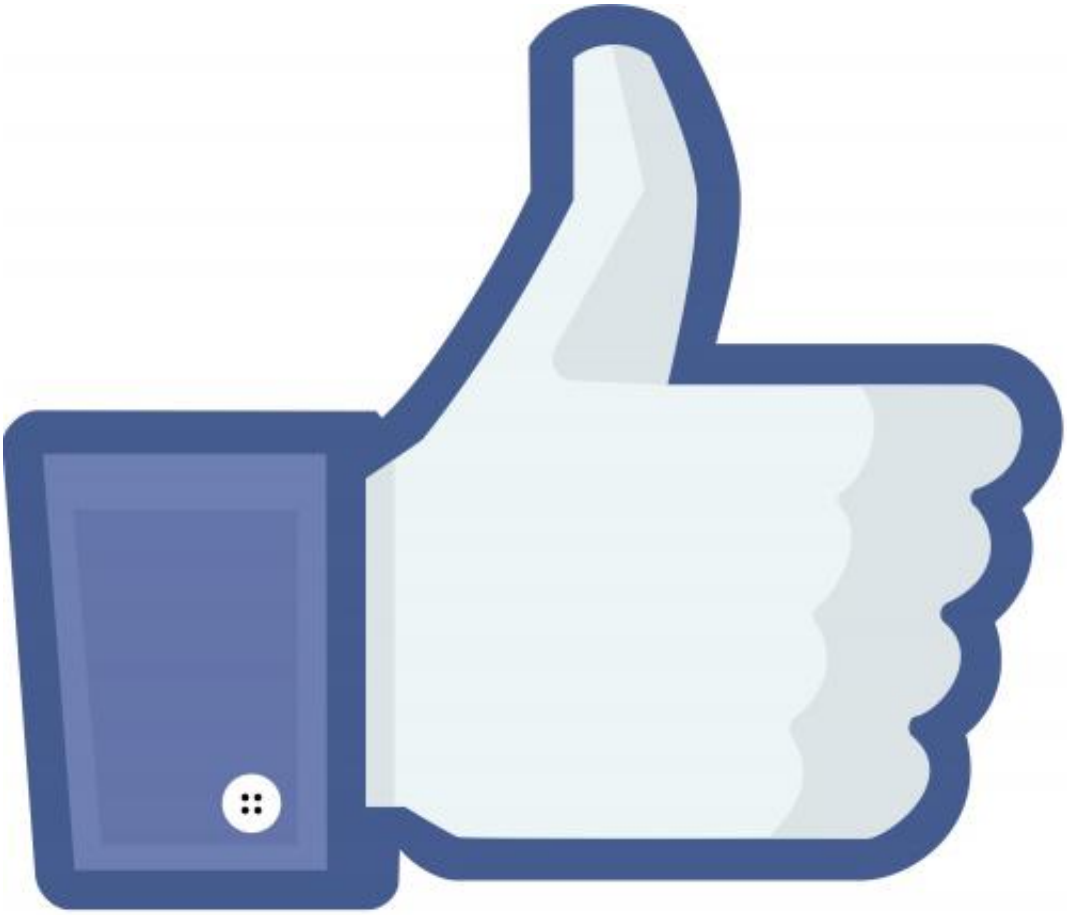
Researcher uncovers inherent biases of big data collected from social media sites

June 23 2015, by Julie Deardorff



With every click, Facebook, Twitter and other social media users leave behind digital traces of themselves, information that can be used by

businesses, government agencies and other groups that rely on "big data."

But while the information derived from social network sites can shed light on social behavioral traits, some analyses based on this type of data collection are prone to bias from the get-go, according to new research by Northwestern University professor Eszter Hargittai, who heads the Web Use Project.

Since people don't randomly join Facebook, Twitter or LinkedIn—they deliberately choose to engage —the data are potentially biased in terms of demographics, socioeconomic background or Internet skills, according to the research. This has implications for businesses, municipalities and other groups who use big data because it excludes certain segments of the population and could lead to unwarranted or faulty conclusions, Hargittai said.

The study, "Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites" was published last month in the journal *The Annals of the American Academy of Political and Social Science* and is part of a larger, ongoing study.

The buzzword "big data" refers to automatically generated information about people's behavior. It's called "big" because it can easily include millions of observations if not more. In contrast to surveys, which require explicit responses to questions, big data is created when people do things using a service or system.

"The problem is that the only people whose behaviors and opinions are represented are those who decided to join the site in the first place," said Hargittai, the April McClain-Delaney and John Delaney Professor in the School of Communication. "If people are analyzing big data to answer certain questions, they may be leaving out entire groups of people and their voices."

For example, a city could use Twitter to collect local opinion regarding how to make the community more "age-friendly" or whether more bike lanes are needed. In those cases, "it's really important to know that people aren't on Twitter randomly, and you would only get a certain type of person's response to the question," said Hargittai.

"You could be missing half the population, if not more. The same holds true for companies who only use Twitter and Facebook and are looking for feedback about their products," she said. "It really has implications for every kind of group."

Hargittai's research group, the Web Use Project, examines how people use the Web in their everyday lives and in particular, how differences in Internet use may contribute to social inequality.

Her latest study focused on issues related to a particular type of big data analysis: Those that draw broad conclusions from data, even when the data is restricted to users of particular sites and services. Though other research has examined the challenges of big data studies, Hargittai's is one of the first to provide empirical evidence suggesting potential biases.

"Many data sets that use so-called "big data" rely on social network sites such as Facebook and Twitter. But studies rarely discuss that people who select into using Facebook and Twitter don't necessarily represent larger populations," said Hargittai, a faculty associate at Northwestern's Institute for Policy Research.

Moreover, what people do on one platform misses potentially important information about how they are using other online services or other means altogether, including face-to-face interactions and phone calls.

Hargittai used two datasets, including one nationally representative sample from the Pew Internet Project (PIP), the high quality, go-to

resource for data on Americans' Internet use. In addition, Hargittai used her own data collected from wired and educated young adults.

The Pew data indicates that demographic factors such as age and gender contribute to what sites people chose; Hargittai's data fills some gaps in the Pew data and suggests people's Internet skills also are related to what services they start using.

"The less privileged are not on these sites so their opinions are not there either," she said. "Even among young adults who are generally thought of as the most active on social network sites, we see socioeconomic differences when it comes to Twitter and Tumblr. We also see gender and skill differences on who is on what site."

Hargittai's data is longitudinal; she followed the same people across several years and found that Internet skills have a lag effect. The skills people learned several years ago were still important for using today's sites.

Careful and thoughtful study design can help alleviate potential biases, Hargittai wrote in the study. It's also critical to seek out additional data sources to supplement what is available through information derived solely from active users of sites like Facebook, she said.

More information: "Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites" *The Annals of the American Academy of Political and Social Science* May 2015 659: 63-76, DOI: 10.1177/0002716215570866

Provided by Northwestern University