SOCIAL SCIENCES

Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By **Derek Ruths**[1]* *and* **Jürgen Pfeffer**[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: “Dewey Defeats Truman” (*1*, *2*). The headline was informed by telephone surveys, which had inadvertently undersampled Truman supporters (*1*). Rather than permanently discrediting the practice of polling, this event led to the development of more sophisticated techniques and higher standards that produce the more accurate and statistically rigorous polls conducted today (*3*).

Now, we are poised at a similar technological inflection point with the rise of online personal and social data for the study of human behavior. Powerful computational resources combined with the availability of massive social media data sets has given rise to a growing body of work that uses a combination of machine learning, natural language processing, network analysis, and statistics for the measurement of population structure and human behavior at unprecedented scale. However, mounting evidence suggests that many of the forecasts and analyses being produced misrepresent the real world (*4*–*6*). Here, we highlight issues that are endemic to the study of human behavior through large-scale social media data sets and discuss strategies that can be used to address them (see the table). Although some of the issues raised are very basic (and long-studied) in the social sciences, the new kinds of data and the entry of a variety of communities of researchers into the field make these issues worth revisiting and updating.

POLICY

REPRESENTATION OF HUMAN POPULATIONS. *Population bias.* A common assumption underlying many large-scale social media-based studies of human behavior is that a large-enough sample of users will drown out noise introduced by peculiarities of the platform’s population (*7*). However, substantial population biases vary across different social media platforms (*8*). For instance, Instagram is “especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents” (*9*) whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

Proprietary algorithms for public data. Platform-specific sampling problems, for example, the highest-volume source of public Twitter data, which are used by thousands of researchers worldwide, is not an accurate representation of the overall platform’s data (*11*). Furthermore, researchers are left in the dark about when and how social media providers change the sampling and/or filtering of their data streams. So long as the algorithms and processes that govern these public data releases are largely dynamic, proprietary, and secret or undocumented, designing reliable and reproducible studies of human behavior that correctly account for the resulting biases will be difficult, if not impossible. Academic efforts to characterize aspects of the behavior of such proprietary systems can provide details needed to begin reporting biases.

The rise of “embedded researchers” (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms, and resources) is creating a divided social media research community. Such researchers, for example, can see a platform’s inner workings and make accommodations, but may not be able to reveal their corrections or the data used to generate their findings.

REPRESENTATION OF HUMAN BEHAVIOR. *Human behavior and online platform design.* Many social forces that drive the formation and dynamics of human behavior and relations have been intensively studied and are well-known (*12*–*14*). For instance, homophily (“birds of a feather flock together”), transitivity (“the friend of a friend is a friend”), and propinquity (“those close by form a tie”) are all known by designers of social media platforms and, to increase platform use and adoption, have been incorporated in their link suggestion algorithms. Thus, it may be necessary to untangle psychosocial from platform-driven behavior. Unfortunately, few studies attempt this.

Social platforms also implicitly target

Reducing biases and flaws in social media data

DATA COLLECTION

- 1. Quantifies platform-specific biases (platform design, user base, platform-specific behavior, platform storage policies)
- 2. Quantifies biases of available data (access constraints, platform-side filtering)
- 3. Quantifies proxy population biases/mismatches

METHODS

- 4. Applies filters/corrects for nonhuman accounts in data
- 5. Accounts for platform and proxy population biases
  a. Corrects for platform-specific and proxy population biases
  OR
  b. Tests robustness of findings
- 6. Accounts for platform-specific algorithms
  a. Shows results for more than one platform
  OR
  b. Shows results for time-separated data sets from the same platform
- 7. For new methods: compares results to existing methods on the same data
- 8. For new social phenomena or methods or classifiers: reports performance on two or more distinct data sets (one of which was not used during classifier development or design)

Issues in evaluating data from social media. Large-scale social media studies of human behavior should i address issues listed and discussed herein (further discussion in supplementary materials).

[1]*Department of Computer Science, McGill University, Montreal, Quebec H3A 0G4, Canada.* [2]*Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15213, USA. *E-mail: derek.ruths@mcgill.ca*

Published by AAAS

and capture human behavior according to behavioral norms that develop around and as a result of the specific platforms. For instance, the ways in which users view Twitter as a space for political discourse affects how representative political content will be. The challenge of accounting for platform-specific behavioral norms is compounded by their temporal nature: They change with shifts in population composition, the rise and fall of other platforms, and current events (e.g., revelations concerning interest and tracking of social media platforms by intelligence services). In the absence of new methodologies, we must rely on assessments of where such entanglements likely occur.

Distortion of human behavior. Developers of online social platforms are building tools to serve a specific, practical purpose—not necessarily to represent social behavior or provide good data for research. So, the way data are stored and served can destroy aspects of the human behavior of interest. For instance, Google stores and reports final searches submitted, after auto-completion is done, as opposed to the text actually typed by the user (*5*); Twitter dismantles retweet chains by connecting every retweet back to the original source (rather than the post that triggered that retweet). There are valid, practical reasons for platforms to make such design decisions, but in many cases these either obscure or lose important aspects of the underlying human behavior. Quantifying and, if possible, correcting for these storage and access policies should be part of the data set reporting and curation process.

Nonhumans in large-scale studies. Despite attempts by platform designers to police accounts, there are large populations of spammers and bots masquerading as "normal" humans on all major online social platforms. Moreover, many prominent individuals maintain social media accounts that are professionally managed to create a constructed image or even behave so as to strategically influence other users. It is hard to remove or correct for such distortions.

ISSUES WITH METHODS. *Proxy population mismatch.* Every social media research question defines a population of interest: e.g., voting preference among California university students. However, because human populations rarely self-label, proxy populations of users are commonly studied instead, for example, the set of all Facebook users who report attending a UC school. However, the quantitative relation between the proxy and original populations studied, typically, is unknown—a source of potentially serious bias. A recent study revealed that this proxy effect has caused substantially incorrect estimates of political orientation on Twitter (*6*).

Incomparability of methods and data. With few exceptions, the terms of usage for social media platforms forbid the retention or sharing of data sets collected from their sites. As a result, canonical data sets for the evaluation and comparison of computational and statistical methods—common in many other fields—largely do not exist. Furthermore, few researchers publish code implementing their methods. The result is a culture in which new methods are introduced (and often touted as being "better") without having been directly compared to existing methods on a single data set. Given platforms' understandable sensitivity to user privacy and the competitive value of their data, the research community will likely improve method and result comparison issues more quickly by focusing on enforcing the sharing of methods at publication time.

There is "...the need for increased awareness of what is actually being analyzed..."

Multiple comparison problems. The body of social media analysis that concerns the development of user/content classification and prediction has unaddressed issues with overfitting. Specifically, when building a computational machine that recognizes two or more classes (of users, for example), it is customary to introduce tens to hundreds of features as the basis for the classifier. At the very least, the performance of the classifier should take into account the number of features being used. Of greater concern, however, is the extent to which the classifier performance is a result of "feature hunting"—testing feature after feature until one is found that delivers significant performance on the specific data set. Standard practices of reporting the P value for classifiers based on the number of features involved, as well as keeping a data set independent of the training set for final classifier evaluation, would work toward addressing these issues (*15*).

Multiple hypothesis testing. In an academic culture that celebrates only positive findings, a meta-issue emerges as multiple groups report successes in modeling or predicting a specific social phenomenon. Without seeing the failed studies, we cannot assess the extent to which successful findings are the result of random chance. This issue has been observed when predicting political election outcomes with Twitter (*16*). We are not the only field struggling with this issue (*17*). Solutions to this problem could involve enabling the publication of negative results or requiring the use of more data sets in a single study (so as to permit the calculation of a significance score within the study itself).

CONCLUSIONS. The biases and issues highlighted above will not affect all research in the same way. Well-reasoned judgment on the part of authors, reviewers, and editors is warranted here. Many of the issues discussed have well-known solutions contributed by other fields such as epidemiology, statistics, and machine learning. In some cases, the solutions are difficult to fit with practical realities (e.g., as in the case of proper significance testing) whereas in other cases the community simply has not broadly adopted best practices (e.g., independent data sets for testing machine learning techniques) or the existing solutions may be subject to biases of their own. Regardless, a crucial step is to resolve the disconnect that exists between this research community and other (often related) fields with methods and practices for managing analytical bias.

Moreover, although the issues highlighted above all have different origins and specific solutions, they share in common the need for increased awareness of what is actually being analyzed when working with social media data. ■

REFERENCES AND NOTES

1. This was not the first or last such erroneous prediction, e.g., the *Literary Digest* on the 1936 U.S. Presidential election.
2. F. Mosteller, H. Hyman, P. J. McCarthy, E. S. Marks, D. B. Truman, *The Pre-Election Polls of 1948* (Bulletin 60, Social Science Research Council, New York, 1949).
3. I. Crespi, *Public Opinion, Polls, and Democracy* (Westview Press, Boulder, CO, 1989).
4. Z. Tufekci, in *ICWSM '14: Proceedings of the Eighth International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media* (AAAI, Palo Alto, CA, 2014).
5. D. Lazer, R. Kennedy, G. King, A. Vespignani, *Science* **343**, 1203 (2014).
6. R. Cohen, D. Ruths, *ICWSM '13: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (AAAI, Palo Alto, CA, 2013), pp. 91–99.
7. V. Mayer-Schoenberger, K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Houghton Mifflin Harcourt, New York, 2013).
8. A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, J. N. Rosenquist, *ICWSM '11: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (AAAI, Palo Alto, CA, 2011), pp. 554–557.
9. M. Duggan, J. Brenner, The demographics of social media users; www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/.
10. 13 'pinteresting' facts about Pinterest users; www.pinterest.com/pin/234257618087475827/.
11. F. Morstatter, J. Pfeffer, H. Liu, *Proceedings of Web Science Track, at the 23rd Conference on the WWW* (Association for Computing Machinery, New York, 2014), pp. 555–556.
12. M. McPherson *et al.*, *Annu. Rev. Sociol.* **27**, 415 (2001).
13. F. Heider, *J. Psychol.* **21**, 107 (1946).
14. L. Festinger, S. Schachter, K. Back, in *Social Pressure in Informal Groups*, L. Festinger, S. Schachter, and K. Back, Eds. (MIT Press, Cambridge, MA, 1950), chap. 4.
15. S. J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson Education, Upper Saddle River, NJ, 2003).
16. H. Schoen *et al.*, *Internet Res.* **23**, 528 (2013).
17. J. P. A. Ioannidis, *PLOS Med.* **2**, e124 (2005).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/346/6213/1063/suppl/DC1

10.1126/science.1257756