User-Generated Big Data and Urban Morphology

Article *in* Built Environment · October 2016

DOI: 10.2148/benv.42.3.396

Arie Croitoru
George Mason University

Ron S Mahabir
George Mason University

User-Generated Big Data and Urban Morphology

A.T. CROOKS, A. CROITORU, A. JENKINS, R. MAHABIR, P. AGOURIS and A. STEFANIDIS

Traditionally urban morphology has been the study of cities as human habitats through the analysis of their tangible, physical artefacts. Such artefacts are outcomes of complex social and economic forces, and their study is primarily driven by traditional modes of data collection (e.g. based on censuses, physical surveys, and mapping). The emergence of Web 2.0 and through its applications, platforms and mechanisms that foster user-generated contributions to be made, disseminated, and debated in cyberspace, is providing a new lens in the study of urban morphology. In this paper, we showcase ways in which user-generated 'big data' can be harvested and analyzed to generate snapshots and impressionistic views of the urban landscape in physical terms. We discuss and support through representative examples the potential of such analysis in revealing how urban spaces are perceived by the general public, establishing links between tangible artefacts and cyber-social elements. These links may be in the form of references to, observations about, or events that enrich and move beyond the traditional physical characteristics of various locations. This leads to the emergence of alternate views of urban morphology that better capture the intricate nature of urban environments and their dynamics.

The world is in the middle of an urbanization wave with worldwide urban population having grown from about one third of the global population in 1950 to over half today, and projected to reach two thirds by 2050 (United Nations, 2014). Accordingly, advancing our ability to capture the form and function of these urban realms is becoming ever more essential for understanding and planning daily life. There are many different perspectives concerning the exact meaning of form and function (e.g. Lynch, 1960; Talen, 2003), and the distinction among these terms is often vague and context-dependent (Levy, 1999). In fact, the term is, to an extent, the mantra of the modern movement in architecture – 'form follows function' – with the term urban morphology often used in conjunction with both terms (e.g. Batty and Longley, 1994; Moudon, 1997; O'Sullivan, 2000). Crooks *et al*. (2015) argue that 'urban form' refers to the aggregate of the physical shape of the city (i.e. its buildings, streets, etc.) whereas the term 'urban function' refers to the activities that take place within these spaces. These activities are not only enabled and driven by form, but also affect it, while providing semantic attribution to it. Form and function are therefore highly interrelated – different sides of the same coin, so-to-speak – as urban spaces emerge and evolve from the coalescence and symbiotic interaction of infrastructures, people, and activities (Besussi *et al*., 2010).

Early efforts to capture urban form and function and their interdependence have been hindered due to the lack of data at sufficiently fine spatial, temporal, and thematic resolutions. This challenge has limited our ability to capture sufficiently the complex dynamics that shape and continuously reshape these urban realms. The emergence of smart cities is bringing big data to the study of cities and their morphology and provides a critical operational link to capture directly and indirectly the content of both form and function. With information and communications technologies (ICT) being merged, coordinated, and integrated with traditional urban infrastructure (Batty *et al*., 2012), we are now able to access dynamic content provided from the field, thus generating *snapshots of the city* as a functioning system.

At the same time as technology has been driving the embedding of ICT in the infrastructure of the cities, it has also been driving the crowdsourcing phenomenon (Howe, 2008). When it comes to spatial information, crowdsourcing has led to the general public contributing information from and about the space within which the city functions (Goodchild, 2007). This information can be harvested to generate snapshots of the city, but these snapshots often differ from the ones generated through more traditional sensor networks. Crowd-generated content often captures one's *impression* of space far better than this information could be deduced from the digital data that are generated through embedded sensors in the smart city. Therefore, one could argue, drawing an analogy from the artistic trends of the mid-nineteenth century, that we have a concurrent emergence of an equivalent photography and impressionism when it comes to capturing the essence of urban morphology. Sensor networks and traditional volunteered geographical information (VGI) are thus providing photographic-like data-driven snapshots of the space, whereas crowd-harvested content produces impressionistic views of the same space, mirroring the general public's impressions of urban space, typically in the form of platial[1] content (Jenkins *et al*., 2016).

In this paper we focus primarily on the latter, namely the impressionistic views of urban spaces, as they can be derived through crowd-generated content. We discuss the relations to the snapshots that one gets from geometrically-driven data sources, and the different types of analysis that are both required for and supported by this newly emerging type of information. In the second section we provide a brief discussion of the emergence of crowd-harvested information and argue that to truly understand urban spaces we need to move beyond just projecting datasets into space towards analyzing the space itself as it relates to people within it. We then discuss how we can derive insights into urban morphology from crowd-contributed data starting with data driven snapshots of urban form. We discuss more impressionistic views of the city via harvesting and analyzing social media data and how these data can be combined via linked spaces in the form of geosocial neighbourhoods in the third section. Finally we conclude the paper with a summary and outlook.

Crowd-Harvested Data and Urban Infoscapes

The emergence of Web 2.0 at the turn of the millennium led to newfound opportunities for the general public to contribute information which can be harvested and analyzed to capture a wide spectrum of content ranging from the mundane to the profound. Fortuitously, at the same time, a proliferation of geolocation capabilities emerged, leading to a large proportion of this crowd-generated content being geolocated, i.e. having a set of geographical coordinates indicating its points of origin. When it comes to the geospatial content of these crowd-contributions, part of it is explicit, such as in the user-generated content of OpenStreetMap (Haklay and Weber, 2008). However, the majority of it is of implicit geographical value: the contributions are not in the form of maps or coordinates, but their content can be harvested and analyzed geospatially in order to reveal valuable patterns (Stefanidis et al., 2013). This is particularly the case for social media contributions.

The term *social media* is typically used to refer to services such as Facebook, Twitter, Flickr and YouTube, which enable the public to communicate with their peers, sharing information with them instantly and constantly in an effortless and intuitive way. Bypassing the need for advanced computing skills to participate, and fostering social interaction in cyberspace, social media has revolutionized information dissemination and presents an alternative means for community formation (Croitoru *et al*., 2015). Today, Facebook has nearly 1.5 billion monthly active users worldwide (exceeding the populations of both China and India), while Twitter and Instagram have in excess of 300 and 400 million users, respectively.[2] These communities contribute massive amounts of crowd-generated data: currently, every minute, there are over 300,000 status updates in Facebook and 450,000 new tweets, while 65,000 new photos are uploaded in Instagram.[3] This is leading to the emergence of a new paradigm involving 'big data' and human social media (see Croitoru *et al*., 2014).

When it comes to urban environments, social media has the potential to foster citizen engagement and participation in urban planning, but recent studies have shown that this potential is only rarely met, and only then is it effective if supported by accompanying actions that engage the community in the physical space (Kleinhans *et al*., 2015). Instead, greater potential is presented through the analysis of crowd-generated data jointly with sensor-derived content and this is leading to a better understanding of the manner in which cities function as systems. City dashboards serve as an effective visualization paradigm to aggregate and communicate such content, harvesting data from the application programming interfaces (APIs) of diverse, accessible databases. For example, the city dashboards shown in figure 1 developed by Gray and O'Brien (2016, this issue) at University College London's Centre for Advanced Spatial Analysis (CASA) aggregates data from distributed sensor feeds (e.g. weather, traffic cameras, pollution sensors), news feeds, and social media contributions (e.g. Twitter trending terms) to offer a snapshot of the city in question at a given time. Such content is visually impressive, and comparable information dashboards are being used to support operations centres around the world (Mattern, 2015), doing so successfully especially in times of crisis. However, one would be hard pressed to come up with arguments for how they are advancing our *understanding* of urban environments. These approaches have been driven by instru-

Figure 1. CASA's city dashboards. (*Source*: http://citydashboard.org/)

mentation rather than science (Kitchin *et al.*, 2015) but we are still unable comprehensively to abstract these feeds and the trends they signify into complementary and integrable elements of knowledge that can be further analyzed and synthesized.

Social media content can help us overcome this limitation as these contributions inherently embed our own understanding of the urban environment and as such, can be analyzed to derive such content. For example, in a non-spatial manner, Bryden *et al.* (2013) have studied the relationship between linguistic term usage and social networking, and showed how term usage in Twitter can be studied to predict community membership. Projecting these patterns onto the physical space, Fischer has revealed the global spatial distribution of different communities in Twitter,[4] and Cheshire[5] has shown how the same approach can be used to identify the spatial distribution of different ethnic communities in London (see Cheshire and Uberti, 2014). Zhong *et al.* (2015) have illustrated how mobility data can be analyzed to study the spatial structure of urban movement (identifying distinct spatial clusters) and their variability using smart card transportation data from Singapore as a test case. In the same manner and direction, Longley and Adnan (2016) show how Twitter data can be analyzed to perform a geo-temporal demographic classification of Twitter users in London, and through that, they are able to identify links among locations. Quercia *et al.* (2015) studied Flickr and Instagram annotations and terms encountered in Flickr contributions from various cities to identify references to various smells, and has compared these to walks in these neighbourhoods to assess how representative these terms are with respect to the corresponding area. Common to the majority of these approaches has been the aggregation of digital content generated within the city by its citizens (i.e. through social media) or embedded sensors (i.e. in a smart city environment) in the form of information landscapes that capture the variation of various properties in this urban space.

These processes are akin to fitting digital landscapes to the information content that is thus generated, and we can refer to these landscapes as *urban infoscapes.* Iaconesi and Persico (2015) have succinctly argued that these processes offer us the potential to move beyond the earlier paradigms of the *City of Bits* (Mitchell, 1995) and DigiPlace (Zook and Graham, 2007). This new paradigm enables a more dynamic view of urban spaces in the light of digital information, which allows us to better observe these spaces as they are made, and constantly re-made through human activities and expressions. Crooks *et al.* (2015) show how social media and VGI can be harvested and analyzed in order to derive urban form and function information. A key issue behind this emerging research direction is to derive information about the space itself, rather than simply projecting datasets upon that space, and this is the main focus of our paper. This will allow us to better understand how we inherently perceive the space in which we move, how we assess the degree to which it matches our needs and wants, and how we assign meaning to it making it a *place* rather than a set of coordinates. It is through such advances that we will attain the level of abstraction that is needed to merge disparate views of cities contributed by man and machine, and eventually use such information to improve participatory urban citizenship, with the public contributing, and receiving information from the space in which it functions. A depiction of such an infoscape is shown in figure 2 where we move from physical space layers (e.g. L_1 and L_2) towards more social and perceptual spaces (e.g. L_3) and then towards a more abstract high-level understanding of locations (e.g. N_1 and N_2) and the relationship between them (as depicted in L_4). In the next section, we provide examples of how crowd-contributed information (from VGI and social media) provide snapshots of different space types within the city.

Urban Morphology Snapshots

Through crowd-generated data, we have new sources of data that not only complement traditional data sources but allow us to study urban morphology in a variety of different ways. As with more traditional data sources (e.g. census, surveys, remote sensing), such data only provide snapshots of the city. What we mean by this is that such data have been collected or contributed over a specific time range and thus only offer a snapshot of the city from the time when they were contributed or harvested, unlike more established methods such as the decadal census in the US or UK, or from national mapping agencies where data can be retrieved back over many years. This new type of data only currently go back a limited number of years, and therefore looking at historical changes within cities is limited. For example, OpenStreetMap only began in 2004 and while Twitter was launched in 2006, and unless one is willing to pay for historical data, one can only go back and retrieve the last 3,200 tweets per user.

In this section, we provide a selection of representative examples from our own work or that of others with respect to how crowd-generated information can provide us snapshots of urban morphology. This enables us to generate the physical and perceptual spaces of cities, as well as showing us how such data can provide us with the ability to study linked spaces which are a combination of physical and perceptual spaces in the formation of geosocial neighbourhoods.

Physical Spaces

We start our discussion of physical space through the geometry of the city, its buildings, and its streets. To some extent this is the most straightforward information to collect in the sense that cities are physical in nature and, since classical times, people have been documenting their physical structure. Furthermore, for at least a century, this has been a serious study with a great deal of fascination with respect to urban form (e.g. Abercrombie, 1933; Batty and Marshall, 2009; Howard, 1902). Ever since the emergence of VGI, researchers have been exploring how such information can be utilized to study

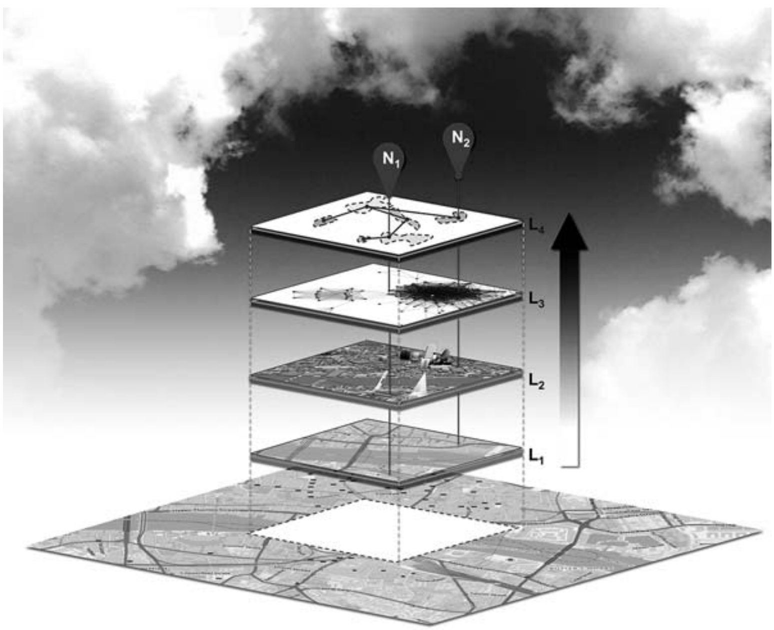

Figure 2. City infoscapes – fusing data from physical (L_1, L_2), social, perceptual (L_3) spaces to derive place abstractions (L_4) for different locations (N_1, N_2).

cities. If we take a building analogy, the most basic approaches to collecting information via the crowd include the use of Google SketchUp and Google's 3D Warehouse, which creates a geographically tagged database of three dimensional objects. For example in figure 3, we show 3D buildings built in SketchUp which can be easily overlaid on Google Earth to provide a sense of the built environment. To some extent this is the digital manifestation of the architect's traditional wooden and paper block models as illustrated in their digital transformation by Batty and Hudson-Smith (2005).

While such data are useful, another line of research has emerged via crowd-generated data, specifically with the growth of open source data. For example, OpenStreetMap (OSM) not only provides information pertaining to roads but much more. For example in figure 4(*a*) we show a screen shot of part of part of New York City from OpenStreetMap which not only shows roads, points of interests, shops, traffic directions but also building footprints. In figure 4(*b*), we show a screenshot from OSM Buildings which takes OpenStreetMap building footprints and information about their height attributes and builds a 3D city model which in essence gives a sense of the characteristics of the urban environment.

As one might expect, initial lines of research have focused on how such crowd-generated information compares to more traditional sources of information (in terms of its accuracy and completeness). This relates to the fact that such crowdsourced data is usually provided with little or no information on mapping standards, quality control procedures, and metadata in general (Feick and Roche, 2013). Recent studies have already started to assess the quality of OpenStreetMap content by comparing it to established authoritative mapping organizations such as the UK Ordnance Survey road datasets (Haklay, 2010) or commercial (e.g. TomTom, NAVTEQ) data providers (see Girres and Touya, 2010; Zielstra and Zipf, 2010). These studies do note that such data are comparable to more authoritative sources. Studies assessing crowdsourced point of interest content such as bars, schools, hospitals (e.g. Hochmair and Zielstra, 2013; Jackson *et al*., 2013; Mullen *et al*., 2015) or building footprints (Fan *et al*., 2014) also indicate that crowdsourced datasets are increasingly becoming comparable to authoritative datasets.

Moreover, such data allow us to explore areas that were not previously digitally mapped, were poorly mapped or where the environment has changed dramatically due, say, to a natural disaster (e.g. Zook *et al*., 2010). For example, in figure 5 we not only show a less developed country but also how crowd-generated datasets are comparable to more authoritative data. Specifically, we show dif-

Figure 3. George Mason University's Fairfax Campus built using SketchUp and visualized in Google Earth.

ferent road coverage datasets per km² for Nairobi and surrounding areas in Kenya. These datasets are either based on authoritative data (e.g. from the Regional Centre for Mapping of Resources for Development (RCMRD)) or non-authoritative crowd-sourced data (e.g. Google MapMaker and/or OSM). As can be seen, coverage varies between

(*a*)

(*b*)

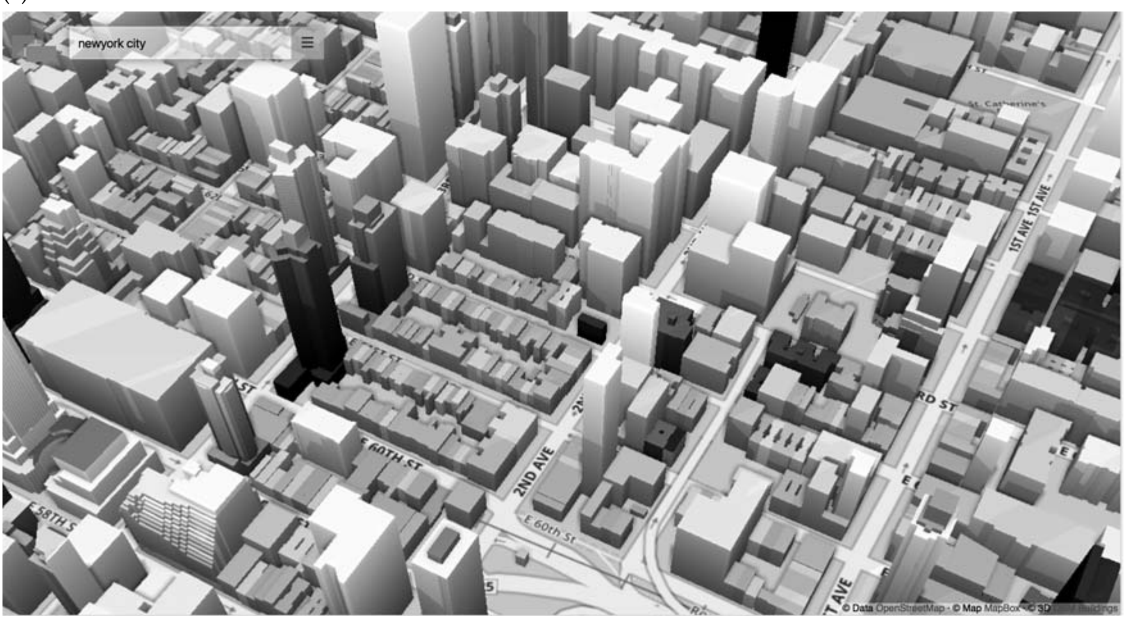

Figure 4. Part of New York City as seen in OpenStreetMap (*a*) and the resulting 3D city model from OSM Buildings (*b*). (*Sources*: http://osmbuildings.org & http://osmbuildings.org)

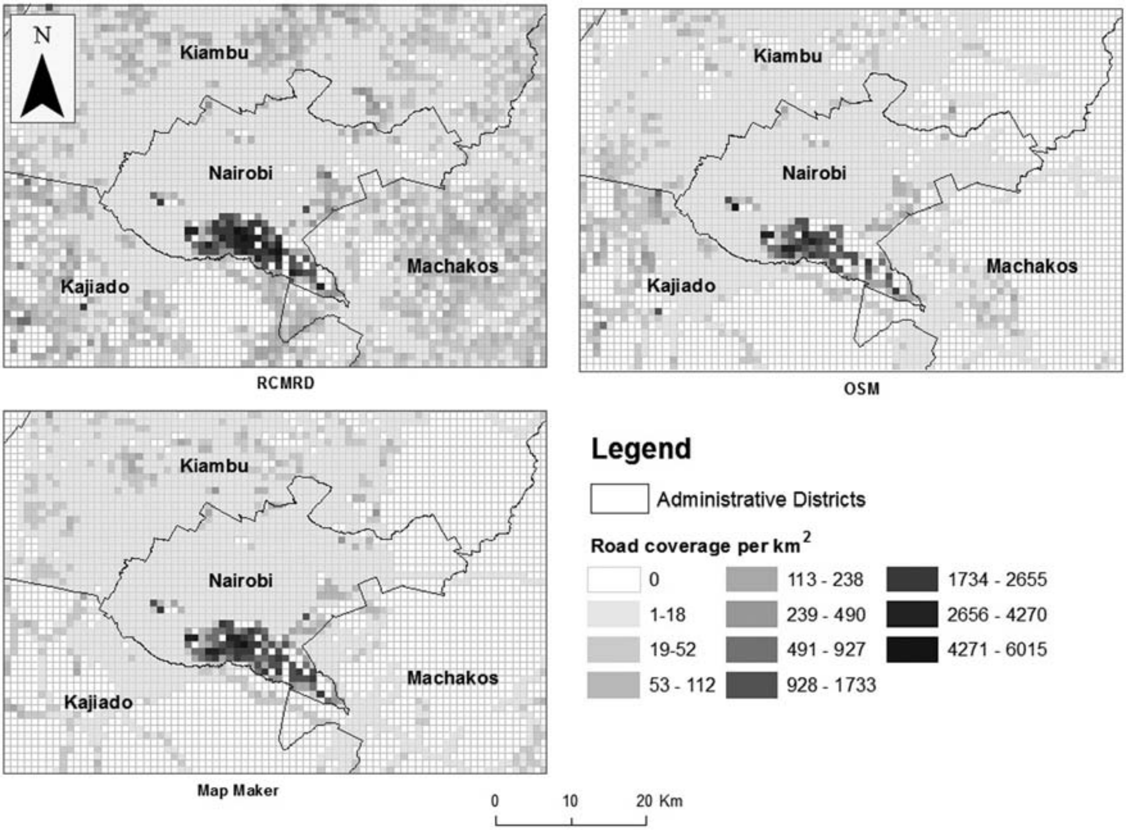

Figure 5. Examples of road coverage morphology in the broader Nairobi area, Kenya as they are articulated though different data sources.

the different datasets, whereby the non-authoritative data provide more coverage in central Nairobi and low coverage in rural areas when compared to authoritative data (i.e. RCMRD). However, such data mentioned above only provide snapshots of the area and the physical properties of areas of interest. In this sense, although we know when the lines and points were edited, we know little else.

That being said, such data on a massive scale do give us new opportunities to study the physical properties of cities. For example, due to data availability, computational resources and copyright issues, research pertaining to the study of urban street networks was often limited to one city (e.g. Masucci *et al.*, 2009); however, with the prolific growth of OpenStreetMap and greater computational resources, researchers now have at their disposal data to explore street networks across different cities within the same country or across continents. Such data allow us to ask questions pertaining to the topological structure of urban street networks and urban development (e.g. Jiang *et al.*, 2014; Louf and Barthelemy, 2014) or offer a new lens to study city boundaries or the scaling properties of cities (Jiang and Jia, 2011).

Perceptual Spaces

While the work presented above captures the physical elements of the city, what it fails to capture is the functional aspects of the social morphology. For this we turn to other forms of crowd-contributed data, specifically that from social media. Initially it was thought that technology, and in particular communication

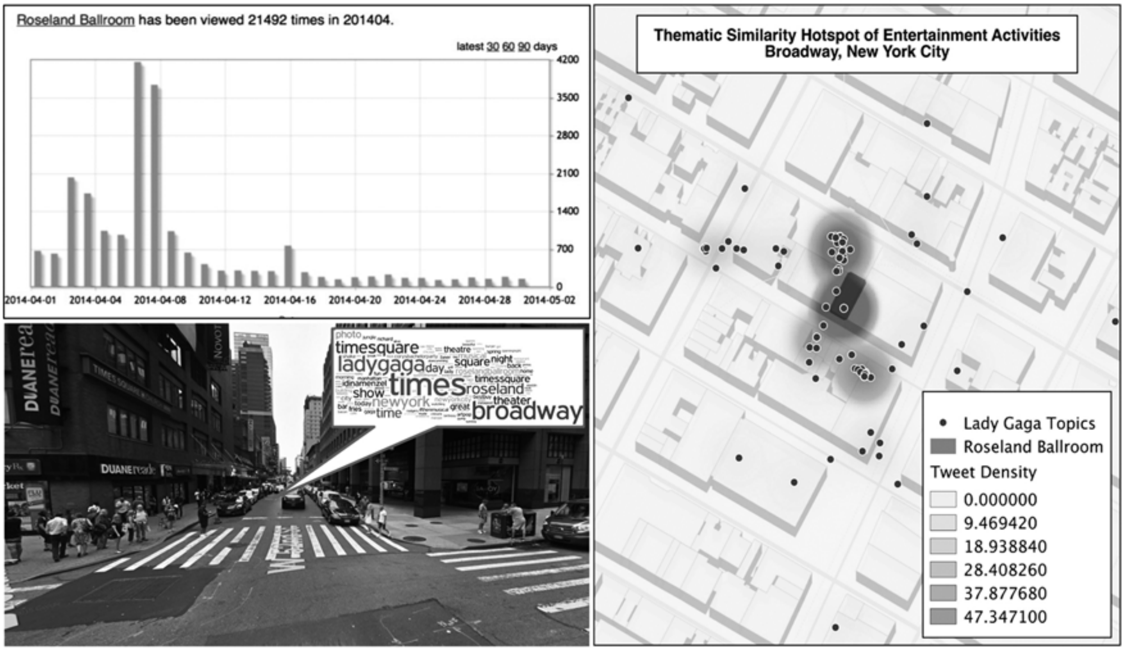

Figure 6. Roseland Ballroom (Broadway, New York City) event featuring Lady Gaga.

technology, would greatly diminish the effects of space on human interactions. However, this was found not to be the case as human activity is largely rooted in physical space, making such activity 'immune' to some extent to the effects of information and communications technologies (ICTs) (Torrens, 2008). In light of this, the challenge is how communication technologies could be utilized to explore interactions between people and place (McCullough, 2005). In this context geography plays a crucial role in bounding the cyber (i.e. social media content) and physical spaces.

In order to showcase the transformation of user-contributed big data into such city infoscapes, we present several vignettes focusing on information that emerges from the spaces themselves. To briefly describe the approach, we gather user-contributed content in the form of geolocated tweets from the Twitter streaming API (approximately 4.5M tweets) and extract (harvest) the textual terms that co-occur with high frequency to reduce the signal-to-noise ratio that is often encountered in such social media. The extracted terms are then semantically linked to higher-level categories using an existing knowledge base (labelling 'theatre' as entertainment, for example). This step abstracts the multitude of individual activities within the spaces into a common semantic group whereby people are experiencing the same event, but using different terms that are ultimately given a similar meaning. Spatial analysis is then performed on the resulting geolocated terms to find statistically significant clusters in each thematic category. This process results in a further reduction of the data volume by finding spatial clusters containing only high semantic values. The Twitter points depicted in the maps in figures 6, 7, and 8 were derived using this methodology.

The first vignette, shown in figure 6, depicts the emergence of an entertainment related event at the Roseland Ballroom in New York City's Broadway district. This event by singer Lady Gaga marked the final show at Roseland Ballroom prior to the building's destruction in 2014. From the space surrounding the event, we observe several artefacts of information in the form of spatial alignment, content analysis, and place-based cyberactivity. Specifically, a spatial cluster of entertainment related tweets that overlap with the building and the surrounding area emerges. In addition, the tweet density depicts the strongest cluster centred on the building itself. This spatial alignment highlights how social media can be semantically linked with the affordances of the building. Using the textual content of the tweets, we generate a word cloud of the most frequent terms to create concordance between Lady Gaga and the Roseland Ballroom. We can also get additional indication of the event by examining the page view statistics of the Roseland Ballroom Wikipedia article to find a significant increase on the date of the concert. Undoubtedly, in this case, the popularity of the performer would have created a spatial cluster regardless of the location. However, the historical precedence of this being the final show at the ballroom and the looming change of landscape caused a parallel emergence of cyberactivity.

The second vignette, shown in figure 7, investigates the use and needs of two different locations that are both characterized by recreational activities. In figure 7, the labelling is as follows: (A) High Line Park; (B) Madison Square Garden. High Line Park is an elevated park constructed on an old railway line through the western portion of central Manhattan, New York City. Madison Square Garden is a multipurpose indoor arena that hosts various sporting and performance events. In this case, we extend the same semantic labelling process as described above by looking at users classified as either 'local' or 'tourist' with respect to a specific area based on the spatiotemporal patterns of their social media contributions over a period of several months. Using this approach, we can gauge the views of the same location by

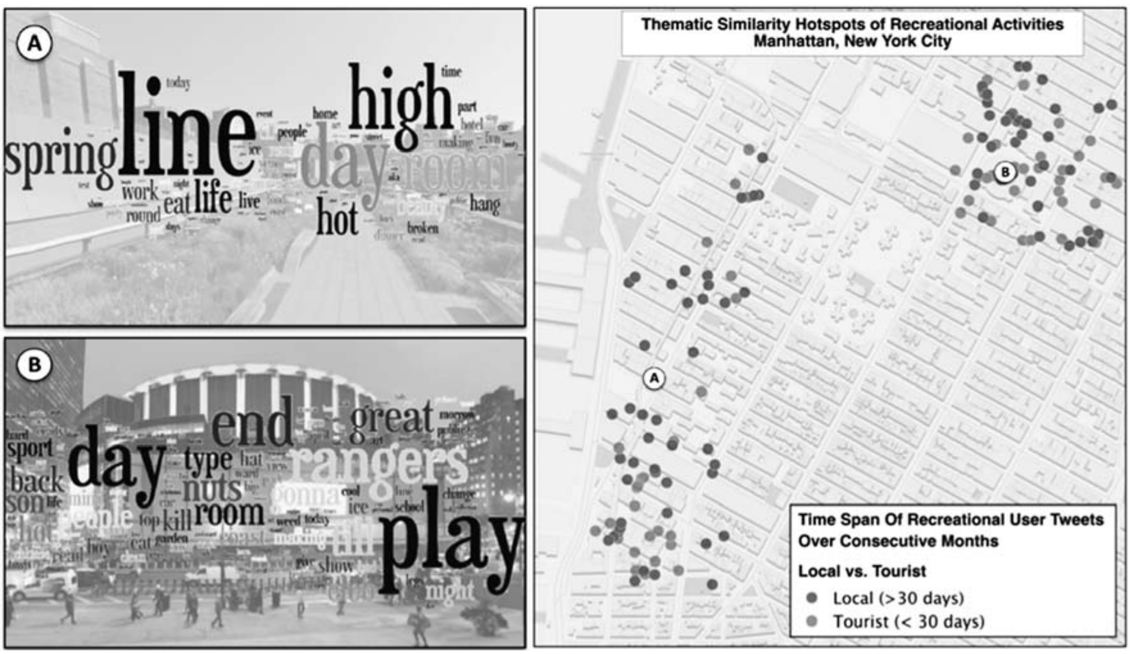

Figure 7. Recreational hotspots composed of 'locals' and 'tourists' with perceived artefacts indicating 'use' and 'need'. (A) High Line Park; (B) Madison Square Garden.

locals and tourists, and discern alignments and discrepancies between them. Through such an analysis, we can see how one's familiarity with the landscape leads to different conceptualizations of the same urban physical space.

In addition to the impact of events on platial signatures, and the potential dichotomy between local and tourist views, we can also have diurnal variations on the perceived character of urban place. In the vignette visualized in figure 8, we show locational changes of politically related tweets within downtown London over the course of a 24-hour period. Once again, by extending the base methodology as described before we temporally bin tweets into 12-hour groups, which we refer to as day and night. During the day, we observe political tweets near government buildings and along the River Thames with its numerous restaurants and bars. Comparatively, political tweets in the night category tend to concentrate to the north of Central London in the Soho area. This simple example conveys a perceived diurnal cycle of daily routines of political discussion within the area, indicating the use of these areas at different times of the day provided through the space itself.

The approach we present above is not only restricted to small areas but can be applied to entire cities or countries, enabling the expansion of the impressionistic view of spaces to any number of thematic categories as we show in figure 9. Here we show education, sport and entertainment hotspots in Singapore. Moreover, we would argue that such hotspots and their alignment to the actual physical infrastructure have important implications for urban planning. For example,

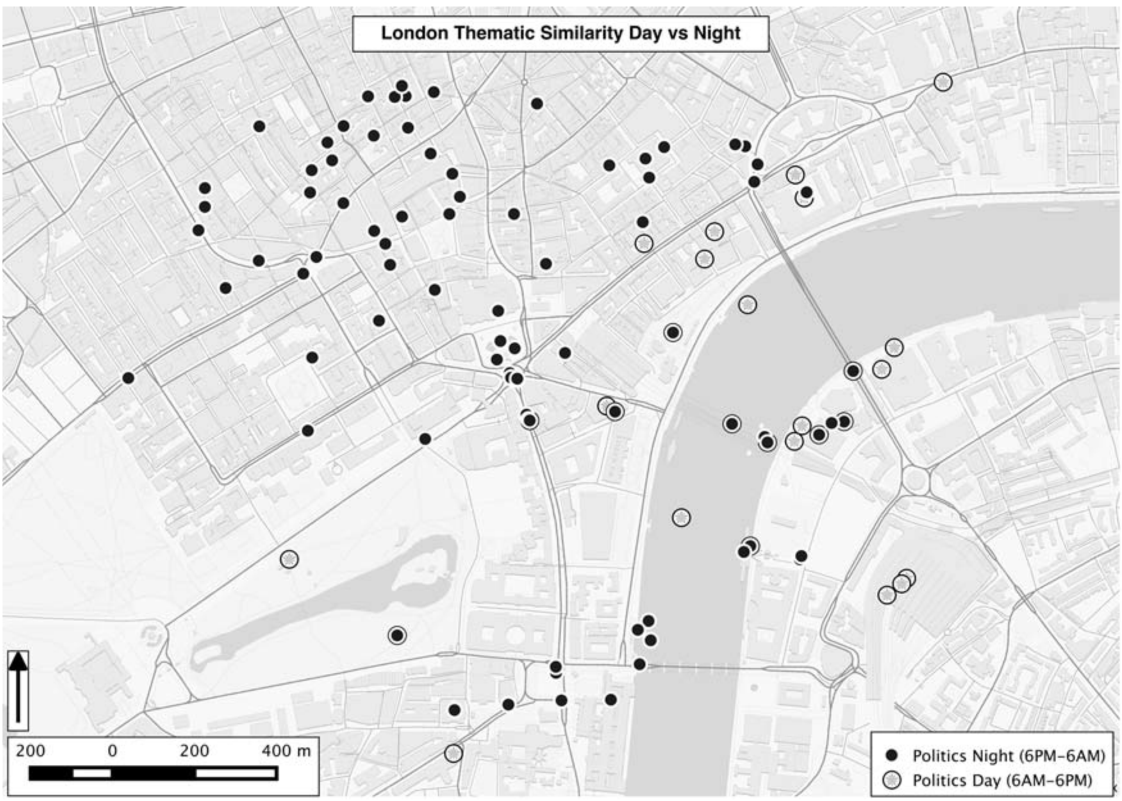

Figure 8. Political hotspots in London temporally separated into 'day' and 'night'.

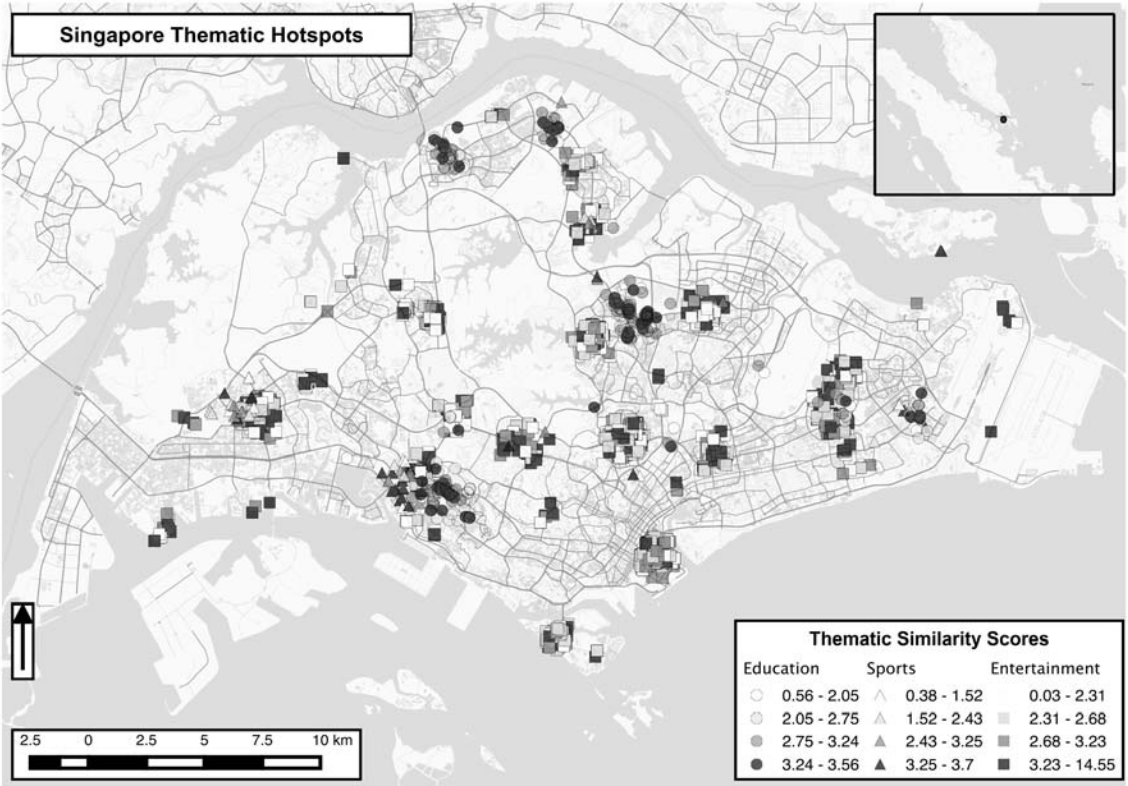

Figure 9. Thematic hotspots for education, sports and entertainment in Singapore.

if the recreational thematic hotspots are aligned with actual physical locations of recreation we could say that such locations are meeting the needs of the people. However, if there is misalignment, for example people are discussing recreational topics and there are no physical recreational resources, this could highlight a need for a population that is not met by the physical space, which could be addressed through appropriate planning.

The above vignettes highlight how we can generate understanding from crowd-harvested content, and, with further work, it is not unimaginable to have a special kind of augmented location-based mashup (Anand *et al*., 2010) from such analysis with other sources of information such as Google Street View as we show in figure 10. Such information could convey an aggregated crowd-sourced view of a place, which could lead to the emergence of immersive augmented reality city dashboards rather than the summary views that currently represent the norm.

Linked Spaces

While the examples above have explored how user-generated content can be used to explore the physical and perceptual spaces of cities to some extent, these ignore the links between places. To explore this we need to turn to networks, and while the ideas of networks and urban morphology is not new (e.g. Haggett and Chorley, 1969; Atkin, 1974; Hillier and Hanson, 1984), much of the research to date has focused on the physical networks (e.g. roads etc.) with much less on the social aspects due to lack of data. While network

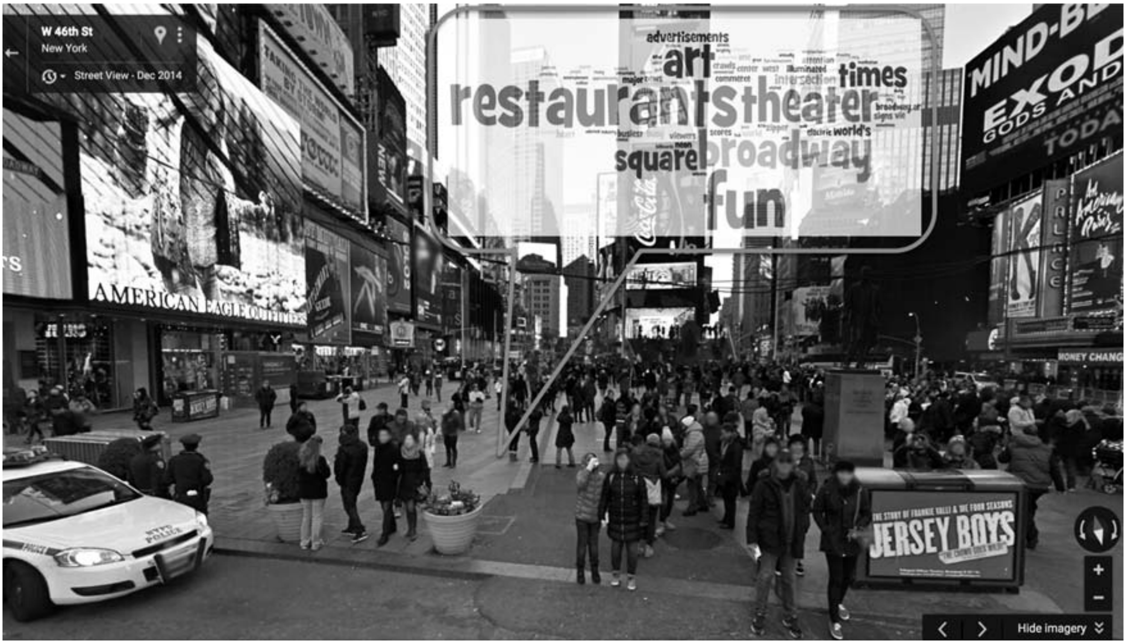

Figure 10. Perceptual information sourced from social media viewed within Google Street View.

analysis is a rapidly growing field (e.g. Lazer *et al*., 2009; Barabási, 2012), it has only recently emerged as a tool in geospatial analysis, and some would argue that it is underutilized (e.g. Ter Wal and Boschma, 2009). Others also suggest that social network analysis too is weakened because of the lack of geographic consideration when exploring social relations (e.g. Bosco, 2006).

The incorporation of links and geographic space can be accomplished by leveraging the inherent interconnected nature of social media. Social media, by its very nature, enables users to form links of various types and significance. Through such links, users can share information of interest, exchange ideas, or identify a social relation with other users (e.g. a 'friend', a 'colleague' etc.). Links between users can be characterized by three primary properties (Kietzmann *et al*., 2011): structure, flow, and strength. *Structure* relates to how a user is positioned and connected in the network of users in the service. *Flow* relates to the type of content and resources that are exchanged or transferred between users. Finally, the *strength* of a relationship between users is often determined by intensity of the flow – both in terms of frequency and content.

Of these properties, the notions of structure and flow are of particular importance in the context of geosocial neighbourhoods. As the users – and their networks – are embedded in geographical space, the structure of the users' network (as defined in social media) defines the location, extent, and shape of the emergent neighbourhood. Such structure, therefore, characterizes the physical morphology of the neighbourhood. At the same time, the flow (and the flow strength) provides a characterization of the type of information (and its intensity) that is exchanged within the network, providing information about the social morphology of the neighbourhood. This notion of geosocial neighbourhoods is depicted in figure 11. This notion expands the idea of a neighbourhood in that it goes beyond the strictly spatial definition of a

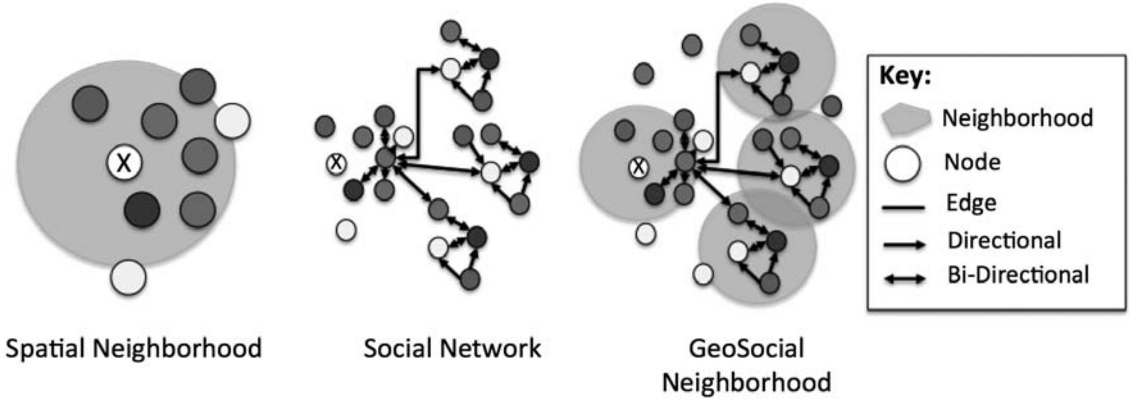

Figure 11. Moving from spatial neighbourhoods to geosocial neighbourhoods via links.

neighbourhood towards a definition that encompasses the links between spatial spaces as defined by the people that act in them. A geosocial neighbourhood is therefore not defined by its administrative boundaries, planning zones, or physical barriers, but rather by its emergence as an organic self-organized social construct that is embedded in geographical spaces that are linked by human activity.

Once a geosocial neighbourhood is formed, it may be of interest to track its evolution over space and time. In particular, as users form new links, or maintain or abandon existing ones, a geosocial neighbourhood may change its composition, structure, and shape. An example of such an evolution process is when a neighbourhood may redefine its footprint or relation with other neighbourhoods between time T1 and T2, or when a geosocial neighbourhood emerges or disappears between time T2 and T3, as shown in figure 12. Such evolution can provide additional information often leading to deeper insights into the underlying processes that govern geosocial behaviour.

The notion of linked spaces and the emergence of geosocial neighbourhoods can be realized by the harvesting and analysis of social media data. Social media feeds can be retrieved from the source data provider through user-driven queries to a service-specific API. This entails sending a query in the form of an http request and receiving data in XML or JSON formats. The query parameters may be, for example, based on location (e.g. a region of interest), time (period of interest), content type, or even by user

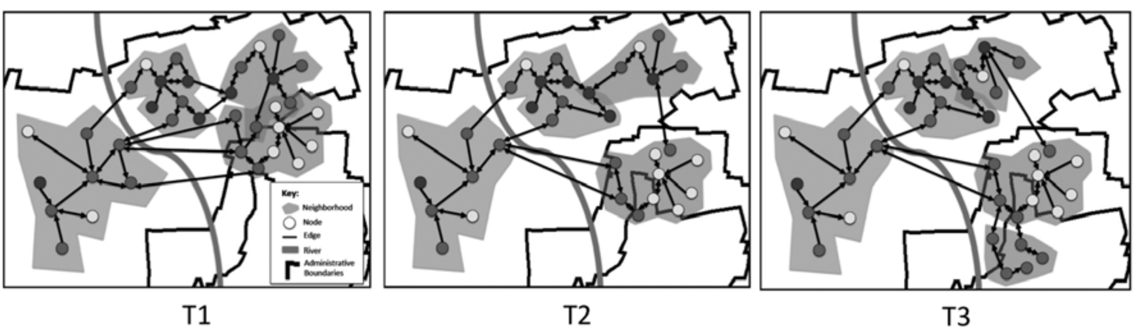

Figure 12. The evolution of geosocial neighbourhoods.

identifiers. Once such data is received, it is processed in order to identify two types of information: user location, and communication flow. In terms of user location, social media data can include location information that was captured directly through a geo-positioning technology (e.g. a GPS sensor), and location information that is implicitly embedded in the social media data (e.g. place names in the body of a message). Communication flow entails revealing links between users that are formed by the flow of information between them. The aggregation of such links results in a network that links across geographic spaces.

These ideas can be demonstrated by the social media activity on Twitter in the aftermath of the Boston Marathon bombing in 2013. Twitter data were collected from the publicly available streaming API using the Geosocial Gauge system (Croitoru *et al*., 2013), and a retweet network was constructed using only geolocated tweets. In order to explore the formation of geosocial neighbourhoods, we analyzed the local retweet networks in the conurbations of urban areas along the northern and central parts of the East Coast of the United States. In particular, the analysis focused on the Greater Boston area, the New York Tri-state area, and the Baltimore-Washington DC area. The results of this analysis over the first four hours after the bombing event are shown in figure 13. In this figure, geolocated users are depicted by circles (the circle radius is proportional to the user degree (i.e. the number of connections in the network) and retweet message exchanges between users are depicted by lines. As can be seen, geosocial neighbourhoods emerge at the conurbation level (for example the Greater Boston area in figure 13C), as well as between conurbations along the US East Coast (for example the connections between Boston, New York and Washington DC in figure 13A).

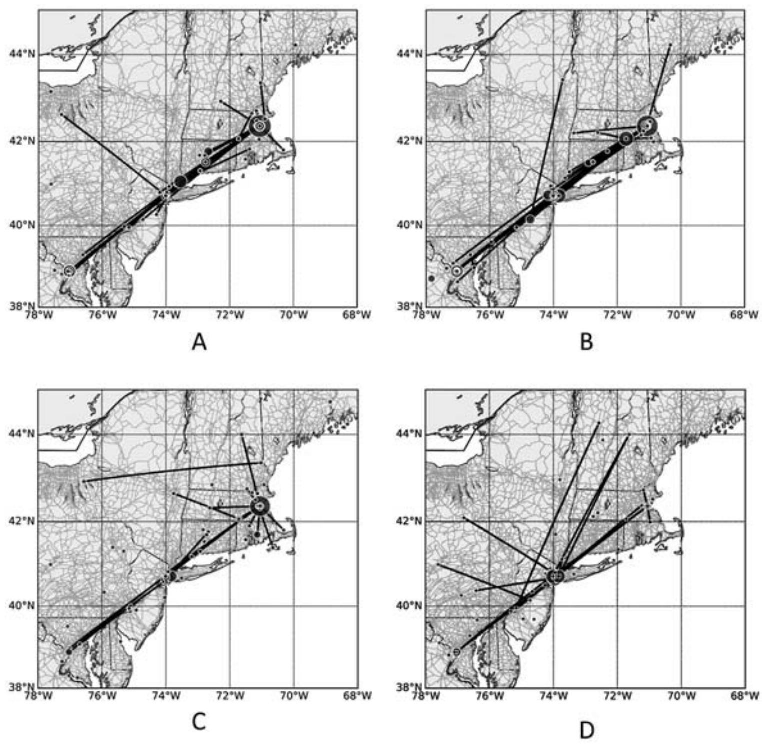

Figure 13. The emergence of geosocial neighbourhoods in the aftermath of the 2013 Boston Marathon bombing, from one hour after the bombing (A) to four hours after the bombing (D).

Summary and Outlook

We are witnessing a data deluge with respect to how data are collected and fused together which can act as a basis to monitor how cities function and this is providing us with many new ways to analyze and model future urban trajectories. This is of crucial importance as the world's urban population continues to grow to the point where most people will be living in cities by the end of this century.

This paper has shown how one can harness crowd-contributed data to develop dynamic views of urban spaces – city infoscapes which move beyond just projecting datasets upon space but define information about the space itself. Such information not only provides us with new ways of knowledge discovery but offers us a way to enhance inference for processes that have a long tradition in urban planning (Miller and Goodchild, 2015). Moreover, the growth in social media allows us to move beyond the photographic snapshots provided by VGI, that of, say, the geometry of the urban environment as provided by OpenStreetMap, to more impressionistic views of the city by the people within it. Taken together, this form and function provides us with a new way to explore spaces (linked spaces). While such data only provide snapshots of processes (e.g. peoples' daily activities) over time, as more data are collected and techniques are developed, we can potentially start exploring daily, weekly monthly patterns of activity and so on.

However we are only at the start of this kind of research. Methods need to be developed to collect, store and analyze such volumes of data and merge them together for cross-source analysis (Croitoru *et al.*, 2013) so that we can move beyond just short-term snapshots which in turn could be used in long-term planning for large areas. Moreover, we need methods to fuse this bottom-up crowd-contributed information to more traditional/authoritative (top-down) data sources to truly grapple with the complex nature of cities. This would therefore allow us to study cities as systems of systems (Batty, 2013), and to build indicators about the functioning of the city. Such indicators would allow us to explore issues which relate to inclusivity, health and safety etc. in the context of smart cities. By doing so we have the potential to move away from general ideas of how cities work to study specific cities and test and refine ideas and theories pertaining to the hierarchies and size of cities, land-use change, economic rent, locational theory and spatial competition at spatial and temporal scales not previously possible, thus providing us new opportunities to study cities and investigate urban problems.

As we have illustrated in this paper, open-source and crowd-contributed data are providing us with multiple portraits of the city, ranging from photographic-style snapshots as they can be derived from sensor feeds, to impressionistic portraits constructed from user-generated contributions such as social media. The latter in particular are especially important as they can be viewed as self-portraits of the city, painted by its citizens, one contribution at a time. Their importance is due to the fact that they encapsulate a level of understanding that is still missing in the traditional processing of sensor feeds. As is typically the case with impressionistic works, the brushstrokes of these city portraits may be broad, often sacrificing the accuracy of the outline, to focus instead on the perception of the scene. Accordingly, the resulting clusters may be imprecise by geographic standards, but are nevertheless invaluable as communications of the views of the city by its inhabitants. They can therefore have great use in the context of urban planning, as they articulate the use of the space, and relevant societal needs and wants.

NOTES

1. In the context of this paper the adjective *platial* is used to refer to the meaning assigned to various locations by humans through their shared experiences. Through these experiences locations

are transformed into places with particular socio-cultural characteristics (e.g. a business district or an entertainment one).

2. http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

3. http://thinktostart.com/how-much-big-data-is-generated-every-minute-on-top-digital-and-social-media-infographic/.

4. https://www.flickr.com/photos/walkingsf/6277163176/.

5. http://spatial.ly/2012/10/londons-twitter-languages/.

REFERENCES

Abercrombie, P. (1933) *Town and Country Planning*, 1st ed. London: Thornton Butterworth.

Anand, S., Batty, M., Crooks, A.T., Hudson-Smith, A., Jackson, M., Milton, R. and Morley, J. (2010) *Data Mash-ups and the Future of Mapping*. Bristol: Joint Information Systems Committee (JISC) Technology & Standards Watch (TechWatch) Horizon Scanning Report 10_01.

Atkin, R. (1974) *Mathematical Structure in Human Affairs*. London: Heinemann Educational.

Barabási, A. (2012) The network takeover. *Nature Physics*, **8**, pp. 14–16.

Batty, M. (2013) *The New Science of Cities*. Cambridge, MA: MIT Press.

Batty, M. and Longley, P. (1994) *Fractal Cities: A Geometry of Form and Functions*. London: Academic Press.

Batty, M. and Hudson-Smith, A. (2005) Urban simulacra. *Architectural Design*, **75**, pp. 42–47.

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. and Portugali, Y. (2012) Smart cities of the future. *The European Physical Journal Special Topics*, **214**, pp. 481–518.

Batty, M. and Marshall, S. (2009) Centenary Paper: the evolution of cities: Geddes, Abercrombie and the new physicalism. *Town Planning Review*, **80** pp. 551–574.

Besussi, E., Chin, N., Batty, M. and Longley, P. (2010) The structure and form of urban settlements, in Rashed, T. and Jürgens, C. (eds.) *Remote Sensing of Urban and Suburban Areas*. New York: Springer.

Bosco, F. (2006) Actor-network theory, networks, and relational approaches in human geography, in Aitken, S. and Valentine, G. (eds.) *Approaches to Human Geography*. London: Sage.

Bryden, J., Funk, S. and Jansen, V.A. (2013) Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science*, **2**, pp. 1–9.

Cheshire, J. and Uberti, O. (2014) *London: The Information Capital*. London: Particular Books.

Croitoru, A., Crooks, A.T., Radzikowski, J. and Stefanidis, A. (2013) GeoSocial gauge: a system prototype for knowledge discovery from geosocial media. *International Journal of Geographical Information Science*, **27**, pp. 2483–2508.

Croitoru, A., Crooks, A., Radzikowski, J., Stefanidis, A., Vatsavai, R. and Wayant, N. (2014) Geoinformatics and social media: a new Big Data challenge, in Karimi, H.A. (ed.) *Big Data Techniques and Technologies in Geoinformatics*. Boca Raton, FL: CRC Press.

Croitoru, A., Wayant, N., Crooks, A.T., Radzikowski, J. and Stefanidis, A. (2015) Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, **53**, pp. 47–64.

Crooks, A.T., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D.A., Karagiorgou, S., Efentakis, A. and Lamprianidis, G. (2015) Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, **29**, pp. 720–741.

Fan, H., Zipf, A., Fu, Q. and Neis, P. (2014) Quality assessment for building footprints on Open StreetMap. *International Journal of Geographical Information Science*, **28**, pp. 700–719.

Feick, R. and Roche, S. (2013) Understanding the value of VGI, in Sui, D., Elwood, S. and Fan, G.M. (eds.) *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. New York: Springer.

Girres, J.-F. and Touya, G. (2010) Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, **14**, pp. 435–459.

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

Haggett, P. and Chorley, R. (1969) *Network Analysis in Geography*. London: Edward Arnold.

Haklay, M. (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B*, **37**, pp. 682–703.

Haklay, M. and Weber, P. (2008) OpenStreetMap:

user-generated street maps. *IEEE Pervasive Computing*, **7**, pp. 12–18.

Hillier, B. and Hanson, J. (1984) *The Social Logic of Space*. Cambridge: Cambridge University Press.

Hochmair, H. and Zielstra, D. (2013) Development and completeness of points of interest in free and proprietary data sets: a Florida case study, in Jekel, T., Car, A., Strobl, J. and Griesebner, G. (eds.) *Proceedings of GI Forum 2013 – Creating the GISociety*, Salzburg: Wichmann-Verlag.

Howard, E. (1902) *Garden Cities of To-Morrow.* London: Sonnenschein & Co., Ltd.

Howe, J. (2008) *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business.* New York: Crown Business.

Iaconesi, S. and Persico, O. (2015) The third infoscape: data, information and knowledge in the city: new paradigms for urban interaction. Available at: http://www.artisopensource.net/2015/06/29/the-third-infoscape-data-information-and-knowledge-in-the-city-new-paradigms-for-urban-interaction/.

Jackson, S.P., Mullen, W., Agouris, P., Crooks, A.T., Croitoru, A. and Stefanidis, A. (2013) Assessing completeness and spatial error of features in volunteered geographic information. *ISPRS International Journal of Geo-Information*, **2**, pp. 507–530.

Jenkins, A., Croitoru, A., Crooks, A.T. and Stefanidis, A. (2016) Crowdsourcing: a collective sense of place. *PLoS ONE*, **11**, pp. e0152932.

Jiang, B. and Jia, T. (2011) Zipf's Law for all the natural cities in the United States: a geospatial perspective. *International Journal of Geographical Information Science*, **25**, pp. 1269–1281.

Jiang, B., Duan, Y., Lu, F., Yang, T. and Zhao, J. (2014) Topological structure of urban street networks from the perspective of degree correlations. *Environment and Planning B*, **41**, pp. 813–828.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P. and Silvestre, B.S. (2011) Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, **54,** pp. 241–251.

Kitchin, R., Lauriault, T.P. and McArdle, G. (2015) Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science*, **2**, pp. 6–28.

Kleinhans, R., Van Ham, M. and Evans-Cowley, J. (2015) Using social media and mobile technologies to foster engagement and self-organization in participatory urban planning and neighbourhood governance. *Planning Practice & Research*, **30**, pp. 237–247.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009) Computational social science. *Science*, **323**, pp. 721–723.

Levy, A. (1999) Urban morphology and the problem of the modern urban fabric: some questions for research. *Urban Morphology*, **3**, pp. 79–85.

Longley, P.A. and Adnan, M. (2016) Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, **30**, pp. 369–389.

Louf, R. and Barthelemy, M. (2014) A typology of street patterns. *Journal of The Royal Society Interface*, **11**, 20140924.

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

Masucci, A.P., Smith, D., Crooks, A.T. and Batty, M. (2009) Random planar graphs and the London street network. *The European Physical Journal B*, **71**, pp. 259–271.

Mattern, S. (2015) History of the urban dashboard. *Places Journal*, Available at: https://placesjournal.org/article/mission-control-a-history-of-the-urban-dashboard/.

McCullough, M. (2005) *Digital Ground: Architecture, Pervasive Computing, and Environmental Knowing.* Cambridge, MA: MIT Press.

Miller, H.J. and Goodchild, M.F. (2015) Data-driven geography. *GeoJournal*, **80**, pp. 449–461.

Mitchell, W. (1995) *City of Bits: Space, Place, and the Infobahn.* Cambridge, MA: MIT Press.

Moudon, A.V. (1997) Urban morphology as an emerging interdisciplinary field. *Urban Morphology*, **1**, pp. 3–10.

Mullen, W., Jackson, S.P., Croitoru, A., Crooks, A.T., Stefanidis, A. and Agouris, P. (2015) Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features. *GeoJournal*, **80**, pp. 587–605.

O'Sullivan, D. (2000) Graph-based Cellular Automaton Models of Urban Spatial Processes. PhD Thesis, CASA, University College London.

Quercia, D., Schifanella, R., Aiello, L.M. and McLean, K. (2015) Smelly maps: the digital life of urban smellscapes, in *Ninth International AAAI Conference on Web and Social Media.* Oxford: AAAI Press.

Stefanidis, A., Crooks, A.T. and Radzikowski, J. (2013) Harvesting ambient geospatial information from social media feeds. *GeoJournal*, **78**, pp. 319–338.

Talen, E. (2003) Measuring urbanism: issues in smart growth research. *Journal of Urban Design*, **8**, pp. 195–215.

Ter Wal, A. and Boschma, R. (2009) Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science*, **43**, pp. 739–756.

Torrens, P. (2008) Wi-fi geographies. *Annals of the Association of American Geographers*, **98**, pp. 59–84.

United Nations (2014) *World Urbanization Prospects: The 2014 Revision*. New York: Department of Economic and Social Affairs.

Zhong, C., Manley, E., Arisona, S.M., Batty, M. and Schmitt, G. (2015) Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science*, **9**, pp. 125–130.

Zielstra, D. and Zipf, A. (2010) A comparative study of proprietary geodata and volunteered geographic information for Germany, in *13th AGILE International Conference on Geographic Information Science*. Guimarães, Portugal, pp. 1–15.

Zook, M.A. and Graham, M. (2007) Mapping digiPlace: geocoded internet data and the representation of place. *Environment and Planning B*, **34**, pp. 466–482.

Zook, M.A., Graham, M., Shelton, T. and Gorman, S. (2010) Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy*, **2**, pp. 7–33.