Remote Access to Micro-data: Transparency in Building Evidence-based Policy

Julia Lane[1]

The new Administration promises to focus considerable attention on evidence based policy. Social science researchers have an unprecedented opportunity to respond to this national imperative given that advances in cyberinfrastructure have created a virtual deluge of new types of data ranging from new data on human interactions through digital imaging, sensors, and analytical instrumentation to new ways of collecting biological and geospatial information from survey respondents and to combining data from different sources, such as surveys and administrative records.

Other disciplines have developed institutions to use the new data collection and analysis capacity provided by cyberinfrastructure advances to respond to similarly pressing needs with great success. Biotechnologists acquired the human genome sequence and used new technologies and analytical methods to identify variations in human DNA that underlie particular diseases; the development of institutional infrastructures, such as the National Center for Biotechnology Information (NCBI), to promote access and analysis has been critical to this response.[2] *In response to the concerns with tsunamis, geoscientists advanced their modeling, mapping and assessment techniques by putting together a tsunami-related data archive*[3]*. Astronomers have developed national and international virtual data observatories of the sky*[4] *to better compare and combine data from different sources.*

Despite the potential recognized and realized by other disciplines, the set of options available to social scientists to access micro-data has remained fundamentally unchanged for decades. It is clear that traditional responses to providing access are unlikely to be sufficient to address the national imperative. Current approaches admit too great a loss of data utility, and too great a risk to confidentiality, to provide the evidence base necessary to guide policy.

A CONCEPTUAL FRAMEWORK FOR DATA ACCESS AND PRIVACY

The basic tension between data access and data confidentiality in the context of studying social science phenomena is well understood (Doyle et al. 2001). The core challenge is balancing the risk of reidentification with the utility associated with data analysis.

The risk[5] *from reidentifying individuals in a micro-dataset is intuitively obvious. Indeed, one way to formally measure the reidentification risk associated with a particular file is to measure the likelihood that a record can be matched to a master file (Winkler, 2005). If the data include direct identifiers, like names, social security numbers, establishment id numbers, the risk is obviously quite high. However, even access to close identifiers, such as physical addresses and IP addresses can be problematic. Such risk of re-identification has been increasing due to the increased public availability of identified data and rapid advances in the technology of linking files*[6]*.*

Access to micro-data generates utility in a number of dimensions (Lane, 2007). Clearly the more information that is provided and the more researchers that have access to the data, the greater the value of the analytical work that can be undertaken. In addition, the more transparent the access, the more likely it is that a body of knowledge will be developed around the dataset, expanding

---

[1] *This is drawn from Balancing Access To Health Data And Privacy: A Review Of The Issues And Approaches For The Future, with Claudia Schur. It does not necessarily reflect the views of the institution I represent*

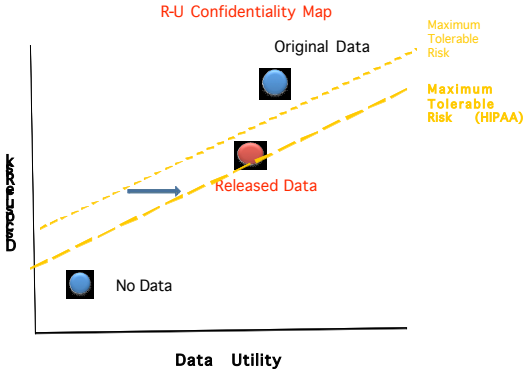
[2] *ncbi.nlm.nih.gov/dbgap*

[3] *http://nctr.pmel.noaa.gov/Dart/dart_home.html*

[4] *NVO: http://www.us-vo.org/; IVOA: http://www.ivoa.net/*

[6] *(http://www.fcsm.gov/working-papers/charlesday.pdf*

FIGURE 1



knowledge about the underlying data quality, the correct uses of the data, and the important data gaps. Finally, data access is essential to ensuring that analytical work is generalizable and replicable, which is the essence of scientific endeavor.

Figure 1 provides a graphical representation of this conceptual tradeoff between data risk and data utility. Here the dashed line identifies the maximum tolerable risk; the core guiding principle should be to generate released data that are as close to the frontier as possible [7] (Duncan et. Al, 2001)

NEW TYPES OF DATA AND NEW APPROACHES

The new demands for micro-data access result from more than the new Administration's emphasis on openness and transparency[8]. Social scientists in many areas of research recognize that new ways of collecting data mean that the traditional ways of providing access are inadequate. The new types of personally identifiable information that are being collected by means of sensors, video imaging, texts and bio-markers cannot be provided by means of public use files, licensing agreements are too insecure and risky, and research data center access is too slow, difficult and costly to be a generalizable solution. Social scientists are also beginning to recognize that the advent of large scale shared datasets in the physical and biological sciences has transformed those disciplines by building scientific communities that share and communicate knowledge. Similar technologies offer a corresponding potential to transform social science research in general and health services research in particular (Lazer et al., 2009)

Just as the new types of data could potentially transform the utility (and risk) associated with access to data on human beings, as indicated by the location of the "new data" element on the R-U map in Figure 2, new approaches to providing access have also evolved (as indicated by the "released data" element on the same map). These include trustworthy computing: models, logics, algorithms, and theories for analyzing and reasoning about all aspects of trustworthiness--reliability, security, privacy, and usability. Protecting databases against intruders has a long history in computer science (Dobkin et al. 1979). Computer scientists themselves are interested in protecting the confidentiality of the data on which they do research (for example, the Abilene Observatory supports the collection and dissemination of network data, such as IP addresses). Cyberinfrastructure advances have the potential to greatly expand the set of access modalities, particularly with

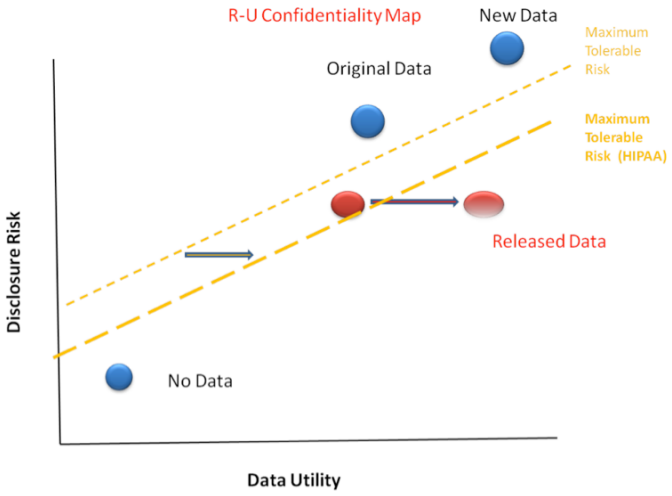


FIGURE 2

[7] *For a good practical implementation of this approach, see Duncan et al., 1004*

[8] *Evidence by data.gov and open.gov*

respect to remote access. The Trustworthy Computing initiative at NSF has created a research community that focuses on developing network computers that are more predictable and less vulnerable to attack and abuse, that are developed, configured, operated, and evaluated by a well-trained workforce, and that educate the public in the secure and ethical operation of such computers. The Department of Defense has developed different levels of web-based access ranging from unclassified (nipr-net) to secret (sipr-net) to top-secret (jwics-net)7 using off the shelf technology.

There are also scientific advances in ways to state, reason about, and resolve conflicts among privacy policies, and between privacy and security policies, particularly understanding the interplay between people and technology and the evaluation of trustworthiness. A good example of this is the PORTIA project which focuses on both the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners and data users. Finally, the recent NSF SBE/CISE workshop on cyberinfrastructure8 outlined a combined computer and social science research agenda for different approaches to access.

REMOTE ACCESS

Indeed, many national and international statistical agencies have moved towards secure remote access as a way to promote researcher access. These entities, often called "data enclaves" have a portfolio approach to protecting confidentiality. This approach combines statistical, technical, legal and operational controls at different levels chosen by the agencies to optimize the combination of confidentiality protection and data utility in their context.

The specific approach can be implemented within a secure data enclave that researchers can access remotely. All access should be in compliance with agency-specific and department-specific data sharing requirements and utilize best practices from the data sharing field as well as state-of-the-art information technologies and applications. The specific approach is implemented within a secure data enclave that researchers can access remotely. In addition, the enclave typically has utilities that permit data archiving, indexing and curation.

A typical data enclave[9] provides an information technology solution using a robust set of data access tools that facilitate high-quality researcher interaction with the data, while at the same time ensuring that data confidentiality is protected through a holistic suite of security and auditing measures. With the remote access mode, the data enclave provides external researchers with the ability to access the data in a controlled manner over the internet. Thus when a researcher needs to remotely access the data enclave's online resources, he/she first initiates an encrypted connection with the data enclave using virtual private network (VPN) technology. VPN technology enables the data enclave to prevent an outsider from reading the data transmitted between the researcher's computer and the enclave's network. Before the VPN connection can be completed, the user must provide a pre-defined user id and password. RSA Smart Card technology can also be used, so that the user must validate his/her identity in real time. Other components of the VPN technology allow the enclave to control which network resources the external researcher can access on the enclave's network. Finally, if it becomes a requirement, the data enclave can also restrict the users to accessing the data enclave from specific, pre-defined IP addresses. So, for example, the researcher would be able to use the remote access tool at work, but not at home or from overseas.

There are typically also statistical protections. Typically data enclaves protect every data set by constructing a set of unique identifiers that can substitute for variables that are explicit

---

[9] See, for example, www.norc.org/dataenclave

personal/organizational identifiers, such as name, address, phone number, Social Security Number and Taxpayer Identification Number. The data enclave is also able to limit researchers' access to the data they need for their specific research questions if necessary. To accomplish this, the data enclave can create custom analytic data files that contain a subset of the columns (and even rows) contained in the master data set.

The utility from such an approach is that the new cybertools could be used to provide an opportunity for health services researchers to develop new modes of analysis, such as virtual organizations that study social science data[10]. The opportunity is clear from the way in which ubiquitous information technologies has transformed many facets of human interaction and organization. Tools such as the Grid, MySpace, and Second Life have changed how people congregate, collaborate, and communicate. Increasingly, people operate within groups that are distributed in space and in time that are augmented with computational agents such as simulations, databases, and analytic services which interact with human participants and are integral to the operation of the organization.

The risk is limited because the enclave access modality relies on multiple approaches to reducing risk rather than one single "silver bullet". There typically legal protections, which can be used to reduce the likelihood of a deliberate breach; researchers are trained and instilled with a culture of confidentiality, to reduce the likelihood of an inadvertent breach; and technical procedures are put in place, through IT technologies, to reduce the likelihood of an external breach. Finally, organizational procedures are put in place, such as audit logs, trails and webcams to monitor behavior and act as a discipline device.

---

[10] is a group of individuals whose members and resources may be dispersed geographically, but who function as a coherent unit through the use of cyberinfrastructure. A virtual organization is typically supported by, and provides shared and often real-time access to, centralized or distributed resources, such as community-specific tools, applications, data, and sensors, and experimental operations.

REFERENCES

Dobkin, D., A. Jones, and R. Lipton, Secure Databases: Protection Against User Influence. ACM Transactions on Database Systems (TODS), 1979.

Doyle, P., J. Lane, J. Theeuwes and L. Zayatz eds. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. 2001, North-Holland: Amsterdam.

Duncan, G., S. KellerMcNulty, and L. Stokes, Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map. 2004, National Institute for Statistical Sciences.

Duncan, G., et al., Disclosure limitation methods and information loss for tabular data, in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, et al., Editors. 2001.

Lazer, D., et al., Computational Social Science. Science, 2009: p. 721-723.

Lane, J., Optimizing the Use of Microdata: An Overview of the Issues. Journal of Official Statistics, 2007. 23(3).

Winkler, W., Overview of Record Linkage and Current Research Directions, in U.S. Bureau of the Census, Statistical Research Division Report. 2005.