Sharing Social Accounting Metadata - Lessons from Netscan

Marc A. Smith[1], Tammara Turner[1], Eric Gleave[2]

[1]Microsoft Research, One Microsoft Way, Redmond, Washington, 98052

[2]University of Washington, Seattle, Washington

masmith@microsoft.com

Abstract. The creation and distribution of data sets is a core practice of the natural sciences. The combined shift of much of social life into computer-mediated channels of communication along with the growing proliferation of mobile devices capable of generating yet more detailed data about social life and behavior beyond the keyboard and mouse is likely to move the social sciences closer to the data centered practices of the natural sciences. These data collection opportunities promise a sea change in the availability and quality of social science data. In the spirit of this trend, the Netscan system has collected and generated descriptions of the social activity within online discussion environments like Usenet newsgroups, web boards, and communities. In support of a growing community of scholarship on computer-mediated interaction, this data set was licensed to a growing collection of universities for research purposes. In the following, the nature of the data, issues in its distribution, and patterns in its use are reviewed.

Introduction

We are on the edge of a major inflection in the social sciences created by the dramatic increase in data about social behavior generated by the proliferation of computationally mediated social interaction. When collective action takes place through a computational medium it leaves behind large scale collections of traces of human behavior in a machine readable format. The opportunity is to collect and mine these streams of data to generate maps and models of collective behavior at a level of fidelity and detail that has yet to be achieved. Challenges, however, must be overcome in terms of accessing, collecting, processing, and analyzing these datasets in ways that comply with essential considerations of privacy and ethics (Enyon et al., forthcoming). Sharing these data among a community of scholarship can help realize these opportunities, overcome these obstacles, and support the process of peer review and re-evaluation of results.

The Netscan project collects large scale data from online communities composed of threaded conversations. Netscan collects messages from a stream of content contributed to Usenet newsgroups along with a more modest collection of content from web boards and discussion spaces from non-NNTP sources. These collections are parsed into "social accounting metadata" catalogs that contain information about every message, message author, message thread, and newsgroup (Smith, 2007). When aggregated, the resulting catalog offers insights

into the patterns, dynamics, and range of variation of individual and collective structures that appear within these social environments.

The aspiration is to model the social study of computer-mediated social activity after disciplines like the meteorological sciences which have been significantly enhanced by the combination of exponentially increasing amounts of data, computational power and successively refined models. Much as the advent of space-based sensing revolutionized weather models, providing, for the first time, a global and holistic view of the phenomena under study, computer-mediated collective action promises to deliver data sets about human collective behavior of similar fidelity, scope and duration.

Starting in 1998, the Netscan project has collected in excess of 2 billion messages from more than 64 million unique author identities drawn from more than 175,000 newsgroups and web boards. The resulting data sets occupy several terabytes in volume and demand even larger working space and resources to manufacture, store, and serve.

The Netscan data process is built around the headers of each message, not its contents. Headers contain information about the date and time the message was sent, the newsgroup or newsgroups to which it was sent, the user name and email address provided by the author, the subject line, and message length. This information is extracted and inserted into a database where they are further aggregated into higher-level models of the activity of each individual and community. The message body itself, while collected and stored for future research, is not yet analyzed in our project nor are message bodies included in our data set.

While we have worked with the data for our own research goals and interests, we have recognized the common interest and value in sharing the data with other scholars. Since 1998, the Netscan data set has been available through a web interface (http://netscan.research.microsoft.com). This web interface provides reports on every newsgroup, author and thread for which data has been collected. The resulting data has been accessed by several hundred thousand people. Often, users of the Netscan system are participants in the communities about which they seek information. This use of Netscan provides support for user tasks like content discovery, selection, and evaluation. It has also been documented acting as a form of reputation system, providing an incentive for some users to make better and more contributions to the collective good (Gleave and Smith, 2007). Users of the Netscan data also include scholars researching online social phenomena. These users may also access Netscan data through the web interface which supports the copying of data published on the site into analysis tools like spreadsheets and statistical analysis packages. The web interface makes the exploration of the Netscan Usenet community data set a simple process of navigating a web site. Analysis of modest numbers of newsgroups, authors or discussion can be easily accomplished via this interface.

Starting in 2003, the project team took the additional efforts to make large subsections of the Netscan data sets available to scholars via a data license program. Under the license, universities and academics request a local copy of a large sample drawn from the complete Netscan data set. Roughly 2% of the complete data set was randomly selected and copied to a set of database tables. Once licensed, each participating university was shipped a copy of the Netscan data sample on a hard disk to avoid bandwidth constraints that make Internet transfers of such large datasets impractical. Once received, institutions copied the data sets to their own database facilities and returned the drive to us in order to have it resent to other licensees.

The Netscan data licensing program is an illustration of the ways the social sciences can be altered by the availability of data generated by computer-mediated collective action. The resulting data sets are more detailed, higher fidelity, and more "natural" than previous social science data sets which have been largely dependent on self-reported survey or interview data sets. Once in machine readable form, data sets can now be shared among a research community, providing a point of shared focus for comparison of results and methods. The data sets studied by each scholar can be larger and more detailed because the costs of data collection and refinement are not borne by each researcher.

Our experiences with the creation and management of the Netscan data licensing program have allowed us to see some of the challenges to doing e-social science as well as its promise. While the process of cultivating a community of empirical scholarship on computer-mediated collective action has begun, and many challenges have been overcome, more must be resolved in order for the full realization of the research value of these approaches.

The single largest issue facing efforts to share data from the process of computer-mediated collective action are the costs and challenges of collecting and creating the data itself. The hardware, bandwidth and software development expenses associated with the collection, aggregation, publication, and updating of a greater than one million message a day flow of content are not trivial. While these expenses are dropping, they remain significant and exceed many organizations resources. This reason alone highlights the importance of sharing such data sets when they are actually realized.

Beyond the costs of data collection and creation are the range of skills and resources needed just to receive and deploy the Netscan data set. Our data is shipped in tables readable by the widely used Microsoft SQL Server database product. While used in many commercial organizations, many social scientists and social science departments lack the infrastructure and technical skills to immediately deploy the Netscan data set. It has been necessary to seed the research community with appropriate skills. Our group has needed to offer guidance in the deployment and utilization of the SQL server product and has supported the development of the basic skills need to use it. A major point for social scientists orienting to this domain of research will be to ensure that graduate training expands or refocuses to ensure the skills needed for the manipulation and management of large scale databases are more widely available.

While technical resources and skills are significant issues, a larger challenge facing the development of this research area is the highly variable process of Human Subjects, IRB, and ethical review board certifications of data for use in academic research. These bodies often focus heavily on the issues of anonymization of the people who created the captured data. The Netscan data sample was additionally processed to bring it into compliance with the requirements set by these institutions. Most notably, all data containing the identification of each user is removed and replaced with a unique identifier that is unrelated to the original user's identifier. This major change to the data set retains its research value while bringing it into compliance.

Redaction may not provide a sufficient level of anonymization, however. Given the existence of public search indexes which contain almost every message contributed to Usenet newsgroups and many web boards, it may be impractical to sever all connections between a piece of data drawn from a computational social space and the identifier of its author. Taking even a fragment of a message, even one with the author's identification removed, for example a few words of the subject line or the date, time and the newsgroup to which the message was

posted, is often sufficient to retrieve the original message, along with its original author identifier. While this second step imposes additional effort in order to pierce the veil of anonymity, such costs remain relatively minor and are amenable to automation.

While these author identifiers are not identical to legal names or other forms of identity that are more tightly linked to physical humans, they can remain sensitive and their owners may value confidentiality and a limited exposure. Whether these expectations remain practical in a changing information landscape, they remain important values that should be respected.

Results

The availability of the Netscan data set has led to a number of research projects, course ware, and an emerging number of publications. Netscan Data Licensees have come from a variety of disciplines and University departments. Human Computer Interaction, Sociology, Internet Studies, Engineering and Technical Communication, Communications, and Economics departments have all licensed the data set. In many cases licensees have generated research, teaching, and publications making use of the data set and related tools. For many scholars, the Netscan data set is an otherwise inaccessible trove of valuable data. Once enabled by the data set, many scholars are able to investigate a range of phenomena of relevance to a diverse number of disciplines.

Social psychologists interested in group processes have seen value in the analysis of Netscan Usenet data. Robert Kraut and his research team at Carnegie Mellon University have published studies based on the Netscan data sample that explore the conditions under which questions and requests raised in newsgroup communities are actually answered and responded to. Kraut et all analyzed Netscan data set to find all authors in several medical support discussion newsgroups who either posted a message that did or did not receive a reply. Integrating the Netscan data set with the Usenet content stored in the Groups.Google system allowed the research team to identify the attributes of messages posted by first time contributors that correlated with the receipt of a reply. They found that messages that make claims about the length of the author's prior participation and the author's substantive interest in the topic discussed were indications of receiving a reply (Burke et al. 2007).

Work further afield from social psychology has leveraged the Netscan data set to explore methods for improving the quality of information retrieval and search systems over catalogs of community content. Search engine researchers Wen Xi and Eric Brill explored the value of social accounting metadata drawn from Netscan analysis of Usenet to assess if search results improved when data about the message author's history of message creation activity was considered (Xi, Lind, and Brill, 2004). The authors found that the fitness of search results returned by a ranking algorithm that contained information about the style of authorship were significantly more fit than those that ignored these features.

Curricular development is another use of Netscan metadata. Jenny Preece at the University of Maryland has integrated the use of Netscan data and reporting tools into a class on the social studies of online communities. Students make use of Netscan data to track the activity of a newsgroup over time and contrast it to itself and other newsgroups in order to grasp the range of variation of online spaces (Preece and Diker, 2005).

Graduate student projects and papers have also emerged from the use of Netscan data. Daphne Raban at the University of Haifa in Israel has made use of Netscan data to support

her dissertation research into the ways many dimensions of computer-mediated user activity have a power law distribution (Raban and Rabin 2007).

At the University of Washington, Blaine Robbins' research into the self-governance of Ultimate Frisbee makes extensive use of participant observation data coupled with Netscan data drawn from the *rec.sports.disc* newsgroup, the online gathering point for many participants in the sport. As a relatively new sport, Ultimate Frisbee competitions are self-refereed affairs which can lead to disputes and disagreements about normative behavior. Once processed by the Netscan system, data from the Usenet newsgroup devoted to the discussion of the sport has been used to document these conflicts and highlight the critical role the newsgroup plays in minimizing conflict and cultivating consensus in the real world.

Discussion

Sharing the Netscan data set has made it far more valuable than if it had been hoarded. Making the data widely available for academic research has broadened the amount of work on the data set while increasing the diversity of approaches brought to the task. This raises the further prospect of inter-disciplinary work as these otherwise diverse studies become comparable based on their share focus on a common data set.

Many licensees asked us for additional data. In some cases this was merely additional data of the same form and type originally provided but with a focus on specific time periods or topical areas of Usenet discussions that were research relevant, such as political discussions or discussions in a particular language. These requests have been relatively easy to fill and raise few additional issues. Other requests, for example for the content of the messages from which our data was drawn, are also fairly common. However, we have yet to provide message content for a number of reasons. First, message content is likely to contain additional individually identifiable information which would require more sophisticated redaction methods as all proper names, phone numbers, addresses and related identifiers must be recognized and removed or obscured. Second, our provision of messages may not be a requirement for those who can access messages themselves via a number of Usenet and community search services, the leading example of which is provided by Google at http://groups.google.com. Further, shifting message retrieval to other services shifts compliance responsibility to the researchers.

Other data has been requested that may have fewer privacy implications. Many researchers have requested data in the form of "edge-lists" from which directed graph models can be constructed. Directed graphs, or "networks", are an increasingly common data structure composed of "nodes" which represent individual and collective entities like people, newsgroups, or communities, and "edges" which represent the relationships that connect the nodes. A relationship can be any form of connection or transaction that links two entities. Being related to someone is a tie, as is working for them, lending them money, or offering them a ride. In the Netscan Usenet research, a reply from one author to another is the main type of relationship tie. Information about the set of all reply relationships among a group of authors is a part of the existing Netscan data set, but it exists in a form that requires manipulation to create. Given the limited SQL and data manipulation skills that many of our data licensees have available, many have requested that we build these tables for them and include them in subsequent releases.

While more papers and studies of the existing data set are likely and valuable, next steps include the development of additional data and the tools needed to analyze and interpret the

results. Additional data that represents social network structures are a likely addition. Other sources of data like web boards and email lists are also attractive possible extensions that would enhance the ability of findings made with the Netscan data set to be generalized to a broader scope of online behavior. Issues related to sharing of message content related to privacy and anonymity must be broadly addressed before message content can be added.

Some licensees have inquired about contributing data they have collected to the data set. This suggests the need for a broader infrastructure than the Netscan system was designed for that would allow for the submission of content from a variety of research projects. The many ownership issues related to the data are a key obstacle, even if contributed under terms that allow for free distribution and use.

Conclusion

We have presented here a vision for a future e-Social Science based on broad access to archives of richly detailed social science data. As social activity is increasingly self-documenting, an opportunity arises for the social sciences to begin to travel the path followed by other disciplines focused on complex dynamic systems, like meteorology, that have benefited significantly by shifts in the costs and availability of data and computation.

Acknowledgments

We would like to thank our many University Licensees and the many participants of Usenet.

References

Burke, M., Joyce, E., Kim, T., Anand, V., and Kraut, R. (2007): 'Introductions and Requests: Rhetorical Strategies That Elicit Response in Online Communities', *Communities & Technologies 2007*. Springer.

Eynon, R., Fry, J. and Schroeder, R. (Forthcoming): 'The Ethics of Internet Research', In R. Lee, N. Fielding and G. Blank (eds): The Handbook of Internet Research, London: Sage.

Gleave, E.P. and Smith, M.A. (2007): 'Reflections and Reactions to Social Accounting Meta-Data' *Communities and Technologies 2007*. Springer.

Preece, J. and Diker, V. G. (2005): "Communities of Practice and Online Communities, Fall 2005 Tentative Syllabus" for course INFM 718J / LBSC 708P at University of Maryland. http://www.wam.umd.edu/vdiker/INFM718J_LBSC708P_05_09/

Smith, Marc A. (2007): Netscan: A Social Accounting Search Engine, Microsoft Research Community Technologies Group, http://netscan.research.microsoft.com.

Raban, D. R. and Rabin, E. (2007): "The Implication of the Power Law Distribution of Activity in Information Sharing Sites for Statistical Inference", July 9, 2007, available at SSRN: http://ssrn.com/abstract=999286

Xi, W., Lind, J., and Brill, E. (2004): `Learning Effective Ranking Functions for Newsgroup Search', *SIGIR '04*, July 25-29, 2004, Shefield, UK. ACM 1-58113-000-0/00/0004.