Study finds obstacles with social media data

December 4 2014, by Deborah M. Todd, Pittsburgh Post-Gazette

Seas of social media posts uncovering the opinions, habits and purchases of billions of people would seem to be a fantasy come true for data researchers, social scientists and businesses. However, finding valuable insights through so-called big data isn't as simple as sifting through a few million Twitter posts.

In fact, anyone analyzing large data sets most likely has to adjust methods in order to find useful and truthful insights surrounding public health, spending and potential political outcomes, according to a study by Carnegie Mellon University researcher Juergen Pfeffer and McGill University researcher Derek Ruths that was published in the Nov. 28 issue of the journal *Science*.

Declaring the Big Data honeymoon over, Pfeffer said researchers must tweak methodologies to account for shortfalls within the data.

For companies such as Twitter, which he said was the most cooperative social media company in terms of providing research data, advances in machine learning that would create more representative samples for study could be a solution.

Social media accounts run by robots, fake accounts and real people lying through their own accounts were among the more obvious issues identified as obstacles to research in the paper, "Social Media for Large Studies of Behavior."

Differences in user demographics - photo-sharing site Instagram is

popular among blacks and Latinos ages 18 to 29 while photo and lifestyle site Pinterest skews toward mostly women ages 25 to 34, according to the study - were also notable factors in study outcomes.

Beyond who populates what social media forum, the notion of how the forum should be used also can affect research.

"The ways in which users view Twitter as a space for political discourse affects how representative political content will be," reads a portion of the study. "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."

One of the more profound insights, according to Pfeffer, is the idea that finding a random sampling of social media posts from sites such as Twitter shouldn't be left to chance.

"People are tweeting around the world, and their data is most likely collected from hundreds of servers around the world. In order to make it possible for Twitter to give you a real-time sample, they need to get data from (at least) one of 10 servers, so you get one of 10 possible tweets or 1 percent of Tweets. But those one of 10 servers may be in different positions of the world so they collect different data."

Without a clear picture of who is in a given study, where they are from or even if they are real people, it's difficult to know if information included in social media data is at all reliable.

"The reality for researchers is they can't be sure the samples they get are representative samples," said Pfeffer.

In any case, Pfeffer said researchers are facing issues as old as the notion

of public opinion polls. Using the infamous, incorrect Chicago Tribune headline "Dewey Defeats Truman" as an example, he said if researchers of yesteryear could use the gaffe to improve their polling methods, then today's researchers should follow their lead and get to work seeking solutions.

"I still think the underlying idea that millions of people create information we can analyze is amazing. My message is, basically, that it's harder than it looked in the beginning."