CHAPTER 1.9

On the Value of Digital Traces for Commercial Strategy and Public Policy: Telecommunications Data as a Case Study

ROB CLAXTON, British Telecommunications plc
JON READES, Centre for Advanced Spatial Analysis, University College London
BEN ANDERSON, Centre for Research in Economic Sociology and Innovation, University of Essex, Colchester

Just as information and communication technologies (ICT) and the digital economy are transforming everyday life, so they are transforming our ways of knowing about everyday life. The breadth of social practices that are mediated by digital infrastructure, and thus recorded by digital traces, has not gone unnoticed in the social sciences.[1] Coupled with technological and methodological advances in large-scale data capture, storage, and analysis, transactional data on communication, consumption, leisure, health, work, and education are now routinely collected and can, in principle, be employed for a wide range of analyses.

Clearly, the increased traceability of social networks can enhance our ability to extract actionable insight by analyzing their form, distribution, and structure through digital media. Consequently, an enormous potential to generate important insights and innovation exists within the social sciences through an improved understanding of spatialized social networks (i.e., place-based analyses of social network structures over time). As we will show, these networks have applications in—at the very least—regional development, market research, and infrastructure planning because the structure and spatial distribution of social networks underpins demand (and, consequently, supply or provisioning) as well as provides indicators of well-being, integration, and cohesion.

Of course, the analysis of social networks has been a key part of the sociological and social psychological analysis of group behavior as well as resource and opportunity identification since at least the 1970s.[2] Social interaction has also been implicated in both the diffusion of innovation and the distribution of power and hierarchy within groups.[3] To date, however, such analyses have typically involved a number of individuals or groups that is tiny compared with the general population and with the magnitude of newly available data from digital sources.

However, the marriage of "big data"—datasets containing billions of records—to social science is enabling us to examine social and economic relationships in a new light, leading to the emergence of what some researchers have termed a *computational social science.*[4] Although Rogers foresaw this direction of travel back in the 1980s,[5] until recently social science has largely lacked the tools to perform this type of research: what had been missing were the tools that had been tested on meaningfully large volumes of data and could be applied in a range of analytic domains.

Thanks to its tractability, telecommunications data are now starting to play a crucial role in the emergence

Ben Anderson and Jon Reades wish to acknowledge the support of British Telecommunications plc for components of this research. Additionally, Jon Reades acknowledges the support of the EPSRC (Grant #EP/I018433/1) and of the European Commission (Complexity-NET/FP6 ERANET).

of these tools. Researchers have, for instance, used mobile and fixed-line telephone calls and Instant Messaging logs, as well as Twitter and Facebook data, to speculate on "universal laws" of human friendship and mobility.[6] In this chapter we present results from four studies of British telephone usage that offer a sense of the ways in which computational social science can be used to expand our understanding of social and economic activity.

We begin with a study of UK regions, comparing the "geographies of talk" to their administrative counterparts, before turning to the ways in which social networks reflect underlying problems of deprivation and of access to opportunity. We will then examine derived indicators of globalization from the United Kingdom's most economically vibrant area, Greater South East England, before finally discussing early work on real-time data-driven household classification systems.

ANALYSIS

Advances in ICT give us more choice about how, when, and where to interact with one another. In particular, these technologies support increasingly complex forms of communication at a distance—voice has been supplemented by video, the letter by email, the local pub or restaurant by Facebook and Twitter. This is, arguably, giving rise to more flexible and extensive forms of social and economic interaction. It is not that space has ceased to be relevant—reports of the "death of distance" have proven wildly exaggerated—but that the traditional ways in which space was categorized, delineated, and managed by governments and firms as "commuting zones" or "sales regions" appear unable to keep pace with the increasingly fluid ways in which people are choosing how and where to work, play, or purchase goods and services.

Dynamic and permeable boundaries: Regions and communities

By analyzing the connections among people, households, and firms, we can derive boundaries that better reflect their interactions with the environment—for instance, we can determine whether people living in Northern Wales interact more with their linguistic cohorts in Central and Southern Wales, or with their English compatriots in the larger cities on the "other" side of Offa's Dyke. For government, this is hardly a trivial issue because social interaction will be reflected in other forms of exchange as well: should economic development in North Wales focus on building links with Manchester or with Cardiff? Should transportation planners prioritize East-West or North-South infrastructure investment? Or should planners work against the trend and try to disrupt the entrenched geographic ties that might reinforce a region's structural weaknesses?

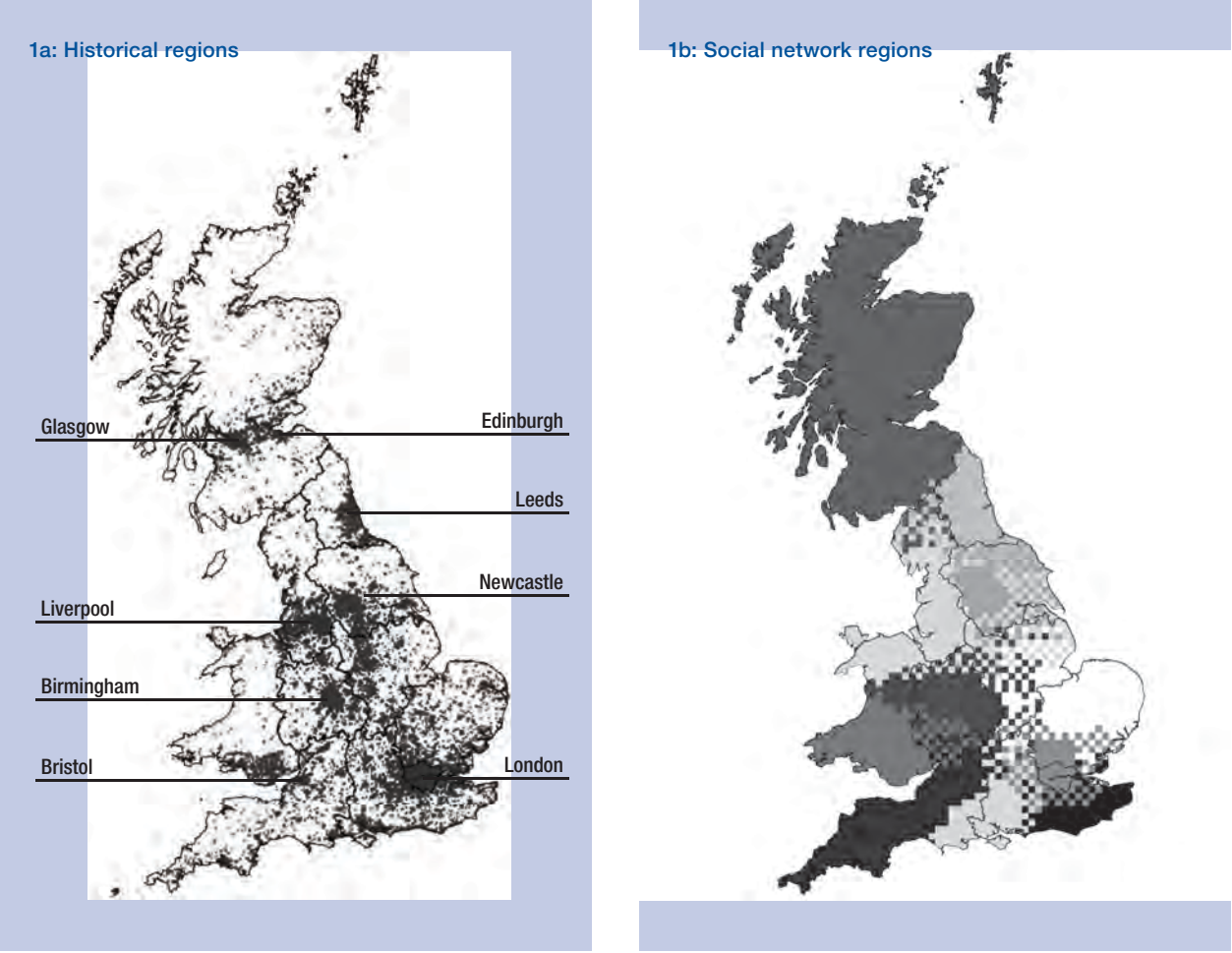
Clearly, it is not within our remit to answer such questions directly, but social network analysis provides us with an important tool with which to investigate the dynamics in play. Ratti et al. built just such a network,[7] deriving 86 million links among more than 20 million numbers made anonymous from an original database of some 8 billion telephone calls. We then examined the resulting network to see if natural communities could be identified in the data, where a community is characterized by relatively dense within-group links and proportionally fewer out-group connections. For example, many social networks fall naturally into two "communities": a group of work colleagues, many of whom will know one another; and a group of friends, many of whom will also know one another, with relatively little overlap between the two.

Our research scaled this approach up to the level of the entire country. The findings appear to capture both deep historical continuities dating back hundreds of years as well as more recent changes in mobility and economic development (see Figure 1). Simply by virtue of its size, the network region that is coterminous with Scotland is particularly visible, but the Devon/Cornwall, Kent, East Anglian, and North East regions are also notable for their overlap with existing administrative regions. But these results are largely to be expected: geography alone dictates that people at the extreme northern and southern ends of the United Kingdom will tend to interact more intensively with others who also fall within these traditional regions. As geographers have often noted: "everything is related to everything else, but near things are more related than distant things."[8]

More intriguing, because they are altogether less expected, are the regions in Wales and in the vicinity of London. Figure 1b highlights areas of overlap: rather than neat lines around geographical or social features, the figure emphasizes areas where these networks seem to pull in more than one direction. The wide belt surrounding London—to which we return later in this chapter—accentuates the extent to which families and firms in these areas are an integral element of a wider "London" that is not visible on administrative maps. The figure also underlines the challenges, faced by the local authorities in these areas, of having one foot in a major world city and the other in a semi-rural economy.

The division of Wales into three distinct subregions, each of which is anchored to a major urban center, further underscores the importance of economic activity and proximity to socioeconomic networks. Thus, although Wales itself has a strong linguistic and cultural heritage, in social network terms it seems to be relatively more important to Northern Wales that it interact with Liverpool and Manchester than with Cardiff, far off in the South. The same applies to Central Wales and Birmingham. In this case, we see little of the "mixing" that exists in the area around London, where several regions overlap geographically. The results for Wales line up nicely with those of Nielsen and Hovgesen,[9] who

Figure 1: Regions of the United Kingdom: Different views



Source: This work is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown. Additionally, the OAC Classification used is subject to Crown Copyright protection.
Note: This shows the centers of non–Countryside Output Areas to give a sense of how population is distributed across Great Britain. The black boundary lines denote the official Government Office Regions.

Source: Based on Ratti et al., 2010.
Note: The shading denotes the core regions of the United Kingdom arising from analysis of communication interactions.

used ward-level commuting data in a similar type of analysis and found clearly demarcated regions whose boundaries were connected both to the accessibility of infrastructure such as the main East-West routes (A5 and A458) and to major employment centers.

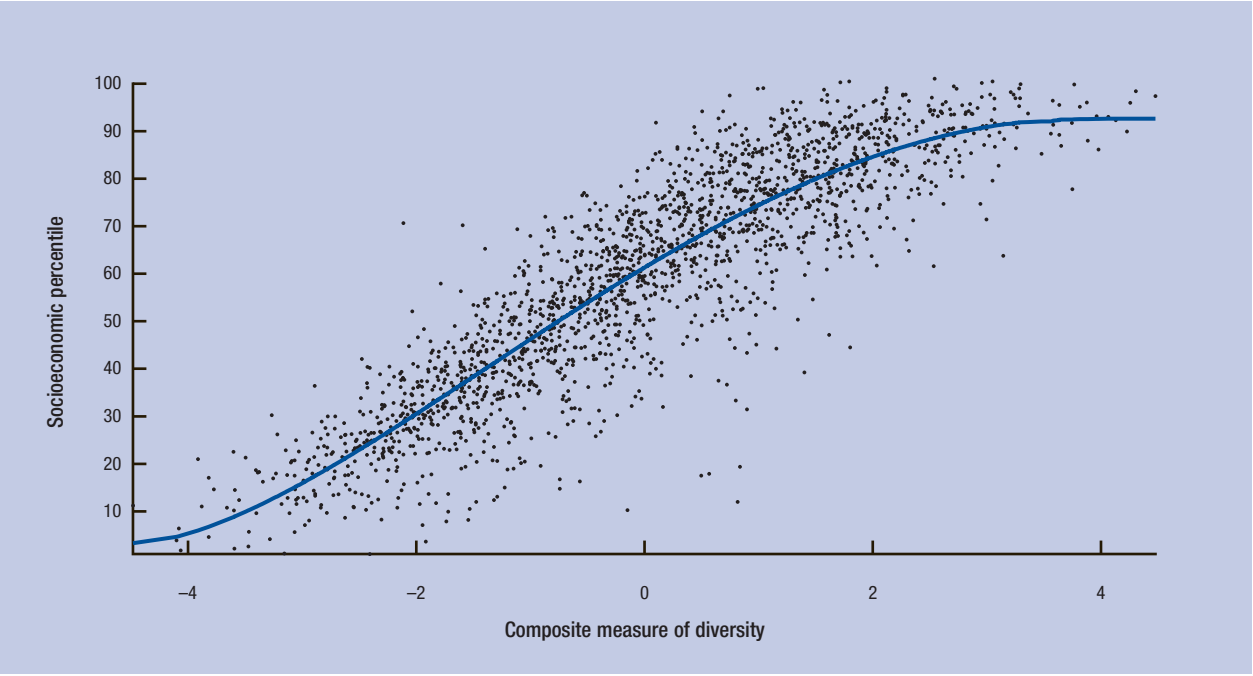
Deprivation and opportunity

We know that social interaction forms the backbone of social and economic life: from finding a good film to landing a deal, who we know and how we know them is a crucial determinant of success. And although the map presented in Figure 1b suggests a strong link on a regional, or even national scale, early work by Granovetter had already established this connection at the individual level.[10] Granovetter's finding—which is self-evident only with hindsight—was that we do not usually uncover novel information through interaction with our close friends. Awareness of a crucial job opportunity or innovation is much more likely to come from acquaintances and those with whom we are only weakly connected.

The simplistic explanation for this weak-ties effect is that we already interact intensively with our close friends and colleagues, and so we come to share the same background knowledge, the same awareness of opportunities, and, ultimately, the same view of the world. This shared perception can blind us to emerging threats—to the firm, community, or country; it can also deprive us of chances to forge new connections and make new discoveries. In contrast, our acquaintances often know people who are not part of our circle of friends; they offer us informational diversity because we are now connected to people who are much less like us socially, economically, and even spatially.

Of course, this is not to suggest that strong ties are less meaningful: they are thought to constitute a major mechanism for social support in hard times. Indeed, the strength of community ties might well be a crucial factor in personal happiness and fulfillment.[11] Eagle et al. sought to test this seemingly simple idea—that the structure of our social interactions can be correlated

Figure 2: Diversity of communications and deprivation



Source: Eagle et al., 2010.

with deprivation and opportunity—on a national scale.[12] Granovetter could test his hypothesis only by using data gathered painstakingly on just two Boston neighborhoods, but with a database of telephone calls encompassing nearly all of the United Kingdom we can examine whether this relationship holds universally or has only local applications for policymaking.

We developed a composite measure of the diversity of calling to and from an area that could be correlated with existing socioeconomic deprivation measures. The results in Figure 2 strongly bear out Granovetter's original work: the wider we cast our social net, the less likely we are to live in a deprived community. More recent work using a much simpler "local-ness" measure—the ratio of local to national calls made from a neighborhood—seems to show a similar effect.[13] Thus it appears that in all cases diversity is correlated with opportunity.

Of course, in reality the picture is a little more complicated because we cannot easily untangle the direction of causality: it is unclear whether people are more deprived because they have less diverse social networks, or they have less diverse social networks because they live in more deprived communities. To put it another way: is it that people who live in deprived areas tend to have made, or have been forced to make, life choices that inhibit their acquisition of more diverse networks, whereas those who live in less deprived areas have been able to take life paths (such as non-local higher education and employment) that tend to lead to more geographically dispersed social networks?

The essential importance of the social dimension makes the answer to this question vital to the planning of appropriate policy interventions. For instance, if it is merely opportunity that is lacking, then we might naively suggest that all that is needed is a job-seekers' forum for enabling introductions. But if the problem is, as seems likely, more deeply rooted in the constrained life-path choices available, then the appropriate policy response is more structural in nature and is unlikely to deliver "quick wins" in the short-term, a circumstance that creates challenges for policymakers looking for 12- to 24-month returns on policy investments.

GLOBALIZATION AND INDUSTRY

As the previous sections have made clear, economic activity is tightly bound up in social interaction. Within the contemporary multi- or transnational firm, these interactions are increasingly global in scope, which indicates the increasingly complex nature of both global supply chains and also knowledge flows between workers in widely separated offices. In order to understand these dynamics in more depth, we need to be able to see the knowledge economy in action, and telecommunications networks remain the best lens through which to do so.

To get at the globalization of knowledge by businesses, we can compare the level of international calling for some small area to the overall level of activity in the region of which that area is a part. The *telecommunications quotient*—named after the classic tool of spatial economic research, the *location quotient*[14]—is a *relative* measure of globalization that gives us a way to compare

different parts of a city or region with one another. And by filtering out individuals, households, and small businesses from the dataset, we can here focus on medium- and large-sized businesses with divergent levels of engagement with the global economy.

The telecommunications quotient is a simple-to-calculate ratio that captures the relative intensity of international calling for a small area within a larger region; it is anonymous, aggregate, and allows us to determine whether an area makes more or fewer calls than we would expect, given the overall behavior of the surrounding region. Thus a quotient of 1 means that the area is "normal" in its international calling behavior, while a quotient of 8 would mean that an area places *eight times* more international calls than expected from the regional average.

This approach enables us to identify differences of behavior both between firms operating in the same industry and between areas with strong specializations in different industries. For example, it is clearly expected that the City of London, which is home to global financial services firms, will be highly internationalized in its calling activity; and it is. But after removing household calling, Figure 3a reveals the rather surprising fact that towns such as Reading, Slough, Sandwich, and Bracknell Forest can match, or even exceed, the city's telecommunications quotient.

Merging the telecommunications quotient results with employment data collected by the government allows us to better understand why this is the case: these smaller cities are home to world-class ICT and defense firms that employ telecommunications to coordinate the activities of developers, designers, and executives in the United Kingdom and the United States. This pattern is rooted in the historical connection between these industries and government procurement, and today the region is an essential part of the United Kingdom's internationally competitive, high-skill service sector.

Figure 3b suggests that, even for finance, there are attractions to moving out of the City: there are back-office sites with extremely high levels of relative international calling visible to the South of Greater London. It is not only highly skilled work in the computing sector—both in software and hardware design—that has left the traditional urban core, but a good deal of work in the media industry has also left Soho to set up around the BBC's facilities in White City. Similarly, in the logistics center, global calling activity is closely tied to major international airports, all of which are located well outside London's core. Most dispersed of all, however, is research and development, with pharmaceutical and high-tech manufacturing lacking any real geographical concentration; major sites are scattered across the region and can be found in smaller towns such as Cambridge, Royal Tunbridge Wells, and—until recently—Sandwich (where Pfizer's facility closed in 2011).

The ability to distinguish between globally and locally interacting industries heralds a step change in our ability to understand the impact of globalization on regional development. The financial industry, because of its effect on government, tends to draw our attention toward the traditional downtowns of Manhattan and the City of London. But our findings point toward the importance of what we could call the "new industrial districts" of the knowledge economy, of which finance is only a part, and their much wider spatial distribution.

CLASSIFICATION AND CODIFICATION: TOWARD A REAL-TIME CENSUS

And yet, even as government and industry face increasing challenges from globalization and mobility, there is increasing pressure to reduce the cost of data collection while accelerating the timeliness of its provision. For example, in the United Kingdom—as in other countries that maintain the practice—the next population census is expected to cost nearly £500 million and to require many thousands of hours of labor to collect and process, with the results taking many months to reach end users. Consequently, the UK Office for National Statistics is already engaged in a program of research to assess whether the data needs of the national and local policy communities, as well as nongovernmental and commercial actors, can be met by integrating existing administrative, commercial, and imputed/modeled data.[15]

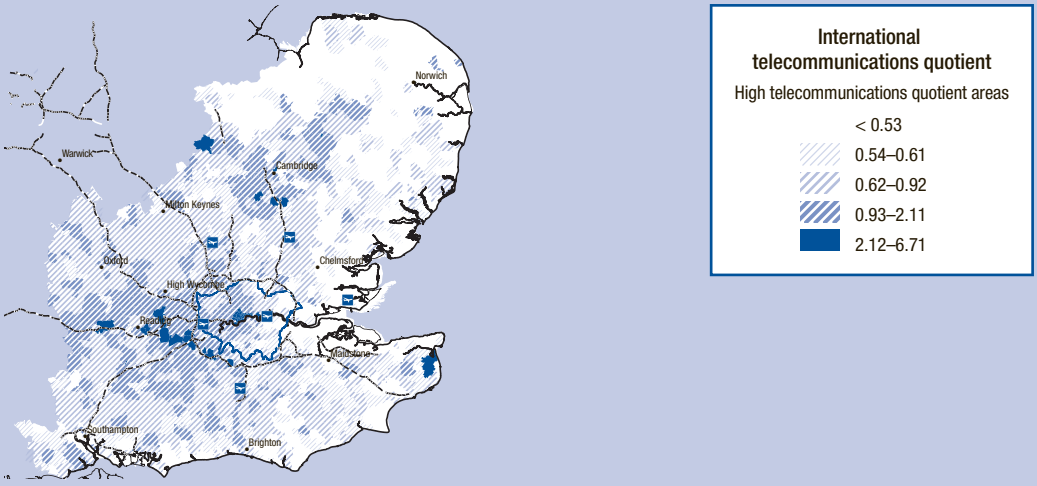
Recently, the number of innovative approaches to this problem has exploded. One of the more notable used names from electoral rolls, telephone books, and related datasets to infer ethnic and demographic characteristics of populations.[16] These characteristics can then be mapped on to neighborhoods with a view to updating a socio-demographic profile whenever someone registers to vote or leaves a forwarding address.

To the extent that this type of data can be collated and quickly associated with people, households, or firms, it provides us with the ability to characterize them at *any* given point in time, and not just every five or ten years. Thus, in addition to being able to provide a form of automated census on specific dates, the data also offer us the potential to understand the flow of life events surrounding an individual or group as well as the changing characteristics of an area. Changes in the telecommunications interactions of an individual, household, or neighborhood could, therefore, act as an early warning system of transitions—such as increased in-migration or changing patterns of work—with significant policy implications.

To illustrate this approach, we surveyed about a thousand households and, with their permission, associated more than a million call records to their responses in order to assess the possibility of classifying households according to their calling networks.[17] The results suggest that some dimensions of social interaction can serve as reasonable predictors of whether a household

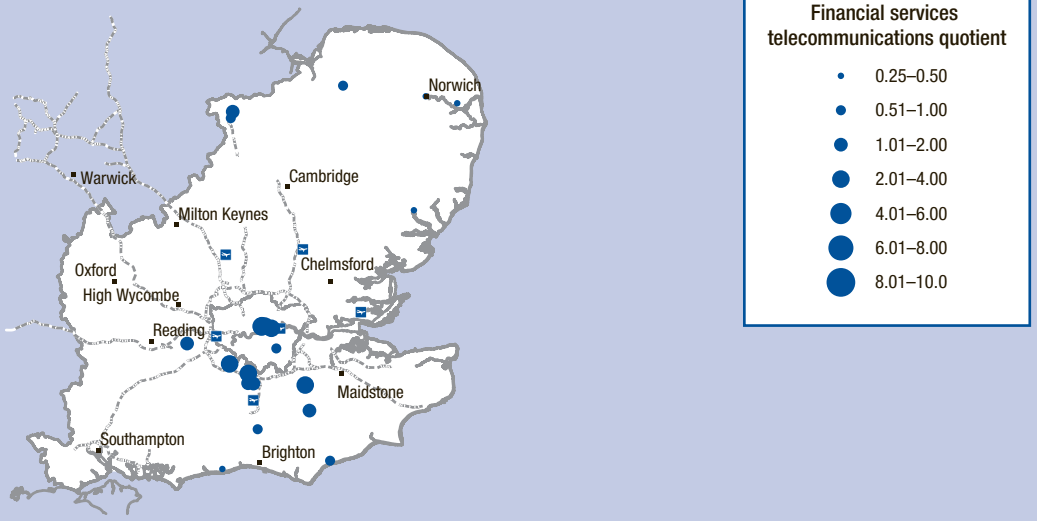
Figure 3: Industrial signatures: The international telecommunications quotient

3a: All business areas calling



Source: Data compiled by authors.

3b: Financial services areas calling



Source: Data compiled by authors.

is comprised of "Alone, over 56," "Couple, both aged over 55 with no cohabiting children," or "Couple, with children aged under 12." However, it has so far proved less effective at predicting other household types.

Although these are only preliminary results, it suggests that research into linked data has the potential to develop templates that could be applied to flows of telecommunication data in order to provide estimates of the local prevalence of different groups at very low cost and with extremely low latencies. One can easily imagine a future in which network operators would supply a government's statistical body with summary metrics that track month-to-month change at the neighborhood level for the entire country.

CONCLUSIONS

In this review we have endeavored to demonstrate that the insights to be gleaned from the marriage of telecommunications data and network-oriented research span the nature of community, the challenges of deprivation, the growth of knowledge-based industry, and the administration of regional economies. However, the real power of this approach to understanding individuals and communities—in their proper context—lies not only in its unprecedented breadth and depth, but also in the radical improvement of the speed with which such data can be collected and processed, and the results delivered to policymakers and strategic planners in both the public and private sectors.

We began by demonstrating one way in which social network analysis could supplement our understanding of regions and large communities. The results from this work suggest that, although many socioeconomic regions are well aligned with administrative units, in others—Wales, for example—the geography of human interaction appears to be diverging from long-standing historical boundaries, suggesting a new dynamic in play on the national scale.[18] This approach to regional delineation could ultimately be used as a monitoring as well as a modeling tool: long-term changes in these patterns could be connected to changes in accessibility or competitiveness, while also permitting planners to simulate the likely effects of social, economic, and infrastructural interventions.

From the bigger picture of regions, we then narrowed our focus to the neighborhood and saw how social network data could be used to investigate deprivation, cohesion, and access to opportunity. We therefore suggest that telecommunications data could provide a timely proxy for multiple aspects of well-being, addressing an increasingly important dimension of government policy. It may well be that by understanding the individual's or group's unique mix of tie strengths, we become able to locally tune policy interventions to suit the community structure, delivering targeted, measurable impacts on the ground.

Our third section examined how a telecommunications quotient can give us a new way to explore the complex web of informational linkages among industrial actors: using a simple, anonymous metric it becomes possible to assess the degree to which firms in a given area are engaged in international communication. In other words, the big data approach to telecommunications allows us to examine the fine-grained variations in how companies interact with one another, and with suppliers or clients around the world. In addition to highlighting important dependencies, we anticipate that this approach will help both firms and governments to monitor a rapidly changing regional economic landscape.

Finally, we noted that the ultimate promise of a real-time census is an environment in which crucial data about people and place are collected regularly and inexpensively, offering governments and researchers new ways to see change on a fine scale, without losing track of dynamics on the scale of cities, regions, or countries. Moreover, when combined with other measures—such as a telecommunications quotient generated from data filtered so as to contain only households, for example—it may also become possible to derive migration and country-of-origin data that could shed further light on neighborhood dynamics with a lag of days or weeks, rather than months or years.

With this as background, we can outline some preliminary opportunities for the private, public, and nongovernmental sectors:

- the collection of transactional data by telecommunications operators not only for the purposes of billing and system engineering but also explicitly for their aggregation and statistical re-use;
- citizen-data auction services for private citizens to aggregate and manage personal data with options for sale of access to commercial analysts—what Pentland called "a new deal on data";[19]
- new approaches to the local, regional, or national assessment of policy interventions through an analysis of changes in local/neighborhood communications behavior over time; and
- the provision of aggregated small area societal well-being indicators by telecommunications service operators for a future real-time census.

Of course, it is early days yet for the emerging computational social science industry, and there remain significant obstacles to the field's success. First, there are reasons to be concerned with data quality and integrity, since a great deal of time and money may ultimately rest on assumptions about the validity of the data that have yet to be systematically verified. Second, as yet there is no mature working model of how industry (which typically generates and collects this type of data) and government can collaborate successfully in a manner that is also trusted by citizens. And third, individuals are right to be concerned about the impact that this emerging

area might have on their own expectations of personal privacy: although the existing protections appear robust enough for one-time work, this dynamic would change with ongoing, linked data transfers to third parties. These questions must be addressed by all stakeholders—in-dustry, government, and citizens alike—if the potential of this field is to be realized.

NOTES

1 Savage and Burrows 2007.

2 Granovetter 1973.

3 Scott 2010.

4 Lazer et al. 2009.

5 Rogers 1987.

6 cf. Gonzalez et al. 2008; Leskovec and Horvitz 2008.

7 Ratti et al. 2010.

8 Tobler 1970.

9 Nielsen and Hovgesen 2008.

10 Granovetter 1973.

11 Putnam 2000.

12 Eagle et al. 2010.

13 Anderson and Vernitski 2011.

14 Florence 1948.

15 Office for National Statistics 2011.

16 Mateos et al. 2011.

17 Anderson and Vernitski 2011.

18 Ratti et al. 2010.

19 Pentland 2009, p. 79.

REFERENCES

Anderson, B. and A. Vernitski. 2011. "On the Social Scientific Value of Transactional Data." Talk presented at an Invited Seminar, July 25. Cambridge Computer Laboratory, Cambridge.

Eagle, N., M. Macy, and R. Claxton. 2010. "Network Diversity and Economic Development." *Science* 328: 1029–31.

Florence, P. S. 1948. *Investment, Location, and Size of Plant.* Cambridge: Cambridge University Press.

Gonzalez, M. C., C. A. Hidalgo, and A. L. Barabasi. 2008. "Understanding Individual Human Mobility Patterns." *Nature* 453: 779–82.

Granovetter, M. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78: 1360–80.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. "Computational Social Science." *Science* 323: 721–23.

Leskovec, J. and E. Horvitz. 2008. "Planetary-Scale Views on a Large Instant-Messaging Network." In *Proceedings of the 17th International Conference on World Wide Web.* New York: ACM.

Mateos, P., P. A. Longley, and D. O'Sullivan. 2011. "Ethnicity and Population Structure in Personal Naming Networks." *PLoS ONE* 6 (9): e22943.

Nielsen, T. A. S. and H. H. Hovgesen. 2008. "Exploratory Mapping of Commuter Flows in England and Wales." *Journal of Transport Geography* 16 (2): 90–99.

Office for National Statistics. 2011. *Beyond 2011.* Available at http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html.

Pentland, A. 2009. "Reality Mining of Mobile Communications: Toward a New Deal on Data." In *The Global Information Technology Report 2008–2009.* Geneva: World Economic Forum. 75–80.

Putnam, R. D. 2000. *Bowling Alone: The Collapse and Revival of American Community.* New York and London: Simon & Schuster.

Ratti, C., S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz. 2010. "Redrawing the Map of Great Britain from a Network of Human Interactions." *PLoS ONE* 5 (12): e14248.

Rogers, E. 1987. "Progress, Problems and Prospects for Network Research: Investigating Relationships in the Age of Electronic Communication Technologies." *Social Networks* 9 (4): 285–310.

Savage, M. and R. Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41: 885–99.

Scott, J. 2010. "Social Network Analysis: Developments, Advances, and Prospects." *Social Network Analysis and Mining* 1: 21–26.

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