Computational Social Science: CSCW in the Social Media Era

Scott Counts
Microsoft Research
Redmond WA 98052
counts@microsoft.com

Munmun De Choudhury
Microsoft Research
Redmond WA 98052
munmund@microsoft.com

Jana Diesner
School of Library & Information Science
University of Illinois
Champaign IL 61820
jdiesner@illinois.edu

Eric Gilbert
School of Interactive Computing
Georgia Tech
Atlanta GA 30308
gilbert@cc.gatech.edu

Marta C. Gonzalez
Massachusetts Inst. of Technology
Cambridge MA 02139
martag@mit.edu

Brian Keegan
Northeastern University
Boston MA 02115
bkeegan@neu.edu

Mor Naaman
Jacobs Technion-Cornell
Innovation Institute
Cornell Tech
New York NY 10011
mor.naaman@cornell.edu

Hanna Wallach
School of Computer Science
University of Massachusetts
Amherst, MA 01003
wallach@cs.umass.edu

Abstract

With the widespread proliferation of social media tools such as Facebook and Twitter, the CSCW community has seen a growing interest among researchers to turn to records of social behavior from blogs, social media, and social networking sites, to study human social behavior. This nascent area, that has begun to be referred to in various research circles as "computational social science", provides an exciting and promising development in the CSCW community. This panel brings together a host of researchers with varied disciplinary perspectives to investigate the potential and current state-of-the-art, as well as forthcoming challenges, practical and methodological issues that engender this emerging topic.

Author Keywords

Computational social science; social media; social networks; data analysis

ACM Classification Keywords

J.4 Computer Applications: Social and Behavioral Sciences.

Introduction

As an area of inquiry, over the past several years, CSCW has investigated the important role of computing technologies in supporting collaborative activities and their coordination in organizations, groups, and communities. Beyond the study of the design and use of tools and techniques of groupware, researchers have also been interested in examining their psychological,

Key Points of Discussion

Methodology. Given the significant reliance on data and methods as well as the cross-disciplinary perspective of the computational social science area, can adequate evaluation standards be developed that allow research in this topic to be open to examination by the broader scientific community?

Online vs. Offline Behavior. Under *what* circumstances does online behavior generalize to offline behavior, or *when* do online platforms manifest a rather idiosyncratic aspect of our interactions with others?

Digital Divide. What are the limitations of population-scale analyses of human behavior, as observed in a notable body of research in computational social science?

Privacy and Ethics. Being essentially human subjects research at scale, what are the risks to user privacy and our ethical responsibilities as researchers? How do IRBs and similar institutions need to adapt their norms to this shifting landscape?

social, and organizational effects—an aspect that has become increasingly predominant with the emergence of social platforms on the web. This emerging research area, often referred to as "computational social science" hold tremendous promise in allowing, for the first time, large-scale measurements of human behavior in the context of use of online social platforms.

A fair number of publications at CSCW in the recent past have focused on a variety of topics in the computational social science: including studies on online communities, on social network structure and ties, on information roles of individuals on social media, on emotional expression of individuals in these platforms, collaboration and coordination in virtual teams and peer-production systems, wellbeing and technology use during personal and societal shifts, and so on.

Perhaps the most stimulating aspect of the area of computational social science is that it allows the opportunity to study populations and collectives and their social interactions, due to the availability, scale, and temporal granularity of the online social data. In order to do so, the area leverages techniques and insights from multiple disciplines (such as sociology, psychology, network science, computer science, statistics, and so on) into making sense of detailed observations of online activities of millions of people in a near real-time fashion and in naturalistic settings. However, there are also pitfalls. The panel will explore some of these issues and concerns beyond opportunities of scientific pursuit, organized by the themes in the following section.

Themes and Key Challenges

Methodology
Computational social science is a relatively new area when compared to other established disciplines. The interdisciplinary aspect of it demands that the scientific

rigor in this area is improved. For this to happen, and to facilitate scientific exchange, models and methods need to be documented when published and presented, as well as samples of data made available for calibration wherever guidelines allow.

Additionally, researchers perhaps need to recognize a common ground in terms of the methodological approach that should be adopted to tackle specific classes of problems. For instance, when is large-scale statistical characterization sufficient to understand a social phenomenon? Related, how can such large-scale statistical approaches adequately filter the "signal" from the profuse "noise" apparently in online social platforms, and still derive meaningful observations? Which problems require experimentation, and which ones benefit most from a grounded qualitative approach? Overall, how can mixed method approaches to computational social science problems be deemed successful and a contribution to the scientific pursuit in their own right?

Online vs. Offline Behavior
Computational social science provides a fascinating environment, in its own right, to study our online social interactions. However perhaps there are distinctive ways in which they may be different from the rich, real-time face-to-face interactions that happen offline.

On the one hand, the ability for computational social science to discover and validate aspects of human behavior in the physical world through studies of online activity, would indicate the emergence of new sources of data to understand known sociological and psychological phenomena. On the other hand, differences would reveal how behavior may deviate under previously uninvestigated conditions. How can we systematically traverse this terrain involving carefully addressing differences and similarities between aspects of online versus offline behavior?

Digital Divide and Generalizability of Findings
Related to the previous theme, is the issue around the generalizability of findings of human behavior as observed on online social platforms. While making population-scale generalizations is an attractive investigation to many computational social scientists (e.g., in politics; finance and economics; public health), and for CSCWers to understand how technology is increasingly closely enmeshed in our lives, it is of paramount importance to understand under *what* conditions online behavior generalizes to behavior in the general population.

Consequently, the findings about our social and psychological behavior derived from online activities need to be interpreted with caution. At the same time, computational social science as an area needs to frame concrete sampling methodologies and selection techniques that can circumvent issues around generalization of findings. The panel will discuss how statistical approaches may be derived using insights from survey approaches as well as statistical methods.

Privacy and Ethics of Human Subjects Research
Spontaneity and persistence of online conversations, richness of information exchange, diversity of social ties and geography make computational social science problems a great environment for studying individual-centric behavior as well, beyond collective phenomena. However, researchers need to careful in understanding the implications of their findings in securing the privacy of the user, and because online environments may stimulate greater levels of self-disclosure. Revealing otherwise non-apparent personal information from online social platforms to undesirable parties can have serious repercussions in how they are used in guiding decisions, ranging from employment, to healthcare—aspects which are questionable from a privacy as well as ethical perspective. The panel will discuss issues

facing IRBs and similar institutions in this kind of research, e.g., privacy, consent, and confidentiality.

Panelists
Scott Counts
Scott Counts is a Senior Researcher at Microsoft Research working in the area of computational social science. Scott studies technology-mediated social phenomena both at large scales like in today's social media, and at small scales, down to the level of neurons. Areas of interest include emotional expression, "cognition of user" in modern information environments, and applications in health and finance.

Munmun De Choudhury (Co-moderator)
Munmun De Choudhury is a researcher at Microsoft Research, Redmond. Her research interests are in computational social science. By combining data mining, HCI, and social science, Munmun's research attempts to decipher social behavior, as manifested online. She has been a recipient of the Grace Hopper Scholarship; recognized with an IBM Emergent Leaders in Multimedia award; a finalist of Facebook Fellowship; and winner of two best paper honorable mention awards from ACM. Earlier, she was a postdoctoral fellow at Rutgers, and obtained a PhD in Computer Science from Arizona State University in 2011.

Jana Diesner
Jana Diesner is an Assistant Professor at the iSchool at the University of Illinois at Urbana-Champaign, and a faculty affiliate at the Department of Computer Science. Jana conducts research at the nexus of network science, natural language processing and machine learning. She develops scalable methods and technologies for extracting information about socio-technical networks from text data and considering the content of information for network analysis. She got her PhD from Carnegie Mellon, School of CS.

Panel Format

Before the panel. Questions around the four key themes would be publicized via social media platforms, e.g., via Twitter's #CSCW2014 hashtag 2-3 days before and during the conference, in order to gain audience participation.

During the panel

(a) ~1 minute introductions of the panelists: about 5-7 minutes of the panel session.

(b) 1-2 minute position statements of panelists. Total 10 minutes of the session.

(c) Moderators pose questions on the four themes to panelists; panelists pick sides and provide argument. A total of 60 minutes of time allocated for the panel session, with approximately 15 minutes per theme.

(d) Audience vote on the panelists whose standpoints they support, using an online survey tool. Panelists summarize the discussion, and open up for questions from audience as well as social media backchannels. This would last for about 30 minutes.

Eric Gilbert (Co-moderator)
Eric Gilbert is an Assistant Professor in the School of Interactive Computing at Georgia Tech. He joined the Georgia Tech faculty in 2011 after finishing a Ph.D. in CS at Illinois. Dr. Gilbert leads the comp.social lab, a research group that focuses on building and studying social media. His work is supported by grants from Yahoo!, Google, the NSF and DARPA. Dr. Gilbert has also founded several social media sites, and has received four best paper awards and two nominations from ACM's SIGCHI. His research has recently been featured in The Wall Street Journal, The Atlantic, MIT's Technology Review and on CNN and NPR.

Marta C. Gonzalez
Marta is Gilbert Winslow Career Development Assistant Professor of Civil and Environmental Engineering at MIT, joint with Engineering Systems (ESD) and the Operations Research Center (ORC). Before coming to MIT she was a research associate at the Center for Complex Network Research (Barabási Lab) at Northeastern University, that she joined right after her PhD in Computational Physics at Stuttgart Universität, Germany, in 2006. Marta's research interests and direction are mainly aimed at advancing the understanding of the laws and principles that characterize human behavior and result in collective social phenomena. Current research explores human mobility patterns using mobile phone communication; data mining combined with geographic information systems (GIS), and urban transportation models.

Brian Keegan
Brian Keegan is a computational social scientist and post-doctoral research fellow at Northeastern University. He received his PhD in 2012 from Northwestern University and his dissertation examined the history, structure, and dynamics of Wikipedia's coverage of breaking news events. He draws upon theories and methods from network science, computer-mediated communication, and organizational studies to understand high-tempo knowledge work, online political communication, and network forms of organization and innovation. His research has been published in the American Behavioral Scientist, CSCW, ICWSM, WWW, and IEEE Social Computing.

Mor Naaman
Mor Naaman is an associate professor at Cornell Tech in New York City, where he is part of the Jacobs Technion-Cornell Innovation Institute. At Cornell, Mor continues to direct the Social Media Information Lab that he founded at Rutgers University, where he served as an assistant professor since 2008. He is also a co-founder and Chief Scientist at Seen.co, a startup founded to make sense of the real-time web and social media. Mor's research applies multidisciplinary methods to gain new insights about people and society from social media data. Previously, Mor worked as a research scientist at Yahoo! Research Berkeley, and received a Ph.D. in Computer Science from Stanford University. He is a recipient of a NSF Early Faculty CAREER Award, research awards from Google, Yahoo!, and Nokia, and three best paper awards.

Hanna Wallach
Hanna Wallach is an assistant professor in the School of Computer Science at UMass Amherst. She is one of six core faculty members involved in UMass's newly-formed computational social science research initiative. Hanna's primary research goal is to develop new machine learning methods for analyzing the structure, content, and dynamics of complex social processes, such as the US political system, the US patent system, and free/open source software development communities. In her not-so-spare time, Hanna (or Logistic Aggression, as she is better known) likes to put on roller skates and hit people really, really hard.