Commentary

BIG DATA & SOCIETY

Big Data & Society
July–December 2015: 1–3

DOI: 10.1177/2053951715604333
bds.sagepub.com

SAGE

Lost in a random forest: Using Big Data to study rare events

Christopher A Bail

Abstract

Sudden, broad-scale shifts in public opinion about social problems are relatively rare. Until recently, social scientists were forced to conduct post-hoc case studies of such unusual events that ignore the broader universe of possible shifts in public opinion that do not materialize. The vast amount of data that has recently become available via social media sites such as Facebook and Twitter—as well as the mass-digitization of qualitative archives provide an unprecedented opportunity for scholars to avoid such selection on the dependent variable. Yet the sheer scale of these new data creates a new set of methodological challenges. Conventional linear models, for example, minimize the influence of rare events as "outliers"—especially within analyses of large samples. While more advanced regression models exist to analyze outliers, they suffer from an even more daunting challenge: equifinality, or the likelihood that rare events may occur via different causal pathways. I discuss a variety of possible solutions to these problems—including recent advances in fuzzy set theory and machine learning—but ultimately advocate an ecumenical approach that combines multiple techniques in iterative fashion.

Keywords

Social media, rare events, causal complexity, automated text analysis, machine learning, cultural sociology

Among the most intriguing aspects of the recent explosion in digital text-based data is that it offers the potential for social scientists to analyze rare events with unprecedented precision. Consider, for example, a longstanding puzzle within the literature on collective behavior and cultural sociology: how do social movements, advocacy groups, or other civil society organizations create sweeping shifts in the way the public thinks about complex social problems such as racism, income inequality, or gender discrimination?

Few social scientists are prescient enough to anticipate such rare events before they occur. As a result, most studies of collective behavior and cultural change employ post-hoc case studies that trace the history of organizations only after they successfully transform the status quo. Yet the overwhelming majority of civil society organizations fail to create any public impact—let alone major shifts in public worldviews (Giugni, 1998; Koopmans, 2004; Summers-Effler, 2010). As a result, extent studies are routinely criticized for using circular reasoning that confuses the characteristics of successful civil society organizations with the causes of cultural change (Bail, 2012; Benford, 1997; Collins, 2001).

The recent explosion of text-based data enables scholars interested in the relationship between collective behavior and cultural change to escape the cardinal sin of selection on the dependent variable. Over the past five years, I developed new app-based technologies that track the influence of hundreds of civil society organizations upon the more than one billion people who use Facebook on a routine basis (Bail, 2015). This new research method not only situates organizations that succeed in shaping public opinion amidst the vast sea negative cases that fail to create broad-scale cultural change, but also collects hundreds of variables that describe these organizations, their audiences, and the broader social context in which they interact. Though important questions about the

Duke University, Durham, NC, USA

Corresponding author:
Christopher A Bail, Duke University, 254 Soc/Psych Building, 417 Chapel Drive, Durham, NC 27708, USA.
Email: christopher.bail@duke.edu

relationship between on and offline behavior remain unanswered, these new data enable fine-grained analysis of the spread of ideas across vast social networks with unprecedented qualitative and longitudinal detail.

Yet the study of rare events with Big Data also presents daunting new challenges. Now that we can see the haystack, we still need to find the needle. Even the most basic forms of descriptive analyses necessary to identify rare events within large datasets require new computing technologies designed to process massive datasets. Once such events are found, it is impossible to compare them to the vast population of negative cases with the naked eye because of their sheer scale. Preliminary analyses such as histograms, cross-tabulations, or scatterplots are of limited utility because the distribution of rare events within such datasets is so heavily skewed.

It is therefore rather tempting to forego basic descriptive analyses of Big Data and proceed with multivariate analyses that might better differentiate rare events from the negative cases that surround them. Yet analysis of variance and multivariate regression are designed to minimize the influence of rare events as "outliers." Indeed, much enthusiasm around Big Data surrounds the improvement of regression estimators because of the Central Limit Theorem, or the tendency for normal distributions to emerge as sample sizes grow.

There are, of course, a number of more nuanced regression models for analysis of rare events (e.g. King and Zeng, 2001), but these techniques suffer from a separate—perhaps more vexing—problem: equifinality, or the likelihood that rare events occur via multiple causal pathways. Consider, again, the study of why certain civil society organizations produce social media messages that reach millions of people, while others go mostly unnoticed. There are not only many different types of organizations with different messages and strategies to publicize them, but also diverse audiences who might receive and distribute them in a variety of different manners. Moreover, there are a variety of broader social conditions that may shape whether and how a message goes viral. Though one campaign may spread organically across dense networks of friends over months—or even years—others may succeed because they are dispatched within the context of a major news story about a social problem.

Even methods that are explicitly designed to capture causal complexity do not yet work well for analyzing rare events within large datasets. Ragin's (2000, 2008) fuzzy-set qualitative comparative analysis (fsQCA), for instance, arrays cases into "truth tables" that describe the frequency of different configurations of necessary or sufficient conditions that produce an outcome of interest. While fsQCA was originally designed to study causal complexity in "small N" studies, it holds considerable promise for studying larger datasets as well (Ragin, 2008). Though state-of-the art algorithms for reducing truth tables can identify multiple sufficient conditions for an outcome, most rare events include multiple *necessary* conditions. A social media message about climate change, for example, is unlikely to go viral in the aftermath of a major terrorist attack or at 3 am on the fourth of July.[1]

Another approach to modeling causal complexity within large datasets is the burgeoning field of machine learning. Random forest techniques, for example, combine conventional regression tree methods with iterative bootstrapping techniques to classify large amounts of data into different branches—or configurations of variables—with sufficiently large samples (Breiman, 2001). These new techniques—which are gradually attracting the interest of social scientists (e.g. Grimmer and Stewart, 2013)—hold great promise for the study of rare events. This is not only because they recognize causal complexity but also because they permit identification of patterns within data that could not be recognized by the naked eye.

Yet machine-learning techniques also introduce an entirely new genre of methodological problems. From the perspective of a computer or other machine, many of the proverbial needles within haystacks look like hay.[2] Therefore, even the most sophisticated machine-learning techniques will produce nonsensical results if humans do not carefully validate them. This is already obvious within the exciting new field of natural language processing. Topic models and other automated forms of content analysis can easily categorize vast corpora with impressive precision—and, some argue, greater efficiency and reliability than human coders (Hopkins and King, 2010). Not unlike cluster analysis, however, topic models require social scientists to specify an expected number of latent topics a priori. Because one cannot possibly read vast corpora, however, many people utilize topic models in an inductive manner that resembles reading tea leaves (Chang et al., 2009). Though such "grounded theory" approaches can be a powerful tool for classifying large corpora when used properly, arbitrary or under-validated topic models become nonsensical if they are combined with the random forest techniques that do not discriminate between variables that were carefully created by a masterful qualitative researcher and those that resulted from the topic model validity measure *du jour*.

The burgeoning field of Big Data visualization provides another exciting opportunity for the analysis of rare events in Big Data. While social scientists typically employ visualization techniques to *represent* data, this new field provides a suite of new visual methods for *analyzing* data as well. These include interactive computer-based tools that enable one to change certain parameters while holding others constant—for

example, "CorrPlots" that spatially encode observations as points on geometric structures that underlay Pearson's correlation (McKenna et al., 2015), or video-based network analysis that may enable identification of sudden shifts in social relationships that create the conditions for rare events (Moody et al., 2005). The major downside of these new methods is that standards do not yet exist for what counts as evidence (Healy and Moody, 2014), and the boundary between aesthetics and empirical evidence currently rests in the eye of the beholder.

At the risk of cliché, this brief review hints at the value of an ecumenical approach to studying rare events with Big Data. Though each of the methods described above has considerable limitations, the rapid diversification of the computational social scientist's tool kit has produced a suite of methods that complement each other rather naturally. For example, random forests or fsQCA might be used to identify interaction terms for negative binomial regression models—or qualitative coding can be used to calibrate large-scale quantitative content analyses as Mohr et al. (2013) have shown. The challenge for studying rare events with Big Data, then, will be to avoid getting "lost" in random forests—or staring too long at any single tree.

Acknowledgement

I am grateful to Ron Breiger, Taylor Brown, John Mohr, and Robin Wagner-Pacifici for comments upon previous drafts of this manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. What is more, establishing causality within truth-table methods requires high levels of consistency—or the regularity with which outcomes occur given different sufficient or necessary conditions—and very rare events may not meet such thresholds because of limited diversity or because they are masked by other forms of causal complexity. These and other challenges could be addressed using non-truth table approaches designed for the study of so-called supersets (e.g. Ragin, 2008).
2. Steve Lohr (2012).

References

Bail C (2012) The fringe effect: Civil society organizations and the evolution of media discourse about Islam, 2001–2008. *American Sociological Review* 77(7): 855–879.

Bail C (2015) Taming Big Data: Using app technology to study organizational behavior on social media. *Sociological Methods & Research*.

Benford R (1997) An insider's critique of the social movement framing perspective. *Sociological Inquiry* 67(4): 409–430.

Breiman L (2001) Random forests. *Machine Learning* 45(1): 5–32.

Chang J, Boyd-Graber J, Gerrish S, et al. (2009) Advances in neural information processing systems. *Reading Tea Leaves: How Humans Interpret Topic Models* 296–299.

Collins R (2001) Social movements and the focus of emotional attention. In: Goodwin J, Jasper J and Polletta F (eds) *Passionate Politics: Emotions and Social Movements*. Chicago: University of Chicago Press, pp. 27–46.

Giugni M (1998) Was it worth the effort? The outcomes and consequences of social movements. *Annual Review of Sociology* 24(1): 371–393.

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*. Epub ahead of print. Available at: http://doi.org/10.1093/pan/mps028.

Healy K and Moody J (2014) Data visualization in sociology. *Annual Review of Sociology* 40(1): 105–128.

Hopkins DJ and King G (2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1): 229–247.

King G and Zeng L (2001) Logistic regression in rare events data. *Political Analysis* 9(2): 137–163.

Koopmans R (2004) Movements and media: Selection processes and evolutionary dynamics in the public sphere. *Theory and Society* 33(3/4): 367–391.

Lohr S (2012) The age of Big Data. *The New York Times Sunday Review*, 11 February.

McKenna S, Meyer M, Gregg C, et al. (2015) s-CorrPlot: An interactive scatterplot for exploring correlation. *Journal of Computational and Graphical Statistics*. DOI: 10.1080/10618600.2015.1021926.

Mohr JW, Wagner-Pacifici R, Breiger RL, et al. (2013) Graphing the grammar of motives in national security strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics* 41(6): 670–700.

Moody J, McFarland D and Bender-deMoll S (2005) Dynamic network visualization. *American Journal of Sociology* 110(4): 1206–1241.

Ragin CC (2000) *Fuzzy-set Social Science*, 1st ed. Chicago: University of Chicago Press.

Ragin CC (2008) *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Summers-Effler E (2010) *Laughing Saints and Righteous Heroes: Emotional Rhythms in Social Movement Groups*. Chicago: University of Chicago Press.

This article is part of a special theme on *Colloquium: Assumptions of Sociality*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/colloquium-assumptions-sociality.