Privacy and Confidentiality

Heng Xu

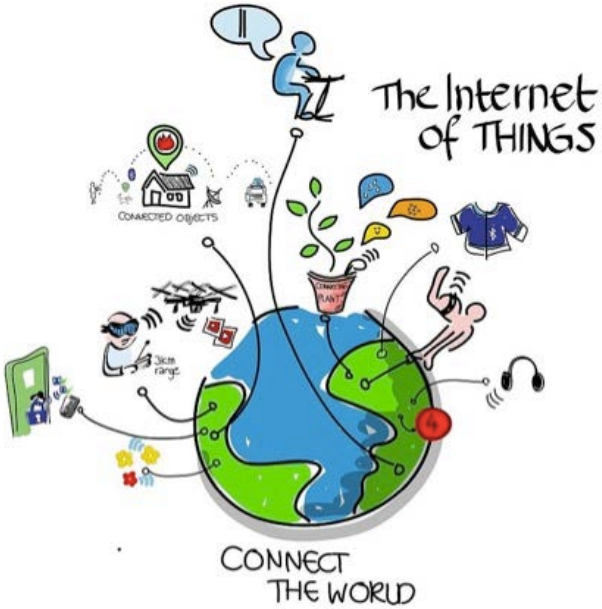
Directorate for Social, Behavioral and Economic Sciences

National Science Foundation

SBE Advisory Committee Meeting
October 30, 2014

Data Intensive Research

- Jeryl mentioned SBE researchers' cyberinfrastructure needs
  - New combinations of "designed data" and "organic data"
    - *Designed data*: survey questionnaire implemented as a smartphone application
    - *Organic data*: location, contact lists, call logs, text messages, emails, browsing histories, usage patterns, calendar, photos, and etc.



For Official Use Only

Security and Privacy

Security

Privacy

- Fair Information Practice Principles (FIPPs)
  - Transparency
  - Control
  - Purpose Specification
  - Data Minimization
  - Use Limitation
  - Data Quality and Integrity
  - Security
  - Accountability and Auditing

Privacy is broader and encompasses permission and use of personal information. Privacy is difficult to achieve without security. But we can successfully secure the personal information in databases but still make bad decisions about subsequent use and distribution of personal information, resulting in privacy problems.

For Official Use Only

Privacy Problems: The Trouble with De-identification

- **Significant risks stem from re-identification:**
  - In 2000, Latanya Sweeney (CS professor) showed that 87 percent of all Americans could be uniquely identified using three types of information: **date of birth**, **gender**, and **ZIP code**.

For Official Use Only

Privacy Problems: The Trouble with Integration

- **Significant risks stem from data integration:**
  - In 2009, Alessandro Acquisti and Ralph Gross (behavioral economics researchers) demonstrated the feasibility of predicting SSNs from public data:
    - the Social Security Administration's Death Master File
    - the widespread accessibility of personal information from multiple sources (e.g., social networking sites).

PNAS

Predicting Social Security numbers from public data

Alessandro Acquisti[1] and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of personal information from multiple sources, such as data brokers or profiles on social networking sites. Our results highlight the unexpected privacy consequences of the complex interactions among multiple data sources in modern information economies and quantify privacy risks associated with information revelation in public forums.

identity theft | online social networks | privacy | statistical reidentification

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within New York state may be assigned any of 85 possible first 3 SSN digits). Within each SSA area, GNs are assigned in a precise but nonconsecutive order between 01 and 99 [RM00201.030] (1). Both the sets of ANs assigned to different states and the sequence of GNs are publicly available (see www.socialsecurity.gov/employer/statewch.htm and www.ssa.gov/history/ssn/geocard.html). Finally, within each GN, SNs are assigned "consecutively from 0001 through 9999" (13) (see also [RM00201.030], ref. 1.)

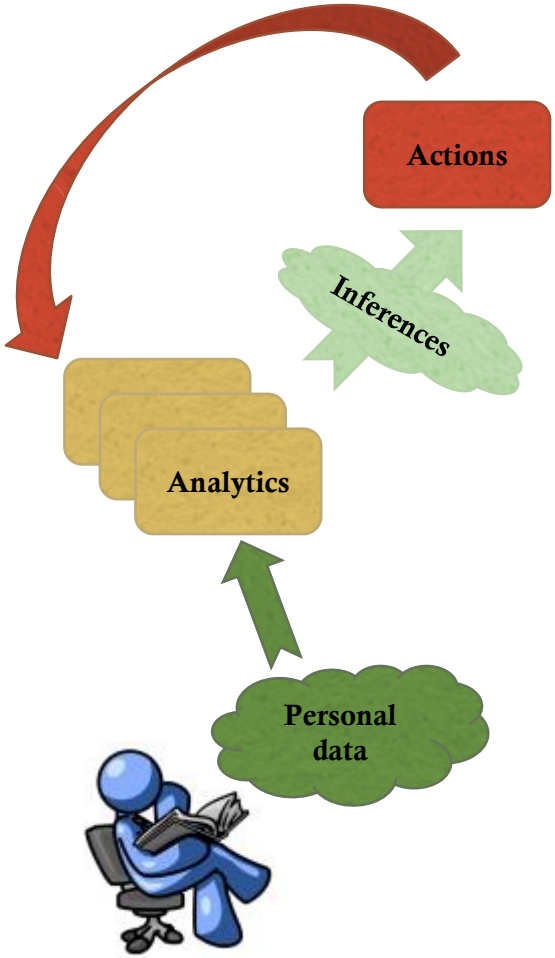
For Official Use Only

Privacy by Design

- **Ensuring Cyberinfrastructure reflects legal obligations and respect for data subjects**
  - Mapping privacy regulations into actionable IT requirements that are re-usable across systems.
  - How, and under what conditions, should we go beyond privacy laws to identify the "right things to do" to ensure privacy?
- **A comprehensive approach needs to be based on a systematic data lifecycle approach.**
  - Compliance from the point of data acquisition through the point of data use, dissemination, maintenance and disposition.
- **Address ongoing legal developments in privacy**
  - Trace design decisions so that, when a policy or law changes, appropriate action can be taken to re-design the systems to return to a state of compliance.

For Official Use Only

Human Data Interaction



- Fundamental research in human factors, psychology of privacy, economics of privacy, regulations of privacy to inform privacy abstractions, privacy models, etc.
- Situational, contextually sensitive, with focus on reasons for use as much as who or what

For Official Use Only

Secure and Trustworthy Cyberspace (SaTC)

- NSF's flagship research program for research in cybersecurity & privacy
- $80+M in FY14 grant cycle
- ~650 currently active grants
- SaTC contains four 'perspectives' under which proposals can be submitted:
  - Trustworthy Computing Systems
  - Social, Behavioral, and Economic sciences (SBE)
  - Cybersecurity Education
  - STARSS (hardware security) – new for FY15

NSF Privacy Research Portfolio

- Formulation of privacy in terms of access to information
- Formulation of privacy as a social-psychological construct
- Privacy policy formulation, specification, enforcement, analysis
- Algorithmic foundations for privacy and tools
- Economics of privacy
- Usability aspects of privacy
- Privacy - Security - Usability trade-offs
- Privacy preserving solutions for data integration, mining, querying
- Privacy preserving solutions for cloud computing

Dear Colleague Letter: Research on Privacy in Today's Networked World

NSF 14-021

Dear Colleague Letter: Research on Privacy in Today's Networked World

December 20, 2013

Dear Colleague:

Privacy is a major issue of the information age. Organizations are increasingly acquiring and storing vast quantities of information about individuals. In addition, advances in big data analytics enable organizations to combine previously distinct information sources and to examine these data to uncover hidden patterns, correlations, and other revealing information. Research on privacy is needed to address how technological change and societal trends are combining to reshape privacy and the implications of such reshaping.

The directorates for Social, Behavioral, and Economic Sciences (SBE) and Computer and Information Science and Engineering (CISE) invite investigators to submit proposals that address the need to develop new and deeper understandings of privacy in today's networked world. Our interest spans both disciplinary and interdisciplinary research in an array of SBE sciences. Proposals for workshops to explore novel and interdisciplinary SBE and SBE/CISE approaches to privacy are also welcome...

Proposals should be submitted through existing programs including:

- SBE Interdisciplinary Behavioral and Social Science (IBSS) Research program
- Secure and Trustworthy Cyberspace (SaTC) program
- Standing SBE programs

Number of Awards	9
NSF SBE Contribution	$1,232,184
NSF CISE Contribution	$2,230,946
Total Funds	$3,463,130
Average Award Size	$384,792
Minimum Award	$49,999
Maximum Award	$1,200,000

Backup slides

Conceptual Underpinnings of Privacy

- **Law:**
  - "right to be left alone" (Warren and Brandeis 1890).
- **Social Psychology, Information S**
  - the ability of individuals t
    personally identif ... shared
    (Westin 1

Privacy "is in disarray and nobody can articulate what it means" (Solove 2006, p. 477)

... ited access to a person" (Schoeman 1984)

- **Computer Science:**
  - access control
  - de-identification via anonymity, unlinkability, unobservability, or pseudonymity

Fair Information Practice Principles (FIPPs)

- **Transparency**
  - Notify individuals regarding collection, use, dissemination and maintenance (CUDM) of Personal Identifiable Information (PII)
- **Individual Participation**
  - Seek consent for CUDM of PII; provide access, correction, and redress mechanisms
- **Purpose Specification**
  - Articulate authority and intent for CUDM of PII
- **Data Minimization**
  - Collect only that PII which is necessary for specific purpose
- **Use Limitation**
  - Use and sharing of PII restricted to the specified purpose
- **Data Quality and Integrity**
  - Ensure PII is accurate, relevant, timely, complete
- **Security**
  - Protect PII in all media from loss, unauthorized access or use, destruction, modification, or unintended or inappropriate disclosure
- **Accountability and Auditing**
  - Accountable for complying, provide training, and demonstrate compliance

Internationally recognized principles for assessing privacy impacts

For Official Use Only

NIST Privacy Engineering Objectives and Risk Model

Illustrative Mapping of Privacy Engineering Objectives to Problematic Data Actions

Data Lifecycle Phase	Normal Data Action	Problematic Data Action	Potential Harms
Predictability			
Collection	Service Initiation	Induced Disclosure	Power Imbalance, Loss of Autonomy
Processing	Aggregation	Unanticipated Revelation	Stigmatization, Power Imbalance, Loss of Trust, Loss of Autonomy
Processing	System monitoring	Surveillance	Power Imbalance, Loss of Trust, Loss of Autonomy, Loss of Liberty
Manageability			
Disclosure	Authorized Attribute Sharing	Distortion	Stigmatization, Power Imbalance, Loss of Liberty
Disposal	Normal Account Deletion	Unwarranted Restriction	Exclusion, Economic Loss, Loss of Trust
Confidentiality			
Use	Authorized Use	Appropriation	Loss of Trust, Economic Loss, Power Imbalance
Retention	Secure Storage	Insecurity	Economic Loss, Stigmatization

Multi-Disciplinary Approach

- 1443014: CIF21 DIBBs: An Integrated System for Public/Private Access to Large-Scale, Confidential Social Science Data
  - This research project develops a pilot of an integrated system for disseminating large-scale data about people.
  - This project addresses critical challenges that have inhibited the wide-spread dissemination of large-scale databases. Among the challenges the dissemination of these data have posed are the unintended disclosures of data subjects' identities and sensitive attributes, thereby violating promises and sometimes laws designed to protect data subjects' privacy and confidentiality.
  - The project will lay the groundwork for a potential transformation in data dissemination, providing data stewards with the infrastructure they need to release data products that advance social science, policy making, and training.

For Official Use Only

NSF Privacy Research Portfolio

- In August 2013 and in February 2014, the White House Office of Science and Technology Policy (OSTP) issued two Requests For Information (RFI) on privacy research activities pursued by the agencies
- NSF: Approximately $25M per year is invested in privacy research activities
  - Approximately 35% of the FY13 SaTC program
  - Ranging from $100K/2 years to $10M/5 years

http://www.nitrd.gov/Pubs/Report_on_Privacy_Research_within_NITRD.pdf

Research on Privacy in Today's Networked World

Privacy DCL (NSF 14-021)

- Social and psychological functions privacy performs
- Relationship between technical and psychological conceptualizations of privacy
- Psychological or social consequences of privacy violations
- Factors encouraging attention or inattention to privacy
- Institutional adaptations to and utilization or avoidance of privacy-invading technologies
- Understanding the contextual nature of privacy or privacy violations
- Examining privacy issues and impacts across different levels of analysis with different technologies

FY14 Privacy Awards

PropID	L/NL	PI	Institution	Title	total funded
1408730		Kapadia, Apu	Indiana Univ	TWC SBE: Medium: Collaborative: A Socio-Technical Approach to Privacy in a Camera-Rich World	$1,200,000
1447589		Jason Dedrick	Syracuse University	Data Privacy for Smart Meter Data: A Scenario-Based Study	$266,101
1450625		Katherine Shilton	University of Maryland Colleg	EAGER: Privacy in Citizen Science: An Emerging Concern for Research and Practice	$181,047
1450600		Patrick Kelley	University of New Mexico	Privacy's Sociocultural Divide across American Youth	$242,832
1450091	L	Yingying Chen	Stevens Institute of Tech	EAGER: Collaborative Research: Towards Understanding Smartphone User Privacy: Implication, Derivation, and Protection	$140,048
1449958	NL	Jerry Cheng	Rutgers, The State University of New Jersey	EAGER: Collaborative Research: Towards Understanding Smartphone User Privacy: Implication, Derivation, and Protection	$140,000
1449446		Keith James	Portland State University	RAPID: Cyberdata Privacy Concerns by Organizations at the Canada-U.S. Border	$68,286
1449619		Paul Pavlou	Temple University	PRIVACY IN AN ERA OF BIG DATA	$49,999
1423629		Alfred Kobsa	University of California-Irvine	TWC SBE: TTP Option: Small: A User-Tailored Approach to Privacy Decision Support	$666,343
1422501		Jaideep Vaidya	Rutgers University Newark	TWC: Small: Privacy Preserving Outlier Detection and Recognition	$508,474