over long periods of time).

Increasingly, information about the inputs, processes, and outputs of innovation is being captured and stored electronically for various purposes. Yet the data is not easily accessed and processed in useful ways. To help overcome these challenges, social scientists need new tools to access innovation-related data.

To address both the collection and linkage issues, computer scientists have been developing a variety of techniques and building new tools that help to collect and manage large datasets. At the same time, the kind of data management that is needed to support innovation activity will pose new and specific challenges that will enrich future computer science research.

Across social science and computer science perspectives, three types of data need to be managed: 1) third-party data such as US Census data, patent databases, NSF funding data, or citation databases; 2) detailed insider data such as internal communications, team video, or team documentation; and 3) broader insider data such as cross-firm surveys. A common challenge is how to make such real-world, large-scale data available to researchers to nurture innovation and perform valid experimentation, while maintaining data privacy. In particular, privacy-preserving techniques will be needed, for example, to integrate data across organizations, to publish statistical data, to perform data mining, to analyze social networks or cooperative work, or to track the location of users and their trajectories.

Despite these common interests and clear interdependence, social science and computer science work frequently takes place in different circles with different goals. These differences create significant barriers to the development of new, transformative approaches.

## Workshop Overview

To better direct its support of innovation and discovery, the US National Science Foundation has an opportunity to fund research that improves our understanding of the factors (including cognitive and social psychological) that improve or increase innovation and discovery. To know how those funds should be profitably directed, a workshop was sponsored with the task of understanding the state-of-the-art and providing a vision for critical future research directions.

Such a workshop was conducted on September 9<sup>th</sup> and 10<sup>th</sup>, 2009 on the Redmond, WA campus of Microsoft Research. Microsoft is an instance of a firm interested in collecting useful data on its own innovation practices. Microsoft also conducts research on how to collect and process many forms of electronic data. Finally Microsoft creates software used by the majority of computer users around the world and already stores relevant electronic data to the study of innovation. Thus, running the workshop at Microsoft Research set a relevant pragmatic context.

**Table 1.** Presenters, Research Areas, and Presentation Titles.

Topic	Name	Title
Detailed insider data	Christian Schunn	Sampling video of innovative engineering design teams
Social Science	Kathryn Shaw	Insider Econometrics: Using Innovative Data to Identify Sources of Productivity Gains
	Susan Finger	What do engineering student project teams talk about when we're not there?
	Nick Bloom	Management as a technology: evidence from India
Detailed insider data	Bhavani Thuraisingham	Confidentiality, Privacy and Trust for Data Analysis and Mining
Computer Science	Philip Yu	Privacy and Mining on Information Networks
	Evelyne Viegas	Breaking down Data Barriers to Drive Open Innovation
Third party external data	Carol Corrado	The Black Box of Intangible Capital: Collecting Data from Deep Within Firms
Social Science	Laurel Smith-Doerr	A Sociological Perspective on Studying Organizations and Innovation: Qualitative and Quantitative Methods and Ethical Issues
	Lee Fleming	Disambiguation of inventors and identification of their mobility and social networks: opportunities for integration with confidential databases
Third party external data	Kevin Fu	Confidential collection of telemetry from medical devices
Computer Science	Cynthia Dwork	Differential Privacy and Pan-Private Algorithms
	Tiziana Catarci	Private Matching of Data for Knowledge Enhancement
Broad insider data	Helen Nissenbaum	Confidentiality and Contextual Integrity
Social Science	Wes Cohen	The Division of Innovative Labor: Features, Determinants and Impacts on Innovative Performance
	Calvin Morrill	Transferring Insights from the Ethnographic Study of Organizational Conflict Management to the Study of Confidentiality and Privacy in Organizational Research on Innovation
Broad insider data	Ravi Sandhu	Application-Centric Security: UCON, PEI and g-SIS
Computer Science	Tim Finin	Confidential Data Sharing and the Semantic Web
	Michael Gertz	Confidentiality in the Context of Spatially Explicit Data
	Isabel Cruz	Data Integration and Context-Aware Applications

At the same time, the workshop was international and broad in flavor, with no direct requirements for any of the presenters to make direct connections to the Microsoft context.

The workshop was led by researchers from computer science (Isabel Cruz and Evelyne Viegas) and social science (Christian Schunn and Nick Bloom).

The workshop included 20 researchers representing the current state-of-the-art in the computer science of privacy and social science of innovation, with approximately half of them from computer science and the other half from social science. The researchers presented their recent contributions in this area in the form of 15-minute talks. Talks were organized into three theme groups based on data type: 1) third-party data such as US Census data, patent databases, NSF funding data, or citation databases; 2) detailed insider data such as internal communications, team video, or team documentation; and 3) broader insider data such as cross-firm surveys (see Table 1).

Between clusters of talks there were three breakout sessions that addressed the following questions:

- What are similarities and differences between how social science and computer science researchers have defined the data problem?
- What are similarities and differences between why social science and computer science researchers are interested in the data problem?
- What are barriers to collaboration between social science and computer science on collecting and using data for analysis of innovation in organizations?
- What are common features of successful social science/computer science collaboration on collecting and using data for analysis of innovation in organizations?
- What are strategies for improved collaboration between social science and computer science on collecting and use data for analysis of innovation in organizations?
- What are open questions that critically depend upon the collaboration between social science and computer science regarding access data for analysis of innovation in organizations?

The sections that follow report and build upon the core insights that came from the talks and breakout sessions.

## **Definitions and Coordination across Disciplines**

### **What is the data problem across social science and computer science?**

The key data challenge is 1) collecting, capturing, and storing critical data 2) private for some purposes/users but open for other purposes/users [Nissenbaum] in a rapidly changing technological environment in which 3) more and more data is being stored, 4) new methods are being developed to mine data while 5) new methods are being developed to keep them private. This five-part challenge applies to the full life-cycle of data: how is data initially obtained, how is it initially processed, with what other data may it be combined/integrated, who has access to it, and when is it disposed of?

Across both social scientists and computer scientists, the volume of the data can be extremely large (e.g., Terabytes) and often requires computational algorithms to manage the data. In both cases, there is a challenge to describing meaningful relationships that must be derived from the raw data, and researchers are continually creating innovative ways to find new relationships in the raw data.

But there are also important differences between computer science and social science perspectives to data privacy. For the social scientists, the top two challenges are firstly, collecting new data, for example by observation within firms, surveys and field work, and secondly, gaining access to currently existing private data by assuring stakeholders (individuals and organizations) that privacy will be maintained. In both cases the stakeholders are often willing to be inconvenienced to provide data if privacy can be maintained, but the methods need to minimize inconvenience of the populations being studied. Further, the privacy-preserving methods cannot disturb critical causally relevant aspects of the data, and thus the key variables (and acceptable privacy-preserving methods) are very much domain or situation dependent. Social scientists are generally unwilling to work with fake datasets.

For computer scientists, privacy-preserving methods are developed in a more context-free or general fashion. It is less important what particular things will definitely be done with the data, and it is more important what could be done with the data by yet unknown third parties. Computer scientists are willing to work on fake datasets because such datasets enable them to test out new methods without having to hassle with obtaining difficult-to-obtain features (e.g., true user identity).

### **Barriers to collaboration on confidential data collection across computer science and social science**

Across any scientists, sharing of confidential data is complex. Computer science studying privacy is diverse and social science making use of confidential data on innovation in organizations is incredibly diverse. Across this diversity, terminology can be a barrier to meaningful collaboration, with common terms in one field being meaningless or defined differently across disciplines. Most importantly, the goals of computer science and social science are different, creating conflict in what research directions are interesting to follow, what kinds of publications are acceptable outlets of the research, etc.

## **State of the Art from the Disciplines**

We begin with a brief summary of the state of the art from each discipline, with pointers to workshop presenters whose presentations elaborate each of those points (for copies of the presentations, visit <http://www.lrdc.pitt.edu/schunn/cdi2009/>).<sup>1</sup>

### **Approaches to Privacy in Social Science Research on Innovation in Organizations**

Some social scientists have very creatively mined existing publicly available data to study factors that influence innovation. For example, [Fleming] and colleagues mined US patent databases to build social network models and track the migration of innovations, innovators, and

---

<sup>1</sup> Participants' presentations relevant to topics are referenced in parenthesis in this section.

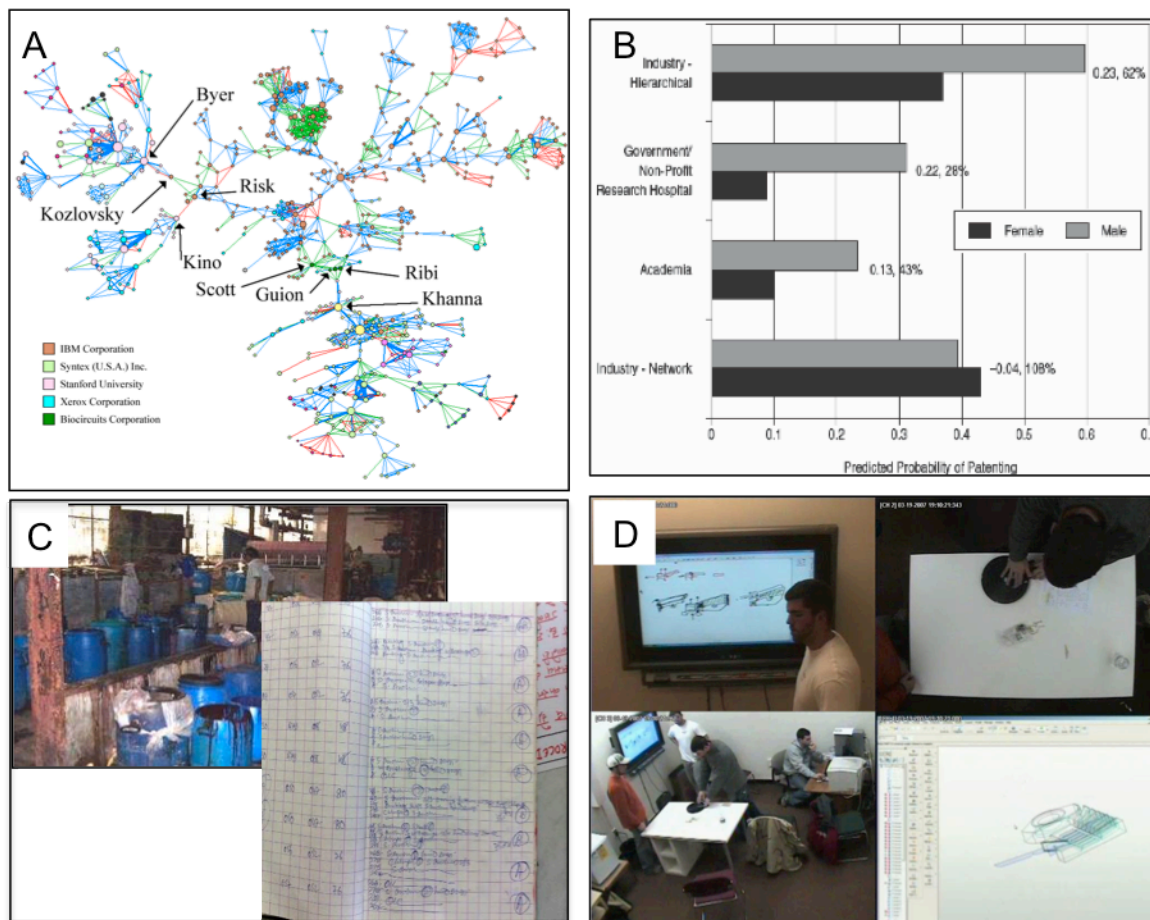
key network innovators following changes in state laws (see the Silicon Valley network in Figure 1A); this work is technically challenging in requiring computer algorithms to uniquely identify individuals on the basis of names, which are very ambiguous, especially for Smiths and Chens.

A related approach is to add interviews to analyses of publicly available large-scale innovation data to help add private causality information to correlational analyses derived from public data (such as determining the factors underlying differences in patenting behavior by males and females in different work sectors; see Figure 1B) [Smith-Deorr].

Others social scientists go to great lengths to assure anonymity of new data collection AND offer rewards for participation in order to obtain more private data on innovation. For example, [Bloom] was able to collect extensive innovation data from within Indian firms on changes in management and organizational practices (see Figure 1C) by offering some kinds of training as a benefit for all firms in the study and considerable face-to-face negotiation with each firm regarding their anonymity in the reports, especially with respect to reporting revenue to government agencies. [Shaw] has very successfully undertaken similar approaches in US and European steel and valve firms, by again collecting new data on innovations in management and organizational practices and how these lead to faster economic growth. A strategy to improve anonymity of particular sources is to use multi-site collection [Morrill].

[Schunn] obtained terabytes of video data on 60 semester-long engineering design projects in a university setting by providing useful design spaces, \$250 to each participant, and assurances to keep identifiable data within the research group. Similarly, [Finger] obtained complete records of email trails from student design teams in many different courses by providing instructors and students with a web-based collaboration tool that facilitates the teams' design work. Privacy concerns create significant limitations on which innovation work can be studied (e.g., in student teams vs. industry teams, or industry teams in which countries) and how widely the difficult-to-obtain data can be shared.

Another common approach is to conduct large anonymous surveys within organizations. However the data collection procedure is itself non-anonymous per se, whether it by mail, Internet, or by phone interview. The respondent must trust the data collector not to store organization or individual identifiers in the data that is shared. Even more difficult, the respondent must trust the data collection and data reporting procedures do not *de facto* identify the respondent through unique combinations of data. One approach is to using organizations with widely known reputations for independents and integrity to collect and house the data—for example the National Opinion Research Center, the Census Bureau or the leading universities. But providing outside scholars access to the raw data is nonetheless complex even if these agencies are able to successfully collect confidential innovation data [Cohen, Corrado].



**Figure 1.** Examples of social science data on innovation practices: A) Silicon Valley innovation network [Fleming], B) gender patenting rates by work sector [Smith-Doerr], C) Indian firm work practices [Bloom], and D) engineering student team video [Schunn].

### Computer Science Research on Privacy

From the data integration perspective, a conflict exists between the methods used for data integration and privacy protection. On one hand, information needed in the datasets must be made explicit in order to integrate data from different datasets. On the other hand, privacy constraints may not allow that information to be made explicit. Approaches using statistical methods may employ datasets in which information is perturbed and anonymized, but perturbing the data may significantly reduce its utility [Catarci, Cruz]. In addition, data perturbation techniques require significant methodological knowledge from the people who make use of the data [Gertz].

The problem of privacy is closely related to the problem of access control, which deals with how to control access to data, how to specify security policies, and how to enforce them (see Figure 2A). The latest developments in the IT world, such as ubiquitous computing and social networks, have characteristics of high dynamicity, and call for new ways for specifying and enforcing access control [Cruz, Finin, Sandhu, Yu]. From this point of view, application-centric security is needed, in which trade-offs between other aspects of the applications, such as performance, cost etc, can be addressed within the objectives of security [Sandhu]. Another direction of research

that tries to take into account the latest developments is the research on Security and Semantic Web [Cruz, Finin]. The Semantic Web is a vision of the web in which data semantics is made explicit and machine-processable (see Figure 2B). It offers good opportunities for solutions to the problems of data integration, protection, and assurance.

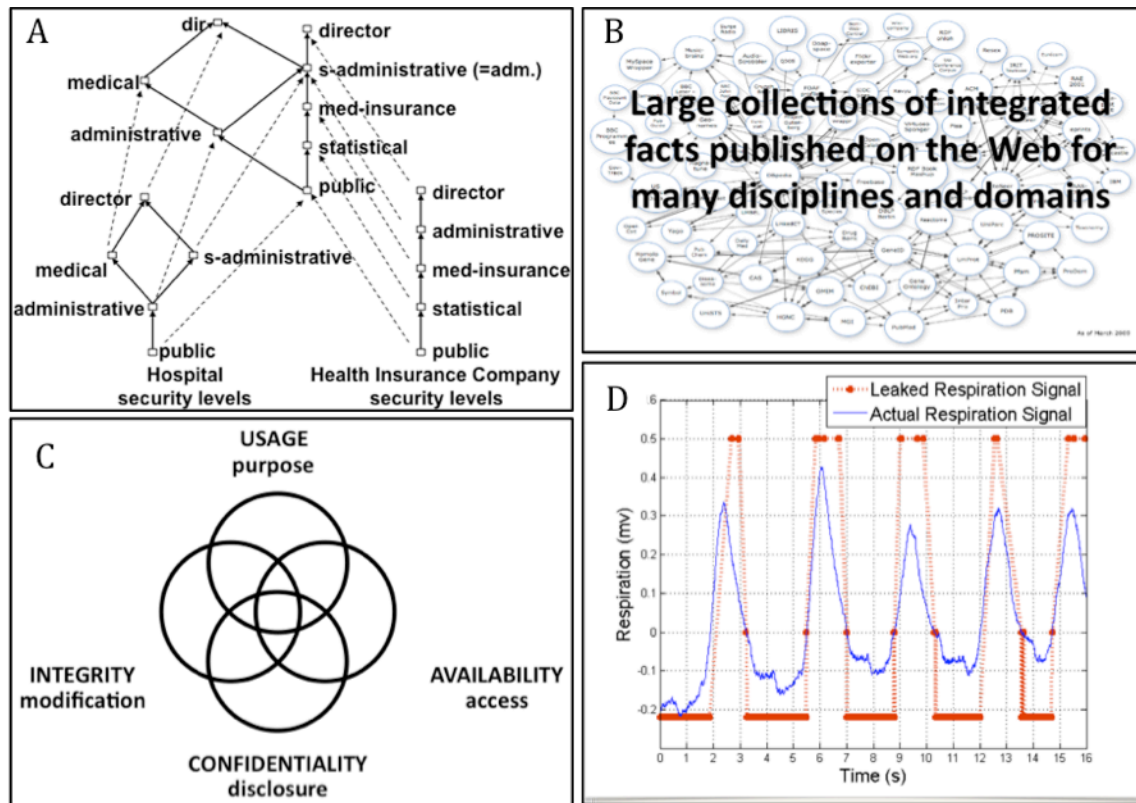
Another notion closely related to access control models is the notion of “Information Assurance”; data protection must involve the whole information-sharing lifecycle [Finin]. All the phases of information sharing, namely advertisement of information, discovery, acquisition, use, and release must be controllable. Different solutions exist to secure each of the phases, but an integration between these solutions is needed [Finin].

One largely agreed upon view is that there is a gap between the policy specification layer (such as legislation, or requirements on privacy that must be respected by applications) and the underlying IT layer, which enforces the policies. This gap needs to be filled in some cases, and there is work on policy languages trying to fill this gap [Gertz]. This gap is also desirable as a way to be able to think about different aspects of the security problem separately, in a divide-and-conquer way [Sandhu] (see Figure 2C). Another important aspect that must be taken into account when sharing data is that risks to data privacy do not come only from people directly using the data, but also from malicious software [Sandhu].

Access to real world data can bring innovation by allowing researchers to unveil *new* analysis or *research directions*, and explore new questions. Data benchmarks are necessary to enable science via *repeatability of experiments* so that results can be confirmed and erroneous results avoided. However, some of these datasets contain sensitive information, making it difficult to share them. An implementation of differential privacy [Dwork] has been applied to search logs, which constitute a good proxy to study present trends and predict new ones, showing that it was possible to do useful and valid research within a reasonable risk-utility trade-off [Viegas].

As the quantity of data continues to grow, calling for new storage and compute solutions, cloud computing is becoming the new paradigm to study data at scale. Today privacy and security threats are not fully understood. [Fu] illustrates the issues in the world of medical telemetry showing that encryption solutions are not sufficient and that new research directions need to be developed. As a simple example, it is possible to eavesdrop and recreate the respiration signal (see Figure 2D) or heart rate data from encrypted medical device bluetooth data [Fu].

Another dimension to consider when studying privacy is the societal dimension. Different organizations may have different definitions of what constitutes privacy; policies may be application specific; different countries have different privacy laws [Thuraisingham]; information ‘consumers’ or producers have yet a different understanding of what constitutes their privacy when they go on social networks and how their data should be used as discussed in [Yu]. As more data is made available by people online, it becomes even easier to cross-reference it with other datasets, increasing the privacy risks, as also shown for instance in [Frankowski].



**Figure 2.** Examples of computer science perspectives on privacy. A) access policies defined by integration of role hierarchies [Cruz], B) semantic web and linked data [Finn], C) different aspects of the security problem [Sandhu], and D) signal detection in spite of encryption [Fu].

A possible solution to keeping data safe and away from any privacy leak may lie in a new research direction on “pan-private” streaming algorithms. Data collected is never stored and is kept “pan-private” with an internal state that is differentially private [Dwork]. This would have the advantage of keeping ‘data snoopers’ at bay and preserving individuals’ privacies. It is still to be understood, at this stage, how data-driven research could be fulfilled in such a framework.

### Recommendations for the Future

We have broken up our recommendations in three broad areas: how to build high quality databases on innovation in organizations, computer science confidentiality research, and how to promote interdisciplinary research between social science and computer science.

#### Research on innovation in organizations

- *Build fewer higher quality databases for public use.* Given the difficulty of confidentiality, the limits on respondent time, and the costs of collection, funding should be focused on fewer high quality databases that can be widely accessed, rather than numerous proprietary databases. Historically there have been numerous individual small databases collected by individuals with limited link-ups. We propose fewer centralized databases with a controlled public access approach. Examples of such databases are the SciSIP funded innovation

database being collected by Wes Cohen, the Kauffman funded firm-survey, the NBER patents database [Fleming], and the Census plant level databases.

- *Collect non-standard data on innovation in organizations.* Historically innovation data has focused on the most easily measured inputs like research and development expenditure, and the most easily measured outputs like patents. Given the increasing economics importance of intangibles, like design practices, organizations, and management [Corrado, Schunn, Shaw, Bloom] we think greater focus should be given to measurement of innovations in these broader areas.
- *Assemble multi-year data.* Given the slow change in innovation practices over time [Cohen, Shaw, Bloom] and the interest in examining the impact of policy, it is important to collect time-series panel data on innovation. This should be built into original research proposals so that they can be easily extended over multiple years.

### **Computer science confidentiality research**

- *Focus confidentiality research on higher quality databases for public use.* Computer science could contribute new tools applied to high value datasets that contributed to the feedback loop in data creation by informing the data creation task.
- *Develop confidentiality approaches that include social scientists as a key user group.* Computer science could provide new ways of processing the data in a secure and private way while allowing for the data to be shared by the community of social scientists involved in the same research topic, and allowing the analysis of the data to be shared, extending the notion of data confidentiality to data derivatives and of privacy to the social network of social scientists.
- *Extend confidentiality approaches to study innovation activities across information technology modalities.* As the world has become largely digital, consumed via different devices (e.g. laptops, phones) and modalities (e.g., emails, wikis, blogs, twitter), focusing a cross-functional effort on information exchange as part of innovation in organizations would place this NSF new program in the 21<sup>st</sup> century of information technology modalities.

### **Promoting interdisciplinary research**

- *Develop sharable problems.* Interdisciplinary collaborations should focus on problem-based issues that are intellectually interesting to both disciplines.
- *Leverage organization input to create sharable testbeds.* Organizations should to be involved in the selection of particular problems to be selected, perhaps creating test-beds that are useful for computer science work and inherently interesting to social science researchers.
- *Invest progressively in interdisciplinary research.* Funding should support multiple PIs, beginning with moderate-size short-term seed funding and continuing with larger funding to groups that demonstrate early success.

## References

- [Bloom] Nick Bloom. Does management matter: Evidence from India. (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Catarci] Tiziana Catarci. Private Data Matching for Knowledge Enhancement (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Cohen] Wes Cohen. The Division of Innovative Labor: Features, Determinants and Impacts on Innovative Performance (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Corrado] Carol Corrado. The Black Box of Intangible Capital: Wanted! Data from Deep within Firms (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Cruz] Isabel F. Cruz. Semantic Web Technologies for Data Integration and Context-Aware Applications (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Dwork] Cynthia Dwork. Differential Privacy and Pan-Private Algorithms (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Fagerberg] Fagerberg, J., Mowery, D., & Nelson, R. (2004). *The Oxford Handbook of innovation*. Oxford: Oxford University Press.
- [Fienberg] Fienberg, S., Anton, A., Bertino, E., Dwork, C., Viegas, E. and L. Zayatz (2007) Data Confidentiality – NSF grant report 0741571. Also available at <http://dcws.stat.cmu.edu/index.html>
- [Finger] Susan Finger. What Are They Talking About When we're not Listening? (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Finin] Tim Finin. Assured Information Sharing (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.

- [Fleming] Lee Fleming. Dataverse Network (DVN) Patent Network Database Project (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Frankowski] Frankowski, D., Cosley, D., Sen, S., Terveen, L. And J. Riedl (2006) You Are What You Say: Privacy Risks of Public Mentions. In Proceedings of the 20<sup>th</sup> Annual International ACM SIGIR, Conference on Research & Development on Information Retrieval, August 6-11, Seattle, WA, USA.
- [Fu] Kevin Fu. Emerging Challenges for Security + Privacy. Confidential Collection of Medical Telemetry (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Gertz] Michael Gertz. Confidentiality in the Context of Spatially Explicit Data (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [K-12] *Engineering in the K-12 classroom. An Analysis of Current Practices & Guidelines for the Future*  
[[http://www.engineeringk12.org/educators/taking\\_a\\_closer\\_look/documents/Engineering\\_in\\_the\\_K-12\\_Classroom.pdf](http://www.engineeringk12.org/educators/taking_a_closer_look/documents/Engineering_in_the_K-12_Classroom.pdf)]
- [Morrill] Calvin Morrill. Transferring Insights from the Ethnographic Study of Organizational Conflict Management to the Study of Confidentiality and Privacy in Organizational Research on Innovation. In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [NAE] *Assessing the capacity of the U.S. engineering research enterprise.*  
[<http://www.nae.edu/nae/engecocom.nsf/weblinks/MKEZ-68HQMA?OpenDocument>]
- [Nissenbaum] Helen Nissenbaum. Confidentiality and Contextual Integrity. In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Sandhu] Ravi Sandhu. Application-Centric Security: How to Get There (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.

- [Shaw] Kathryn Shaw. Insider Econometrics: Using Innovative Data to Identify Sources of Productivity Gains (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Smith-Doerr] Laurel Smith-Doerr. A Sociological Perspective on Studying Organizations and Innovation: Qualitative and Quantitative Methods and Ethical Issues (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Thuraisingham] Bhavani Thuraisingham. Data Mining, Security and Privacy (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.
- [Viegas] Evelyne Viegas. Microsoft Research. Breaking down Data Barriers to Drive Open Innovation (slide presentation). In NSF/Microsoft Workshop on Confidential Data Collection for Innovative Analysis in Organizations, Redmond, Washington, September 2009.