Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

What Big Data May Mean for Surveys

Mick P. Couper[1]

Abstract

Two converging trends raise questions about the future of large-scale probability surveys conducted by or for National Statistical Institutes (NSIs). First, increasing costs and rising rates of nonresponse potentially threaten the cost-effectiveness and inferential value of surveys. Second, there is growing interest in Big Data as a replacement for surveys. There are many different types of Big Data, but the primary focus here is on data generated through social media. This paper supplements and updates an earlier paper on the topic (Couper, 2013). I review some of the concerns about Big Data, particularly from the survey perspective. I argue that there is a role for both high-quality surveys and big data analytics in the work of NSIs. While Big Data is unlikely to replace high-quality surveys, I believe the two methods can serve complementary functions. I attempt to identify some of the criteria that need to be met, and questions that need to be answered, before Big Data can be used for reliable population-based inference.

Key Words: Big Data; Surveys; Social Media

1. Introduction

1.1 Meaning and Types of Big Data

This paper is a brief update and extension of an earlier published paper on the topic (see Couper, 2013). In that paper, I raised a number of concerns about big data from a survey researcher's perspective, but argued that we cannot ignore the Big Data phenomenon, and must work towards finding ways to engage with big data scientists to evaluate the quality of big data and understand the strengths and weakness of various types of big data, as we have done for decades with surveys.

One of the first issues we have to face when talking about "Big Data" is that there are many different types of big data (as indeed, there are many different types of surveys). The tendency to capitalize Big Data (as I have done thus far) in fact suggests that it is a unitary phenomenon. We should be wary about making broad claims (either positive or negative) about big data, in the same way we should avoid claiming that all surveys are equally good or equally bad. Such collapsing of many different approaches into a single category inevitable leads to over-generalization.

First, what do we mean by big data? The three criteria most commonly used to identify big data are volume, velocity, and variability (see Daas et al., 2002; TechAmerica Foundation, 2012). While one feature that characterizes big data is size (necessitating different computation resources), the other features also require different statistical processing and analysis approaches.

There are many different types of big data, as there are many different types of surveys. These types vary in quality and information content. The inferential challenges posed are different for different types of big data. Some of the varieties of big data are the following:

- Administrative data – provided by persons or organizations for administration of a program (e.g., electronic medical records, insurance records, bank records, tax records, registers)
- Transaction data – generated as an automatic by-product of transaction and activities (e.g., credit card transactions, online transactions)

[1]Survey Research Center, University of Michigan, P.O. Box 1248, Ann Arbor, MI 48106, USA (mcouper@umich.edu).

- Sensor data (e.g., satellite imaging, road sensors, climate sensors)
- Tracking device data (e.g., GPS, mobile phones)
- Behavioral data (e.g., online searches, page views, cookie data)
- Social media data – created by people with the express purpose of sharing with (at least some) others (e.g., Facebook, Twitter)

Groves (2011) used the term "organic data" to contrast with "designed data" produced by surveys. But this appellation is more applicable to some types of big data than to others. For example, administrative big data is largely designed data, albeit often for a different purpose than survey data. But administrative data are often highly structured and, in the case of register data (for example) may not suffer the same selection biases as other types of big data. Similarly, sensor data such as road sensors are designed to collect certain types of information, but may be put to different uses for which they were not originally intended. On the other hand, behavioral data generated by online activities and social media data have the characteristic features of velocity and variability but are not particularly large compared to big data from the natural or physical sciences. For instance, Twitter generates about 10 gigabytes (1000^3 bytes) of data per day and the World Wide Web contains about 4 zettabytes (1000^7 bytes) of data, while a small fraction (0.001%) of sensors on the Large Hadron Collider generate about 25 petabytes (1000^5 bytes) of data annually. By this standard, social media data could be considered very large data but not exceptionally big data, and samples of Twitter data (e.g., the garden hose or 10% sample) could be analyzed on powerful desktop computers.

The primary focus of this paper is on the last two types of big data, those that can be described as organically generated by the online activities of humans whether directly and intentionally (in the case of Twitter) or indirectly (in the case of Internet searches and browsing behavior), and what Lampe and colleagues (2014) refer to as "big social data." These kinds of "data in the wild" present the biggest inferential challenges.

1.2 The "Threat" of Big Data

Many proponents of big data argue that big data will replace surveys. For example, Mayer-Schönberger and Cukier (2013, p. 31) write "Reaching for a random sample in the age of big data is like clutching at a horse whip in the era of the motor car." This claim is reminiscent of similar claims in earlier eras of survey research. For example, in 1999 Gordon Black (then CEO of Harris Interactive) pronounced: "Internet research is a 'replacement technology'—by this I mean any breakthrough invention where the advantages of the new technology are so dramatic as to all but eliminate the traditional technologies it replaces: like the automobile did to the horse and buggy" (cited in Couper, 2000). I suspect that – much like Internet surveys have become an important part of survey research without fully replacing other modes – big data will similarly find a place in statistical production without making surveys redundant. While survey research is facing a number of challenges, to paraphrase Monty Python, "it is not dead yet."

Part of the problem with these claims is that they typically focus on size to the exclusion of other criteria such as quality or representation. While big data have undoubted value, they are not the same as survey data. Each type of data has strengths and weaknesses, and a careful understanding of the limitations of big data is critical to being able to fully exploit their potential.

2. Limitations of Big Data

The proponents of big data – as with the proponents of any other new method – tend to focus on the promise of the new data sources. It is up to the researchers that follow to undertake the careful evaluation of the benefits and drawbacks of big data. As probability-based surveys matured, the field has developed a healthy interest in understanding the weaknesses of the method (for a classic paper, see Deming, 1944). Similarly, the initial innovators of Internet survey were focused on gaining acceptance for the new method, rather than pointing out its flaws. Subsequent research has focused on the inferential challenges of online panels and other types of Internet surveys (e.g., American Association for Public Opinion Research, 2013; Callegaro et al. 2014). Now is the time (I believe) to turn a similar critical eye on big data.

In the earlier paper, I identified a number of limitations of big data. These included:

- **Selection bias**
- Measurement bias (self-presentation)
- Limited covariates
- **Volatility or lack of stability**
- Privacy issues
- Access issues
- **Opportunity for mischief**
- Size is not everything (bigger is not necessarily better)
- **File drawer problem**
- **Correlation is not causation**

In this paper, I focus on a selected subset of these – those highlighted in bold. I address these in the sections below, and refer readers interested in the other limitations to Couper (2013).

2.1 Selection Bias

The issue of selection bias – whether through non-coverage or non-response – is well known to survey researchers. It is a major concern for big social data, but those analyzing social media data do not pay much attention to the issue. There are two elements of importance to inference from social media data. First, not everyone uses social media. While the number of users of Facebook worldwide is impressive, in terms of proportions of the population of any country, use of Facebook is far from universal. Similarly, only about 15% of U.S. adults were estimated to use Twitter in 2013 (Pew Research Center), while in Canada this was estimated to be about 21% by the end of 2014 (http://www.itbusiness.ca/news/canada-becoming-one-of-worlds-top-twitter-strongholds-emarketer-finds/49190).

But it is not just the number (or proportion) of social media users that is important – it is how different users (and particularly "power" users) differ from those not using social media. A number of papers have looked at how social media users are different from non-users (see, e.g., Blank, 2014; Couper, 2014). But even defining "users" is tricky. For example, according to http://twopcharts.com/, about 44% of Twitter users create an account but never use it, and only about 13% of registered accounts have sent a tweet in the past 30 days. Similarly, Arnaboldi et al. (2013) estimate that about 68% of Twitter accounts are human, the balance being corporate accounts or automated systems. A number of studies have demonstrated the highly-skewed nature of social media activity (e.g., Bakshy et al., 2011; Cha et al., 2010; Lorince et al., 2014). This reflects Jakob Nielsen's oft-cited 90-9-1 rule for participation inequality, which states "In most online communities, 90% of users are lurkers who never contribute, 9% of users contribute a little, and 1% of users account for almost all the action" (http://www.nngroup.com/articles/participation-inequality/). Few (if any papers) using social media data have accounted for the highly skewed nature of the data in their analyses, typically treating each tweet or posting as independent observations with equal weight.

2.2 Volatility or Lack of Stability

A key characteristic of national statistics produced by NSIs is the consistency of measurement over time, permitting analysis of time trends on key indicators in society. In contrast, a defining characteristic of big data is volatility, and this is especially true of social media data.

While this may make big social data useful for short-term trends or understanding a particular event (such as an election), they may not be suitable for measuring time trends that last years or even decades. The social media that are popular today may not be around (at least in the same form) tomorrow. Google was founded in 1998, making it still a teenager. The first tweet was sent in 2006, and Twitter has grown 5000% in the space of five years. On the other hand, other social networking sites have lost ground or ceased operation (e.g., MySpace, Second Life, Hyves in the Netherlands) and others are fighting for attention (e.g., Google+). With the success of the early sites, we are seeing a rise of alternatives (Instagram, Pinterest, Reddit, Snapchat, Tumblr, Yammer, etc.). The popularity of these sites – and the kind of people who use them – is likely to vary over time, presenting challenges for longer-term time trends. Even Google (used for Google Trends) may serve a different audience than (say) Bing or Yahoo or other

search engines. It would be useful to know if the findings from Google Trends (for example) replicate using a different search engine.

While the immediacy of big data is an attractive feature for understanding the here and now, social media may not be a suitable platform for building long term statistical indicators.

2.3 Opportunity for Mischief

One of the challenges facing the secondary uses of social media data is that the analyst (unlike the survey designer) has very little control over how the medium is being used. Temporal trends in who uses which sites for what purposes (discussed above) are one source of volatility or uncertainty. But another relates to the possible manipulation of social media data.

Social media are being used for particular purposes. A large part of social media revolves around popular culture, and popularity or influence is sometimes a goal in itself. This raises the risk of manipulation. For example, in February 2014, it was estimated that between 5.5% and 11.2% of all accounts on Facebook are duplicate, malicious or otherwise "fake" (see https://nakedsecurity.sophos.com/2014/02/10/). While this is similar to the ballot-stuffing problems of call-in polls, there are several factors that increase the likelihood of mischief with social media relative to other media:

- The relative anonymity of the Internet facilitates such behavior
- It is virtually costless to register and post to social media
- Automated systems (bots) can be written to generate content (e.g., astroturfing)
- There is a possibility of personal gain (notoriety, attention, financial reward, etc.) from such behavior

Online polls share the first three characteristics, while the last is a peculiarity of social media.

In their discussion of Google Flu Trends, Lazer et al. (2014) make a distinction between blue team and red team dynamics, borrowing terms from the military. Blue team dynamics are where the algorithm producing the data (and thus user utilization) has been modified by the service provider in accordance with their business model. Service providers are constantly changing their business model and revising or remaking their products, and this may have implications for those intending to use the data for other purposes.

Red team dynamics occur when research subjects (e.g., social media users) attempt to manipulate the data-generating process to meet their own goals, such as economic or political gain. As Lazer et al. (2014) note, "Ironically, the more successful we become at monitoring the behavior of people using these open sources of information, the more tempting it will be to manipulate those signals." While examples of influencing open-access online polls exist (see, e.g., the Ubermotive Guide to Media Influence, at http://www.ubermotive.com/?p=68), it is harder to find evidence of social media manipulation. However, *The Guardian* revealed in 2011 that the U.S. government was manipulating social media (see http://www.guardian.co.uk/technology/2011/mar/17/us-spy-operation-social-networks). In popular culture, fake accounts are used to promote celebrities and causes. For example, it is estimated that about half of Justin Bieber's 37 million Twitter followers were either fake or inactive (see http://www.digitalspy.com/music/news/a471915/justin-bieber-twitter-followers-50-percent-are-fake-says-report.html). This is relevant because many of the top Twitter accounts are those of celebrities or media companies – suggesting that popular culture is a primary focus of social media. Of the top 10 Twitter accounts in June 2014, seven (Katy Perry, Justin Bieber, Taylor Swift, Lady Gaga, Britney Spears, Rihanna, Justin Timberlake) are musicians, while two (YouTube and Instagram) are company accounts; President Obama's Twitter account was ranked third in terms of number of followers (see http://twopcharts.com/twoplist?source=atlan). Twitter's own company account is ranked 12th in terms of number of followers. A *New York Times* article (see http://bits.blogs.nytimes.com/2014/04/20/friends-and-influence-for-sale-online/?_r=0) describes how friends and influence can be bought online.

The ease with which social media can be manipulated raises question about the use of such data for official government estimates, especially once it becomes known that the data are being used in that way. Again, this is not meant to invalidate such efforts, but rather to suggest that caution is needed to separate the wheat from the chaff – or signal from noise – when analyzing such organic data sources.

2.4 File Drawer Problem

Much of the early focus of big data analytics, especially those focused on social media, has been on the successful "predictions". Failures of such approaches are likely to get less attention, giving us a possibly distorted view of the value of the new methods and data sources. This is often true of any new development in research, as proponents of the new approach try to generate interest in their methods. Two examples of the issue relevant to big social data will suffice.

The first relates to Google Flu Trends. The first paper showing that Google searches on influenza-like symptoms tracked closely with Centers for Disease Control and Prevention data (Ginsberg et al., 2008) received a lot of attention, both in the scientific literature and popular press. However, later papers (see, e.g., Cook et al. 2011; Butler, 2013; Dugas et al., 2013; Lazer et al., 2014) suggest the predictions were not as successful in later years. Furthermore, the early results were achieved through fitting the data to the existing trends, with the initial algorithm finding the best matches among 50 million search terms to fit 1152 data points. The algorithm has been modified several times subsequently to fit more recent data. Similar results are found for Twitter flu trends, with the early results showing a correlation of 0.96 with Centers for Disease Control and Prevention data (Paul and Drezde, 2011) not holding up in subsequent analyses (e.g., Murphy, 2013).

A second example comes from the 2009 German election. Using Twitter data, Tumasjan et al. (2011) reported a mean absolute error of 1.65 on vote share among political parties, and claimed that "the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls" (Tumasjan et al., 2011, p. 15). However, a replication of the analyses by Jungherr and colleagues (2012) also included the Pirate Party (a party with a big online presence but small vote share) and concluded that the Pirate Party would have been predicted winners of the election. Furthermore, varying the time period chosen by Tumasjan et al. (2011) for their prediction resulted in different conclusions.

The kinds of secondary analyses and replications described above are critical for a careful evaluation of the value of social media data. One prerequisite for such analyses is that the original data and algorithms be made available to others. If either is withheld for proprietary or other reasons, the ability to replicate is severely hindered. The fields of survey methodology and big data analytics must be open to publication of null findings or replication failures. If only the "successful" findings are made public, we will get a distorted view of the value of big social data. The file drawer problem (first identified by Rosenthal, 1979) is not unique to big data analytics. For a recent example, see Franco, Malhotra, and Simonovits (2014).

2.5 Correlation is Not Causation

A final challenge of big data discussed here is related to the file drawer problem. With the size of the datasets being analysed, almost any relationship may reach statistical significance. An amusing example can be found in Leinweber (2007), who "predicted" the Standard and Poor 500 stock market index in the U.S. with an R^2 of 0.99, using just 5 variables: butter production in Bangladesh and the U.S., cheese production in the U.S., and sheep population in Bangladesh and the U.S. Similar examples of spurious correlations can be found at http://www.tylervigen.com/. While these don't use big data, with the large number of observations used in big data analytics and the data mining techniques employed, significant correlations are easy to find. We need to find ways to account for the likelihood of false positives, and we need to remember to separate statistical significance from substantive meaning. Almost all analyses are over-powered in big data, relative to surveys. Both replication and sensitivity analyses may help protect us from what Lazer and colleagues (2014) call big data hubris.

3. Discussion and Conclusions

This brief review of some of the issues facing the analysis of big social data is not meant to be a blanket criticism of big data. Before we raise too many complaints about big data, we need to remember that surveys vary in quality too, and there are many bad surveys. Rather, my intent is to caution that the analysis and use of big social data – especially to replace or supplement survey data – is still in its infancy, and there is much we still need to learn. The field of survey research has a long tradition of self-criticism, and a focus on various sources of error that may cause

one to question the inferences drawn from a survey. The notion of statistical uncertainty, so much part of the probability sampling paradigm, remains an important concept even in the world of N=All, to use Mayer-Schönberger and Cukier's (2013) term. While the frequentist paradigm may not hold for big social data – and for much of survey research too – we need to develop and apply indicators of uncertainty using Bayesian approaches. As Hal Varian of Google (2014, p. 24) notes, "In this period of 'big data,' it seems strange to focus on sampling uncertainty, which tends to be small with large datasets, while completely ignoring model uncertainty, which may be quite large."

Big Data – in all of its varieties – offers enormous potential. While the proponents of Big Data often argue that it is transformative, and will make existing methods (especially surveys) obsolete, I believe this is more true in some areas than others. While Big Data is unlikely to make the survey method obsolete, this does not mean we can be complacent. The exploration of big data offers exciting opportunities, and I welcome the research attention. Some of the methods and frameworks that are well-known to survey researchers (such as the total survey error framework) may well apply to big data, while other methods may need to be developed further. I believe that survey research has a lot to offer the world of big data analytics, and rather than viewing such developments as competition, we should find ways to collaborate, in order to improve our understanding of the potential benefits and limitations of both big data and survey research. There's also a lot we have to learn from the big data world, especially on the challenges of inference from data that may not be fully representative, and over which we have little control.

References

AAPOR (2013), *Report of the AAPOR Task Force on Non-Probability Sampling*. Deerfield, IL: American Association for Public Opinion Research.

Arnaboldi, V., Conti, M., Passarella, A., and Dunbar, R. (2013), "Dynamics of Personal Social Relationships in Online Social Networks: a Study on Twitter", In *COSN '13 Proceedings of the First ACM Conference on Online Social Networks*, pp. 15-26, doi: 10.1145/2512938.2512949.

Bakshy, E., Hofman, J.M., Mason, W.A., and Watts, D.J. (2010), "Everyone's an Influencer: Quantifying Influence on Twitter", *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 65-74.

Blank, G. (2014), "Who Uses Twitter? Representativeness of Twitter Users", Paper presented at the General Online Research Conference, Cologne, March.

Butler, D. (2013), "When Google Got Flu Wrong: US Outbreak Foxes a Leading Web-based Method for Tracking Seasonal Flu", *Nature*, 13th February 2013, http://www.nature.com/news/when-google-got-flu-wrong-1.12413.

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A., Krosnick, J.A., and Lavrakas, P.J. (Eds.) (2014), *Online Panel Research: A Data Quality Perspective*. New York: Wiley.

Cha, M., Haddidi, H., Benevenuto, F., and Gummadi, K.P. (2010), "Measuring User Influence in Twitter: The Million Follower Fallacy", *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Cook, S., Conrad, C., Fowlkes, A.L., and Mohebbi, M.H. (2011), "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic", *PLoS ONE*, 6 (8): e23610. doi:10.1371/journal.pone.0023610.

Couper, M.P. (2000), "Web Surveys: A Review of Issues and Approaches", *Public Opinion Quarterly*, 64 (4): 464-494.

Couper, M.P. (2013), "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys", *Survey Research Methods*, 7 (3): 145-156.

Couper, M.P. (2014), "Social Media and Surveys: Collaboration, Not Competition", Paper presented at the General Online Research Conference, Cologne, March.

Daas, P.J.H., Roos, M., van de Ven, M., and Neroni, J. (2012), "Twitter as a Potential Data Source for Statistics", Den Haag/Heerlen, The Netherlands: Statistics Netherlands, discussion paper 201221.

Deming, W.E. (1944), "On Errors in Surveys", *American Sociological Review*, 9 (4): 359-369.

Dugas, A.F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., et al. (2013), "Influenza Forecasting with Google Flu Trends", *PLoS ONE*, 8 (2): e56176. doi:10.1371/journal.pone.0056176.

Franco, A., Malhotra, N., and Simonovits, G. (2014), "Publication Bias in the Social Sciences: Unlocking the File Drawer", *Science*, published online 28 August 2014, DOI:10.1126/science.1255484.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2009), "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature*, 457 (7232): 1012-1014.

Groves, R.M. (2011), "Three Eras of Survey Research", *Public Opinion Quarterly*, 75 (5): 861-871.

Jungherr, A., Jürgens, P., and Schoen, H. (2012), "Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T.O., Sander, P. G., & Welpe, I.M. 'Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment'", *Social Science Computer Review*, 30 (2): 229-234.

Lampe, C., Pasek, J., Guggenheim, L., Conrad, F., and Schober, M. (2014), "When Are Big Data Methods Trustworthy for Social Measurement?" Paper presented at the annual meeting of the American Association for Public Opinion Research, Anaheim, CA, May.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014), "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, 343 (6176) (March 14): 1203–1205.

Leinweber, D.J. (2007), "Stupid Data Miner Tricks: Overfitting the S&P 500", *The Journal of Investing*, 16 (1): 15-22.

Lorince, J., Zorowitz, S., Murdock, J., and Todd, P.M. (2014), " 'Supertagger' Behavior in Building Folksonomies", *Proceedings of the 2014 ACM conference on Web science (WebSci '14)*. ACM, New York, NY, USA, 129-138, DOI: 10.1145/2615569.2615686.

Mayer-Schönberger, V., and Cukier, K. (2013), *Big Data; a Revolution That Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin.

Murphy, J. (2013), "10 Things Every Survey Researcher Should Know about Twitter", Paper presented at the FedCASIC workshop, Washington, DC, March.

Paul, M.J., and Dredze, M. (2011), "You Are What You Tweet: Analyzing Twitter for Public Health", *Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, July 17-21*. Palo Alto, CA: AAAI Publications, pp. 265-272.

Rosenthal, R. (1979), "The File Drawer Problem and Tolerance for Null Results", *Psychological Bulletin*, 86 (3), 638-641.

TechAmerica Foundation (2012), *Demystifying Big Data: A Practical Guide to Transforming the Business of Government*. Washington, DC: TechAmerica Foundation.

Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. (2011), "Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape", *Social Science Computer Review*, 29 (4): 402-418.

Varian, H.R. (2014), "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, 28 (2): 3-28.