1-1-2013

Big Data and the Fabric of Human Geography

Sandra González-Bailón
University of Pennsylvania, sgonzalezbailon@asc.upenn.edu

Big Data and the Fabric of Human Geography

Abstract
Digital data tracking what we do, the time and place of our actions, and the chains of interdependence that link those actions together, help us draw a richer picture of human geography as it unfolds in its multiple layers. This commentary briefly illustrates the type of maps and models we can build with that data as well as some important challenges that arise from their complexity and unsolved validity concerns.

Keywords
big data, digital technologies, human geography, validation

Disciplines
Communication | Social and Behavioral Sciences

Big data and the fabric of human geography

Sandra González-Bailón
University of Pennsylvania, USA

Abstract

Digital data tracking what we do, the time and place of our actions, and the chains of interdependence that link those actions together, help us draw a richer picture of human geography as it unfolds in its multiple layers. This commentary briefly illustrates the type of maps and models we can build with that data as well as some important challenges that arise from their complexity and unsolved validity concerns.

Keywords

big data, digital technologies, human geography, validation

Digital technologies have transformed the ways we observe and experience the world. Every time we are online (communicating, buying, or playing) we leave behind digital traces, the source for the vast amounts of information we have come to call 'big data'. These two words offer a convenient, although somewhat misleading, label: the size of the data, often aggregating millions of records, can indeed overwhelm usual storage capacities and analytical frames; but rather than the size, it is the breadth and depth of those records (the level of detail they reveal about human activity) that has transformed the way we can approximate the world and measure its constant pulse.

This change is particularly relevant for the study of human geography. We can now build better models and more dynamic maps of how people interact with places, how those places are perceived, and how they come to be. In much the same way as topographic maps give us a sense of altitude, the social spaces we inhabit have an invisible relief carved by the constant stream of social interactions: some urban areas rise in visibility and popularity while others sink in the valleys of forgotten places; some cities become the hubs in the flow of human mobility, others glitter weakly in their role as tributaries; even whole continents shrink in their capacity to compete with an assignment of prominence that has less to do with the underlying geography and more with social arrangements.

The data we generate with our digital transactions and online interactions can help us build better maps of the world—in the cartographic sense but also (and especially) in the more metaphorical sense of maps as models that can go beyond description and help us scrape the surface of appearances. Digital data (i.e. data generated through the use of information and communication technologies) are helping us identify the drivers of human activity as well as infer which patterns make us predictable and which result from randomness. This brief commentary aims to outline a few examples of how big data—understood as the richer set of observations tracking what we do, when and where we do it—are helping us advance our theories of human geography and improve our maps and models to better disentangle the multiple, nested layers of social life.

Blending perceptions with realities

Social topographies are more intangible than the physical ones, but they are becoming more visible and easier to track down. Big data offer a new scale and compass to map the unchartered territories of social life. They are also offering a fertile middle ground where different disciplines (spanning from the social to the physical sciences) can converge to try to climb outside the box of conventional thinking. Fomenting this sort of inventiveness is necessary because the fabric of human geography is becoming increasingly complex: it results from a constantly active pattern of mobility and communication, including content creation that feeds back into the pattern. To capture this complexity, models and maps need to be less static than they used to be; they have to blend perceptions with ontologies, merge in meaningful ways the structure of the world as it is with how it is perceived and portrayed.

Minimizing the bias of subjectivity without undermining its relevance in how we perceive the spaces that surround us is not an easy task; it is probably one of the greatest challenges that geographers face today. The world exists as it is, but we navigate it on the basis of subjective interpretations, as when we opt to go to a holiday destination because of what other visitors said about it before. Social distance is not measured in miles or even money; it is often measured in the currency of influence and peer pressure. When Jane Jacobs, the classic urban sociologist, said that swarming streets help cities self-regulate, she was tapping into the effects that social interactions have in how we perceive spaces and behave in them, encouraging others to perform in similar ways (Jacobs, 1961). And when William Whyte, another classic urban sociologist, noted that what attracts people most is other people, he was also tapping into the importance of social interactions to understand what makes certain public spaces more successful than others (Whyte, 1980). These two approaches, now intrinsic to how cities are conceived by many urban planners, suggest that social interactions weave an invisible network that makes the places we live in more vibrant. Social interactions both reinforce and erode spatial and social boundaries: people might not venture beyond certain borders because the network of social activity falls behind—and with it, the social capital it helps sustain.

Digital technologies have but accelerated these dynamics and the cumulative effects that accompany them—after all, much urban life (like fashion) is governed by herding behavior. Urban mobility patterns are so prone to routine that they even allow prediction (Gonzalez et al., 2008). At the very least, the paths of our daily movements (more traceable now because of mobile technologies) offer evidence that our social world is smaller than the geography on which we erect it. Popular places that bring most people together (the peaks in the landscape of urban life) are few, at least according to Foursquare check-in behavior (Brown et al., 2013). Perhaps it is that the places we want the world to know we frequent are few, and we hide revelations about the less glamorous places we also visit. This sort of behavior helps us uncover a relevant, yet elusive, layer of information: perceptions of space and its significance to public life.

Mapping, modeling, and harnessing diversity

Cities that were once captured by similar representations (i.e. street maps) can now be unfolded in myriad layers, each representing the biographical choices (and constraints) of the many residents, and all giving a sense of how diverse a city can be when looked through different eyes. Although capturing this heterogeneity was always possible in principle, in practice adding the impressionistic layers with which we all dress physical spaces was difficult to do at scale. The city

details brought up by mental representations has long intrigued social psychologists (Milgram, 1977) and inspired urban planners (Lynch, 1960). Mental representations of cities are maps with overlaps and omissions, a window to the common grounds that we all recognize and to the urban pockets that only a few are familiar with: a visual manifestation of the silent dialogue in which, unconsciously, most residents engage with the spaces they inhabit. But only now, with digital data, can cities become smarter and integrate those perceptions into a grander (albeit bubbly, designed from the bottom-up) scheme of things. This means that researchers can analyze on a larger scale how perceptions match (or not) planned urban interventions (Quercia et al., 2013). Above all, digital data helps find holes in what would otherwise be depicted as a continuous, seamless reality: these are holes in attention, in visibility, in the actual (i.e. how people make use of opportunities) as opposed to the potential (i.e. how those opportunities distribute according to some grand plan). Omissions and misconceptions in planning can now find their place in maps with improved accuracy.

The analysis of large-scale communication networks, for instance, reveals a spatial distribution of economic opportunities that cannot be reduced to geography and socioeconomic variables. Access to economic opportunities is shaped by location but also (and mostly) by the structure of social interactions. This idea has been at the core of economic sociology for a while (Granovetter, 1973, 1983) but only now can be tested at the level of entire populations (Eagle et al., 2010). The models and maps that networks draw of social interactions reveal that diversity matters: social and economic prospects depend on being able to tap into diverse networks of acquaintances and friends, people who can open the doors to novel information and opportunities. Big data shows that having that sort of diversity embedded in social networks is positively related to economic development (Eagle et al., 2010). This finding has implications not only for how the distribution of wealth and resources are mapped, but also for how redistributive policies are targeted.

Diversity can benefit communities in other ways: it sits at the foundation of the so-called ‘wisdom of crowds’, that is, the aggregation of opinions or decisions that, taken individually, can be biased or off target but which, when averaged, help reach better decisions because they draw from specialized expertise (Page, 2007). The idea is that, as long as mistakes are independently made and distributed randomly, they will cancel each other out; what is left is the signal of good judgment. The implication is that the accuracy of geographic scholarship can be improved by drawing from the massive inputs digital users produce; the flip side is that this will only be the case if there is no systematic bias in the characteristics of those entering the inputs, for instance, in their socioeconomic backgrounds or locations. As long as diversity exists, however, the sum will tend to be greater than the parts; that is, the aggregated outcomes will be better and irreducible to the linear aggregation of the individual decisions that made those outcomes emerge in the first place.

Diversity and crowdsourcing can also help model the world in other ways. For example, when search and rescue operations need to comb vast geographical areas in search of specific data points, digital technologies can help deconstruct those maps into more manageable pieces and put them back together, having been thoroughly scanned by multiple people. That reconstruction will contain an additional layer of information based on a compound measure of human judgment that will be less prone to error. When time is critical, speed is as important as accuracy. Again, screening geospatial data can be completed much faster if digital technologies

are used to mobilize a large number of people and effectively use them as a distributed network of sensors (Goodchild, 2011; Pickard et al., 2011). This bottom-up approach relies on local information and mechanisms for its aggregation, and it has been used successfully in crisis situations to react in near-real-time to fast-evolving events (Meier, 2012). The same digital tools that can result in distorted maps if data input is systematically biased can also yield the best representation of events that would otherwise be invisible.

Validity, potentials, and challenges

Being able to model and map subjectivity as a function of time and space, as recent methods to mine opinions and sentiments from online communication do (Dodds and Danforth, 2010; Dodds et al., 2011;Kamvar and Harris, 2011; Kramer, 2010), is in itself a great achievement, blending two worlds that not too long ago fell short of being irreconcilable. Validity, though, looms as the great challenger of the depictions of the world that big data can yield (Goodchild, 2007). What models and visualizations of big data really tell us about social life is a crucial question that is not always explicitly answered. For instance, mapping the spatial distribution of positive emotions, or the frequency with which certain words are mentioned in online communication, does not tell us much about the correspondence of those patterns with the generative mechanisms. This is related to the problem of data bias (i.e. who is being over- or underrepresented) but also to the issue of construct validity. Measuring things just because they can be measured does not make them interesting or relevant.

To show that the patterns identified with digital traces help understand the off-line world, systematic validation strategies need to be put in place. Search data, for instance, have been linked to influenza epidemics by comparing the spatial distribution of search volumes with data on physician visits (Ginsberg et al., 2009); likewise, self-declared voting behavior in online networks has been compared with actual voting records, identifying the spillover effects of social influence through online interactions (Bond et al., 2012). Validation often requires linking online data with off-line information on actual behavior. This helps prevent offering a substantive interpretation of patterns that in fact are an artifact of data collection. Metrics and maps devised on the basis of digital activity should be used as a starting point for further investigation, not as an end product.

Every time the tools we have to observe the world improve, theory needs to catch up, either to rectify mistakes made because of imperfect observations or to blaze new trails that before were out of scope. In this sense, it is difficult to argue that big data are not improving our knowledge of important dimensions of the social world: we have a better viewpoint now of social dynamics that span from the individual to the collective. The relationship between mental maps and perceptions of space; between mobility patterns and urban life; or between social networks and economic development are but three examples of areas where the analysis of big data is already shedding new light. There are many more aspects of human geography, of how people interact with space and with each other, that can be uncovered or painted in better light, thanks to large-scale digital data.

In delivering that promise, there are three big challenges that the analysis of large data sets will present to geographers and, by extension, to all researchers interested in modeling social life. The first is finding the right scale for analysis, both temporal and spatial: better data granularity improves the level of detail in observations but does not solve the problem of finding

the best resolution for the question at hand. Zooming in too close to the highest level of resolution in the data will make patterns disappear; zooming out too much to capture aggregated trends might hide the mechanisms that made those dynamics emerge in the first place. Because most digital data is time-stamped, deciding how to aggregate that activity should be driven by theory, or by a good empirical sense of how patterns vary when the rule for temporal aggregation changes.

The second challenge is to specify the right boundaries for data collection. To make data analysis meaningful, much information needs to be disregarded, as when messages exchanged in online networks are filtered according to key words. Excluding content is akin to adjusting the scope of focus, but if not done carefully, representations of the processes being analyzed will be biased and misleading. When data are so abundant, applying a filter that uses the right parameters can be more important (and difficult) than indulging in data mining.

Finally, a third challenge is how to make a coherent assembly of perceptions (as expressed in the myriad forms of content people publish online) and reality (as mapped using objective standards). Perceptions do not change reality, but they change behavior, and public spaces are particularly malleable scenarios. What one day is a peaceful public square, the next day becomes the epicenter of a battleground with political colors, allegiances, and red lines, as the recent wave of political protests has shown. Space mediates social dynamics in nontrivial ways not just because of the physical characteristics of the space but because of the social significance and consequent use of that space.

Conclusion: big data is here to stay

Even though the richness of big data raises legitimate concerns about privacy and surveillance, the aphorism 'know thyself' could never reach such collective heights before. By helping us trace and reconstruct the fabric of human geography, big data will improve our vantage point to map where we are and where we would like to be. Knowledge is agnostic about use: it can benefit marketing campaigns trying to maximize profit as much (or as little) as policy makers fighting poverty and deprivation. As far as creating knowledge is concerned, we were never in a better position to understand the complexities of social life, of how we relate to each other, and to the environment we shape. This is not an opportunity we will miss.

Funding

This work was partially supported by the JISC project 'Big Data: Demonstrating the value of the UK Web Doman Dataset for Social Science Research' while the author was at the Oxford Internet Institute.

References

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow CA, Settle JE, et al. (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489: 295–298.

Brown C, Noulas A, Mascolo C, and Blondel V (2013) *A Place-Focused Model for Social Networks in Cities*. Paper presented at the IEEE/ASE Conference on Social Computing, Washington, DC.

Dodds PS and Danforth CM (2010) Measuring the happiness of large-scale written expression: songs, blogs, and Presidents. *Journal of Happiness Studies* 11: 441–456.

Dodds PS, Harris KD, Kloumann IM, Bliss CA, and Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PloS ONE* 6(12): e26752.

Eagle N, Macy MW, and Claxton R (2010) Network diversity and economic development. *Science* 328: 1029–1031.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, and Brilliant L (2009) Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–1014.

Gonzalez MC, Hidalgo CA, and Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453: 779–782.

Goodchild M (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211–221.

Granovetter M (1973) The strength of weak ties. *American Journal of Sociology*, 78: 1360–1380.

Granovetter M (1983) The strength of weak ties: a network theory revisited. *Sociological Theory* 1: 201–233.

Jacobs J (1961) *The Death and Life of Great American Cities.* London, UK: Pimlico.

Kamvar SD and Harris J (2011) *We Feel Fine and Searching the Emotional Web*. Paper presented at the WSDM'11, 9–12 February, Hong Kong, China.

Kramer ADI (2010) An Unobtrusive Model of 'Gross National Happiness'. In: *Proceedings of SIGCHI conference on human factors in computing systems*, New York, NY, ACM, pp. 287–290.

Lynch KA (1960) *The Image of the City*. Cambridge, MA: MIT Press.

Meier P (2012) Crisis mapping in action: how open source software and global volunteer networks are changing the world one map at a time. *Journal of Map and Geography Libraries* 8(2): 89–100.

Milgram S (1977) *The Individual in a Social World: Essays and Experiments*. London: Pinter & Martin.

Page SE (2007) *The Difference. How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton, NJ: Princeton University Press.

Pickard G, Pan W, Rahwan I, Cebrian M, Crane R, Madan A, et al. (2011) Time-critical social mobilization. *Science 334*(6055): 509–512.

Quercia D, Pesce JP, Almeida V, and Crowcroft J (2013) *Psychological maps 2.0: a web engagement enterprise starting in London*. Paper presented at the World Wide Web Conference, 13–17 May, Rio de Janeiro, Brazil.

Whyte WH (1980) *The Social Life of Small Urban Spaces*. New York, NY: Project for Public Spaces.

Corresponding author:
Sandra González-Bailón, Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: sgonzalezbailon@asc.upenn.edu