AAAI Technical Report WS-13-04
When the City Meets the Citizen 2

The Limited Usefulness of Social Media and Digital Trace Data for Urban Social Research

Robert Goodspeed
MIT Department of Urban Studies and Planning
Room 9-415
77 Massachusetts Ave.
Cambridge, MA 02139
rgoodspe@mit.edu

Abstract

Drawing on the literature on social science methodology, this paper argues the potential research contributions from studies using social media data are limited. These data sources are conceptualized as unobtrusive measures, a longstanding category of information for social research. Although useful in some respects, digital unobtrusive measures are limited by their content poverty, focus on espoused theory, and positivist assumptions about social reality. To conclude, the paper describes the limitations of all forms of social research, and the need for mixed methods and action research.

Introduction

Social media researchers have argued that social media data such as photos, tweets, and check-ins provide useful new sources of information about city life. The call for this workshop argues new methods allow researchers to "extract local insights" from such data and "better understand cities and their residents." A related strand of research has conducted analysis of proprietary digital network data, such as the location and activity of cell phones recorded from all phone owners (e.g., Calabrese et al 2010). This paper argues both types of data are unobtrusive measures, and both present epistemological limitations to researchers.

The argument proceeds as follows. First, the concept of trace measures from the methodological literature is briefly described, and it is argued both social media data and network data constitute forms of trace measures. Second, several limitations of this data as trace measures are presented: content poverty, focus on espoused theory, and limitations shared by all forms of non-interpretative social research. As a consequence, social media data alone has limited potential to contribute to new theoretical understanding of social life in cities. In order to overcome the limitations of trace measures, social media researchers must design and deploy methods that allow for interpretive research, and conduct mixed methods research.

Unobtrusive Measures

Unobtrusive measures include any data about human behavior that can be collected without the subjects' knowledge (Webb 2000; Kellehear 1993). These measures, similar to Silverman's (2008) "naturally occurring material," minimize the reactive effects created by more intrusive methods, such as interviews and surveys. Typical examples of unobtrusive measures include physical traces, written and archival records, and simple observation. The advent of wide adoption of social media websites and other digital systems has created new sources of trace measures: social media data. This form of data has been justified as ethically appropriate for analysis because of its purported "voluntary" and public character, although the lack of informed consent raises concerns of research ethics not addressed here.

Since social media participation is voluntary, this data by definition cannot be easily used to generalize to larger populations. New statistical methods that allow for generalization from unrepresentative samples (such as matching, weighting and raking) require demographic variables, generally not available from social media (e.g., Liu et al 2010).

A related type of data, network data or spatiotemporal digital traces, are derived as an incidental output from use of mobile or hardline telephones, taxicab meters, or other digital sensors in the urban environment. These datasets have been viewed as potentially useful because of their apparently broader coverage of the urban population. However, the privatized nature of most infrastructure casts a cloud on the representativeness of most digital traces. As an example, long-distance telephone datasets generally involve one company, but in most markets a portion of

calls are placed through rival companies and online calling. Even seemingly more complete trace datasets, such as taxicabs, generally only include cabs regulated by one municipality in the metropolitan region, and by definition omit illegally operated cabs and other transit modes needed for comprehensive mobility studies.

Limitations

While useful in some respects, beyond the straightforward concerns about representativeness above, three fundamental limitations of findings possible from the exclusive use of unobtrusive measures are described here: content poverty, focus on espoused theory, and positivist assumptions.

Content Poverty

Building rich social theory requires multivariate data, but too often social media data has an acute content poverty problem. For example, in the realm of transportation, while social media can be used to measure relative popularity, or even the locations of individuals in the city, their behavior cannot be evaluated in the absence of contextual information about their choices, goals, and activities. Traditional transportation research methods involve highly detailed travel surveys, collecting not only travel routes but also information about costs, perception, and alternatives. New methods of activity-based modeling require even greater detailed information of the specific activities people perform at each location (Shiftan and Ben-Akiva 2011). Such content limitations are examples of why analysis of novel datasets is often limited to pattern recognition and description from data mining, not social theory building. For example, a novel analysis of taxicab traces resulted in the relatively commonsense findings that taxi drivers avoid congested areas and some are more efficient than others (Liu, Andris, and Ratti 2010).

Espoused Theory vs. Theory-in-use

While content poverty is certainly widespread among social media data, it is by no means universal. Certain social media datasets present rich datasets amenable to a variety of analysis methods. For example, Tweets can be subjected to sentiment analysis (Evans-Cowley and Griffin 2012; Bollen, Pepe, and Mao 2011), blog posts and other texts can be analyzed using qualitative textual analysis methods, and photos or multimedia content analyzed for their color, content, or locations.

However, rich qualitative data does not ensure validity. Organizational researchers have emphasized the division between espoused theory and theory-in-use (Argyris and Schön 1974). In short, it holds that people's explanations about their actions may not align with the assumptions reflected by their actual actions. Importantly, this does not only describe misrepresentation, but also the idea that people's actions are guided by tacit knowledge they cannot easily explain. Social media data generally contains either what people say (espoused theory) or what they do (theory-in-use), but not a rich record of both. In the few cases where social media data contains both, datasets have the previous problem, knowing only limited dimensions of revealed behavior, not enough information to build novel theoretical understanding of the behavior.

Positivist Assumptions

Finally, qualitative researchers have argued that alternative research methods are associated with assumptions about the nature of reality, human nature, and nature of knowledge (Morgan and Smircich 1980). This perspective is used to justify the use of qualitative methods to develop valid forms of social knowledge. The argument for alternative research methods is a generalization of the previous limitation, which introduced a distinction between the actual structure of behavior and stated preferences and behaviors. This perspective views human behavior as shaped not only from universal impulses, but also by culture. At the extreme is ethnography, which argues that understanding the nature of culture requires a holistic, close observation of members of a culture, not merely voluntary public (or semi-public) social media data. From an ethnographic viewpoint, social media data is expected to be unrepresentative, avoid taboos, and itself filtered by cultural expectations for appropriate interactions. In the field of human-computer interaction, Dourish has argued for a "technomethodology" as a means for sociotechnical system design (Dourish and Button 1998).

Limitation of All Social Research Strategies

McGrath has argued that all social research seeks to maximize three dimensions: generalizability with respect to populations, precision and control in measurement, and realism (McGrath 1981). However, any specific social research method involves inevitable trade-offs between these desiderata, introducing methodological dilemmas. In particular, digital trace data maximizes realism, but is not generalizable as it concerns a specific place and time, and lacks measurement flexibility. McGrath argues for mixed-methods research. Perhaps social media data could be combined with other methods, as some studies due to a limited extent (Cranshaw et al 2013). Ensuring research relevance may require participatory research methods (Greenwood and Levin 2007; Hearn et al. 2009).

Conclusions

The aim of this paper is to introduce ideas from the broad field of social research methods into the discussion of about using social media data and digital trace measures. Three interrelated limitations of these data sources are described: content poverty, limited ability to explore espoused theory versus theory-in-use, and the need for subjectivist research. In conclusion, a useful framework describing the limitations of all social research is described and researchers are urged to adopt mixed-methods research designs.

References

Argyris, Chris, and Donald A. Schön. 1974. *Theory in practice : increasing professional effectiveness*. 1st ed. San Francisco: Jossey-Bass Publishers.

Bollen, Johan, Alberto Pepe, and Huina Mao. 2011. Modeling Public Mood And Emotion: Twitter Sentiment And Socio-Economic Phenomena. Paper read at Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

Calabrese, Francesco, Francisco Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. 2010. The Geography Of Taste: Analyzing Cell-Phone Mobility And Social Events. *Pervasive Computing* 22-37.

Cranshaw, Justin, Raz Schwartz, Jason I Hong, and Norman Sadeh. 2012. The Livehoods Project: Utilizing Social Media To Understand The Dynamics Of A City. In Proceedings of the 6th International AAIA Conference on Weblogs and Social Media. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, Inc.

Dourish, Paul, and Graham Button. 1998. On' Technomethodology': Foundational Relationships Between Ethnomethodology and System Design. *Human-Computer Interaction* 13 (4):395.

Evans-Cowley, Jennifer S, and Greg Griffin. 2012. Microparticipation with Social Media for Community Engagement in Transportation Planning. *Transportation Research Record: Journal of the Transportation Research Board* 2307 (1):90-98.

Greenwood, Davydd J., and Morten Levin. 2007. *Introduction to Action Research : Social Research for Social Change*. 2nd ed. Thousand Oaks, Calif.: Sage Publications.

Hearn, Gregory, Tacchi, Jo, Foth, Marcus, and Lennie, June. 2009. *Action Research and New Media: Concepts, Methods and Cases.* Cresskill, NJ: Hampton Press.

Kellehear, Allan. 1993. *The Unobtrusive Researcher*. St. Leonards, NSW, Australia: Allen & Unwin.

Liu, Liang, Clio Andris, and Carlo Ratti. 2010. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems* 34 (6):541-548.

Liu, Honghu, David Cella, Richard Gershon, Jie Shen, Leo S Morales, William Riley, and Ron D Hays. 2010. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *Journal of Clinical Epidemiology* 63 (11):1169-1178.

McGrath, J.E. 1981. Dilemmatics: The Study of Research Choices and Dilemmas. *American Behavioral Scientist* 25 (2):179-210.

Morgan, G, and L Smircich. 1980. The Case For Qualitative Research. *Academy of Management Review*:491-500.

Shiftan, Y., and M. Ben-Akiva. 2011. A Practical Policy-Sensitive, Activity-Based, Travel-Demand Model. *The Annals of Regional Science* 47 (3):517-541.

Silverman, D. 2008. *A Very Short, Fairly Interesting And Reasonably Cheap Book About Qualitative Research*. London: Sage Publications.

Webb, Eugene J. 2000. *Unobtrusive Measures*. Rev. ed, Sage classics series. Thousand Oaks, Calif.: Sage Publications.