

Characterizing Social Data Sets: Why So Hard to Share?

Jennifer Thom-Santelli & David Millen

IBM TJ Watson Research

1 Rogers Street

Cambridge, MA 02142

{jthomsa, david_r_millen}@us.ibm.com

ABSTRACT

Data sharing is an essential foundation of the scientific process for reasons of transparency and replication. In CSCW and HCI, large-scale studies of social interaction hold great promise for research. Unfortunately, to date, few data sets from the social web have been shared among the research community. To better understand why, we propose typology that characterizes datasets in order to describe what makes data sharing so difficult. We then discuss the possible implications of such categorization for the field of CSCW.

Author Keywords

Data sharing, social networks, social software

ACM Classification Keywords

H5.3. Information interfaces and presentation (e.g., HCI): Group and Organization Interfaces.

INTRODUCTION

As a larger proportion of our social interactions occur in networked mediated environments, there is enormous potential for gaining insight into human behavior through social science research. From a researcher perspective, this work could be best accomplished through open data sharing. From the perspective of the data providers, however, open data sharing presents a challenge for a variety of ethical, legal and technical reasons.

Data sharing is a fundamental tenet in the scientific process, particularly in fields where large datasets can be mined to answer a variety of research questions. Funding agencies, such as the NIH and the NSF in the United States, require grant proposals to include a plan for data sharing and dissemination [1,2]. Journals, such as Nature, post the datasets and protocols for published papers for reasons of transparency and replication [3]. This transparency has not necessarily translated into sharing datasets of a socially

networked nature, many of which are owned by entities whose primary goals are not necessarily research. As a result, it may be more difficult to for the field of CSCW to develop a cohesive area of work in this area. In the case of social media datasets, an interdisciplinary group of researchers speculate that the slower uptake of computational social science may be, in part, because of the difficulty in obtaining these datasets [16].

One socio-technical impediment to data sharing is the necessity to protect user privacy. Social data contains highly personal information, such as age, marital status, health conditions and financial holdings. When such datasets are improperly prepared for distribution, this information can be linked to users' real identities in a substantial breach of privacy, as in the case of the release and quick recall of AOL and Facebook datasets [4,23]. Even when providers sufficiently anonymize datasets, its networked properties also allow researchers to make inferences about user activities in a variety of contexts, such as one's sexual orientation [14].

While anonymization and privacy remain a serious technical and social challenge, we suggest, however, that there may be other underlying factors contributing to the difficulties in data sharing. In this essay, we present an initial typology to describe and characterize large social datasets in an attempt to move the discussion about data sharing beyond its current state. We then discuss some implications about what this might mean for CSCW design and research.

THE DIFFICULTY OF DATA SHARING IN THE SCIENCES

CSCW, HCI and computational social science are not the only disciplines to have wrestled with these challenges. Researchers in the physical and life sciences have had to deal with the difficulties of preparing large datasets for sharing with other researchers. [11] have suggested that the factors in hindering data sharing in laboratories are social in nature, particularly as self-collected datasets serve as currency that leads to increased reputation for scientists. This is, in part, because the quality of the dataset, in terms of how one collects it, is an important marker of status for the researchers. A survey of geneticists reveals the reasons behind declined requests for sharing datasets -- to ensure the best findings for the junior scholars who have collected the data [12].

Table 1. Data characteristics of various social media

	Scale (members)	Identity	Transparency
Wikipedia	10M+	Pseudonym	Public (content)
Facebook	300M+	Real name	Network/apps
LinkedIn	50M+	Real name	Network
Beehive (IBM-internal)	65K	Real name	Limited/firewall

We suggest that there are a few key differences between data sharing in the contexts of laboratories and in social software. In the case of social software, researchers are not necessarily concerned with creating their own datasets for reasons of scientific reputation. Instead, research success stems from the findings observed by naturalistic cases that are influenced by real-world usage. However, in many instances, these datasets are controlled and managed by third parties that are governed by a set of policies that are not necessarily aligned with the goals of researchers.

PROPOSING A TYPOLOGY FOR SOCIAL DATA

Social media datasets themselves are diverse, ranging from social networking sites to collaboratively edited wiki-based encyclopedias. Their characteristic social and technical features may influence how easy or difficult it is for user data to be shared for behavioral research. Table 1 describes how various social media systems would fall onto each proposed dimension, which we now describe in more detail.

Scale

The enormous popularity of many social media sites has resulted in social interaction data on an unprecedented scale. Popular web-based sites have membership in the hundreds of millions. In 2003, one of the first HCI research studies of Wikipedia use was completed using a Wikipedia corpus that contained 130,596 content pages [21]. Six years later, the currently reported size of the English Wikipedia is 3,077,725 [5].

The very large scale of social interaction observable on these sites makes them especially attractive for social scientists, who are interested in understanding these novel forms of technologically enabled human interaction. At the same time, the very size and rapid change of these social web applications presents challenges of data management and analysis for both the provider of the data and the research community.

Providers of large data sets need to schedule, process and host data sets, and provide associated data dictionaries and data set processing instructions. For example, the Wikimedia foundation provides web access to Wikipedia

data, as well as pages of instructions about data downloading (including sample code) and user assistance from various community forums [6]. To allow for researchers to make better sense of these large amounts of data, these types of research programs will continue to need more efficient data processing algorithms to speed up analysis, and novel forms of data visualization [13].

Identity disclosure by the user

Social media data also varies considerably on how online identity is managed, disclosed and revealed [17]. For instance, social web applications found in business settings require use of real-world names and are linked to corporate directories with other personal information such as email, phone, work address, etc. This traceability seems appropriate in a business setting and it has been argued that it helps to promote appropriate conduct and use (see, for example, [18]). Other social web applications, however, allow participants to create their own online identity, enabling anonymous or pseudonymous participation.

The varying degree of identity disclosure is important in any discussion of social media data sharing. Social web participants adjust online behavior based on personal assumptions about how traceable their online actions are. It is here that issues of anonymization are most important. As a result, the ease of data sharing is highly dependent on how tightly real-world identity is linked to information found within the dataset.

Transparency of user actions

Participants in social web applications are often explicitly informed about the types of data that is collected from them and who has access to collected data. The EULA (end-user license agreement), terms of use or the application privacy policy are the documents that spell the specifics out to the user. The kinds of data that are collected most commonly include the information contributed to the site (e.g., photos, comments, profile information), as well as login information and user action logs, including site navigation.

More complex are the various parties with which data can be shared. In some circumstances, online social behavior is viewable only to a restricted or “privileged” group of colleagues. Social applications in business settings are often behind a corporate firewall, which limits the audience to colleagues [13]. Some social sites require subscriptions for access, which also limits the scope of viewing somewhat.

Participants in social web applications make assumptions about the visibility (and discoverability) of their online actions. While most authoring actions are intended to be visible to some audience, the social “network” often mediates who can read what content. In social web applications, fellow participants are another audience that users must manage. Privacy policies then must state how data will be shared with fellow participants with some information (such as viewing behavior) never visible to other participants on the site. For example, the Facebook privacy policies (as of November 2009) indicate that

specific user behaviors such as profile and photo *views*, will not be shared with others [7].

In other cases, social web sites clearly state the degree of data transparency to others beyond fellow participants. For example, Facebook describes what data that has been collected will be shared with third-party application developers or legal entities. The Wikimedia Foundation also explicitly describes who is allowed access to user data in their privacy policy [8]. Included in this list are Wikipedians with access to various site operation positions (e.g., maintainers of Open Source Ticket Request System), Wikipedia foundation employees, and selected system developers.

Less often, there is a description of the data that will be shared with researchers and how it may be used. For example, in one of our research projects, the “terms of use” indicates

To support our research, we will be recording selected usage data, which we will analyze to understand which features work well, and which need more attention. We may publish reports based on your comments, ideas, and suggestions and the usage data. We will protect any personal information you provide us in accordance with IBM’s privacy practices.

The explicit nature of this contract is one such approach in a research-oriented application but it is less obvious how such language should be implemented in a more commercial site.

DISCUSSION

Our proposed typology is a first attempt to describe why and how social media datasets can be difficult to share. We now describe initial implications that emerged from this exercise.

Scale matters. What distinguishes empirical social science research on these datasets from other more micro-level approaches is the sheer size of these databases. Large datasets have particular hosting requirements, and secure data storage of sensitive information is not a trivial matter. As a result, those who have the capability to manage immense amounts of information are the ones who are best equipped to attempt this type of networked analysis.

This barrier to entry, however, may dissuade researchers from different disciplinary perspectives and institutional affiliations from this area of study, which may ultimately prove to limit the quality of work that is produced. First, collaborative visualization sites, like the tools available on Many Eyes [20], are necessary for researchers to make better sense of the large amounts of information available through social media datasets. Open data repositories, such as Dataverse [15], are a good first step in providing access and storage for user-contributed social science databases. However, motivating data owners to actually share is an ongoing issue for scientific fields, CSCW included.

Anonymity remains a stumbling block. Social media datasets contain a large amount of private information that

can identify many facets of user identity and behavior. This property of the data is precisely what makes the datasets so attractive to researchers who are interested in social interactions. If datasets are to be legally and ethically shared, they must be properly anonymized to protect the unauthorized access of user data. However, datasets thought to sufficiently mask the link between identity and behaviors have are easily upended through a variety of means, from decryption techniques to simple Google searches [10,23]. Until the issue of anonymization is settled, it will be extremely difficult to easily share datasets ethically without facing the wrath of the user, IRBs and the legal system.

Encourage granular user opt-in agreements. Our survey of a number of end user licensing agreements from popular social media sites suggests that users are informed that data is collected but not necessarily for research purposes. What if, however, the research process might be made transparent for users of these sites? Taking this a step further, what if users could be motivated to contribute to social research, similar to [22]’s notion of human computation? Such motivation to participate may appeal to a number of levels, such as a sense of altruism and/or that of individual benefit. If users were to opt into data collection in an informed manner on a wide scale, the burden of identity protection might shift somewhat.

It is important to note though that users may still require a high level of anonymity as a condition for participation in research. For instance, in the case of National Geographic’s Genographic Project, participants submit to DNA tests that reveal personal genetic information that is then aggregated but each individual contribution is anonymized [9]. That is not the surprise – what is notable is the upfront nature of the contract that the researchers have presented to the user.

CONCLUSION

We have presented observations on why data sharing remains difficult but we would like to conclude by emphasizing why it is vitally important to the field.

First, to echo [16], empirical research of a social nature on these large datasets allows us to monitor collective action. This may lead to real societal benefit, such as disease prevention or environmental action, through better knowledge of how phenomena occur in a networked environment. We want to further extend this argument to suggest that CSCW researchers should become part of the conversation by better articulating how our research might be able to answer these large questions. If users and dataset managers perceive CSCW research as high value, we may be able to influence how end-user opt-in agreements are crafted such that users might be able to consent to data collection for research’s sake.

Second, we want to encourage researchers in the field of CSCW to continue towards open data sharing for the sake of scientific breakthrough. The potential for

groundbreaking insight into human behavior through empirical analysis of these datasets is enormous, particularly if researchers from a diversity of perspectives and affiliations can get into the game. We note that many of our colleagues have acknowledged a debt to Elinor Ostrom's [19] work. We then propose that research in the field of CSCW can potentially have the same impact, particularly if we can provide transparency and support so that those who aim for Nobel Prize-caliber work can achieve it.

REFERENCES

1. NIH Data Sharing Information - Main Page. http://grants.nih.gov/grants/policy/data_sharing/.
2. NSF Policy from Grant General Conditions. <http://www.nsf.gov/pubs/gc1/jan09.pdf>.
3. Availability of data & materials : authors & referees @ npg. http://www.nature.com/authors/editorial_policies/availability.html.
4. AOL Proudly Releases Massive Amounts of Private Data. <http://www.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/>.
5. Wikipedia:Size of Wikipedia - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.
6. Wikipedia:Database download - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Wikipedia_database#Latest_complete_dump_of_english_wikipedia.
7. Facebook | News Feed and Wall Privacy. <http://www.facebook.com/privacy/?view=feeds>.
8. Privacy policy - Wikimedia Foundation. http://wikimediafoundation.org/wiki/Privacy_policy.
9. The Genographic Project - Human Migration, Population Genetics, Maps, DNA - National Geographic. <https://genographic.nationalgeographic.com/genographic/index.html>.
10. Backstrom, L., Dwork, C., and Kleinberg, J. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. *Proc.WWW2007*, ACM Press.
11. Birnholtz, J.P. and Bietz, M.J. Data at work: supporting sharing in science and engineering. *Proc.GROUP03*, ACM Press, 339-348.
12. Blumenthal, D., Campbell, E.G., Gokhale, M., et al. Data withholding in genetics and the other life sciences: prevalences and predictors. *Academic Medicine: Journal of the Association of American Medical Colleges* 81, 2 (2006), 137-145.
13. van Ham, F., Schulz, H., and Dimicco, J. Honeycomb: Visual Analysis of Large Scale Social Networks. *Proc.INTERACT 2009*.
14. Johnson, C.Y. Project 'Gaydar'. *The Boston Globe*, 2009. http://www.boston.com/bostonglobe/ideas/articles/2009/09/20/project_gaydar_an_mit_experiment_raises_new_questions_about_online_privacy/?page=full.
15. King, G. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods & Research* 36, 2 (2007), 173.
16. Lazer, D., Pentland, A., Adamic, L., et al. SOCIAL SCIENCE: Computational Social Science. *Science* 323, 5915 (2009), 721-723.
17. Marx, G.T. What's in a Name? Some Reflections on the Sociology of Anonymity. *The Information Society* 15, 2 (1999), 99-112.
18. Millen, D.R., Feinberg, J., and Kerr, B. Dogear: Social bookmarking in the enterprise. *Proc.CHI2006*, ACM Press, 111-120.
19. Ostrom, E. *Governing the commons: The evolution of institutions for collective action*. Cambridge Univ Pr, 1991.
20. Viegas, F.B., Wattenberg, M., Van Ham, F., Kriss, J., and McKeon, M. Many Eyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1121.
21. Viégas, F.B., Wattenberg, M., and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. *Proc.CHI2004*, 575-582.
22. Von Ahn, L. Games with a purpose. *Computer* 39, 6 (2006), 92-94.
23. Zimmer, M. 'But the Data is Already Public': On the Ethics of Research in Facebook. Presented at *Association of Internet Researchers Annual Conference 2009*.

Appendix 1 – Case Study

Several large-scale social software projects have been undertaken in recent years by IBM Research, including the *beehive social network (now known as SocialBlue)* [1], *dogear social bookmarking service* [2], and *cattail social file sharing* [3]. These applications have several common characteristics: 1) required authentication using a personally identifiable intranet ID, 2) instrumentation to monitor most important user actions, including create, read/view, update and delete, 3) articulated personal networks (i.e., friend relationships), 4) a combination of personal (private) and public information, and 5) application wide full text search.

While the “terms of use” vary somewhat across the three applications, the terms for the *beehive (SocialBlue)* application are generally representative. Included in the terms of use for *beehive* is the following:

To support our research, we will be recording selected usage data, which we will analyze to understand which features work well, and which need more attention. We may publish reports based on your comments, ideas, and suggestions and the usage data. We will protect any personal information you provide us in accordance with IBM’s privacy practices.

Also included in the terms is an agreement to comply with the IBM “Code of Conduct,” the “IBM Social Computing Guidelines,” and US Export Regulations.

The research teams have been approached by other researchers from both within and outside of IBM for access to the content of the applications and the data that has been collected. In most cases, the release of the data has been shared with other IBM researchers, with a Document of Understanding (DOU) created for each request.

Questions for discussion:

1. *What are the key topics to be covered in the DOU for internal (within IBM) data sharing?*
2. *Should there be restrictions on use? Are there limits to what data should be shared?*
3. *What level of research “transparency” is desirable (i.e., what should be communicated to the users of the applications?)*
4. *What should be done the same or differently for data sharing with researchers outside of IBM?*

References

1. Millen, D. R., Feinberg, J., and Kerr, B. 2006. Dogear: Social bookmarking in the enterprise. In Proceedings of *SIGCHI* (Montréal, Québec, Canada, April 22 - 27, 2006).
2. DiMicco, J., Millen, D. R., Geyer, W., Dugan, C., Brownholtz, B., and Muller, M. 2008. Motivations for social networking at work. *Proceedings of the ACM 2008 Conference on CSCW* (San Diego, CA, USA, November 08 - 12, 2008).
3. Muller, M., Millen, D., & Feinberg, J. Information Curators in an Enterprise File-Sharing Service. *ECSCW '09*. Vienna, Austria. September 7-11, 2009.