How to draw a neighborhood? The potential of big data, regionalization and community detection for understanding the heterogeneous nature of urban neighborhoods

Forthcoming in *Geographical Analysis*

Ate Poorthuis, Singapore University of Technology and Design

Abstract

How to draw neighborhood boundaries, or spatial regions in general, has been a long-standing focus in Geography. This paper examines this question from a methodological perspective, often referred to as regionalization, with an empirical study of neighborhoods in New York City. I argue that methodological advances, combined with the affordances of big data, enable a different, more nuanced approach to regionalization than has been possible in the past.

Conventional datasets often dictate constraints in terms of data availability and spatio-temporal granularity. However, big data is now available at much finer spatio-temporal scales and covers a wider array of aspects of social life. The emergence of these datasets supports the notion that neighborhoods can be fuzzy and highly dependent on spatio-temporal scales and socio-economic variables. As such, these new datasets can help to bring quantitative analysis in line with social theory that has long emphasized the heterogeneous nature of neighborhoods. This paper uses a dataset of geotagged tweets to demonstrate how different ‘sets’ of neighborhoods may exist at different spatio-temporal scales and for different algorithms. Such varying neighborhood boundaries are not a technical problem in need of a solution but rather a reflection of the complexity of the underlying urban fabric.

Introduction

When we think of the city we live in, we often tend to think in neighborhoods. We split up an otherwise large and abstract city into a series of smaller areas or neighborhoods, each with a unique or defining set of characteristics, that together help us understand and make sense of the social and spatial fabric of the city. Similarly, urban researchers also use strategies that consists of dividing the city into smaller areas to better understand urban socio-spatial processes. This abstraction, from larger geographies into more tightly defined spatial areas, helps to clarify otherwise complex and multidimensional processes. However, researchers interested in the city often only have a choice between using relatively small census blocks and tracts (about 4,000 people in each tract) or much larger administrative areas (e.g. Brooklyn's Kings County has 2.6 million people), both of which might have boundaries that do not adhere to cohesive underlying 'neighborhoods' that define the common notions of how a city is delineated in the minds of residents and researchers alike.

Perhaps because of the practical implications, this issue has seen extensive attention within the discipline of Geography, both from a more theoretical or ontological point of view ('what is a neighborhood?') as well as a methodological perspective ('how do you define a neighborhood?'). The latter question, often referred to as *regionalization*, has seen sustained attention going back all the way to the 1930s (Christaller 1933). While practical constraints have historically forced social science to analyze society at an aggregated regional scale, we now see the emerging potential of new computational resources and methods that allow for the analysis of large, detailed data sets without resorting to aggregation, but rather analyze individual people, their characteristics and their socio-spatial behavior and networks – all of which has relevance to the question of regionalization.

In this article, I take these advances (in both methods and data) and show that they might enable an important change in how we define and think of regions and neighborhoods. That is, by

combining individual-level social media data with methodological advances from both regionalization and social network analysis, we can define regions in a more flexible way, where the shape and scale of regions can adapt based on context. To do so, I will start out with a short review of existing work and current direction of regionalization in Geography and community detection in Social Network Analysis. I then show the affordances and limitations of these methods in a case study in which a dataset of geotagged tweets is used to establish the kaleidoscopic character of regions and neighborhoods in New York City.

Regionalization and Community Detection

There are, quite literally, an infinite number of ways in which regions can be delineated and geographers have devised many techniques to do so in the last few decades. The interest in this process, often referred to as regionalization in Geography, took off during the Quantitative Revolution (Bunge 1966; Grigg 1965; Johnston 1968). As a sign of the times, many researchers devised ways to find an objective, replicable set of regions (Bunge, 1966) or an ‘exact solution’ to the problem (Berry 1961; Berry 1968; Berry & Kasarda 1977). On the other hand, other contemporaries theorized that there is no such thing as the perfect or absolute classification of regions. For them, each set of regions is but one representation of the world (Grigg 1965, p.471), in very much the same way as the map is not the territory (Korzybski 1933). Following this line of thought, since we inherently cannot derive a singular optimal classification, we should instead seek out many different classifications that can subsequently be evaluated for appropriate use depending on the purpose, context or specific academic study (Grigg 1965; Johnston 1970).

Additionally, in urban sociology a related, but more ontological, discussion took place on the nature of, and difference between, neighborhoods and communities. As Wellman and Leighton (1979) point out, much early work in sociology had been ‘neighborhood sociology’. The

neighborhood here is a territorial constraint or proxy for studying the actual object under study: communities. As such, they propose a solution to the ‘community question’ and break from the territorial constraint through the adoption of network analysis (Wellman 1979). Of note here is that network analysis and neighborhood analysis in sociology have recently been rejoined in work that combines the two together. For example, Hipp et al. (2012) use social ties to construct urban neighborhoods and, as will be discussed in the next section, there is a resurgence of work that uses ‘new’ data sources (e.g. mobile phone data) to do the same.

Regardless of one’s stance in this debate, a clear distinction can be made between two different methodological approaches to regionalization (Berry 1968; Brown & Holmes 1971). The first is that of the so-called formal region. Formal regions are uniform or homogenous regions that group similar locations together based on a certain set of variables. The foundation of this approach was laid by Eshref Shevky and collaborators and their work on social area analysis (Shevky & Bell 1955; Shevky & Williams 1949), which was subsequently enhanced through factor analysis (Bell 1955; Anderson & Bean 1961) (Berry 1961) (Berry & Rees 1969). The Modifiable Areal Unit Problem (MAUP) is an important problem within Geography because different aggregations can yield a very different view or results based on the exact same underlying data (Coombes & Openshaw 1982). In other words, drawing the boundaries of a region in a different way (shape) or changing the size of regions (scale) will affect the results and conclusion of the analysis performed on such regionally aggregated data. As such, we find Openshaw grappling with the problem in the domain of regionalization (Openshaw 1973; Openshaw 1977; Openshaw 1978) before actually formally defining it as MAUP (Openshaw 1984; Openshaw & Taylor 1979). In Openshaw’s approach — and the computational methods that followed — regions are often called ‘zones’. Zones are the modifiable areal unit, each of which consists of a number of immutable basic spatial units (e.g. these could be census blocks, or whatever the smallest spatial unit for which data is available). The

classification of these basic spatial units into zones is what Openshaw refers to as the automatic zoning problem (AZP). To find the optimal zone configuration in a fully automatic way, Openshaw (1977) first employs a heuristic hill-climbing algorithm is based on Ward's (1963) hierarchical clustering technique – later followed by a simulated annealing approach (Openshaw & Rao 1995). More recently, self-organizing maps (SOM) have been employed in the same fashion (Spielman & Thill 2008; Arribas-Bel et al. 2011).

In contrast, functional regions are based on the interactions or links between locations. Functional regions group together locations that interact more with each other than with locations outside the region. These interactions could be anything from commuting trips (Hemmens 1966; De Montis et al. 2013), public transportation (Green 1955; Ducruet & Lugo 2013), phone calls (Nystuen & Dacey 1961; Ratti et al. 2010), migration (Hägerstrand 1957; Brown & Holmes 1971; Slater 1981; Guo 2009) or even newspaper data (Haughton 1950; Kariel & Welling 1977). This could theoretically mean that those locations are similar in certain aspects, but more it can also be the opposite; two locations can be dissimilar but complementary economically or socially and thus generate high amounts of activity between them. Although some techniques can be applied in both approaches, the distinction between formal and functional regionalization is an important one.

Similar to the interactions in functional regions, the field of social network analysis has long been interested in finding, describing and analyzing cohesive subgroups within a region (Scott 1992), often referred to as community detection. Just like their counterparts in regional science and geography, community detection can be used to understand the structure of a social network.

Most of the early work on finding cohesive subgroups was conducted by sociologists and other social scientists (Luce 1950; Alba 1973; Borgatti et al. 1990). However, with the advent of increased

computing power and the contributions of physicists, the sub-field of community detection has seen a burgeoning of activity in the last fifteen years.

Historically, the results of community detection algorithms were compared with prior, qualitative knowledge about the network under study. However, in 2004, Newman and Girvan made a landmark contribution and introduced the concept of modularity (Newman & Girvan 2004). In short, modularity (Q) is a measure that calculates the fraction of edges within a community minus the expected fraction of edges with in a community. Modularity yields a score between 0 and 1. If Q approaches 0, it means that the strength of the community ties are no better than in a random network. The closer to 1 Q gets, the stronger the community structure. The concept of modularity provides a straightforward way to assess and compare the performance of a community detection algorithms. It can be applied to the resulting communities of any algorithm. In many ways, this enables new community detection procedures to strive towards maximizing modularity. In subsequent years, physicists and network analysts introduced a wide array of algorithms that fall in this class of ‘modularity optimization’. For an extensive review of different community detection algorithms see Lancichinetti and Fortunato (2009) and Fortunato (2010). They have also discussed some of the limits to modularity optimization (Lancichinetti & Fortunato 2011; Fortunato & Barthelemy 2007).

Modularity optimization has also seen an uptake in regional science and geography, especially in the realm of functional regionalization. Recent work of both network analysts and geographers working with spatial data has focused primarily on the use modularity optimization algorithms (De Montis et al. 2007; De Montis et al. 2013; Gao et al. 2013; Austwick et al. 2013; Manley 2014; Y. Liu et al. 2014; Amini et al. 2014). Despite the wide variety of alternative approaches, modularity is adopted as somewhat of a standard. For some regional scientists, modularity can be used as an objective standard with which to measure how ‘good’ resulting regions are or where to cut off

iterative, hierarchical algorithms (Thomas et al. 2012). In other work, the focus on modularity has become so universal that the actual community algorithm applied is not mentioned. In such cases, ‘community detection’ is performed without detailing any of the specifics on algorithm use or parameter settings (Kang et al. 2013; Y. Liu et al. 2015).

Case Study

As discussed in the previous section, how to define spatial regions and neighborhoods has a long history within Geography, with a connected line of inquiry to be found in the domain of community detection within social network analysis. However, the commuting patterns that underpin the labor market areas and metropolitan statistical areas that are so often the focus of functional regionalization are of little use on the smaller neighborhood scale. Similarly, the socio-economic variables available in census and large-scale survey data might not connect well with how we think and perceive urban neighborhoods. Because of this, this case study combines community detection methods from social network analysis with data derived from geosocial media to draw urban neighborhoods.

A single, objective rendering of neighborhoods is not the goal of this study. Rather, we try to gain insight into how the city is used by its inhabitants – in other words, what are the *lived* neighborhoods in a city? Naturally, there is no such thing as ‘*the*’ definition of such a neighborhood; this will differ from person to person and from group to group. Additionally, geosocial media data are produced by a subgroup of the general population that might be biased in multiple ways. The fact that these data are not a random, representative sample matters less, as we are not trying to generalize to an all-encompassing objective set of neighborhoods.

In any case, I argue that regionalization methods focused on urban neighborhoods would need to be able to reflect the heterogeneity of such neighborhoods and not hide behind a veil of

objectivity. If successful, neighborhoods derived from geosocial media can then help us better understand the underlying fabric of a city. They can also, in turn, function as subsequent input for urban research and policy; such neighborhoods can be potentially be used as spatial zones that more closely reflect underlying socio-spatial processes than the arbitrary and administrative boundaries of census tracts or the counties.

Data and Operationalization

In this case study, I use a dataset that consists of all geotagged tweets sent between July 2012 and July 2015 from the larger New York City metropolitan area¹. The total dataset consists of 149 million tweets sent by a total of 1.4 million different users (see Poorthuis and Zook (2017) for details on data collection). Twitter is, of course, called a social network because users can choose to ‘follow’ other users — thus building a network of follower-followed relationships that are well-studied within social network analysis (Golder & Yardi 2010; Gruzid et al. 2011; Kivran-Swaine et al. 2011). However, here we disregard those explicit user networks in favor of a more implicit spatial network. We are not interested in the direct relations between Twitter users, but rather how those individuals are connected to different spaces within the city. When a user tweets from a specific location within the city, that tweet can be seen as a trace of activity that person performed at that specific location. The more often a person tweets from a location, the more likely it is that the tie between the person and the location bears some significance.

In this sense, tweets are used as a source of interaction data that reflect (however partially) each user’s activity space (Schönfelder & Axhausen 2003) in much the same way as can be done with mobile phone data (Ahas et al. 2010) or other data sources that contain traces of people’s spatio-

¹ Defined as a bounding box where latitude is between 40.3455 and 41.3335 and longitude is between -74.5697 and -73.0371.

temporal activities (Axhausen 2008). Social media data has already seen extensive use in this vain, from tweets (Shelton et al. 2015; van Meeteren & Poorthuis 2017) to Foursquare check-ins (Hristova et al. 2016) and Instagram photos (Boy & Uitermark 2016).

As such, this data can be used to map the connections an individual has to the different parts of the city – not only their home and place of work, but also a variety of other places and experiences that may include nights out at the bar; picnics in the park; or morning strolls along the boulevard. Tweets can be considered an appropriate data source for such an operationalization, as the act of tweeting, especially when opting to geotag tweets, is a conscious and explicit act that links an activity or experience specifically to the current location.

The dataset used for this case study represents Twitter users on the one hand and their ties to specific locations within New York on the other hand. Within social network analysis, such a network is called a two-mode network (Borgatti & Everett 1997). A two-mode (also known as affiliation or bipartite) network always consists of two different sets of nodes (in our case, both people and locations) that have ties between them but not among them. The analysis of two-mode networks is more complicated than standard one-mode networks (although certainly not impossible, cf. (Borgatti & Everett 1997), (Agneessens & Everett 2013), and (Everett & Borgatti 2013)). For this reason, and simply because we are only interested in relations between one set of nodes, two-mode networks can be ‘projected’ to a one-mode network. A projection of a two-mode matrix can be as simple as transposing the original matrix and then multiplying the original matrix with the transposed matrix (Neal 2012).

A first complication arises because Twitter users display different levels of activity. This can range from sending only a handful of tweets in a year to tweeting thousands of times over the course of just a one single month (Poorthuis & Zook 2014). If left unaddressed, active users potentially

carry a larger weight in the resulting location-to-location network. To alleviate this issue, rather than performing a straightforward matrix multiplication, we weigh each user's connection to a specific place based on the total amount of activity that a user displays:

$$W_{ij} = \sum_{u=1}^U \frac{T_{iu} + T_{ju}}{(L+1) \sum_{l=1}^L T_{lu}}$$

The weight between location i and j (W_{ij}) is the sum of the weights of every user's (u) connection between i and j. The weight of a user's connection is defined by the sum of tweets (T) in both location i and j, relative to the total weight of all tweets between all locations for that user. The latter can be calculated by multiplying the total number of tweets for the user by the number of locations + 1 — as this is an undirected graph that allows for loops. This process is graphically illustrated in Figure 1 for just a single user. Ultimately, all ties for a single user always add up to 1. As such, each Twitter user is carrying exactly the same amount of 'vote' within the network no matter how active or inactive that user has been on the platform.

Figure 1 Projection of two-mode to one-mode network

A second complication arises because a projection from a two-mode to a one-mode network can introduce structural artifacts that are not a result of the underlying social process but rather the of the projection method (Neal 2014). To determine the significance of the resulting one-mode weighted edges, I adopt an approach based on Fixed Degree Sequence Methods to construct an appropriate null model (Stegbauer & Rausch 2012; Horvát & Zweig 2013; Derudder & X. Liu 2015). Specifically, a random network is constructed while ‘fixing’ the degree of both users and locations (row and column marginals are identical to the observed network). Patefield’s (1981) algorithm, as implemented in the `r2dtable` method in R, is used to create 1,000 of such random networks (see Fosdick et al. (2016) for an overview of fixed degree sequence methods). Each random network is then processed through the same weighting procedure outlined above and an edge in the observed network is only deemed significant if its edge weight is larger than the edge weight in 990 of the random network ($\alpha = 0.01$).

It is important to note here that the definition of a ‘location’ is arbitrary. Geotagged tweets may have a very precise latitude/longitude coordinate pair attached to them but the accuracy of this coordinate depends on a multitude of factors including the accuracy of the underlying GPS, privacy settings and user-entered geo-location (Crampton et al. 2013; Shelton et al. 2015). It is sensible to aggregate these individual point locations up to at least a somewhat larger area to not give an impression of accuracy where little exists in practice. In other words, a tweet that appears to be coming from 600 Main Street might have also been sent from 580 Main Street, 660 Main Street or even 600 Broadway. Binning tweets together in slightly larger areas thus not only decreases the complexity of the data set, it also more appropriately reflects the underlying characteristics of the data. The shape and size these larger areas or neighborhoods is completely arbitrary and can potentially have a large effect on the outcome of analysis or, in this case, the resulting network structure: this is the Modifiable Areal Unit Problem (MAUP) exactly (Openshaw 1984).

The end goal of delineating neighborhoods and the regionalization process is itself related to the same MAUP problem, as the resulting areas can form the basis of an aggregation that is potentially more closely related to the actual underlying social process under study. However, to get to that stage, it is necessary to make the data slightly more manageable and decrease the complexity by aggregating individual point location up to a reasonably small ‘building block’ spatial unit. Without this step, the location-by-location matrix could potentially measure millions of nodes and billions of ties between them, which still poses considerable computational challenges despite advances in computational power in the last decades.

As becomes clear from Figure 2, the size of these cells has a significant impact on the ‘view’ on the city seen through this data. Ultimately, it would be sensible to take the smallest cell size as the initial building block (just large enough to account for the underlying issues around accuracy of geotags) to then find larger neighborhoods within the city. However, for the purposes of this case study, a balance between detail and feasibility had to be made in order to keep computation time

reasonable.

Figure 2 Grid with resolutions of 0.05; 0.025 and 0.01 degrees

Description of the Network

A grid with cell sizes of 0.0075 degrees (roughly 0.5 mile) yields a total number of 27,060 cells in the New York City metropolitan area. There are 16,668 cells or locations (from now on: nodes) that contain one or more tweets with a total of 3.5 million ties between those nodes. Each individual tie between two nodes has a weight that indicates the strength of that connection, based on the

weighting procedure outlined above. After only retaining ties that are significant, we are left with a network consisting of 2.1 million ties and 15,048 cells. As a first validity check, we can look at the distribution of centrality for each individual node. Although there are different ways to look at how central a node is within a network, the most straightforward way is to look at the number of incoming and outgoing connections a node has. Since the network has weighted edges in this case, I calculate a weighted degree centrality by adding up the total weight of the ties for each node (Scott 1992). As is clear from Figure 3 (with log scale on the x axis), the distribution of degree centralities is skewed with most nodes having a relatively low centrality, while only a few nodes are on either end of the spectrum in being either hardly connected at all or extremely central. The network is lumpy, which accords with how we know people to live in cities. Some locations — parks, airports, stations, landmarks — form nodal points in how people use the city, while others — most — are more peripheral in the sense that they are only part of the activity spaces of a few people.

Figure 3 Degree Centrality

Of course, this particular network, does not operate completely independent of its underlying geography. From previous studies, we already know that Twitter’s social network as a whole does not exist independent of our location, as most ties are local (Takhteyev et al. 2012; Stephens & Poorthuis 2014). This seems to be the case within this network as well. Most strong ties between locations are across a very short distance of less than 5 miles (see Figure 4). Again, this makes sense; we connect to and visit mostly places that are physically close to us. There are a few notable exceptions, but they are nodal points such as airports, stadiums and a variety of tourist attractions across the city (e.g. Bronx Zoo, Flushing Meadows Park, Liberty Island etc.).

Figure 4 Relational Distance versus Physical Distance

With 16,000 nodes and millions of connections, it is difficult to make sense of such a large network in a single birds-eye view. We can describe various properties of the network as a whole; or even attempt to visualize it. But, in this case, we are not trying to understand the network in itself. We want to gain insight in the underlying socio-spatial processes through the analysis of the activity-space network that we have constructed. We could, of course, look at individual nodes instead of the entire network. Many nodes hold interesting stories: for example, a node in the heart of The Bronx has 6 out of the 8 strongest connections to nodes in its direct spatial vicinity. As such, this neighborhood perhaps can be construed as very ‘local’, linked to its direct neighbors primarily. On the other hand, a node in the gentrified neighborhood of Williamsburg, Brooklyn, only has three

connections to neighboring nodes with the other links to nodes further west in Manhattan: the East Village, New York University, and the Museum of Modern Art. While such insights may be interesting in and of themselves, it is hardly feasible to examine every single node in the network, let alone repeat a study with a different data set, methods or cell sizes. As such, community or neighborhood detection within interaction networks like this can help strike a balance between the very general network-level overview and the detailed, but impractical, node-level perspective.

Drawing Neighborhoods

As discussed previously, recent approaches to community detection often employ an optimization approach, without always detailing the intricacies of the algorithm employed. In this section, I will illustrate the importance of those intricacies and the potential influence the choice of algorithm and parameters can have on the resulting neighborhood boundaries. I start out by comparing two modularity optimization approaches to community detection. One of the aims of this section is to show how a slightly different algorithm can give significantly different results and ask the question whether modularity optimization algorithms should indeed be used as a ‘one-stop shop’ to regionalization and neighborhood detection within Geography.

To do so effectively, the rest of this section will zoom in on the borough of Brooklyn in order to more easily compare and contrast the resulting neighborhoods. We also switch to a slightly smaller cell size of 0.005 degrees (~ 0.25 mile) that yields a total of 1337 significant nodes with 252,564 ties in the borough of Brooklyn.² After a review of modularity-based approaches, the Infomap algorithm is used as a contrasting approach, specifically because it gives more control to the researcher. To aid the interpretation of the maps in the next section, Figure 5 is a reference map of Brooklyn that includes approximate references to different parts and neighborhoods of the city.

² Using a 2 kilometer buffer, including parts of lower Manhattan, to take into account possible edge effects.

Since there is no objective, or base, set of neighborhoods (after all, that is main premise of this paper), these neighborhoods are not visualized as polygons but just as approximate text labels in their general location with larger labels for neighborhoods that extend to a larger area.

Figure 5 Map of Brooklyn Neighborhoods

The first modularity optimization algorithm used here is Clauset et al.'s (2004) fast greedy approach (CM from now on), as implemented in the R library igraph by Csardi and Nepusz (Csardi & Nepusz 2006). The CM approach is iterative and keeps merging nodes and communities until, at the last step, all nodes are within one single community. To determine the best step at which to stop, convention dictates that the step with the highest modularity is chosen. However, as can be seen in Figure 6, the modularity plateaus around step 1000 and keeps at a very similar level for the next 300-

or-so steps. The highest modularity is at step 1326 with a value of 0.5489. But at step 1025 the modularity is very close at 0.5479 and at step 1331 it still is 0.5425. How can we determine that the resulting communities at step 1326 are really ‘better’ if the modularity values are so close to each other?

Furthermore, when we look at the difference in the actual communities found at each of these steps, another issue emerges. The resolution limit of modularity optimization (Lancichinetti & Fortunato 2011) yields that, even at step 1078, larger communities have already formed and the ‘extra’ communities at that step are mostly isolated nodes in the periphery of the boundaries of the borough, and not a set of cohesive, smaller communities that would actually be useful to understand Brooklyn’s neighborhoods. As such, the fast greedy approach cannot be used to find neighborhoods much smaller than what we see in Figure 7. More importantly, the researcher has little control over the results. By choosing this algorithm, the only choice left to the researcher is which step to choose as the final solution, which does not substantially affect the outcome of the procedure. As such, the CM approach to finding neighborhoods or communities seems incompatible with a how urban neighborhoods function.

Figure 6 Modularity for CM approach from step 1 to last step

Figure 7 Brooklyn neighborhoods at three different steps of the CM algorithm

We see a similar issue with other modularity optimization approaches. For example, the Louvain algorithm is, after the fast greedy approach, perhaps the most often used modularity optimization procedure (Blondel et al. 2008). Although still optimizing the same modularity equation, it does not use the same approach as CM. Rather, every single vertex starts out in its own community. Vertices are assigned to communities, based on a local, greedy modularity optimization. When no improvements can be made to the global modularity, the vertices in each community are merged — the communities in itself then considered to be a vertex and the procedure is repeated — as such, creating a hierarchy of nested communities.

The Louvain algorithm finds two of these hierarchies within the Brooklyn network with modularity values of 0.5421 and 0.5510. The solution of the second hierarchy that the Louvain algorithm finds is thus slightly ‘better’ than the CM approach. However, as can also be seen from Figure 8, the solution is topologically quite different. For example, the fast greedy community (3) is now split in two separate neighborhoods (4 and 9), while the boundary of the large South Brooklyn neighborhood (2) in the CM algorithm is much different in the Louvain version, and so on. This problem is inherent to the modularity optimization approach in which solutions with similar modularity values can represent very different community segmentations (cf. (Fortunato 2010; Good et al. 2010)).³

³ We can also quantify this topological difference by looking at the normalized mutual information (NMI) between the two community structures. The NMI (Danon et al. 2005) reflects how much can be learned about one community structure from another community structure. If the value is 0, then the two communities divisions are completely different, if the value is 1, they are completely identical. In this case, the NMI between the Louvain solution and the CM method is only 0.72.

Figure 8 Brooklyn neighborhoods at the two different steps of the Louvain algorithm

The key limitation with both of these approaches is that the researcher has very little control over the resulting communities. If quantitative methods are to be employed in a way that is compatible with what social theory tells us about social space (e.g. different neighborhoods exist for different people, with overlapping boundaries, at different scales etc.), then other methods might be more appropriate. In simpler words, the straightforward application of modularity optimization approaches might not adequately answer the question ‘What are the possible sets of neighborhoods in Brooklyn?’, as, from theory, we know to expect many different, overlapping and heterogeneous neighborhoods to exist — and not just one set of neat and objective communities.

To illustrate this further, I also employ the Infomap algorithm that is based around the map equation (Rosvall et al. 2010). The argument here is not that this algorithm yields better results, closer to some form of ground truth. We know that different, overlapping neighborhoods exists and need algorithms that allow us to gain insights into these differences and overlaps, rather than an algorithm that presents us with a single canonical set. Here, the Infomap algorithm might have potential as it gives more control to the researcher and allows for hierarchical, overlapping communities with multiple sizes. I use the C++ implementation made available by the authors

themselves (Edler & Rosvall n.d.). Infomap is a non-modularity approach that comes from information theory and uses the metaphor of the map (Rosvall & Bergstrom 2008). In making a map, the cartographer has to reach a balance between showing detail and generalizing. Oversimplify things and the resulting map does not give the reader much information. Too much detail and the resulting map is hardly legible. Rosvall and Bergstrom (2008) generalize this to a compression problem: good maps have a good compression technique. They define a ‘map equation’ that measures how efficient this compression is: how much detail is lost in generalizing the map. Naturally, here we are not interested in cartography or maps per se — this is just a metaphor. Our ‘map’ is a network of nodes and edges and, using that information, we try to compress individual nodes into communities to make the network more ‘legible’. The Infomap algorithm uses a flow-based approach in which the route of a random walker through the network, a proxy for information flow, is described using Huffman code. To find communities, we thus try to minimize the length of this description.

One of the most relevant features of the Infomap algorithm is that it allows the researcher to directly influence the size of the resulting communities. This is done by using the Markov-time as a resolution parameter (Schaub, Lambiotte, et al. 2012; Schaub, Delvenne, et al. 2012). Increasing this parameter, increases the number of steps the walker takes before its position is encoded, which in turn yields larger communities and vice-versa. This is a very salient feature for the implementation of community detection algorithms in social science. After all, depending on the social process under study or the specific research questions, we might be specifically interested in neighborhoods at a certain scale or size. The graph in Figure 9 shows how, by varying the Markov-time from 0.05 to 2.00, we can directly control the number of communities found. In particular, the range of Markov times between 0.6 and 1.1 (1.0 being the algorithm’s default) yield many useful divisions in the context of drawing urban neighborhoods.

Figure 9 Number of communities at various Markov-times for Infomap algorithm

As we can see from the four maps in Figure 10, each different Markov-time not only changes the number of communities found, but also yields communities of different sizes. The resolution parameter thus gives us direct control over the scale on which we find neighborhoods. This is in contrast with the different steps in the CM approach, where choosing a different step-size only increased the noise present in the solution, without significantly affecting the size of the resulting communities. Depending on the research context, we can vary the size of neighborhoods from only a few blocks to much larger meta-neighborhoods that potentially span large parts of the city. This approach complements more theoretically-derived notions around the multi-scalar nature of urban neighborhoods. After all, at one scale level we might think of downtown Manhattan as one single

neighborhood; at another we recognize the smaller neighborhoods downtown constitutes off such as Tribeca, Chinatown and the East Village.

If we pick a specific scale to take a closer look at the resulting neighborhoods (Figure 11), we see a set of neighborhood boundaries that has the potential to give us some new insights into Brooklyn's neighborhoods, as well as confirm local knowledge about the borough. At this scale level, eastern Manhattan is one cohesive neighborhood (1), while split up in the sub-neighborhoods of the Financial District (1), SOHO (4), Lower East Side (2), Gramercy (5), and Midtown (14) at lower Markov times. The trendy area of Williamsburg (2), is separate from the distinct neighborhood around the Navy Yard (25) to the south and Long Island City (5) to the north. The neighborhood of Red Hook (21) is an interesting example as directly to its southwest we find one single isolated node that is actually part of the Manhattan neighborhood (1). This non-contiguous part of the Manhattan neighborhood is home to Brooklyn's Ikea, which is directly connected to lower Manhattan by a ferry ride that specifically services this Ikea store. It is difficult to reach the store by other public transit. As such, it is much more connected to Manhattan than it is to the rest of Red Hook. Similarly, the southern tip of Coney Island, covering its boardwalk and the beach, is more connected to Manhattan (1) than its contiguous neighborhood of Coney Island (16). Although the Infomap algorithm does not have a contiguity constraint like many geographic regionalization procedures, apart from these exceptions, almost all resulting communities are indeed contiguous areas. Another exception to this rule can be found in some boundaries between neighborhoods, reflecting the fuzziness of such boundaries. For example, the boundary between Kensington (18) and South-West Brooklyn (4) has two nodes that are 'switched'.

Figure 10 Brooklyn neighborhoods at the four different resolution parameters for Infomap algorithm

Figure 11 Brooklyn neighborhoods for Infomap algorithm with Markov-time 0.85

Although, the results of the Infomap community detection indeed seem to resemble neighborhoods that Brooklynites would instantly recognize, that is not to say that these are *the* neighborhoods of Brooklyn, particularly given the data source on which they are based. And that is exactly the point. We know that urban space is a complex, variegated, heterogeneous process and the different results from different algorithms, with different parameters, shown here confirm that. The neighborhoods in Figure 11 are just one particular perspective on those neighborhoods. The boundaries and size of the neighborhoods can change by using different data, a different algorithm, or by adjusting the resolution parameter of the Infomap algorithm ever so slightly.

In addition, even the neighborhoods found with one particular resolution parameter are by no means binary or entirely cohesive – location within each ‘neighborhood’ still have connections to locations *outside* the neighborhood to varying degrees. In Figure 12, I calculate the degree of belonging for each node with a neighborhood, dividing the weight of the total amount of ties of that node by the weight of the connections that are completely within the neighborhood. The darker the node, the more tightly it is connected to its own neighborhood. Lighter nodes, even though they ‘belong’ to the neighborhood they are currently in, have large share of their total, weighted, connections to other nodes *outside* the neighborhood. This shows the fluid and flexible nature of neighborhoods. A small parameter change, a slightly different algorithm, a slightly different weighting scheme, a different dataset, or a slightly different timeframe — they might all change to which neighborhoods these nodes belong

Figure 12 Brooklyn neighborhoods for Infomap algorithm with degree of ‘belonging’

Neighborhoods are not static over time either. Their boundaries and size shift and change as different people move in and out of them; businesses and public spaces disappear and are re-created; and people build new connections to different places. The city and its neighborhoods are never at rest but always changing. As it does, the neighborhoods reflected in these geosocial media data do too. For example, we can compare the neighborhoods found based on connections made in 2012, with the same set of neighborhoods based on connections made in 2015 (all other parameters kept equal). Figure 13 shows some clear temporal changes in the neighborhood structure. Hip and trendy Williamsburg (2) was still separated from Manhattan (1) in 2012 — reminiscent of *Sex and the City*

scene where Miranda Hobbes' move into Brooklyn is deemed unthinkable by her Manhattanite girlfriends. But Brooklyn has undergone substantial gentrification in recent years and the neighborhood of Williamsburg in 2015 has become an extended part of the Lower East Side of Manhattan, which also includes the revamped Brooklyn Bridge Park under the Brooklyn Bridge. Of course, activity spaces are hardly the only aspect relevant to identifying or understanding the process of gentrification, and this article does not pertain to gentrification in particular. However, if one were interested in studying urban gentrification, such insights into the perception of neighborhoods can be a research objective in and of itself.

Figure 13 Brooklyn neighborhoods for Infomap algorithm in 2012 versus 2015

Conclusion

The case study illustrates the potential of studying the variegated, heterogeneous nature of urban space through the application of quantitative methods from social network analysis to geosocial big data. The network-based approach employed here is can provide us with new and alternative understandings of neighborhood processes. The Brooklyn neighborhoods identified in this article can give us insight into the how Twitter users interact with their city. By further disaggregating such a dataset into specific subgroups of people, an even deeper understanding of

this aspect of social space can be attained. Such sub-groups can for example be constructed based on home location (e.g. showing how people from Williamsburg experience the city versus people from Sunset Park), based on day versus night-time activity (does social space have a different structure at night?), or based on the tweet content (e.g. supporting different political candidates, tweeting about specific social or cultural behavior), etc. We have also not fully explored the affordances of the Infomap approach to community detection in this regard, as it also allows for explicit fuzzy boundaries and overlap (Viamontes Esquivel & Rosvall 2011), hierarchical clustering (Rosvall & Bergstrom 2011) and multi-layered communities that include different types of ties (e.g. positive and negative connections) (De Domenico et al. 2015).

Thinking of the city in this way is bringing quantitative geography much closer to the cognitive mapping of behavioral geographers (e.g. Gould & White 1974; Golledge et al. 1976), which can be a useful addition to more traditional interest in functional labor market regions or the formal regions in geodemographics. As such, it is a nice example of some of the affordances of big data for new types of quantitative analysis that go beyond the more generalized and limited datasets available to researchers in the previous century. In this way, we can use regionalization algorithms without obscuring the variegated and fluid nature of regions and neighborhoods, instead exposing and highlighting their complexities. This is not a weakness or un-scientific approach to research: embracing the multiplicity of neighborhoods, rather than looking for a single objective neighborhood delineation, potentially yields much greater insight into the urban fabric.

The comparison of several algorithmic approaches in this article has shown that the recent focus on modularity optimization techniques and the neighborhood boundaries they yield warrants more attention. Small changes in algorithms, parameters and input data have a large potential impact on resulting boundaries, which needs to be considered in subsequent work on urban neighborhoods. New (big) data sources, computing resources and convenient algorithms have certainly helped push

the boundaries of the field of regionalization and community detection. We can leverage these possibilities without losing sight of the limitations and idiosyncrasies of specific data sets and algorithms. In doing so, we have the opportunity to study and understand neighborhoods in ways that were previously impossible.

References

- Agneessens, F. & Everett, M.G., 2013. Introduction to the special issue on advances in two-mode social networks. *Social Networks*, 35(2), pp.145–147.
- Ahas, R. et al., 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), pp.3–27.
- Alba, R.D., 1973. A graph-theoretic definition of a sociometric clique†. *The Journal of Mathematical Sociology*, 3(1), pp.113–126.
- Amini, A. et al., 2014. The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1), pp.6–20.
- Anderson, T.R. & Bean, L.L., 1961. The Shevky-Bell Social Areas: Confirmation of Results and a Reinterpretation. *Social Forces*, 40(2), pp.119–124.
- Arribas-Bel, D., Nijkamp, P. & Scholten, H., 2011. Multidimensional urban sprawl in Europe: A self-organizing map approach. *Computers*, 35(4), pp.263–275.
- Austwick, M.Z. et al., 2013. The Structure of Spatial Networks and Communities in Bicycle Sharing Systems J. Gomez-Gardenes, ed. *PLOS ONE*, 8(9), p.e74685.
- Axhausen, K.W., 2008. Social Networks, Mobility Biographies, and Travel: Survey Challenges. *Environment and Planning B: Planning and Design*, 35(6), pp.981–996.
- Bell, W., 1955. Economic, Family, and Ethnic Status: An Empirical Test. *American Sociological Review*, 20(1), p.45.
- Berry, B.J., 1968. A synthesis of formal and functional regions using a general field theory of spatial behavior. In B. J. L. Berry & D. F. Marble, eds. *Spatial Analysis A Reader in Statistical Geography*. Englewood Cliffs, NJ: Spatial analysis.
- Berry, B.J.L., 1961. A Method for Deriving Multi-Factor Uniform Regions. *Przegląd Geograficzny*, 33(2).
- Berry, B.J.L. & Kasarda, J.D., 1977. Contemporary urban ecology. New York, McMillan.

- Berry, B.J.L. & Rees, P.H., 1969. The Factorial Ecology of Calcutta. *American Journal of Sociology*, 74(5), pp.445–491.
- Blondel, V.D. et al., 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), p.P10008.
- Borgatti, S.P. & Everett, M.G., 1997. Network analysis of 2-mode data. *Social Networks*, 19(3), pp.243–269.
- Borgatti, S.P., Everett, M.G. & Shirey, P.R., 1990. LS sets, lambda sets and other cohesive subsets. *Social Networks*, 12(4), pp.337–357.
- Boy, J.D. & Uitermark, J., 2016. How to Study the City on Instagram T. Preis, ed. *PLOS ONE*, 11(6), p.e0158161.
- Brown, L.A. & Holmes, J., 1971. THE DELIMITATION OF FUNCTIONAL REGIONS, NODAL REGIONS, AND HIERARCHIES BY FUNCTIONAL DISTANCE APPROACHES*. *Journal of Regional Science*, 11(1), pp.57–72.
- Bunge, W., 1966. *Theoretical geography*, Royal University of Lund, Dept. of Geography; Gleerup.
- Christaller, W., 1933. *Die zentralen Orte in Süddeutschland: eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*, Jena: Fischer.
- Clauset, A., Newman, M.E.J. & Moore, C., 2004. Finding community structure in very large networks. *Physical Review E*, 70(6), p.066111.
- Coombes, M.G. & Openshaw, S., 1982. The use and definition of travel-to-work areas in Great Britain: Some comments. *Regional Studies*, 16(2), pp.141–149.
- Crampton, J.W. et al., 2013. Beyond the geotag: situating “big data” and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2), pp.130–139.
- Csardi, G. & Nepusz, T., 2006. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), pp.1–9.
- Danon, L. et al., 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09), pp.P09008–P09008.
- De Domenico, M. et al., 2015. Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems. *Physical Review X*, 5(1), p.011027.
- De Montis, A. et al., 2007. The structure of interurban traffic: a weighted network

- analysis. *Environment and Planning B abstract*, 34(5), pp.905–924.
- De Montis, A., Caschili, S. & Chessa, A., 2013. Commuter networks and community detection: A method for planning sub regional areas. *The European Physical Journal Special Topics*, 215(1), pp.75–91.
- Derudder, B. & Liu, X., 2015. How international is the Annual Meeting of the Association of American Geographers? A social network analysis perspective. *Environment and Planning A*, 48(2), pp.309–329.
- Ducruet, C. & Lugo, I., 2013. *Structure and Dynamics of Transportation Networks: Models*, The SAGE handbook of transport studies.
- Edler, D. & Rosvall, M., mapequation.org - code. *mapequation.org*. Available at: <http://www.mapequation.org/code.html> [Accessed August 26, 2015].
- Everett, M.G. & Borgatti, S.P., 2013. The dual-projection approach for two-mode networks. *Social Networks*, 35(2), pp.204–210.
- Fortunato, S., 2010. Community detection in graphs. *Physics Reports*, 486(3-5), pp.75–174.
- Fortunato, S. & Barthelemy, M., 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), pp.36–41.
- Fosdick, B.K. et al., 2016. Configuring Random Graph Models with Fixed Degree Sequences.
- Gao, S. et al., 2013. Discovering Spatial Interaction Communities from Mobile Phone Data. *Transactions in GIS*, 17(3), pp.463–481.
- Golder, S. & Yardi, S., 2010. Structural predictors of tie formation in twitter: Transitivity and mutuality. *Social Computing (SocialCom)*.
- Golledge, R.G., Rivizzigno, V.L. & Spector, A., 1976. Learning about a city: Analysis by multidimensional scaling. *Spatial choice and spatial behavior*, pp.95–116.
- Good, B.H., de Montjoye, Y.-A. & Clauset, A., 2010. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4), p.046106.
- Gould, P. & White, R., 1974. *Mental maps*, New York: Penguin Books.
- Green, H.L., 1955. Hinterland Boundaries of New York City and Boston in Southern New England. *Economic Geography*, 31(4), p.283.
- Grigg, D., 1965. The Logic of Regional Systems. *Annals of the Association of American Geographers*, 55(3), pp.465–491.

- Gruzd, A., Wellman, B. & Takhteyev, Y., 2011. Imagining Twitter as an Imagined Community. *American Behavioral Scientist*, 55(10), pp.1294–1318.
- Guo, D., 2009. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer*
- Haughton, J.P., 1950. Irish local newspapers: A geographical study. *Irish Geography*, 2(2), pp.52–57.
- Hägerstrand, T., 1957. *Migration and area*, Migration in Sweden.
- Hemmens, G.C., 1966. *An analysis of urban travel and the spatial structure of urban activities* (Doctoral Dissertation).
- Hipp, J.R., Faris, R.W. & Boessen, A., 2012. Measuring “neighborhood”: Constructing network neighborhoods. *Social Networks*, 34(1), pp.128–140.
- Horvát, E.-Á. & Zweig, K.A., 2013. A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs. *Social Network Analysis and Mining*, 3(4), pp.1209–1224.
- Hristova, D. et al., 2016. A multilayer approach to multiplexity and link prediction in online geo-social networks. *EPJ Data Science*, 5(1), p.24.
- Johnston, R.J., 1968. Choice in Classification: The Subjectivity of Objective Methods. *Annals of the Association of American Geographers*, 58(3), pp.575–589.
- Johnston, R.J., 1970. Grouping and regionalizing: some methodological and technical observations. *Economic Geography*, 46, p.293.
- Kang, C. et al., 2013. *Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages*, New York, New York, USA: ACM.
- Kariel, H.G. & Welling, S.L., 1977. A NODAL STRUCTURE FOR A SET OF CANADIAN CITIES USING GRAPH THEORY AND NEWSPAPER DATELINES*. *The Canadian Geographer / Le Géographe canadien*, 21(2), pp.148–163.
- Kivran-Swaine, F., Govindan, P. & Naaman, M., 2011. The impact of network structure on breaking ties in online social networks. In the 2011 annual conference. New York, New York, USA: ACM Press, p. 1101.
- Korzybski, A., 1933. A non-Aristotelian system and its necessity for rigour in mathematics and physics. *Science and Sanity*, 1933, pp.747–761.
- Lancichinetti, A. & Fortunato, S., 2009. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5), p.056117.

- Lancichinetti, A. & Fortunato, S., 2011. Limits of modularity maximization in community detection. *Physical Review E*, 84(6), p.066122.
- Liu, Y. et al., 2015. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, 105(3), pp.512–530.
- Liu, Y. et al., 2014. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data P. Csermely, ed. *PLOS ONE*, 9(1), p.e86026.
- Luce, R.D., 1950. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15(2), pp.169–190.
- Manley, E., 2014. Identifying functional urban regions within traffic flow. *Regional Studies, Regional Science*, 1(1), pp.40–42.
- Neal, Z., 2012. Structural Determinism in the Interlocking World City Network. *Geographical Analysis*, 44(2), pp.162–170.
- Neal, Z., 2014. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39, pp.84–97.
- Newman, M.E.J. & Girvan, M., 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), p.026113.
- Nystuen, J.D. & Dacey, M.F., 1961. A graph theory interpretation of nodal regions. *Papers of the Regional Science Association*, 7(1), pp.29–42.
- Openshaw, S., 1977. A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling. *Transactions of the Institute of British Geographers*, 2(4), p.459.
- Openshaw, S., 1973. A regionalisation program for large data sets. *Computer Applications*, 3(4), pp.136–147.
- Openshaw, S., 1978. An empirical study of some zone-design criteria. *Environment and Planning A abstract*, 10(7), pp.781–794.
- Openshaw, S., 1984. The modifiable areal unit problem. In Geo Abstracts University of East Anglia.
- Openshaw, S. & Rao, L., 1995. Algorithms for reengineering 1991 Census geography. *Environment and Planning A abstract*, 27(3), pp.425–446.
- Openshaw, S. & Taylor, P.J., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical applications in the*

- spatial sciences*, 21, pp.127–144.
- Patefield, W.M., 1981. Algorithm AS 159: An Efficient Method of Generating Random $R \times C$ Tables with Given Row and Column Totals. *Applied Statistics*, 30(1), p.91.
- Poorthuis, A. & Zook, M., 2014. Artists and Bankers and Hipsters, Oh My! Mapping Tweets in the New York Metropolitan Region. *Cityscape*, 16(2).
- Poorthuis, A. & Zook, M.A., 2017. Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. *Journal of Urban Technology*, 36, pp.1–21.
- Ratti, C. et al., 2010. Redrawing the Map of Great Britain from a Network of Human Interactions O. Sporns, ed. *PLOS ONE*, 5(12), p.e14248.
- Rosvall, M. & Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), pp.1118–1123.
- Rosvall, M. & Bergstrom, C.T., 2011. Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems F. Rapallo, ed. *PLOS ONE*, 6(4), p.e18209.
- Rosvall, M., Axelsson, D. & Bergstrom, C.T., 2010. The map equation. *The European Physical Journal Special Topics*, 178(1), pp.13–23.
- Schaub, M.T., Delvenne, J.-C., et al., 2012. Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit O. Sporns, ed. *PLOS ONE*, 7(2), p.e32210.
- Schaub, M.T., Lambiotte, R. & Barahona, M., 2012. Encoding dynamics for multiscale community detection: Markov time sweeping for the map equation. *Physical Review E*, 86(2), p.026112.
- Schönfelder, S. & Axhausen, K.W., 2003. Activity spaces: measures of social exclusion? *Transport Policy*, 10(4), pp.273–286.
- Scott, J., 1992. *Social Network Analysis (A Handbook)*, London: Sage.
- Shelton, T., Poorthuis, A. & Zook, M.A., 2015. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, pp.1–14.
- Shevky, E. & Bell, W., 1955. *Social area analysis*, Stanford: Stanford University Press.
- Shevky, E. & Williams, M., 1949. *The social areas of Los Angeles, analysis and typology*, Pub. for the John Randolph Haynes and Dora Haynes Foundation by the

Univ. of California Press.

- Slater, P.B., 1981. Comparisons of aggregation procedures for interaction data: An illustration using a college student international flow table. *Socio-Economic Planning Sciences*, 15(1), pp.1–8.
- Spielman, S.E. & Thill, J.-C., 2008. Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, 32(2), pp.110–122.
- Stegbauer, C. & Rausch, A., 2012. How International Are International Congresses? *Connections* (02261766), 32(1), pp.1–11.
- Stephens, M. & Poorthuis, A., 2014. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems*, pp.1–9.
- Takhteyev, Y., Gruz, A. & Wellman, B., 2012. Geography of Twitter networks. *Social Networks*, 34(1), pp.73–81.
- Thomas, I. et al., 2012. Revisiting the extension of the Brussels urban agglomeration: new methods, new data... new results? *Belgeo*, (1-2).
- van Meeteren, M. & Poorthuis, A., 2017. Christaller and “big data”: recalibrating central place theory via the geoweb. *Urban Geography*.
- Viamontes Esquivel, A. & Rosvall, M., 2011. Compression of Flow Can Reveal Overlapping-Module Organization in Networks. *Physical Review X*, 1(2), p.021025.
- Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), p.236.
- Wellman, B., 1979. The Community Question: The Intimate Networks of East Yorkers. *American Journal of Sociology*, 84(5), pp.1201–1231.
- Wellman, B. & Leighton, B., 1979. Networks, Neighborhoods, and Communities: Approaches to the Study of the Community Question. *Urban Affairs Review*, 14(3), pp.363–390.