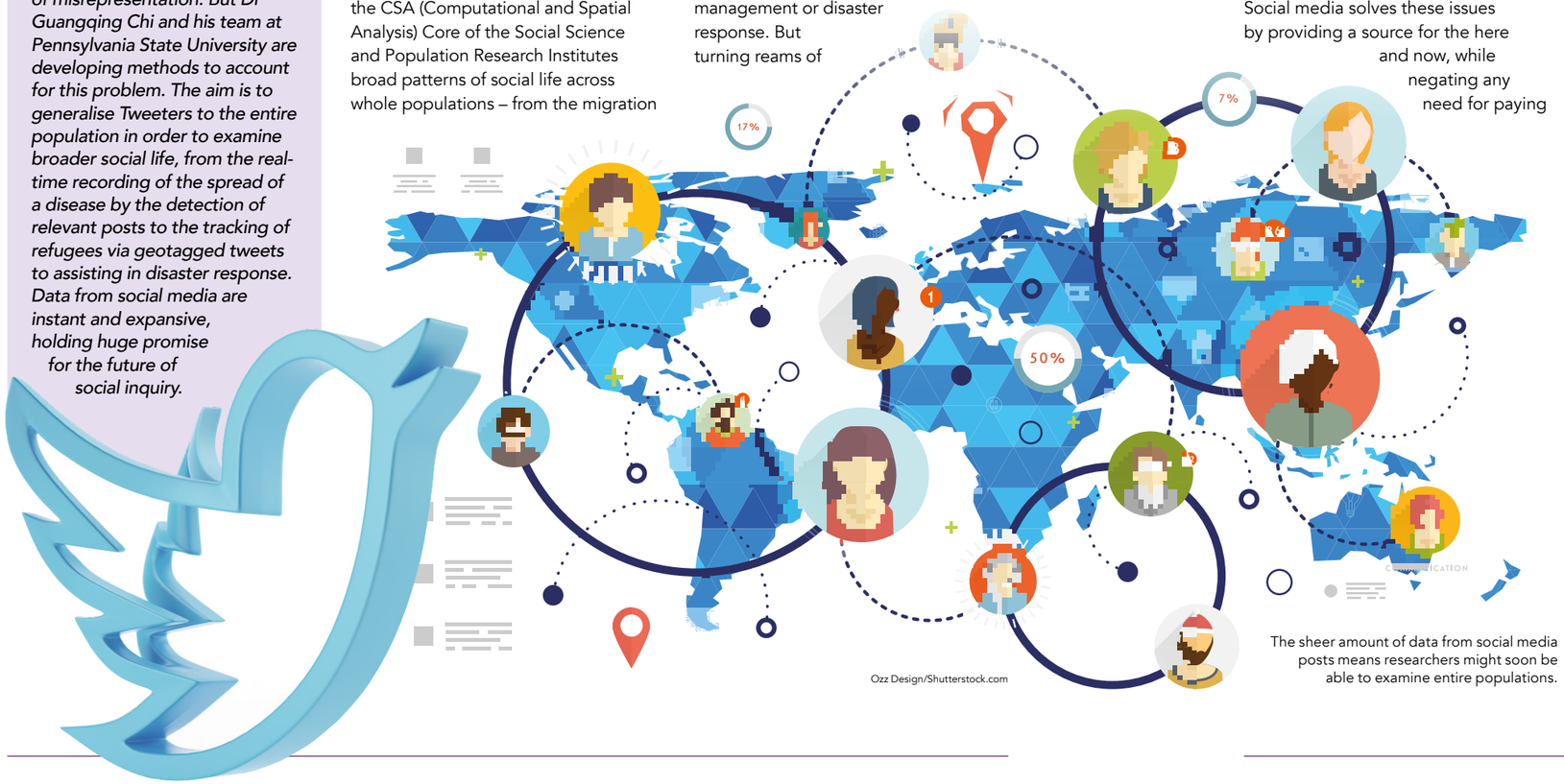
# #SocialScience: Mining Twitter for Social and Behavioural Research

*A disaster strikes, somewhere in the world. Within minutes there are millions rushing to their screens, letting out their thoughts and feelings. This makes Twitter, along with other social platforms, a rich and timely resource for social and behavioural research. Surprisingly, however, it is a lot less utilised in the social sciences compared with other fields. Tweeters don't reflect the general population, making social scientists wary of misrepresentation. But Dr Guangqing Chi and his team at Pennsylvania State University are developing methods to account for this problem. The aim is to generalise Tweeters to the entire population in order to examine broader social life, from the real-time recording of the spread of a disease by the detection of relevant posts to the tracking of refugees via geotagged tweets to assisting in disaster response. Data from social media are instant and expansive, holding huge promise for the future of social inquiry.*

A tweet has a limit of 280 characters, enough for an average of three sentences. But for Dr Guangqing Chi, Professor of Rural Sociology and Demography at Penn State University, tweets can say a thousand words. They can tell a story of family relocation when a user posts from a new region after a natural disaster. Another can show belief in a fake news story, along with the age, gender, race and ethnicity of the person misled.

When collected en masse, such tweets reveal to Dr Chi and his colleagues at the CSA (Computational and Spatial Analysis) Core of the Social Science and Population Research Institutes broad patterns of social life across whole populations – from the migration of people to the makeup of cities to the spread of misinformation through particular groups. This is why the CSA Core sees social media as a window into society, and like the windows of the real world, it is a view of the present. Twitter users react as events unfold, making data direct and immediate. This conquers the constraints of the traditional survey, where participants can forget details about the time in question. Instead, real-time data from online social networks allow for real and timely action to meet social need, whether in emergency management or disaster response. But turning reams of posts into a picture of wider societal phenomena requires sophisticated methods of social research. The CSA Core is at the forefront of building these tools, and the projects they pursue are as engaging as they are effective.



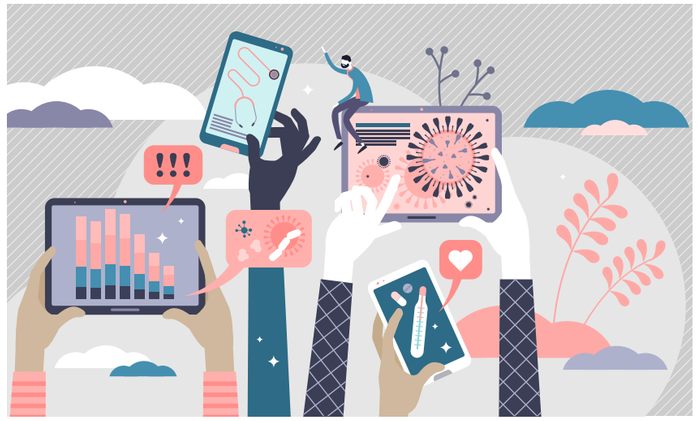

Ozz Design/Shutterstock.com

The sheer amount of data from social media posts means researchers might soon be able to examine entire populations.

## TRACKING THE TRENDS

At the click of a button, social media turns our lives into data for research. Minutes after an event erupts, people race onto Twitter and other online platforms, to share with the world their emotions and experiences. What was unfeasible with the conventional survey becomes a real possibility in the age of the Internet. If we want to know the percentage of the population in fear as news breaks out about a terrorist attack, it would be far too expensive to do fieldwork on foot. Likewise, a survey of popular sentiment on a highly publicised trial will cease to be reliable if respondents vaguely recount their reaction to the verdict. Elsewhere, a probe into public attentiveness towards official health advice will fail to fulfil its purpose if by the time of its completion, the research is rendered obsolete by the spiralling scope of the pandemic.

Social media solves these issues by providing a source for the here and now, while negating any need for paying participants or funding legwork. But it doesn't stop there. Unlike 'data by design', where researchers pick participants and plan questions, the Twitter posts and the Facebook profiles are 'found in the wild'. The organic nature of online communities makes for spontaneous ties and tensions, whether in friendship, political debate, or criticism of the culture. This banishes the bias where subjects pander to the researcher, which can otherwise distort data in the process.




VectorMine/Shutterstock.com

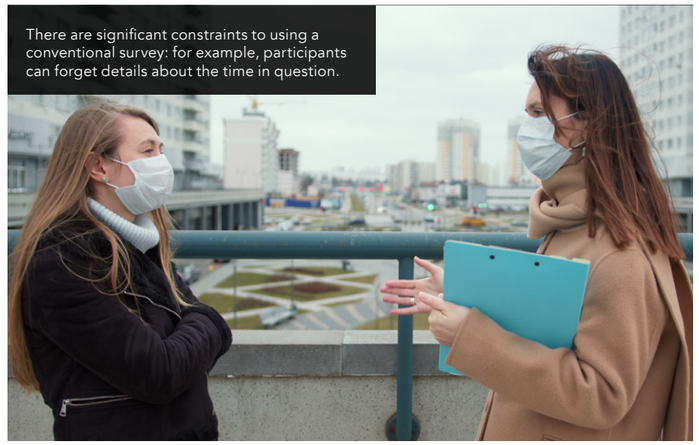
How do people react to official health advice regarding Covid-19? Social media can reveal this in real time, informing how public messages may be improved.

Social media is more than the sum of its posts. Every user has a unique path of online activity, which can be traced through time in social research. The unfolding information can uncover how people are primed for certain content, or their movement across the world from their changing tagged location. The immediacy of online engagement turns the tracking of single users into a living document. Indeed, tomorrow's measurements of social media could display the dynamics of migration even before the release of official records. But the biggest and most basic value of mining these platforms is the sheer amount of data put into the hands of the social scientist. There are 2.5 billion users of online social networks, and with the right methods to wrestle with this massive material, they can be used to examine entire populations. The reach of this research is therefore vast in scope, so it's unsurprising to have drawn such eager attention – from the sphere of public health to political science. But to see clearly through this window into our society, an institution must first wipe the mist that obscures our view. The CSA Core at Penn State is committed to this task, refining statistical techniques and ways of weighting data, to reveal how Twitter represents the wider world.

***There are 2.5 billion users of online social networks, and with the right methods to wrestle with this massive resource, they can be used to examine entire populations.***

## MAKING WAVES IN METHODOLOGY

Scrolling through a top trend on Twitter does not reflect the general population. Users are skewed towards the young, the middle-class, and the male, which is far from representative of an entire town or city. Meanwhile, many members of social platforms are a demographic mystery, and without a way to unmask their identities, such data becomes useless and is omitted from research. For social scientists, these difficulties have


There are significant constraints to using a conventional survey: for example, participants can forget details about the time in question.

KayStock/Shutterstock.com

***With real-time social media data, official messages could be swiftly improved through immediate intelligence on public reaction.***

diminished confidence in social surveys as a means of drawing conclusions about the broader public. But for researchers at the CSA Core, these studies are not a lost cause. The group seeks to build on their success in detecting the 'who' and 'where' of Tweeters, improving techniques for assigning gender, age, race, and location. Analysts can then find bias that is buried in the data, where certain groups are under or over-represented. Findings can be weighted and brought into the right balance, making Twitter samples representative of whole populations. However, the terrain of Twitter research is scattered with mines. Researchers must deal with the complexities of slang, sarcasm, and esoteric communication – from hashtags to emoticons to acronyms. Meanwhile, some Twitter users are not human but 'bots' – automated programs designed to advertise and misinform. Finding what is real and what is fake in these massive datasets adds another hurdle to an already arduous challenge.

The research group uses multiple methods to identify Twitter users. One approach asks, "what's in a name?", as researchers detect gender and ethnicity in usernames and screen names. Entries are compared against public name databases, from phonebooks to Facebook, to find the most common characteristics. With a match rate of up to 96%, name measures have made huge strides in the investigative process. However, the group is refining the method by adding several rounds of data cleaning. Initials can now be taken into account and special characters removed like emojis. Another technique employed by the group resembles a scene from a sci-fi film. Profile pictures are scanned using facial recognition technology, picking up gender and race, or making age estimates. This helps to identify users when names provide few clues. Moreover, tracking users' movements is made all the more achievable with the geotag. These mark tweets with the place from which they were sent, but only when users select it. Barely 1% of tweets are currently being geotagged, seemingly drawing the limit on researching user location. But the group engages with ways to infer user whereabouts, by probing public profiles and social connections. The group has been using every geotagged tweet from January 2014 to test its representativeness. This is just part of the fifty terabytes of data already collected by the CSA Core.

**PRESENT PROJECTS AND FUTURE PROMISE**

A pandemic has been declared as a virus sweeps the globe. Information is now invaluable for saving people's lives. The number of people diagnosed with the coronavirus disease (COVID-19) has been widely recorded since the start of the outbreak. However, effects on human behaviour and social dynamics have received little study so far. This is a glaring gap when such findings would allow us to track public awareness of the disease, the following of social distancing, attitudes toward lockdown, and anti-Asian sentiment, among other important factors. Indeed, with real-time data from social media platforms, official messages could be swiftly improved through immediate intelligence on public reaction. Furthermore, geotagged tweets reporting symptoms and confirming cases have been used to predict new contractions of the virus, up to five days in advance. The size of the sample is also so large that researchers can zero in on specific neighbourhoods, and the finest of time intervals, when scouring social networks for mentions of the disease. The Core is breaking new ground in this area with their online dashboard (https://webgis.pop.psu.edu/COVID-19), which can track these geotagged tweets across time and space.

The capacity of the CSA Core to gain immense insight from tracking Twitter is clearly demonstrated by their range of research projects. The team is currently compiling geotagged tweets from a six-year period to renew existing methods of detecting demographics. These techniques will be tested for reliability by estimating internal migration within Puerto Rico after Hurricane Maria and comparing findings against records from the Internal Revenue Service of the U.S. Moreover, grants have been obtained for other exciting studies – from an inquiry into inclination towards fake news to the food, energy, and water choices of Alaskans to the regional differences in opinion towards climate change. Social science stands at the foot of a mountain, where an expanse of evidence could be opened for analysis. The CSA Core is scaling the summit and seeking new opportunities to apply and extend their expertise. After all, if tweets from one user can say a thousand words, a million can make for research that effectively examines our social world.



# Behind the Research

## Dr Guangqing Chi

**E:** gchi@psu.edu **T:** +1 814 826 4686 **W:** https://aese.psu.edu/directory/gfc5047

## Research Objectives

Prof Chi gathers social media data in order to analyse real-time responses to social problems and phenomena.

## Detail

Guangqing Chi
112E Armsby
University Park
PA 16802
USA

**Bio**

Dr Guangqing Chi is Professor of Rural Sociology and Demography and Director of the Computational and Spatial Analysis Core at Pennsylvania State University. His research is focused on socio-environmental systems, aiming to understand the interactions between human populations and built and natural environments.

**Funding**

National Science Foundation (#1823633 and #1541136), and Social Science Research Institute, Population Research Institute, and Institute for Computational and Data Sciences, Pennsylvania State University

**Collaborators**

- Dr Junjun Yin, Computational and Spatial Analysis Core, Pennsylvania State University
- Prof Jennifer Van Hook, Department of Sociology and Criminology, Pennsylvania State University
- Prof Eric Plutzer, Department of Political Science, Pennsylvania State University
- Prof Heng Xu, Kogod School of Business, American University

## References

Auman-Bauer, K. (2020). Penn State researchers evaluating Twitter data during pandemic. [online] PennState. Available at: https://bit.ly/3go5wZA [Accessed 26 May 2020].

Auman-Bauer, K. (2019). Seed grants awarded to projects using Twitter data. [online] PennState. Available at: https://bit.ly/2M3jySo [Accessed 26 May 2020].

Chi, G., Yin, J., Hook, J.V., Plutzer, E., Xu, H. (2019). The Generalizability of Twitter Data for Population Research. Population Association of America. Available at: https://bit.ly/2X0xR0y [Accessed 26 May 2020].

## Personal Response

***Do you think your work, particularly the 'global dashboard', could be useful for public health researchers?***

"Twitter data can be greatly useful for public health research, if used correctly. By following individual Twitter users over space and time, the data have a potential to allow researchers to track public health-related phenomena such as the COVID-19 outbreak and public awareness, the effectiveness of policy interventions, and more. As the data can be streamed in near real time, they provide timely information to help decision makers and planners evaluate the situation and make data-informed precision decisions. That said, the data need to be used in a rigorous and careful way. For example, the demographic characteristics of the Twitter users versus the targeted population must be considered."

metamorworks/Shutterstock.com