P3 - Insuring Privacy and Proprietary in Big Data for Population Informatics

Hye-Chung Kum (Texas A&M University)
Ari Kahn (UT Austin, TACC)
Kristin Jenkins (DFWHC)
Charles Schmitt (UNC-CH, RENCI)

General Information at South BD Hub (Click Here)

General Information about NSF National Efforts in Big Data (Click Here)

NSF Spoke CFP

MISSION

The ***social genome*** is the collective footprints of our society captured in ever-larger and ever-more complex databases (e.g., government administrative data, EHR data, etc) about people in the digital society. ***Population informatics*** applies data science to social genome data to answer fundamental questions about human society and population health much like bioinformatics applies data science to human genome data to answer questions about individual health. It is an emerging research area at the intersection of SBEH (Social, Behavioral, Economic, & Health) sciences, computer science, and statistics in which quantitative methods and computational tools are used to answer fundamental questions about our society. Population informatics has emerged much more slowly than other computational fields in part due to the lack of the data infrastructure required. Data-driven studies of our society require a well-integrated sustainable data infrastructure of data collected at different times of our lives (e.g. hospitals, schools, jobs) for a large population. This type of comprehensive data infrastructure can support broad SBEH research that cannot be answered in any other way. The transformative benefits to SBEH science are obvious (e.g., studies on the impact of policy in one domain on another domain such as the impact of policies in juvenile justice on the child welfare system, comparative effective studies, longitudinal studies); as are the potential violations of privacy, both for people and organizations, from a poorly designed system. This proposal addresses the challenging privacy and proprietary issues that are inherent when working with person level big data by developing a ***virtual social genome data library*** that can support safe analysis for population informatics. We propose to evaluate the design by prototyping a model and developing a business model to understand the sustainability in collaboration with the existing NSF funded cyberinfrastructure high performance computing (HPC) community and data custodians of valuable data assets to support diverse SBEH research.

CORE TEAM

Hye-Chung Kum, PhD, MSW (Texas A&M University-TAMU, SPH & TEES): Project lead. Responsible for coordinating and supporting all required activities including participating in the design of the cyberinfrastructure, co-leading the data governance effort, leading the outreach activities working with RENCI, and acting as the data custodian for the Texas state data asset use case.

In addition, Texas A&M University will develop an ethics panel composed of national experts in human subject research using big data where obtaining informed consent is impracticable. All use cases will have to submit an IRB-like application to the ethics panel, who will evaluate the risk of privacy violation in the research taking into account the data, the software/algorithm required, the computing infrastructure used, and who needs access to what data including what data is released to the public (e.g., in publications). We will minimize the risk by assigning the required secure computing for a given use case. The panel will provide a written risk assessment to each project that they can submit to their IRB at the home institution to facilitate better IRB processes for human subject research using big data (e.g., start to build a common vocabulary about privacy risk in data).

Ari Kahn, PhD (UT Austin, TACC): Cyberinfrastructure lead. Responsible for (1) providing secure computing and storage (HIPAA, FISMA moderate etc.) via virtualization and secure network access for all data assets and use cases, (2) developing a business model for such computing by measuring required metrics such as cost, storage need etc., and (3) replicable best practice guidelines for how to run such a secure data center at HPC centers.

Kristin Jenkins, JD (DFWHC): Data governance and regulation co-lead. Together with Dr. Kum, they will (1) assist each data custodian to stand up an efficient data governance model for the given data asset and automate much of the data governance process, so that human review and approval can be focused on the steps that require human review, (2) review legal documents such as DUA and MOU and develop common template documents.

Charles Schmitt (UNC-CH, RENCI): Facilitate interaction with the Hub at RENCI. RENCI has experience running testbed secure computing environments for population informatics and will participate in transferring and hardening the technology for full operation at TACC. Also, RENCI will work with Odum Institute and CTSA for outreach and education to the broad SBEH science community.

Three ways to participate in the spoke

For those interested in participating in this P^3 Big Data Spoke, there are multiple levels of participation based on your interest, expertise, and existing access to data assets about people. Below we list the different ways to participate.

ONE, Test Drive the Pilot System: Data Custodians or Delegates

For those of you who meet the following criteria, we invite you to sign up to pilot the system in phase 2 and 3, as you are further along in the data pipeline than many researchers and in a position to evaluate the system.

1. You have access to a population data asset, and would like to facilitate wider access (e.g. collaborators at other institutions), AND
2. You have an interesting domain research question you want to answer using the data asset in the pilot system. (If you don't have a data scientist who is the user, we can discuss how our team might match you up with someone. We can also help you define the domain research question depending on the domain.)

Phase 1 (Year 1) Data custodians who have signed on to be use cases for **developing the pilot system**. We will provide access to the secure data infrastructure.

1. Texas State Health and Human Services Commission (Cecilia Cazaban, UT Houston & Hye-chung Kum, TAMU)
2. Dallas Fort Worth Hospital Association (Kristin Jenkins, DFWHC)
3. Texas Department of Transportation (Eva Shipp, TTI at TAMU)
4. Coastal Atlas: regional population data (Walter Peacock at TAMU)

Phase 2 (Year 2) Data custodians who have signed on to **test the pilot system**. We will provide access to the secure data infrastructure.

1. North Carolina, Division of Social Services (Dean F Duncan, UNC-CH)
2. PHAST (Public Health Activities & Services Tracking): five statewide local Health Department data asset (WA, FL, NY, OH, MN) (Betty Bekemeier, Univ. of Washington)
3. Washington State Association of Local Health Department Directors (Betty Bekemeier, Univ. of Washington)
4. Four National data assets: Approved data users will submit an application to the corresponding agencies to obtain approval to take the data to the pilot system for research. This will allow us to evaluate if the pilot system will meet the requirements of these agencies or not, and if not, why not.

    a. DOD Tricare data (Todd Leroux, Uniformed Services Univ. of the Health Sciences)
    b. VA data (David Gutman, Emory)
    c. AHRQ (Agency for Healthcare Research and Quality) Data (Alva Ferdinand, TAMU)
    d. SEER-Medicare (NIH NCI) Data (Hye-Chung Kum, TAMU)
5. UAB: Medicaid data, NIST data (Meredith L Kilgore)

Phase 3 (Year 3) Data custodians who have signed on to **evaluate the replicability of the system on their local HPC cyberinfrastructure** in the final phase. We will provide technical assistance.

1. South Carolina (Kevin McKenzie, Clemson)
2. Virginia (James Harrison, UVA)
3. Maryland (Rick Barth, UMB)
4. Georgia (TENTATIVE discussions)

TWO, Join the Expert Ethics Panel on Human Subject Research using Big Data:

For those of you who have expertise in evaluating the privacy risk of human subject research using big data (e.g. access control, FISMA/HIPAA compliant computing, training, deidentification, pseudonymization, disclosure limitation methods, privacy enhanced algorithms including record linkage), we are recruiting experts to serve on the ethics panel to develop the application and review use cases to provide a written evaluation of the risk. ***We especially encourage teams pilot testing the system to have someone serve on the board in the spirit of peer review.***

We will collaborate with the Hub to make the ethics panel accessible for other projects (i.e., those projects that are not using our pilot system) who would like to submit an application to obtain a written evaluation of the privacy risk from the expert panel.

Current Status of Human Subject Research using Big Data:

In terms of human subject protection in big data research and IRB, there is still much to work through with national efforts by the National Research Council (NRC), IOM, and US-DHHS. In the 2014 report by NRC, “Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences”, there are even recommendations for making a new category for human subject research using only existing data as “excused” research when proper measures are taken to reduce the risk of harm from disclosure to minimum levels. These and other major changes to the Common Rule have been proposed as a Notice of Proposed Rulemaking (NPRM) in

September 2015. The public commenting period is now over, and it remains to be seen whether this proposal will become part of the revised Common Rule. Such changing landscape in human subject research are difficult to maneuver as a researcher using big data about people as well as those responsible for the infrastructure to support such research. The P3 spoke will build a peer support network to facilitate these activities.

source:

http://www.hhs.gov/ohrp/humansubjects/regulations/nprm2015summary.html

http://bdes.datasociety.net/council-output/human-subjects-protections-and-big-data-open-questions-and-changing-landscapes/

In the P3 project, we will focus on building capacity in local IRBs at each institution to keep up with the cutting edge changing regulations and better assess the risk of information disclosure and harm for research involving big data about people. These efforts will ultimately support better decisions made by IRBs based on proper risk assessments and understanding of the Common Rule, whatever they may be. Risk assessment for big data about people is a complex task requiring expertise in computer security, information privacy and disclosure, as well as social implications of disclosures that result in harm. Many institutional IRBs do not have sufficient expertise in these areas to do a good risk assessment.

THREE, Become a member: Interested in population informatics, but have not obtained access to real data about people yet

Participate by giving us constructive input. In collaboration with the Hub, we expect to host a workshop at end of year 2, to share the best practice models for (1) cyberinfrastructure, (2) data governance, and (3) IRB processes and seek input. We invite you to sign up with the listserv at the south BD Hub, and look out for call for participation to the workshop. We plan to make all materials available online and gather as much input as possible, and incorporate it into the final best practice guidelines in year 3.

Participate by giving us encouragement and positive feedback. Become a member of the spoke and monitor our progress. We will maintain a mailing list for members who are interested in this work to share our successes and failures. Please join the mailing list, monitor our progress, and let us know if you think our work is moving in a direction to meet, or not meet, your domain need for privacy and data science. This will give us the best chance of having real impact in your domain. Tackling privacy is like having a war with an invisible monster: we will need all the encouragement and guidance after the inevitable battles we will lose to ultimately win the war!

You are also welcome to contact us when you have obtained access to a population data and would like to test drive the system. We will do our best to accommodate all requests within the limits of the resources available.

Desired Outcomes: Three year immediate goals

1. [TACC, TAMU, UNC-CH] Pilot an operational system and harden the technology for secure computing via virtualization
2. [TACC, TAMU] Develop a business model for operating a virtual social genome data library
3. [TACC, TAMU] Best practice guidelines to replicate the secure computing environment at a HPC center
4. [TAMU, DFWHC, TACC] Best practice guidelines for data governance models
5. [TAMU, DFWHC, TACC, UNC-CH] Best practice guidelines for IRB applications and risk assessment for population informatics with big data
6. [TAMU, DFWHC, UNC-CH] At least 10 data assets, with data custodians from multiple states and varied types of organizations (e.g. local agency, state agency, federal agency, non-profit, private company) and varied domain (e.g. health, city planning, coastal hazard, transportation), available on the pilot system with varying levels of access
7. [TACC, TAMU] Evaluate the replicability of the system at 2 other HPC sites
8. [TAMU, DFWHC, UNC-CH] 5 or more projects that involve students to provide opportunity for students to learn via real data science projects
9. [TAMU, DFWHC, UNC-CH] Make at least 3 connections on siloed data assets that were not previously connected (e.g., conduct on analysis using an integrated dataset from two or more sources that were previously difficult to combine)

Desired Outcomes: Long term goals

1. Develop a business model to sustain the pilot system to provide the service to the research community
2. Maintain and disseminate updates to guidelines, both security and regulatory, to keep the systems current
3. Provide technical assistance to other HPC centers who want to replicate the system
4. Provide the secure platform to the research community, connect SBEH scientists to HPC, and build a data ecosystem for population informatics research

5. Develop common tools/software (e.g. record linkage) critical to population informatics
6. Maintain an active student pool to learn via real data projects

Related Readings

- Our Vision paper in IEEE Computer: includes summary for other papers below
    - Kum, H.C., Krishnamurthy A., Machanavajjhala A., and Ahalt S. Social Genome: Putting Big Data to Work for Population Informatics. IEEE Computer Special Outlook Issue. Jan 2014
    - IOM. Beyond the HIPAA Privacy Rule:Enhancing Privacy, Improving Health Through Research: The National Academies Press; 2009.
- Data Access Models
    - Kum, H.C., and Ahalt, S. (2013). Privacy by Design: Understanding Data Access Models for Secondary Data, American Medical Informatics Association (AMIA) joint summits on translation science: clinical research informatics
    - RENCI White paper on VISR
- Data Ethics and Governance Models
    - Legal Framework: Contextual Integrity
        - http://bdes.datasociety.net/council-output/human-subjects-protections-and-big-data-open-questions-and-changing-landscapes/
        - Nissenbaum H. Privacy as Contextual Integrity. Washington Law Rev. 2004:79(1):19-158.
        - Barth, A.; Datta, A.; Mitchell, J.C.; Nissenbaum, H., "Privacy and contextual integrity: framework and applications," in *Security and Privacy, 2006 IEEE Symposium on* , vol., no., pp.15 pp.-198, 21-24 May 2006, doi: 10.1109/SP.2006.32
    - New funded work by PCORI
    - K. Davis and D. Patterson. Ethics of Big Data: Balancing Risk and Innovation. 2012. O'Reilly Meida.
    - Not directly related, but IRB for ICT (information, communication, and technology) research: Menlo Report. M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, The Menlo Report, IEEE Security & Privacy, vol. 10, no. 2, pp. 71--75, Mar 2012.
- Disclosure Limitation: What information is allowed to be taken out to public
    - S. Fienberg. Confidentiality, privacy and disclosure limitation, Encyclopedia of Social Measurement, Academic Press 2005;1:463-9.

- Privacy Preserving Record Linkage Tool to integrate the different datasets
    - Kum, H.-C., Krishnamurthy, A., Machanavajjhala, A., Reiter, M. K., & Ahalt, S. (2014). Privacy preserving interactive record linkage (PPIRL). Journal of the American Medical Informatics Association : JAMIA, 21(2), 212–220. http://doi.org/10.1136/amiajnl-2013-002165
- International Efforts
    - International Population Data Linkage Network (IPDLN)