Biases in platform data and their implications for communication research and other social sciences

In the debate about big data, scholars have raised the issue of a potentially widening gap between "data haves" and "data have nots" (boyd & Crawford, 2012; Driscoll & Walker, 2014). This refers to the considerable amount of resources required for big data research, with regard to data access, storage, handling, or analysis. Likewise, concerns about the unequal distribution of power between platform providers on the one and users (including researchers) on the other hand have been expressed (Andrejevic, 2014). There are thus fears in the research community that big data could further inequality—to our own disadvantage. What is less often addressed, and sometimes even overlooked, is how we ourselves as researchers could contribute to social inequality and biases. I would like to discuss this at the workshop with a focus on studies that use data from social media platforms, a rather common type of big data analysis in communication research.

Data from social media (or other platforms) carry a number of social biases by default. First, Internet access is not distributed evenly around the globe, nor within countries (International Telecommunication Union, 2017). In industrialized countries like Germany or the US, about ten percent of the respective populations currently do not use the Internet (Anderson & Perrin, 2018; Frees & Koch, 2018). And even among users, there are, secondly, diverse ways of making use of online offerings. Research based on data created through such usage thus necessarily reflects social inequalities in access and use. This has been acknowledged before and some commentators accordingly suggest to complement online with offline data (e.g., Murthy, 2008; Orgad, 2009). Yet such research is far from being the norm in big data studies. It seems that there is still enough excitement about research exclusively using data from digital platforms that these are collected, analyzed, and then reported to the research community as well as the wider public and policy makers.

However, when we actually compare big and small data from different sources, we sometimes come to very different conclusions about the social phenomena we investigate. In a study on digital fragmentation (discussed, among others, by Pariser, 2011; Sunstein, 2007, as potentially leading to "filter bubbles" or "echo chambers," respectively), I combined different methods to analyze usage of online content. Time permitting within the workshop format, I would be happy to illustrate how vastly results on the popularity of YouTube videos differ

when we compare YouTube's own scores of popularity with data from a survey or clickstream data.

What a platform first counts and in a second step announces as "popular right now" or "trending" is notoriously intransparent (Gillespie, 2012). As can be expected from skews in online use in general, popularity scores contain biases due to social differences in usage of the respective platform (as reflected, for instance, in my survey and clickstream data on YouTube for age, gender, and education). But respective datasets of platform scores gathered through an API may also be biased for other reasons. Since popularity on a platform may be worth money, numbers of clicks or "likes" may be generated through bots or otherwise manipulated (Karpf, 2012; Lazer, Kennedy, King, & Vespignani, 2014). And platforms may favor content creators that are already popular, leading to a Matthew effect where initial success begets more success (Gillespie, 2010; Shields, 2009).

There are thus numerous reasons to approach data from digital platforms with caution. And I think critical reflection and systematic study of potential biases in big data are essential for the future of this research field. Especially in the social sciences, we usually want to further *understanding of social processes* through our research. But if our data only capture online use, we exclude a considerable part of the population (which is in many countries already disadvantaged due to old age, low income, poor health, disability, or other factors). And even for the onliners, we can often not be sure what our data will over- or undercover because what big platform data actually represent is usually unclear (Mahrt & Scharkow, 2013).

In addition to complementing big data with more traditional research approaches from the social sciences, as suggested above, a fruitful way forward could be to adopt methods from computer science on how to detect and avoid algorithmic discrimination. That data and algorithms may contain "hidden biases" (Crawford, 2013) is, luckily, more and more acknowledged as a major problem in areas such as credit scoring (Chander, 2017) or selection of future employees (boyd, Levy, & Marwick, 2014). I would like to see a similar movement emerge in other fields as well. Ultimately, how well we are able to address the implications of biases in digital data will, in my opinion, decide what contribution big data research in the social sciences can bring to understanding social issues—and to society in general. And it is my hope that we would want to avoid contributing to social inequality through our research.

References

Anderson, M., & Perrin, A. (2018). *11% of Americans don't use the internet. Who are they*. Retrieved from http://www.pewresearch.org/fact-tank/2018/03/05/some-americans-dont-use-the-internet-who-are-they/

Andrejevic, M. (2014). The big data divide. *International Journal of Communication, 8*, 1673-1689. Retrieved from http://ijoc.org/index.php/ijoc/article/view/2161

boyd, d., & Crawford, K. (2012). Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662-679. doi:10.1080/1369118x.2012.678878

boyd, d., Levy, K., & Marwick, A. (2014). The networked nature of algorithmic discrimination. In S. Peña Gangadharan, V. Eubanks, & S. Barocas (Eds.), *Data & discrimination: Collected essays* (pp. 43-57). n. p.: Open Technology Institute.

Chander, A. (2017). The racist algorithm? *Michigan Law Review, 115*(6), 1023-1045. Retrieved from http://repository.law.umich.edu/mlr/vol115/iss6/13

Crawford, K. (2013, April 1). The hidden biases in big data. *Harvard Business Review*. Retrieved from https://hbr.org/2013/04/the-hidden-biases-in-big-data

Driscoll, K., & Walker, S. (2014). Working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication, 8*, 1745-1764. Retrieved from http://ijoc.org/index.php/ijoc/article/view/2171

Frees, B., & Koch, W. (2018). ARD/ZDF-Onlinestudie 2018: Zuwachs bei medialer Internetnutzung und Kommunikation. *Media Perspektiven, n. v.*(9), 398-413. Retrieved from https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2018/0918_Frees_Koch_01.pdf

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), 347-364. doi:10.1177/1461444809342738

Gillespie, T. (2012). Can an algorithm be wrong? *Limn, n.v.*(2). Retrieved from http://limn.it/can-an-algorithm-be-wrong/

International Telecommunication Union. (2017). *ICT facts and figures 2017*. Retrieved from: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society, 15*(5), 639-661. doi:10.1080/1369118x.2012.665468

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203-1205. doi:10.1126/science.1248506

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20-33. doi:10.1080/08838151.2012.761700

Murthy, D. (2008). Digital ethnography. An examination of the use of new technologies for social research. *Sociology, 42*(5), 837-855. doi:10.1177/0038038508094565

Orgad, S. (2009). How can researchers make sense of the issues involved in collecting and interpreting online and offline data? In A. N. Markham & N. K. Baym (Eds.), *Internet inquiry. Conversations about method* (pp. 33-53). Los Angeles, CA: Sage.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London, UK: Penguin.

Shields, M. (2009). YouTube plays partner. *MediaWeek, 19*(11), 7. Retrieved from http://www.adweek.com/news/technology/youtube-plays-partner-111638

Sunstein, C. R. (2007). *Republic.com 2.0.* Princeton, NJ: Princeton University Press.