Getting Data for CSCW Research

Shagun Jhaver
shagun.jhaver@rutgers.edu
Rutgers University
New Brunswick, NJ, USA

Kiran Garimella
kg766@comminfo.rutgers.edu
Rutgers University
New Brunswick, NJ, USA

Munmun De Choudhury
munmund@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Christo Wilson
cbw@ccs.neu.edu
Northeastern University
Boston, MA, USA

Aditya Vashistha
adityav@cornell.edu
Cornell University
Ithaca, NY, USA

Tanushree Mitra
tmitra@uw.edu
University of Washington
Seattle, WA, USA

ABSTRACT

This panel will bring together a group of scholars from diverse methodological backgrounds to discuss critical aspects of data collection for CSCW research. This discussion will consider the rapidly evolving ethical, practical, and data access challenges, examine the solutions our community is currently deploying, and envision how to ensure vibrant CSCW research going forward.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

data collection; ethics; data access

ACM Reference Format:
Shagun Jhaver, Kiran Garimella, Munmun De Choudhury, Christo Wilson, Aditya Vashistha, and Tanushree Mitra. 2023. Getting Data for CSCW Research. In *Computer Supported Cooperative Work and Social Computing (CSCW '23 Companion), October 14–18, 2023, Minneapolis, MN, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3584931.3608440

1 OVERVIEW

In recent decades, Computer Supported Cooperative Work and Social Computing (CSCW) researchers have made extensive efforts to empirically understand how individuals, communities, and societies engage with online technologies. A crucial aspect of such efforts is to obtain high-quality data that allows researchers to extract valid contributions. CSCW researchers use a wide variety of data collection methods such as interviews, surveys, workshops, system deployments, and social media log queries. This panel will reflect on what drives researchers to make these data collection choices, the pros and cons of each approach, the new challenges we are facing in data collection, and how we, as a community, should address such challenges.

Some questions that may animate this panel discussion are:

- What data do we use today in our research? Has this changed over time? If so, why? Have our ways of collecting this data changed over time? If so, why?
- A lot of CSCW research has relied on easily available or obtainable datasets. For example, much of the research using social media logs has focused on Reddit and Twitter as their datasets were traditionally more easily available. Does this compromise the usability of our research outputs or bias our field's view? What avenues do we have to study platforms that have been traditionally understudied because of data access issues?
- One crucial way to obtain useful CSCW data is to collaborate with the industry. What makes such data-focused collaborations successful? What are the ethics of collaborating (or not) with the industry? How do we independently do critical audit work while not being completely antagonistic to our industry partners?
- Social media sites such as Twitter and Reddit are making it increasingly challenging to easily and/or affordably collect social media data by deprecating official Application Programming Interfaces (APIs) or making them completely unavailable. Are researchers changing the way they work and the research questions they ask because of these developments? How can we navigate this challenge? For example, how practical is it to consider developing inter-university data repositories to incentivize CSCW research? How could or would such efforts run in parallel with social media platforms' own Terms of Service-related contractual agreements that bar scraping data?
- What are the ethical, legal, and practical ways to future-proof our research for a post-API future? User-driven data donations for research have been on the rise recently. Are these models scalable to replace our existing API-based models of research?
- Can we find ways for data donation models to complement different data collection methods, such as interviews, surveys, and workshops, in providing a comprehensive understanding of people's engagement with digital technologies?
- How can the CSCW community collaborate to develop standardized approaches for data collection, ensuring comparability and generalizability across studies?
- What potential biases or limitations are associated with different data collection methods in CSCW research, and how can researchers mitigate them?

CSCW '23 Companion, October 14–18, 2023, Minneapolis, MN, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0129-0/23/10.
https://doi.org/10.1145/3584931.3608440

- How can we foster convergence between the qualitative and quantitative 'sides' of CSCW researchers? How do we enable systems where they can benefit from each other's work?
- CSCW researchers are increasingly working with individuals from marginalized and under-represented groups as research participants. How can we improve CSCW researchers' access to these groups? What special responsibilities do we have as researchers when relying on these groups for our data?
- The current CSCW scholarship is heavily skewed toward people in the West. How do we build datasets that authentically represent people and interactions in non-Western contexts?
- Since what we have outlined above are not just problems in CSCW, how can interdisciplinary collaborations and knowledge sharing contribute to addressing the challenges in data collection for CSCW research?

2 PANEL STRUCTURE

The panel discussion will follow a structured format designed to facilitate insightful conversations and engage both the panelists and the audience.

- It will commence with brief five-minute talks by the panelists, providing an opportunity for each expert to share their valuable insights and experiences within this domain. These talks will introduce the panelists' work, establishing a foundation for further discussion.

- Following the talks, the panel participants will pose thought-provoking questions to the other panelists. This phase aims to identify common themes among the panelists' work and address any critical aspects that may have been overlooked. This interactive exchange will encourage collaboration and exploration of synergies among the panelists' perspectives.

- Subsequently, the panel will open the floor to the audience, allowing them to actively participate by posing their own questions. This interactive session will provide a platform for diverse viewpoints and foster a dynamic dialogue between the panelists and the audience.

3 RELATED EVENTS

- The Post-API Conference: Social media data acquisition after Twitter
- The D.A.R.E. Workshop, a part of the AAAI ICWSM 2023 conference at Limassol, Cyprus

4 BIOGRAPHICAL SKETCHES

Shagun Jhaver is an Assistant Professor in the School of Communication and Information at Rutgers University. Before joining Rutgers, he was a postdoctoral scholar at the University of Washington. In 2020, he received his PhD in Computer Science from the Georgia Institute of Technology. Shagun's research focuses on improving content moderation on digital platforms. He studies how internet platform design, technical affordances, and moderation policies can address societal issues such as online harassment, misinformation, and the rise of hate groups. His work aims to instill fairness and transparency in platforms' communications with end-users.

Kiran Garimella is an Assistant Professor in the School of Communication and Information at Rutgers University. Dr. Garimella's research deals with using large-scale data to tackle societal issues such as misinformation, political polarization, or hate speech. Prior to joining Rutgers, Garimella was the Michael Hammer postdoc at the Institute for Data, Systems and Society at MIT. Before joining MIT, he was a postdoc at EPFL, Switzerland. He received his PhD at Aalto University in Finland and worked in the industry prior to his PhD.

Munmun De Choudhury is an Associate Professor of Interactive Computing at Georgia Tech where she directs the Social Dynamics and Well-Being Lab. Dr. De Choudhury is best known for laying the foundation of a new line of research that develops computational techniques towards understanding and improving mental health outcomes, through ethical analysis of social media data. To do this work, she adopts a highly interdisciplinary approach, combining social computing, machine learning, and natural language analysis with insights and theories from the social, behavioral, and health sciences.

Christo Wilson is an Associate Professor in the Khoury College of Computer Sciences at Northeastern University, a faculty associate at the Berkman Klein Center for Internet & Society at Harvard University, and an affiliate member of the Center for Law, Innovation and Creativity at Northeastern University School of Law. His research seeks to investigate the sociotechnical systems that shape our lives using a multi-disciplinary approach. Along with co-PIs from Northeastern, Christo is currently working to launch the National Internet Observatory. This NSF-funded project seeks to gather data about the online habits of a large, representative panel of US residents and then make it available to qualified researchers around the world.

Aditya Vashistha is an Assistant Professor of Computing and Information Science at Cornell University. His research focuses on the design and evaluation of technologies that contribute to the socioeconomic development of underserved communities in low-resource environments. His current work aims to combat misinformation and hate speech in low-income communities and design responsible AI systems in high-stakes settings. He received a Ph.D. in Computer Science and Engineering from the University of Washington, where his dissertation was recognized with the William Chan Memorial Dissertation Award and the WAGS/ProQuest Innovation in Technology Award.

Tanu Mitra is an Assistant Professor at the University of Washington, Information School. Her research focuses on studying and building large-scale social computing systems to understand and counter problematic information online. Her work spans auditing online systems for misinformation and conspiratorial content, understanding digital misinformation in the context of the news ecosystem, unraveling narratives of online extremism and hate, and building technology to foster critical thinking online. Her work employs a range of interdisciplinary methods from the fields of human computer interaction, data mining, machine learning, and natural language processing.