## Bringing the Neighborhood In: Exploring the Inequalities Within and Outside of Big Data

*A position paper for Specialist Meeting on Human Dynamics in the Mobile Age, August 11-12 2015*

Author: Joseph Gibbons, Department of Sociology, San Diego State University

Big data presents tremendous promise for the study of social networks. However, an enduring challenge lies with the impact of the physical world on these data. For example, while much has been said about how one uses big data to measure and scale networks, there has been insufficient attention placed on the intersection of "hyper local" networks captured by big data with more physical networks. In this position paper, I turn the discussion towards the neighborhood effects that impact the formation of the networks found in big data. Such an exploration can reveal much as to how biases in physical space translate into cyberspace.

My own work has focused extensively on the implications of residential segregation and class division between neighborhoods, including onto what I term 'community connection' - a composite of mutual trust with neighbors, working collaboratively with neighbors, and a feeling of belongingness to a neighborhood -- all key underpinnings of social networks. I argue that the divisive character of segregation carries class implications which inhibit residents from forming community connection in their neighborhoods. One example of this effort includes research I conducted with Professor Tse-Chuan Yang at SUNY Albany recently published in the journal *Urban Affairs Review*. Using data from the Southeastern Pennsylvania Household Health Survey and the American Community Survey, we found evidence suggesting deep biases in community connection by race and class. For example, a person who is black will likely have stronger community connection if they reside in a mostly black or nonwhite neighborhood than they would living in a white neighborhood. My subsequent work has also found evidence to suggest that nonwhite neighborhoods have weaker community connection overall compared to white communities.

The division in the networks in physical communities by race and class, as personified by community connection, raises questions as to the divisions that may exist within big data. First, there is the well established, nagging issue of the 'digital divide'. My own analysis of the 2010 Southeastern Pennsylvania Household Health Survey found that 24% of those surveyed in the Philadelphia metropolitan area 'never' use the internet; of that, around 47% do not do so because of lack of access or cost barriers. Drawing on 2006-2010 American Community Survey Data, as well as the Philadelphia data, I found that of those who report not having internet, 20.62% more lived in neighborhoods with at least 25% poverty than those who did not, suggesting a geographic component to this divide. These stark figures do not necessarily mean these neighborhoods lack networks, instead, networks of the disadvantaged may not be 'picked up' by big data. Second, among those that do have the internet, there are questions as to the impact that physical environment carries onto the generation of these data. Recent research by Shelton, Poorthuis, and Zook, to be published in an upcoming issue of *Landscape and Urban Planning*, took a random sample of tweets originating in the racially divided city of Louisville, KY. They found that those who tended to tweet in the white part of town were very unlikely to ever tweet in the predominately black neighborhoods, while those who tended to tweet in the black side of town were comparatively far more likely to tweet in the white neighborhoods. This raises important questions for how the information collected by big data is being disseminated among new media users. Take for example the center of Human Dynamics in the Mobile Age's (HDMA) maps of tweets pertaining to wild fires in San Diego: are San Diegans in the predominately white northern part of the city reading the tweets from their peers in the mostly Latino southern sections of the city?

How then do we study big data in ways which meaningfully account for these physical divisions? For starters, it would be of great benefit to look more closely at the people generating these data. With the HDMA center's wild fire research, where do these twitter users reside? is the information from their tweets disseminated within their neighborhoods? are communities across racial or class lines reading their tweets? On the surface, addressing these questions presents a challenge as we often know little about the people who either use twitter or don't use twitter. This is where traditional social science methods can supplement big data analysis. One straightforward way would be to interview a random sample of twitter users. In-depth interviews would do much to unpack how the information from a users tweets spread across their digital and physical networks. More ambitiously, it would be of great use to conduct a random sample survey of both twitter users and non twitter users in the San Diego area to better understand this diffusion. Along similar lines, I am in the planning stages of a project using the 2014 wave of the Southeastern Pennsylvania Household Health Survey which has questions on new media usage as it pertains to social networks and health service usage.

Another worthwhile way to build on big data research is to change the questions that we are asking. We need to go beyond just looking at where people are talking about things, or what they talk about. Instead, we also need to ask how the location affects the nature of the conversation itself. One especially pertinent area that could be explored is with mapping racist attitudes captured by a platform like twitter. This itself is not new, as maps of the recent #Ferguson twitter campaign demonstrated. However, with the technology HDMA has at its disposal, there is great potential for studies which directly examine the dispersion of tweets at the neighborhood level. This could allows us to see how neighborhood characteristics directly shape the content of big data. Are we more likely to see racist tweets in neighborhoods that are far from nonwhite areas or near nonwhite areas? In addition, what is the nature of the discourse of these tweets as it pertains to geography? Shelton, Poorthuis, and Zook's paper for example found deep racial bias in the discourse of tweets across Louisville. Those who tended to tweet in the white side of Louisville were likely to make racist comments about the black part of town in ways not seen from those who tended to tweet in the black part of town.

In summary, an incorporation of neighborhood networks to big data studies can reveal the full potential of these data. In particular, closer attention must be paid to how neighborhood divisions in community connection, as well as access to new media, shape how this 'big data' is created and in turn disseminated across these communities. Answers to these questions would tells us much as to how the big data reflects the social behavior of the local populace, as well as how we the researchers should study it.