Engineering a Common Good: Fair Use of Aggregated, Anonymized Behavioral Data

Nathan Eagle
The Santa Fe Institute
1399 Hyde Park Rd, Santa Fe, NM USA
Email: nathan@santafe.edu

***Abstract*—This position paper makes the case that aggregate behavioral data generated by a society has the potential to be of significant public good. While this potential has been demonstrated in one-off academic research projects, to fully realize the benefits of the public good, we urgently need a set of standardized protocols for behavioral data acquisition and usage. These protocols must deal with the details associated with the de-identification process and deductive disclosure - problems analogously handled by the medical research community decades ago. While legislation may ultimately be required, it is our hope that the academic community can design a set of requirements may provide both companies and researchers guidance about how to proceed with data sharing in the near-term.**

I. The Behavioral Data Potential

Last year Google turned disease surveillance on its head, enabling the detection of influenza outbreaks ten times faster than previously possible by using the petabytes of search queries generated by millions of Americans[1]. However, while most people today have never used Google, the majority of the human race is currently carrying a mobile phone a device that generates massive amounts of movement and communication data. The unprecedented coverage and detail provided by these new types of data have enormous potential for informing social policy, particularly for issues of urban planning, economic development, and public health. During the recent outbreak of H1N1, for example, researchers attempted to access data from mobile phone operators in Mexico to gain insight into the human contact networks underlying the dynamics of influenza transmission. However, even the most forward-thinking operators during the peak of the global concern about swine flu were reluctant to provide access to anonymized mobile phone records to researchers (both employees and academics). The liability associated with sharing private data, even if anonymized and aggregated, was perceived to out-weigh the potential societal benefit.

The inadvertent generation of vast behavioral datasets is a fact of life in the 21st century, however, and provides extraordinary opportunities for improving the lives of people across the world. In particular, it is the underserved and understudied societies that have the most to gain from the study of societal-level data and with the majority of mobile phone subscribers living in the developing world, there is suddenly a massive amount of behavioral data to study. The data is being used by epidemiologists modeling human movement to support informed decisions about allocation of malaria eradication resources in Kenya. Developmental economists are attempting to quantify a society's reactions to exogenous events, such as the collapse of crop prices in local markets or the onset of severe drought. Bayesian anomaly detection algorithms developed within the machine learning community are now being implemented to identify behavioral signatures associated with outbreaks of cholera, enabling health officials within the developing world a sophisticated disease surveillance system with little additional cost. Regional communication data throughout the Dominican Republic is being studied to uncover patterns associated with the spread of HIV and regional contraception norms. Using data from every mobile phone in Rwanda over the last four years, the city planners of Kigali are able to quantify the dynamics of slums and the social impact of previous policy decisions ranging from road construction to the placement of latrines[1]. These projects represent the tip of the iceberg in terms of the utility of aggregate data in benefiting society, and they illustrate the importance of preventing corporations who own the data from withholding it from researchers. Addressing the serious privacy concerns associated with this type of research is paramount, however. Rigorous data-sharing protocols and appropriate legislation are urgently required to protect the privacy and rights of the individuals who are often unwittingly generating this wealth of behavioral data.

II. Privacy Concerns

The reluctance of mobile phone operators to collaborate with the academic community is understandable. Current data sharing and privacy protection measures are ad-hoc, so sharing mobile phone data carries a real risk of violating the privacy of customers. Two issues in particular are a cause for concern. First, even if data are anonymized, the nature of behavioral data is such that very few observations are required to deduce the identity of an individual. To address this problem of deductive disclosure, strict data sharing protocols must be developed, similar to those that have been used by the medical community

[1] References for these and related projects are on the Artificial Intelligence for Development (AI-D) site: http://AI-D.org

for decades. A second major concern is the inability of individuals to remove their data from these aggregate datasets. This second issue will likely require specific legislation about individual ownership of personal data.

A. Addressing The Issue of Deductive Disclosure

Medical research involves the analysis of the type of personal information about individuals that is in many cases more sensitive than behavioral data. The Add Health database, for example, contains intimately personal details of 90,000 human subjects ranging from sexual encounters to intellectual aptitude to genetic predisposition for disease[2]. As with behavioral data, very few - in this case as few as five - variables are necessary to identify an individual participant, and yet the data has been used by thousands of researchers working on health issues such as HIV and depression. Deductive disclosure is a widespread problem with many types of personal data sets, but one that has been overcome to some extent by strict data sharing protocols that ensure the data cannot be released to the general public, and that researchers are required to register before accessing the data. Comparable protocols will be needed to share behavioral data sets containing information susceptible to deductive disclosure, such as purchasing decisions, friendship networks or movement patterns. In their absence, it is likely that there will be blunders similar to the AOL Data Valdez, where the data released about personal search terms were sufficient to identify many individuals[3]. As the AOL disaster illustrates, regional aggregation is not adequate to prevent deductive disclosure of individual identities. As these datasets continue to grow, a similar blunder has the potential to violate the privacy of billions of people.

B. The Urgent Need for Standards

Despite these problems, we have found many companies are eager to share the anonymized behavioral data with researchers in effort to explore uses that may ultimately lead to alternative sources of revenue. We currently have collaborations with mobile phone operators in dozens of countries, and behavioral data from over 250 million people globally. We have found that each operator has asked for the data sharing agreement signed by other operators. This agreement has evolved over time into a set of protocols that may be more broadly applicable to the research community and minimize the risks of individual identification. Due to federal regulations, many corporations silo their data internally, making it difficult (and sometimes impossible) to gain access to it across different corporate divisions. Sharing this data with external researchers is also often a foreign concept that requires long discussions with the companies. In our experience, the negotiations for data sharing typically take between six months and one year to complete, a time span that clearly prevents a rapid response to scenarios like epidemic outbreaks of disease. With a universal set of protocols for sharing data with registered researchers, however, it may be possible to expedite data sharing between industry and academia without jeopardizing individual privacy.

First, the security of the data prior to anonymization must be ensured. It must be stored on a secure machine that does not have direct connection to the internet, making it exclusively accessible only within the research institute. Additionally, the data needs to undergo adequate encryption for further security. The research team that has access to the data needs to be defined and any changes to the personnel of the team needs to be approved by the company. With respect to lifespan, the data cannot exist indefinitely; typical contracts allow the researchers access to the data for between 12-18 months after the beginning of the agreement. At the end of the agreement, the data should be destroyed. These measures for the safe handling of sensitive data are a critical prerequisite to any additional data sharing protocols. As with the Add Health data, the university affiliated with the researchers is responsible for the enforcement of these protocols and will be held legally liable for any breach.

All public results and publications are required to be approved by the corporation. The company has a vested interest that output of the project is not damaging to them either competitively or with respect to PR. The details within these legal contracts are the first step at building a relationship between academia and industry. However the contracts also rely on several less tangible factors. A relationship between the individual researchers entrusted with the data is critical, for example. Additional factors include demonstrating precedence to assuage fears of liability and demonstrating the benefits to the company, typically in terms of positive PR.

Once the security of the original data has been established, the risks of deductive disclosure must then be minimized through rigorous anonymization techniques. Random hashing can provide significant protection by substituting each piece of indentifying information (name, phone number, social security number) with a randomly generated string. In the past this hashing technique was not computationally efficient for large datasets, and required the translation of the information into a hashed string, making it vulnerable to reverse engineering. It is now possible to acquire a computer with 128GB of RAM for under $10,000, however, which makes it possible to keep a hash table of billions of people in memory affordably, so no algorithm is needed, and the hashes can be truly random and therefore more secure. Even if an attacker is able to compromise the identity of a single individual in the network, the majority of a randomly hashed large-scale network is still not compromised.

However, no behavioral data is truly invulnerable to malicious attack. If the researchers decide to breach their legal contract, it will always be possible to violate an individual's privacy, irrespective of the hashing technique. There has been a significant amount of work demonstrating how personal data (movement, medical records, social networks) supposedly 'de-identified', is vulnerable to malicious attack by identifying behavioral signatures associated with a target individual, irrespective of that individual's hashed identifier[4], [5], [6].

III. Individual Ownership of Behavioral Data

A major difference between sensitive medical records and behavioral data is the ability of subjects to 'opt-out' of a medical research study. When faced with prospect of federal regulation, online advertisers are attempting to develop self-regulatory techniques, involving informed consent and yielding the control of the data to consumers. Many envision a future where companies provide substantial discounts to individuals consenting to the sale of their data to third parties, or its use in targeted marketing campaigns. Just as there is a market for data about a potential employee's medical condition, however, without appropriate legislation similar markets are forming for a prospective job applicant's purchasing, communication, or movement data, leading to scenarios in which less wealthy individuals could no longer afford their own privacy. Within the medical community there is a push for legislation enabling individuals to own their personal health records to prevent this type of exploitation. Similarly, there is also pressure for legislation on the ownership of personal behavioral data, providing individuals with the right to access and remove their data from corporate databases enabling them to 'opt-out' from any type of analysis[7].

Advocates of individual data ownership make a compelling case, and a vocal minority has suggested that this pervasive behavioral data should be categorically deleted. However, the issue is complex, and extremists are in danger of jeopardizing the potential of this data to benefit society. Although it is important not to understate these privacy concerns, hundreds of corporations will continue to store extremely personal data about our behavior without the mandate to use it for positive social change. In all likelihood, adequately protecting individual rights while harnessing behavioral data's inherent potential for positively impacting societies, and in particular societies in the developing world, will require legislation in addition to data-sharing protocols.

While the protocols above have been useful as a stop-gap, the current system for behavioral data study and sharing needs to be formalized. A first step towards ensuring the rights of the individual over his or her own data would be the adoption of licensing procedures similar to those found in the medical research industry. While random hashing is a first step at developing methods for studying individual-level data, this sort of data cannot be anonymized to adequately protect the individual identities of the individuals who generated it. As such, not only should the companies vet potential researchers who want access to the data, but also independent ethics organizations, similar to the NIH or the Better Business Bureau may be necessary to ensure the proposed data usage is appropriate. Before making the data public, for example, AOL should have been required to ensure the impossibility of deductive disclosure by only making counts of specific (non-identifiable) search terms available on a regional level.

Although leaving digital traces is a necessity of living in this century, companies should have no more of a right to sell this personal information without consent than hospitals do to sell patient medical records. While most people may consent to sharing most data with adequate compensation (lower service charges, extra features, etc), individuals should always have the legal right to opt out at any time. Once in place, both individual and societal-level data may be considered a form of intellectual property. The behavioral IP of an individual should be owned by that individual, and licensed to third-parties for a fee if desired. The behavioral IP of a society should be considered as a valuable public good. Existing fair use policies on the use of intellectual property may directly apply to these types of data. However, without the infrastructure connecting academic researchers to companies generating the data, this public good will never benefit the individuals who created it.

A society has the right to benefit from the data it generates. A vast amount of global data is continuously being collected that has the potential to dramatically improve millions of lives, and we urgently need to develop the protocols and regulations necessary to facilitate its appropriate and ethical use. Draconian restrictions on this data will not change the fact that it is being collected, and will do little to mitigate privacy advocates' concerns about privacy violations by government or industry. Instead, I believe that standardization of protocols and the development of legislation concerning individual ownership of data should be welcomed as a means to use behavioral data for positive global change. These measures will allow for the preservation of individual privacy while providing immense value to the community through the aggregation of personal data into societal-level statistics.

Acknowledgment

The authors would like to thank Yves-Alexandre de Montjoye, Jameson Toole and other members of the Santa Fe Institute for useful discussions that proved instrumental in the writing of this position paper.

References

[1] J. Ginsberg, M. Mohebbi, R. Patel, and L. Brammer, "Detecting influenza epidemics using search engine query data," *Nature*, 2008.

[2] J. Udry, "The national longitudinal study of adolescent health (Add Health), waves i & ii, 1994-1996," *Data Sets*, 1996.

[3] M. Lawless, "The third party doctrine reducx: Internet search records," *UCLA Journal of Law and Technology*, 2007.

[4] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography" *Proceedings of the WWW 07*, 2007.

[5] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," *WPES 05*, 2005.

[6] W. Xu, X. Zhou, and L. Li, "Inferring privacy information via social relations," *ICDEW 08*, 2008.

[7] A. Pentland, "Reality mining of mobile communications: Toward a new deal on data," *The Global Information Technology Report 2008-2009*, 2008.