Big Data Ethics and Politics

Special Issue of *Social Science Computer Review (SSCR)*

***SSCR* Volume 38, No. 1**

February, 2020 (print version)*
* All articles published in 2018 on Sage Online First.

Guest Editors

Wenhong Chen
University of Texas at Austin, USA

Anabel Quan-Haase
Western University, Canada

The two guest-editors contributed equally to the editing of the special issue and the introduction. Names are listed in alphabetical order.

Table of Contents

science research such as data representativeness and sampling biases. Many big data studies rely on traces of people's behavior on social media platforms such as opinions expressed through Twitter posts. How representative are such data? Whose voices are most likely to show up on such sites? Analyzing survey data about a national sample of American adults' social network site usage, this paper examines what user characteristics are associated with the adoption of such sites. Findings suggest that several sociodemographic factors relate to who adopts such sites. Those of higher socioeconomic status are more likely to be on several platforms suggesting that big data derived from social media tend to oversample the views of more privileged people. Additionally, Internet skills are related to using such sites, again showing that opinions visible on these sites do not represent all types of people equally. The paper cautions against relying on content from such sites as the sole basis of data to avoid disproportionately ignoring the perspectives of the less privileged. Whether business interests or policy considerations, it is important that decisions that concern the whole population are not based on the results of analyses that favor the opinions of those who are already better off.

Big Data and the illusion of choice: Comparing the evolution of India's Aadhaar and China's Social Credit System as technosocial discourses / Saif Shahin & Pei Zheng

Abstract: India and China have launched enormous projects aimed at collecting vital personal information regarding their billion-plus populations — and building the world's biggest datasets in the process. However, both Aadhaar in India and the Social Credit System in China are controversial and raise a plethora of political and ethical concerns. The governments claim that participation in these projects is voluntary, even as they link vital services to citizens registering with these projects. In this study, we analyze how the news media in India and China — crucial data intermediaries that shape public perceptions on data and technological practices — framed these projects since their inception. LDA topic modeling suggests news coverage in both nations disregards the public interest and focuses largely on how businesses can benefit from them. The media, institutionally and ideologically linked with governments and corporations, show little concern with violations of privacy and mass surveillance that these projects could lead to. We argue that this renders citizens structurally incapable of making a meaningful "choice" about whether or not to participate in such projects. Implications for various stakeholders are discussed.

Artificial Intelligence and inclusion: Formerly gang-involved youth as domain experts for analyzing unstructured Twitter data / William R. Frey, Desmond U. Patton, Michael B. Gaskell. & Kyle A. McGregor

Abstract: Mining social media data for studying the human condition has created new and unique challenges. When analyzing social media data from marginalized communities, algorithms lack the ability to accurately interpret offline context, which may lead to dangerous assumptions about and implications for marginalized communities. To combat this challenge, we hired formerly gang-involved young people

as domain experts for contextualizing social media data in order to create inclusive, community-informed algorithms. Utilizing data from the Gang Intervention and Computer Science Project—a comprehensive analysis of Twitter data from gang-involved youth in Chicago—we describe the process of involving formerly gang-involved young people in developing a new part-of-speech tagger and content classifier for a prototype natural language processing system that detects aggression and loss in Twitter data. We argue that involving young people as domain experts leads to more robust understandings of context, including localized language, culture, and events. These insights could change how data scientists approach the development of corpora and algorithms that affect people in marginalized communities and who to involve in that process. We offer a contextually-driven interdisciplinary approach between social work and data science that integrates domain insights into the training of qualitative annotators and the production of algorithms for positive social impact.

Inferring public opinion from social media, the citizen's perspective. Authors: Elizabeth Dubois, Anatoliy Gruzd, & Jenna Jacobson

Abstract: Journalists increasingly use social media data to infer and report public opinion by quoting social media posts, identifying trending topics, and reporting general sentiment. In contrast to traditional approaches of inferring public opinion, citizens are often unaware of how their publicly available social media data is being used and how public opinion is constructed using social media analytics. In this exploratory study based on a census-weighted online survey of Canadian adults (N=1,500), we examine citizens' perceptions of journalistic use of social media data. We demonstrate that: (1) people find it more appropriate for journalists to use aggregate social media data rather than personally identifiable data; (2) people who use more social media are more likely to positively perceive journalistic use of social media data to infer public opinion; and (3) the frequency of political posting is positively related to acceptance of this emerging journalistic practice, which suggests some citizens want to be heard publicly on social media while others do not. We provide recommendations for journalists on the ethical use of social media data and social media platforms on opt-in functionality.

Constructing a public narrative of regulations for Big Data and analytics: Results from a community-driven discussion / James Popham, Jennifer Lavoie, & Nicole Coomber

Abstract: This paper reports on community perspectives about the regulation of municipality-led Big Data initiatives developed through an exploratory, deliberative democracy-informed approach. While analytics hold great promise for policy design and service delivery improvements, their mythologized nature may elicit a blind faith in empirical outcomes, leading to misrepresentation or omission of marginalized populations. Scholars have begun pointing to public consultation as a means of avoiding these challenges, suggesting that a truly "smart city" should vet potential Big Data polices through the community in order to identify locally-relevant concerns. The Big Data in cities: Barriers and benefits symposium, held in May of 2017, took a deliberative democracy approach designed to contribute toward a mid-sized southern Ontario city's regulatory framework for data aggregation and mobilization. Approximately 100 self-

selected participants (primarily public advocates) attended a two-day symposium that featured a series of presentations designed to introduce critiques to and strategies for the implementation of Big Data initiatives. Participants also engaged in several facilitated roundtable discussions during the symposium, and their transcribed conversations served as the data for this study. Thematic analysis identified three recurrent concerns: publicly vetted data ethics; consultation and literacy practices; and regulatory frameworks. The public consultation process employed by this study produced results that reflect critiques raised in other academic papers.

When digital trace data meets traditional communication theory: Theoretical/methodological directions / Sujin Choi

Abstract: This study suggests one direction of theoretical and methodological coupling of communication research with the digital trace data, utilizing its differences from the traditional social science approach (e.g., sampling vs. population, normal distribution vs. power–law distribution, generalization vs. simulation, deductive vs. inductive, and perceived vs. actual). We propose specific examples of i) combining communication research with trace data methodologically and theoretically ; ii) collaborating with linguistic psychology complemented with the automated content analysis and natural language processing techniques; and iii) creating new theoretical inquiries by configuring the granular level of interactivity and underlying dynamics, observing the longitudinal change of interactions, and discovering the neglected presence of outliers and the invisibles. We expect the direction suggested by this study contributes to deepening our understanding of human communication behavior.