The Role of Human Geography in Collective Intelligence

BRENT HECHT, Northwestern University
LOREN TERVEEN, University of Minnesota

1. INTRODUCTION

Collective intelligence paradigms were initially developed to address problems that do not have an explicit geographic footprint (e.g. image labeling [von Ahn and Dabbish 2004]). However, in recent years, researchers and practitioners have used collective intelligence approaches to tackle geographic challenges. This is manifest in particular in *physical crowdsourcing systems* (e.g. TaskRabbit and related sharing economy platforms like UberX), *peer production communities* dedicated to describing geographic phenomena (e.g. "geowikis" [Priedhorsky and Terveen 2008] such as OpenStreetMap), and the consideration of *geotagged social media* (e.g. geotagged tweets and photos, check-ins).

Geographers and those in the spatial computing and geographic information science (GIScience) communities have long known that "spatial is special" (e.g. [Longley et al. 2010; Goodchild 2001]), or that geographic processes have unique properties that are important to consider in both research and practice. In this abstract, we discuss our recent work that shows that collective intelligence is no exception to the "spatial is special" rule. That is, for a given collective intelligence system, the geography of the application area can have a dramatic effect on the operation and ultimate effectiveness of the system, i.e. *same system + different geographic context = different result.*

We focus below on a set of spatial principles that emerge from the sub-field of geography known as *human geography* and these principles' role in geographic collective intelligence systems. Human geography, broadly speaking, seeks to gain an understanding of how human processes shape and are shaped by the Earth's surface (e.g. [Fellmann et al. 2007]). We discuss three human geographic principles that our work has found can play an important role in collective intelligence: (1) *spatial homophily* (i.e. "the Big Sort"), (2) *distance decay*, and (3) *structured variation in population density.*

2. HUMAN GEOGRAPHY PRINCIPLES

2.1 Spatial Homophily (i.e. "The Big Sort")

A large percentage of human geography research relates in some way to *spatial homophily*, or the notion that people of specific demographic characteristics – e.g. race, ethnicity, income, religion – tend to live close to one another. As has been highlighted in the popular science book "The Big Sort" [Bishop 2008], this principle has had a large effect on the quality of public schools, the outcomes of elections, income inequality, and any number of other major societal issues.

In our work looking at physical crowdsourcing markets, we have seen evidence that spatial homophily can have a similarly substantial effect in the domain of collective intelligence. For example, in Thebault-Spieker et al. [2017; 2015], we observed that spatial homophily is a major conduit by which demographic biases in TaskRabbit worker recruitment result in analogous demographic biases in service quality for customers. Specifically, we saw that there were substantially fewer 'taskers' (i.e., workers) in neighborhoods with certain demographic characteristics (e.g. low-socioeconomic status, large racial and ethnic minority populations), likely due in part to sign-up requirements (e.g. background checks, the need for a bank account). Coupled with distance decay (see below), this has resulted in TaskRabbit providing substantially worse and more expensive service in these neighborhoods. In other words, in physical crowdsourcing systems like TaskRabbit, spatial homophily means that worker recruitment practices that are more successful with one demographic are likely to lead to more effective service where that demographic tends to live (and in nearby areas). We also observed a related trend in the ride-hailing platform UberX [Thebault-Spieker et al. 2017].

2.2 Distance Decay

Distance decay is a human geographic principle related to spatial homophily (and one that often acts in concert with it) that describes the phenomenon in which, all things being equal, *spatial interaction tends to decrease as distance increases*. This principle is often considered through the use of spatial interaction models, specifically gravity models (e.g. [Stewart 1948; Anderson 2010]). However, this principle need not be modeled to be understood: it explains why, holding all else constant, trade decreases as the distance between two countries increases (e.g. [Koo and Karemera 1991]), why you are more likely to visit a coffee shop closer to your residence than further away (e.g. [Erlander and Stewart 1990; Rodrigue et al. 2017]), and a large variety of other similar phenomena.

In the geographic collective intelligence domain, we have observed that distance decay is a critical principle to consider for both researchers and practitioners. In the physical crowdsourcing context, we and others have seen that distance plays a key role in worker decision-making processes like willingness to complete a task and the price a worker charges for a task (e.g. [Alt et al. 2010; Musthag and Ganesan 2013; Thebault-Spieker et al. 2015]). The spatial interaction patterns seen in physical crowdsourcing are, however, simply a manifestation of physical-world spatial interaction patterns that have been observed for decades. More surprising are results that show that these physical world patterns persist in collective intelligence processes *that require no physical world movement*. Hecht and Gergle [2010] defined a spatial interaction spectrum on which all geographic crowdsourcing processes exist that is defined by two endpoints: (1) "you have to be there" processes, in which travel to the site about which one is contributing is required (e.g. taking a geotagged photo) and (2) "flat Earth" processes, in which no travel is required at all (e.g. making a spelling correction on a geotagged Wikipedia article). Hecht and Gergle [2010] and Hardy et al. [2012] showed that even for mostly "flat Earth" processes like editing Wikipedia, spatial interaction still dramatically decreases with increasing distance, with people much more likely to edit Wikipedia articles about places near to them than those that are far away. Moreover, in ongoing work, we recently identified that for non-local geographic crowdsourcing contributions, "flat Earth" processes may even have roughly the same "friction of distance" (i.e. strength of negative association with distance) as "you have to be there" processes.

2.3 Structured Variations in Population Density

Human geographers frequently study phenomena associated with structured variations in population density. These variations are most noticeable in the form of the urban/rural "divide", but also occur within metropolitan areas. Across a number of research projects, we have observed substantial effects on collective intelligence processes due to population density variation. By and large, this has resulted in collective intelligence systems that work well in urban areas, are less effective in North American suburban areas, and fail in rural areas. We found this to be the case in a wide variety of contexts including peer production communities [Johnson et al. 2016], sharing economy platforms [Thebault-Spieker et al. 2015; Thebault-Spieker et al. 2017], collective intelligence components of location-based games [Colley et al. 2017], and geotagged social media [Hecht and Stephens 2014].

For example, with regard to peer production, we have observed that Wikipedia articles describing rural areas are of lower quality, have less content written by humans (rather than bots), and likely contain a smaller percentage of local knowledge than their urban counterparts [Johnson et al. 2016]. We have also observed the same phenomenon in OpenStreetMap [Johnson et al. 2016]. In fact, our results suggest that much of the “peer produced” content describing rural areas is not actually peer produced at all: it is bot-generated.

We have hypothesized that, in certain cases, it may be *nearly impossible* for collective intelligence approaches to succeed in rural areas. Our urban-rural findings above can likely be attributed to at least two causes: (1) under-participation in collective intelligence communities in rural areas and (2) the ratio between the amount of work that needs to be done and the number of people local to the work region. Research on increasing participation can address the first cause (e.g. [Halfaker et al. 2014]), but a solution to the second is much less tractable.

More generally, we believe that collective intelligence may fail in rural areas for what we call “per-area” processes, and that this is also true of research projects and applications that seek to use data from collective intelligence systems [Johnson and Hecht 2016]. We have described geographic processes associated with the number of people in a place as “per-capita” processes, e.g. the number of sports teams in each city (many local sports teams require large local audiences). Per-area processes are on the other side of the spectrum: they are independent of the number of people in a location, e.g. natural history topics. In order for collective intelligence to succeed in rural areas in the case of per-area processes – for example, writing good Wikipedia content about rural areas’ natural history – at least one of two highly unlikely events must occur: (1) rural participation must be orders of magnitude higher than in urban areas and/or (2) urban users must contribute information about rural areas at tremendous scales, which, as per distance decay, is quite implausible.

Lastly, while more work needs to be done examining geographic collective intelligence systems in suburban geographies, we have completed early research on this topic. Specifically, we found that for UberX, population density is a significant negative predictor of wait times in the Chicago metropolitan area [Thebault-Spieker et al. 2017]. The primarily implication of this result is that suburbs have much longer wait times. We also saw a similar effect in TaskRabbit [Thebault-Spieker et al. 2015].

3. DISCUSSION

The above research has focused on identifying and explaining problematic geographic variation in the effectiveness of collective intelligence systems using core human geographic principles. However, a small body of research also has recognized the potential for improving performance by actively leveraging these principles into core system design. In particular, Sen et al. [2015] found that incorporating key geographic principles into their models resulted in peer-production crowdsourcing being more effective for a specific natural language processing task (semantic relatedness estimation). We are hopeful that future work that takes similar approaches will lead to similar outcomes. Along the same lines, while we have focused on three geographic principles here, many others are worthy of consideration. For instance, in Thebault-Spieker et al. [2017], we discuss how the notion of *mental maps* can play an important role in physical crowdsourcing.

Finally, we note that although our work has focused on collective intelligence contexts, geographic principles like those considered here will likely affect most systems in “geographic human-computer interaction” [Hecht et al. 2013] in a similar fashion. This has led us and our colleagues Johannes Schöning and Isaac Johnson to discuss a tentative “First Law of Geographic HCI”: *An algorithm or sociotechnical system that works one way in one area will not necessarily work the same way in another area* (akin to the First Law of Geography [Tobler 1970], this is more of a “rule of thumb”).

4. ACKNOWLEGEMENTS

The authors would like to thank our many co-authors on the work we have discussed above, in particular Jacob Thebault-Spieker, Shilad Sen, and Johannes Schöning.

REFERENCES

Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. New York, NY, USA: ACM, 319–326.

Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, Urs Kramer, and Zahid Nawaz. 2010. Location-based Crowdsourcing: Extending Crowdsourcing to the Real World. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. NordiCHI '10. New York, NY, USA: ACM, 13–22.

James E. Anderson. 2010. *The Gravity Model*, National Bureau of Economic Research.

Bill Bishop. 2008. *The Big Sort: Why the Clustering of Like-minded America is Tearing Us Apart*, Houghton Mifflin Harcourt.

Stanley D. Brunn, Jack F. Williams, and Donald J. Zeigler. 2003. *Cities of the World: World Regional Urban Development* Third Edition., Rowman & Littlefield Publishers.

Ashley Colley et al. 2017. The Geography of Pokémon GO: Beneficial and Problematic Effects on Places and Movement. In ACM *SIGCHI '17: 35th International Conference on Human Factors in Computing Systems*. CHI '17. Denver, CO.

Sven Erlander and Neil F. Stewart. 1990. *The Gravity Model in Transportation Analysis: Theory and Extensions*, VSP.

Jerome D. Fellmann, Arthur Getis, and Judith Getis. 2007. *Human Geography: Landscapes of Human Activity* 9th ed., McGraw-Hill.

Eric Gilbert, Karrie Karahalios, and Christian Sandvig. 2008. The network in the garden: an empirical analysis of social media in rural life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08.

Mike Goodchild. 2001. A Geographer Looks at Spatial Information Theory. In *COSIT '01: 5th International Conference on Spatial Information Theory*. Morro Bay, CA.

Aaron Halfaker, R. Stuart Geiger, and Loren G. Terveen. 2014. Snuggle: Designing for Efficient Socialization and Ideological Critique. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: ACM, 311–320. DOI:https://doi.org/10.1145/2556288.2557313

Darren Hardy, James Frew, and Michael F. Goodchild. 2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science* (2012), 1–22. DOI:https://doi.org/10.1080/13658816.2011.629618

Brent Hecht et al. 2013. Geographic human-computer interaction. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. ACM Press, 3163. DOI:https://doi.org/10.1145/2468356.2479637

Brent Hecht and Darren Gergle. 2010. On The "Localness" of User-Generated Content. In *CSCW '10: 2010 ACM Conference on Computer Supported Cooperative Work*. Savannah, GA, 229–232.

Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *ICWSM '14: Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, Michigan.

Isaac Johnson et al. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. In *CHI '16: 34th International Conference on Human Factors in Computing Systems*.

Isaac Johnson and Brent Hecht. 2016. Structural Causes of Bias in Crowd-derived Geographic Information: Towards a Holistic Understanding. In *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.

Won W. Koo and David Karemera. 1991. Determinants of World Wheat Trade Flows and Policy Analysis. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* 39, 3 (November 1991), 439–455.

Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 1603–1612.

Paul A. Longley, Mike Goodchild, David J. Maguire, and David W. Rhind. 2010. Cartography and Map Production. In *Geographic Information Systems and Science*. Wiley.

Mohamed Musthag and Deepak Ganesan. 2013. Labor dynamics in a mobile micro-task market. In *CHI '13*. Paris, France.

Reid Priedhorsky and Loren Terveen. 2008. The Computational Geowiki: What, Why, and How. In *CSCW '08: 2008 ACM Conference on Computer Supported Cooperative Work*. San Diego, CA.

Jean-Paul Rodrigue, Claude Comtois, and Brian Slack. 2017. *The Geography of Transport Systems* 4 edition., New York: Routledge.

Shilad Sen et al. 2015. Towards Domain-Specific SR: A Case Study from Geography. In *IJCAI '15: 24th International Joint Conference on Artificial Intelligence*. Buenos Aires, Argentina.

John Q. Stewart. 1948. Demographic Gravitation: Evidence and Applications. *Sociometry* 11, 1/2 (February 1948), 31–58.

Jacob Thebault-Spieker, Loren G. Terveen, and Brent Hecht. 2015. Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. New York, NY, USA: ACM, 265–275.

Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Towards a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. *Transactions on Computer-Human Interaction (In press)* (2017).

Tobler, Waldo R. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (1970): 234–40.