Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data

Junjun Yin1, 2, 3, Aiman Soliman1,2, 4, Dandong Yin1, 2, 3 and Shaowen Wang1, 2, 3, 4

1. CyberGIS Center for Advanced Digital and Spatial Studies

2. CyberInfrastructure and Geospatial Information Laboratory

3. Department of Geography and Geographic Information Science

4. National Center for Supercomputing Applications

University of Illinois at Urbana-Champaign, IL, 61801, USA

\* To whom correspondence should be addressed. Email: shaowen@illinois.edu

Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data

Existing urban boundaries are usually defined by government agencies for political and administrative purposes. However, it is not clear whether the boundaries truly reflect human interactions with urban space in intra- and inter-regional activities. Defining urban boundaries which consider socio-economic relationships and citizen commute patterns is important for many aspects of urban planning. In this study, we presented a method to redraw urban boundaries based upon human interactions with physical space. Specifically, we depicted the urban boundaries of Great Britain using a mobility network of Twitter user spatial interactions that was inferred from over 69 million geo-located tweets. We redrew the non-administrative anthropographic boundaries in a hierarchical fashion based on different physical movement ranges of users inferred from the collective mobility patterns of Twitter users in Great Britain. The results of strongly connected urban regions in the form of communities in the network space yield geographically cohesive, non-overlapping urban areas, which provide a clear delineation of the non-administrative anthropographic urban boundaries of Great Britain. The method was applied to both national (Great Britain) and municipal scales (the London metropolis). While our results corresponded well with the administrative boundaries, many unexpected and interesting boundaries were identified. More importantly, as the depicted urban boundaries exhibited a strong instance of spatial proximity, we further employed a gravity model to connect human mobility research to understand and justify the distance decay effects in shaping the delineated urban boundaries. This well-fitted gravity model explains how geographical distances found in the mobility patterns affect the interaction strength among different non-administrative anthropographic urban areas, which provides new insights into the interactions between human activity and urban space.

Keywords: mobility pattern, urban boundary, spatial interaction, spatial network, community structure

# Introduction

Official urban boundaries are defined by government agencies for political and administrative purposes. Urban environments are conceptualized as spaces that are recreated and formed by human activities (Schliephake 2014). A fundamental question when using the administrative, “top-down”, approach to defining urban boundaries is whether the outcome respects the spatial interactions of humans. These interactions can take the form of trade, commerce, social connections, and political activity across borders. Urban boundaries that respect the human interaction space are important to city planning, traffic management and resource allocation (Gao et al. 2014; Jiang and Miao 2015; X. Liu et al. 2015; Long et al. 2015). Many studies adopt a “bottom-up” approach to urban boundary delineation, where the geographic space is partitioned into small units and each unit is represented as a node within a network structure. A suitable community detection algorithm is applied to partition the network and associated geographic space based on the strength of human interaction among the nodes (Lancichinetti and Fortunato 2009). Different social and spatial human interactions were considered to establish the edges of the network connecting the nodes. For example, a large set of telephone call records were used to represent the network of human interaction across space to delineate urban boundaries in Great Britain (Ratti et al. 2010). Extending the previous method to different countries (Sobolevsky et al. 2013), the authors argue that this method yields cohesive geographic divisions that follow the socio-economic boundaries. While other researchers use social ties of Twitter users to identify cohesive regions for different countries across the world (Kallus et al. 2015), they found evidence for dividing the urban space due to local conflicts and cross-country unifying trends that further support the “bottom-up” approach to mapping non-administrative anthropographic boundaries.

A common finding from the mentioned studies is that the strongly connected urban regions in the form of communities in the network space yield geographically cohesive areas, in spite of different community detection methods and various forms of social and spatial human interactions were used. A general consensus is that those geographically cohesive areas are instances of the spatial proximity effects, where the interaction strength between two urban regions decreases as the geographical distance between them increases (Fotheringham 1981). Spatial proximity is closely related to Tobler's First Law of Geography: “everything is related to everything else, but near things are more related than distant things” (Miller 2004). While it is perhaps intuitively logical, few research efforts, which seek to quantitatively understand and explain how the spatial interactions shape the forms of connected geographical areas (i.e., urban boundaries), have been carried out. One of the major reasons is that geographical distance may affect the interaction strength, it is not an explicitly expressed constraint in the “virtual” human interactions, such as social ties or phone call initiation. In addition, there is a general lack of exploration regarding the linkages between the spatial proximity effects and the characteristics of the underlying spatial interactions.

In this study, we describe a novel approach to delineating non-administrative anthropographic urban boundaries from a mobility network of spatial interactions. Specifically, the spatial interactions refer to the actual movements of Twitter users (i.e., the reallocation across the geographical space), which were extracted from more than 69 million Twitter messages from June 1st to December 31st, 2014. Geo-located Twitter data is proven to be a useful source for studying human mobility patterns at large spatial scales (e.g. the national level) (Hawelka et al. 2014; Jurdak et al. 2015). In addition, Twitter data are not as sensitive to user privacy issues and do not exhibit spatial granularity that is limited to the postal code level (Thiemann et al. 2010). We argue here that by investigating Twitter user mobility patterns, we can provide a different view of non-administrative units based on physical commutes rather than social ties or phone call initiation. A unique advantage is that non-administrative anthropographic urban boundaries can be delineated in a hierarchical fashion based upon different ranges of physical movement, which are inferred from the collective mobility patterns of Twitter users in Great Britain.

We delineated the geography of urban boundaries in Great Britain by imposing a virtual fishnet over the islands of Great Britain. Twitter user movements were used to establish the connections between the fishnet cells to form a connectivity network, where each cell acts as a node within the network. We applied the map equation algorithm (De Domenico et al. 2015) to partition the network and associate geographic regions. We found that the collective mobility patterns of Twitter users in Great Britain are divided into several distance ranges ranging from short, intra- to inter-city movements with clear distinction points. The identification of connected regions at each of these distance ranges yielded hierarchical boundaries of urban spaces in Great Britain. As the depicted urban boundaries exhibited a strong instance of spatial proximity, we further employed a gravity model to connect human mobility research to understand and justify the distance decay effects in shaping the delineated urban boundaries. The well-fitted gravity model explains how geographical distances found in the mobility patterns affect the interaction strength among different non-administrative anthropographic urban areas. Our study provides a first-step in connecting human mobility research with the delineation of non-administrative anthropographic urban boundaries based on Twitter user spatial interaction. This provides a new understanding of the interactions between human activities and urban space.

# 2. Background and Related Work

Urban regions are discrete components in a greater set of regions, with or without physical boundaries separating them (Jiang and Miao 2015). For political and administrative purposes, government agencies define various sets of boundaries to partition the geographical space into spatial units at different scales, for instance: states, counties, census tracts, and electoral districts. However, the spatial extents of these units often overlap and agglomerate depending how citizens perceive, organize their image of a city, and interact with the urban environments (Lynch 1960). As connections are made between these units via various human activities crossing borders, such as social-economic relations and commute patterns of citizens, certain groups of units become more strongly connected than others. The boundaries of the agglomeration of these units are argued to reflect how people naturally interact with their geographical environment, which is important for city planning (Hollenstein and Purves 2010), urban growth evaluations (Jiang and Miao 2015; Long et al. 2015), and traffic management (Gao et al. 2014).

Empirical studies have attempted to delineate such boundaries with different methods and data sets. In general, the methods from existing literature can be summarized into two classes: spatial clustering and network based approaches. Spatial-clustering based approaches determine the boundaries based the intensity of geographic locations related to human activities, for instance: locations of social media check-ins (Cranshaw et al. 2012; Jiang and Miao 2015; Sun et al. 2016), place descriptions from crowd-sourced Web content (Vasardani, Winter, and Richter 2013), and geo-tagged Flickr data (Stefanidis et al. 2013; Hu et al. 2015). While notable boundaries of urban areas were identified and delineated, the dynamic connections between different spatial units were neglected in the spatial clustering based approaches, where the results are discrete and independent areas reflect a high intensity of human activities.

On the other hand, network based approaches delineate urban boundaries based on the intensity of human interactions between different spatial units, where each spatial unit is a node and the edge is modeled by human interactions between two nodes. Such human interactions can take physical or virtual forms, such as trade, commerce, social connections, and political activity across the borders. In terms of networks of virtual human spatial interactions, the connections between nodes are formed by virtual human relations, for example: social ties of Twitter users are used to identify cohesive regions for different countries across the world (Kallus et al. 2015), and a map of Great Britain is redrawn based the communication network of phone call initiations (Ratti et al. 2010). In contrast, physical human spatial interactions form the connections between nodes by the collections of individuals physically allocating from one node to another, which is referred to as a mobility network of spatial interactions in this study. Many existing studies have attempted to extract the mobility network from various data sources, such as census migration data (Rae 2009), GPS recorded vehicle (Rinzivillo et al. 2012) and taxi trip records (X. Liu et al. 2015), mobile phone call data (Sobolevsky et al. 2013; Zhong et al. 2014), social media check-ins (Y. Liu et al. 2014), and geo-located Twitter data (Hawelka et al. 2014; Gao et al. 2014). These networks of human spatial interactions are then further explored to reveal clusters regarding the intensities of the interaction strength, for example, by applying visual analytics methods (Rae 2009) or community detection methods (Coscia et al. 2011).

Researchers argue that those geographically cohesive areas taking the form of communities in the network space, are related to the distance decay effect, which implies that the interaction strength between two urban regions decreases as the geographical distance between them increases. However, few research efforts explored the linkages between the spatial proximity effects and the characteristics of the underlying spatial interactions, which is critical for explaining how the spatial interactions affect the shapes of connected geographical areas (i.e., urban boundaries). While geographical distance is not explicitly expressed constraint in the “virtual” human interactions, we argue that seeking answers from the mobility network of spatial interactions with the characteristics of underlying mobility patterns can help to explain how distance decay effects affect the interaction strength and the shape of depicted urban boundaries.

## 2.1. Large-scale mobility network from geo-located Twitter data

To construct a large-scale mobility network of human spatial interactions, the capability of capturing human movements with fine-grained spatial and temporal granularity is critical. The low-resolution mobility data collected from census records (Rae 2009) is estimated and aggregated at census tract level and does not necessarily reflect movements of the same individuals. In terms of collecting detailed mobility data of individuals, using GPS trackers tends to produce the most accurate records of individuals’ movements, which means a high degree of recording accuracy of user locations and update frequency (Zheng et al. 2008). However, the data is often limited in spatial scale (e.g. within a specific city or region) with a small group of people, for example, 182 and 226 volunteers participated in collecting such mobility data in (Zheng et al. 2008) and (Rhee et al. 2011) respectively. Other than tracking people directly, the vehicle-based GPS data is often tied to a specific vehicle (e.g. taxi), which may only be accessible to a limited group of people (Kung et al. 2014). Another popular mobility data source found in academic literature is the mobile phone call data in the form of Call Detail Records (CDR), where the locations of mobile users are estimated by cell tower triangulation with accuracy in the order of kilometers (González et al. 2008; Kung et al. 2014; Zhong et al. 2014). Such a dataset can cover relatively large spatial scale (e.g., national level) (Sobolevsky et al. 2013) and a large portion of the population in the study region (Kung et al. 2014). However, due to the concerns of infringement on individual privacy, the mobile phone call data is not publicly accessible, which is not ideal for the replicability in scientific findings or comparisons across different regions.

On the other hand, it becomes increasingly popular for researchers to exploit the publicly accessible mobility data captured from the Location Based Social Media (LBSM) platforms (e.g., Foursquare and Twitter). This is also evident from the related studies mentioned above. However, there are some limitations and complexities in directly extracting and using the mobility data from the LBSM data sets. For example, comparing to GPS traces, the update frequency of an individual’s location varies depending on when a user is posting a new geo-located tweet or check-in at a new place. LBSM data have also been criticized for lacking of representativeness of the population, as not all people use social media or send geo-located messages (Kung et al. 2014). Another research challenge is to identify the users, as a social media account is not equivalent to a real person in the physical world. Many studies have started to look into the demographic aspect of LBSM data, in particular Twitter data (Steiger et al. 2015; Luo et al. 2016). While the used methods vary, these studies suggest that the mentioned issues require us to pose stricter criteria in filtering and extracting individual movements.

In this study, geo-located Twitter data are chosen as the source for constructing large-scale mobility networks of human spatial interactions and studying detailed mobility patterns. A geo-located tweet is a Twitter message with an additional geo- tag expressed as a pair of geographical coordinates that represent the location from which the tweet was sent. Twitter is one of the most popular platforms and is been actively used in many countries. It provides a publicly accessible streaming API (http://dev.twitter.com/streaming) for easy data access. The geo-located Twitter data present some unique advantages regarding the purpose of this study, for example, the high-resolution location information enables to identify multiple travel modes in user mobility patterns (Jurdak et al. 2015); the large spatial coverage enables to study global mobility patterns (Hawelka et al. 2014), which is almost impossible for other mobility datasets. Also, it provides the opportunities for reproducing this study in other countries.

# 3. Materials and Methods

## 3.1. Geo-located Twitter Data and Data Processing

For this study, the geo-located tweets were collected using the Twitter Streaming API by supplying a geographical bounding box to retrieve all the geo-located tweets within an area of interest. To ensure complete coverage of Great Britain, we set the bounding box to the British Isles using the lower left and upper right coordinates (49.49, -14.85), (61.18, 2.63) respectively. We implemented a data crawler to continuously collect 7-months of data (June 1st – December 31st, 2014) resulting in over 101.8 million tweets with a total data volume of 60 GB. During the data collection phase, the data crawler did not encounter any issue regarding whether it exceeds the data quota by the 1% policy mentioned in (Hawelka et al. 2014). It means we have managed to download all the geo-located tweets for the given bounding box. To showcase the overall spatial coverage of the collected geo-located tweets, the geo-locations of all the collected tweets are shown in Fig. 1. The collected point visualization reveals the geography of cities. Notice the clusters with higher densities of tweets correspond to the locations of major cities.

**Figure 1. The spatial coverage of collected geo-located Tweets in Great Britain (left) and London (right).** Each point corresponds to an individual geo-located tweet collected for this study. Note that Twitter activities are most apparent in urban areas.

The original location information embedded in the geo-tag is given in units of latitude and longitude. We examined the “geo” attribute in each raw tweet and kept the one with location information derived from GPS receiver (with high resolution) rather than from geocoding process. We projected the points into the British National Grid (EPSG: 27700) coordinate system to reduce the complexity of the required distance calculations. The geographical boundary of Great Britain is derived from Office for National Statistics (ONS) of UK (http://www.ons.gov.-uk/ons) to further restrict the remaining tweets to be “domestic”. Based on these restrictions, the filtered dataset contains 69,847,497 tweets made by 1,153,891 Twitter users. To remove tweets from non-human users, the raw tweets were further filtered using the following steps. First we removed the duplicated messages. We then removed non-human users based on unusual relocation by examining all of the consecutive locations of each user and excluded those with relocating speeds in excess of the threshold of 240 m/s as used by (Jurdak et al. 2015). Finally, to reflect the spatial interactions of residents rather than tourists, we impose a condition that the time interval between a user’s first and last recorded tweets should be more than 30 days. In other words, a user that is identified to have stayed in the study region more than 30 days is considered as a resident. The filtered dataset for the following study contains 60,209,778 tweets made by 824,712 Twitter users.

At this stage, each geo-located tweet is represented as a tuple , where is an anonymous Twitter user ID; is the recorded location of the tweet as a coordinate pair; is the timestamp of the tweet’s post; and is the actual content of the tweet. We constructed a trajectory for each Twitter user by appending all the recorded locations (with the same ) in the chronological order (based on the timestamps). To protect Twitter user’s privacy, the ID field was replaced with a randomly generated unique number and the content of the message was removed. The actual location of each geo-located tweet is only used for distance calculation and determining the geographic unit it falls in. Our simplified geo-located tweet dataset can be shared with other researchers upon request.

## 3.2. Mobility network of Twitter user spatial interactions

A Twitter user’s movement is defined as the individual’s geographic relocation or displacement (González et al. 2008). This is not equivalent to a “trip” taken by an individual, because, displacement includes situations where the time interval between two consecutive recorded locations is one month. To identify the clusters of urban regional connectedness, Twitter user movements are used to establish a connectivity network, where two urban regions connect when a Twitter user’s movement begins in one and ends in another. These connections can be represented by an origin-destination (OD) matrix based on the collective Twitter user displacements within the dataset. This OD matrix is essentially a mathematical representation of a weighted directed graph , where is a set of spatial nodes corresponding to the underlying urban regions, is a set of edges representing the connections between a pair of nodes, and the weights are assigned by the accumulated volume of Twitter user movements.

To build the spatial network at a national level, we had to determine the basic units to serve as spatial nodes of the connectivity network. Previous studies have suggested equidistant spatial tessellation to generate nodes, which uses Voronoi polygons to partition the space based on the collected points (Rinzivillo et al. 2012; Zhong et al. 2014). This approach demonstrates improvements in estimating the locations of mobile phone records from cell tower triangulation (González et al. 2008; Qian et al. 2013). However, equidistant tessellation decreases the spatial resolution of aggregated geo-located tweets, because the location information is usually derived from the embedded GPS within mobile devices and tends to provide greater accuracy (Zandbergen 2009). Another approach is to partition the space into a grid of spatial pixels (Ratti et al. 2010; X. Liu et al. 2015). However, the cell size can potentially lead to biases due to the Modifiable Area Unit Problem (MAUP) (Openshaw 1984; Wong 2009), where different choices of unit size can lead to significant variant findings. To compare our investigation with the findings from similar studies, and avoid subjectively deciding the cell size, we performed statistical analysis of Twitter user mobility patterns in Great Britain and measured the distribution of collective Twitter user displacements and the radius of gyrations of individuals (González et al. 2008; Jurdak et al. 2015). The radius of gyration is a metric to distinguish mobility patterns of individuals (González et al. 2008), which is defined as Eq. (1):

xx

This measures the accumulated distances of deviation from the center of mass of an individual user’s trajectory, where is one of the user’s locations and is the center of mass of the user’s trajectory. By examining the probability distributions of the radius of gyration, also known as the spatial dispersal kernel (Brockmann et al. 2006), we chose 10 km as the cell size at the national level of Great Britain (Fig. 3 - c, with details shown in the next section). More importantly, as 10 km is the distinct geographic distance for separating two main groups of Twitter users in terms of the spatial coverage in Great Britain, a 10-km size cell serves as a mask to partitioning the space. In this way, we focus on the inter-connections among different urban regions with less attention to movements around a user’s neighborhood (i.e., within 10 km radius), such as home or work places. Thus, we created a fishnet with 2784 10-km size cells. The cells of the fishnet act as proxies to represent individuals’ spatial coverage areas to focus more on the inter-connectivity among cells and identify strongly connected cell clusters.

## 3.3. Community structure of the network of spatial interactions

Based on the derived mobility network of spatial interactions, which is a directed weighted graph, we further determined clusters of strongly connected spatial nodes, known as communities in the graph space. There are a variety of community detection algorithms that produce different results depending the definition of community within the network (Coscia et al. 2011). A common community detection method is based on modularity maximization (Newman 2006), seen in previous studies (Ratti et al. 2010; Song et al. 2012; Hawelka et al. 2014). However, such an approach is often problematic: it is found to have an inherent resolution problem, where small communities are either ignored (Fortunato and Barthélemy 2007) or assigned with high modularity scores (Guimerà et al. 2004); and it is found to produce less informative partitions in many empirical networks (Good et al. 2010). Since our graph is a directed weighted graph, an alternative community detection method from the literature is Infomap (Rosvall and Bergstrom 2008; De Domenico et al. 2015), which is considered to produce better community detection results (Lancichinetti and Fortunato 2009).

Infomap uses the map equation (Rosvall et al. 2010) to represent the probability of flow of random walks within information systems (Rosvall and Bergstrom 2008). It identifies communities by minimizing the expected description length of the trajectory of a random walker, which is shown below:

ccc

In Eq. (2), consists of two terms: is the entropy of the movement among clusters and is the entropy of movement within clusters. Specifically, is the probability that a random walker jumps from one cluster to another, while pi is the probability of the in-cluster movement of cluster . This algorithm can be intuitively tailored to describe strongly connected clusters of urban regions based on Twitter user movement. The detailed literatures and implementations of Infomap can be found on this website (http://mapequation.org). Note that Infomap can perform multi-level community detection (De Domenico et al. 2015), but we only use it to produce the most detailed community structures to examine groups of strongly connected urban regions.

## 3.4. Distance decay effect and gravity model

As mentioned above, the clusters of urban regions in the form of communities in the network space often yield geographically cohesive urban areas. This phenomenon is speculated to be related to the distance decay effects, where the interaction strength between two urban regions decreases as the geographical distance between them increases. A gravity model is often used to express such relations, as is shown in Eq. (3), where and denote the interaction from to and distance between two places, is a constant, and and are the population size of place and respectively. The interaction strength decreasing with respect to increasing geographic distance is expressed by the distance decay function, , where the parameter reveals the distance impact on interaction strength. A greater indicates stronger decay and the interaction strength is more influenced by distance. While it is suggested that population size may not be an accurate indicator to describe the repulsion or attractiveness between places, the gravity model is usually fitted by using observed interaction strength and the distance between geographical entities (Y. Liu et al. 2014).

xx

In this study, the main purpose for adopting the gravity model is not to find the best value to estimate the potential interaction strength among depicted urban areas. Interestingly, the distance decay effects are also found in human mobility patterns (Zhao et al. 2016), the authors argue that it is due to the constraints of complex urban structure. In line with the idea that urban environments are conceptualized spaces that are recreated and formed by human activities (Schliephake 2014), we speculate that the distance decay effects in affecting the interaction strength of two geographic regions and ultimately depicting the urban structures (e.g., urban boundaries), is contributed by (or related to) the distance decay parameters found in the underlying mobility patterns. Considering we have used a mobility network of spatial interactions, if this hypothesis stands, it will provide strong support that the depicted urban boundaries are not random artifacts but indeed reflect how naturally people move across the geographical regions.

# 4. Results

## 4.1. Collective Mobility Patterns of Twitter Users in Great Britain

We modeled different aspects of the collective mobility patterns of Twitter users. These patterns include: the number of visited locations per user, the collective user displacements, and the radius of gyration of individuals to identify distinct distance ranges in user travel patterns. We utilized these distance ranges within the mobility patterns to partition the geographic space of Great Britain into fine-grained cells and established the connectivity among these cells to redraw non-administrative anthropographic urban boundaries.

We found that the cumulative distribution function of the number of locations visited by each Twitter user follows a two-tier power law distribution (Fig. 2). Majority of the data follow a truncated power-law distribution , where ; and the tail part (less than 2% of the whole population) follows a power-law distribution with α value is 3.2. The distribution was found to be consistent over each month examined (June to December, 2014), which has a slight offset in the truncated power-law distribution (the mean α value is 1.26 ± with a and the mean value as 0.00134 ± with a 0.0002 ).

The two-tier power law distribution indicates that the collective behavior of Twitter users visiting different locations can be well approximated with a (truncated) Lévy Walk (a random walk) model (Rhee et al. 2011; Reynolds 2012), which has also been identified in many human mobility studies using different mobility data (Zhao et al. 2016). The similarity among the distributions suggests that the mobility data collected from geo-located tweets is temporally stable, at least at monthly intervals, which indicates that our approach using Twitter user mobility to delineate urban boundaries is viable. In addition, the Lévy Walk model reveals the diversity regarding the number of visited locations per user, which indicates a level of “randomness” in Twitter user movement across space. It, in turn, justifies our choice of using the map equation community detection algorithm (Rosvall and Bergstrom 2008) to identify the clusters of urban regional connectedness using large-scale Twitter user movement data.

We then studied two aspects of the Twitter user mobility patterns: the distribution of Twitter user displacement and the radius of gyration. Twitter user displacement refers to the distance between two consecutive locations in a user’s trajectory using a straight-line distance metric. The radius of gyration describes the deviation of distance from the center of mass in a user’s trajectory. The probability distributions of the collective user displacement and radius of gyration are presented in Fig. 3, where the fitting method for identifying different distance ranges is derived from (Jurdak et al. 2015). The probability distribution of the collective displacements can be approximated by from [10 m, 70 m] (accounting for 3% of the population), from [100 m, 70 km] (93% of the population), and [> 70 km] (4% of the population). The displacement distance between 70 m and 100 km can be further approximated by two power law distributions with a cut-off point at 4 km (55% distances are less than 4 km and 40% distances between 4 km and 100 km), which indicates the urban movement captured by the geo-located Twitter data to reveal two different modes: inter-city and intra-city movement. In short, these fitting functions suggest the existence of multi-scale or multi- modal urban movements captured from Twitter users in Great Britain, which means the geographically cohesive, non-overlapping urban areas identified in the next section are not just a result of short distance movement but emerge naturally from the broader Twitter user mobility pattern. Note that a similar multiphase pattern was observed in Twitter user displacements in Australia, but with slightly different distance ranges (Jurdak et al. 2015).

Further, we analyzed the distribution of radius of gyration to understand the collective movements of individual Twitter users rather than separate displacements. The of Twitter users in Great Britain can be approximated through a combination of three functions: from [10 m, 30 m], from [50 m, 10 km], and from [10 km, 100 km], where these three functions account for 92% of all the users. This suggests that there are three primary types of users that: (1) tend to stay at one location or at nearby locations when they tweet, or (2) tend to move at the intra-city scale when they tweet, or (3) tend to exhibit a large spatial coverage. (1) and (2) account for approximately 53% of all users. Note that the accuracy of these values for defining the distance bound depends upon the accuracy of the location information of each geo-located tweet. These findings are consistent with the ones in the literature on human mobility, where the radius of gyration of human movement is bounded to different distance ranges (Brockmann et al. 2006; González et al. 2008). Interestingly, the over the greater London region can be fitted by similar functions. However, as it reflects intra-city level mobility patterns, there is no distinct distance range to indicate large spatial coverage. The distance decay effects found in both user displacements and the radius of gyration shows evidence of spatial proximity in Twitter user movement. It explains that the communities of urban regions within the graph space are geographically close but can be separated from other groups, which results in the delineation of urban boundaries based on the collective spatial interactions of Twitter users.

## 4.2. Redrawing Great Britain’s Urban Boundaries

The mobility network of Twitter user spatial interaction was constructed by nodes representing 10 km by 10 km fishnet cells, where 10 km is the distinct geographic distance for separating two main groups of Twitter users in terms of the spatial coverage (i.e., radius of gyration) in Great Britain (see Fig. 3 - c). The cells of the fishnet act as proxies to represent individuals’ spatial coverage areas to focus more on the inter-connectivity among cells and identify strongly connected cell clusters. It provides an adequate resolution for a country wide investigation (Ratti et al. 2010). The edges of this network were de- rived from the number of directed Twitter user displacements between each pair of cells. We used this connectivity network as a proxy to partition the space associated with its nodes. Coherent geographic regions were identified as individual fishnet cells showing more internal user movements compared to user movements across the cell boundaries to neighboring cells. To help readers who are not familiar with the geographic context in Great Britain better interpret the derived boundaries, two additional layers (i.e., locations of airport fields and population-weighted-centroids of workplace zones in the Great Britain) are added in the background of the figures. Fig. 4 presents the delineated urban boundaries based on Twitter user displacement distance less than 4 km, greater than 4 km, greater than 10 km, and using all available displacements together compared to the administrative boundaries of Great Britain. One clear observation in both the coarse and fine delineations is that most of the geographic divisions are centered around big urban cores with relatively high populations. These results are expected given that most of the tweets originate in urban centers. However, what is remarkable is the performance of this approach in dividing the remaining space between cities. We found that restricting the trip distance results in different delineations of the catchment area around these centers. For example, one could explain these effects as a manifestation of the underlying gravity law (Simini et al. 2012) and the distance decay effect on attracting movers (González et al. 2008). Remarkably, our approach performed well in terms of dividing the entire space with minimum gaps. Empty cells were found in regions where no, or few, Twitter users had visited (e.g. forests, agriculture) especially when restricting the analysis to short distance Twitter user displacements.

Regional boundaries inferred from short distance Twitter user displacements (less than 4 km) exhibit very small and fragmented regions, which is probably related to daily commuting around a user’s home location. Redrawing the boundaries based on longer distance displacements produces more cohesive, large regions. For example, by partitioning the space based on displacements greater than 10 km created regions that are comparable to the NUTS (Nomenclature of Territorial Units for Statistics - 1) regions (Fig. 5 - a). However, the power of this novel mapping technique is not to reproduce the partitions already known, rather it is to point out some of the unexpected boundaries. For example, the boundaries between England and Wales were found to be more diffuse compared to the abrupt boundary of England and Scotland. Moreover, the city of London has a wider visitor catchment area that extends beyond the authoritative boundaries of the city. Increasing the displacement distance results in revealing the large region connected to London (Fig. 5).

Using Twitter user mobility to delineate non-administrative anthropographic boundaries enables the researcher to redraw the city at different mobility ranges inferred objectively from the user’s collective distribution. In addition, the distance range of the movements is usually explained by local socio-economic factors (e.g., work commuting) that provide for a specific interpretation of the apparent patterns. The patterns obtained from Twitter user mobility are comparable to the patterns produced by those of the network of landline phone calls (Ratti et al. 2010). For example, the region of Wales appears to consist of three communities as found in the connectivity of both phone calls and long distance movements. However, the regions extracted from the mobility network seems to be more spatially consistent with minimal spatial gaps compared to the partitions extracted from land-line call networks in Great Britain (Ratti et al. 2010).

A more detailed study was conducted over the greater London region revealing the intra-city spatial interaction patterns. Since the captured Twitter user movements were on intra-city level (in comparison to the national level in Fig. 3-c), there was no distinct distance range to separate Twitter user spatial coverage in terms of radius of gyration (see Fig. 3-d). We chose 1-km cell size as referred from the related literature (X. Liu et al. 2015), which suggested 1-km cell size is detailed enough to depict urban structures.

The spatial partitions derived from a fine grid of 1 km used all available Twitter user trips without any restriction on trip distances yields geographic boundaries comparable to some of London’s boroughs (Fig. 6). However, some areas are shown to be more cohesive and display greater spatial interactions across the administrative boundaries, for instance, central London. Although, these results suggest that travelers seem to be localized over certain areas of the city most of the time, some regions do exhibit long distance interaction patterns. For example, the separate geographic areas in the south of Hillingdon which includes Heathrow Airport exhibits more connectivity to central London than its surrounding areas, which is explained by the usual flight passenger routes. The technique also reveals some of the emerging communities around the borders due to the spatial intermingling of both communities. For example, East Barnet and West Enfield seem to have higher interactions than those resulted from in the emerging cohesive zone between the two boroughs.

## 4.3. Explaining the distance decay effect with a gravity model

The results of strongly connected urban regions in the form of communities, which are derived from the mobility network of Twitter user spatial interactions, yield geographically cohesive, non-overlapping urban areas. While it provides a clear delineation of the non-administrative anthropographic urban boundaries of Great Britain, the reasons on why they are geographically cohesive and non-overlapping or why the boundaries stop/emerge at certain spatial extent that leads to different size of the urban areas are not clear. As the depicted urban boundaries exhibit a strong instance of spatial proximity, a gravity model is employed (Eq. (3)) to explain how distance decay effects found in the mobility patterns affect the interaction strength between the derived non-administrative anthropographic urban areas.

In this model, the distance between two derived urban areas is measured by the geodetic distance between the centroid locations of the two. As it is mentioned above, and are the observed interaction between urban area and , which are measured by the aggregation of movement flows in each urban area. In particular, we set the distance decay parameter value as 0.8: (1) As we hypothesize that the distance decay parameters found in the underlying mobility patterns potentially contribute to β in the gravity model (2) and we chose the 10 km cell-size based on the collective Twitter user mobility pattern regarding radius of gyration, where the distance decay parameter is 0.8 when radius of gyration (Fig. 3(c)). We found that the gravity model indicates strong linear correlation between the observed versus the estimated interaction strength with and . This confirms that the depicted urban areas are instances of spatial proximity effects, where the strength of human (in this case, Twitter user) spatial interaction between two urban regions decreases as the geographic distance between them increases. The well fitted gravity model provides support that the depicted urban areas are not just random artifacts but reflect how naturally move across geographic regions. More importantly, since we have used a mobility network to delineate the boundaries, the distance decay effects are well related and explained by the distance decay parameters found in the underlying mobility patterns. To elaborate, the spatial interaction strength decreases along with the decay for the probability of longer distance Twitter user movements, and eventually stops at certain spatial extent, which leads to more geographically cohesive cluster of urban regions.

# 5. Discussion

It is worth noting that constructing a mobility network of spatial interaction using geo- located tweets has some potential concerns to limit the results from this study. First, the geo-located Twitter data is not able to generalize to the entire population. As the demographic information of the Twitter users cannot be easily identified, the results of delineated urban boundaries may not reflect a complete real-world image from human movements. The demographics of Twitter users in this study can be an under/over- representation of the all overall population in Great Britain. Related studies have been carried out to examine the demographic information in geo-located Twitter data (Steiger et al. 2015; Luo et al. 2016; Huang and Wong 2016), which should be carefully considered in future studies. Second, regarding the concern of spatial sparseness of geo-located Twitter data (Steiger et al. 2015), the urban regions that do not have any, or limited, Twitter coverage can be missed during the delineation process. To investigate whether this limited the ability to capture the connections made through Twitter user movements between urban regions, we visualized the flows of Twitter user movements using the method mentioned in (Rae 2009) (see Supplement Materials section 1). The outcome showed that the Twitter user movements in this study connected most urban areas in Great Britain and clearly exhibited long and short distance movements, where was essential for investigating the connection strength between urban regions. Third, since the collective radius of gyration was used to determine the cell-size in the network, we examined the temporal stability for the measurement in the geo-located Twitter data. The probability distributions of the radius of gyration for Twitter users in Great Britain are verified to be consistent across different monthly time span (see Supplement Materials section 2), which indicates the stability of using such measurements in this study. Finally, as the geo-located Twitter data offers easy data accessibility, our methods can be reproduced for other countries. Along with the large sample size and spatial coverage, geo-located Twitter data demonstrated its applicability for this study.

A major component in this study was imposing a virtual fishnet to partition the geographic space over Great Britain. Alternatively, we had used the ward divisions as spatial units for aggregating Twitter user movements, which is the finest administrative boundaries of Great Britain (see Supplement Materials section 3). Without looking into the details of the results, the derived communities in the network space are similar to the ones from using fishnet approach. The strongly connected communities also yield geographically cohesive, non-overlapping urban areas. However, as the ward division is still defined by administrative purpose, the polygonal units tend to be geographically continuous. It causes problems to aggregate regions that do not have Twitter coverage into certain clusters. Aggregating Twitter user movements at the ward level also imposes more apparent concerns of the mismatch of the overall population, where less populated areas were overly represented and connected into large areas. These are the reasons why we chose a virtual fishnet approach for this study.

We should be aware that using different fishnet cell-size to partition the space will produce different mobility networks, and can potentially lead to different delineations of the urban boundaries. As it is discussed above, we chose 10 km as the cell-size of the fishnet at national level in this study. We also carried out an experiment by arbitrarily setting the cell-size to 5 km at the national level (see Supplement Materials section 4). Note that the cell-size could be set to any value, such as 4.9 km or 5.1 km. The fishnet with smaller cell-size (i.e., 5 km) produced more and smaller strongly-connected communities within the network space. It indicates that the spatial resolution of the fishnet cells does affect the outcome from the community detection method employed in this study, where fishnet with smaller cell-size leads to more discrete and locally connected (i.e., smaller) clusters of urban areas. Such an effect can be explained by the probability distributions of the radius of gyrations of individual Twitter users. The probability of distance that deviates from a user’s center location decays with a stretched-exponential function from [50 m, 10 km], which means the movements from Twitter users with smaller spatial coverage dominate the delineation of the connected urban areas. To avoid arbitrarily deciding the cell-size, we studied the probability distributions of the radius of gyrations of individual Twitter users and selected 10-km as the cell-size, which is the distinct geographic distance for separating two main groups of Twitter users in terms of their spatial coverage at the national level. In addition, this choice enabled us to focus on the inter-connections among different urban regions with less attention to movements around a user’s neighborhood (i.e., within 10 km radius), such as home or work places.

To apply the map equation algorithm, the mobility network should be organized as a weighted and directed graph, on which we confirmed an undirected graph cannot lead to a meaningful result (see Supplement Materials section 4). Since we used different physical movement ranges of users in delineating the urban boundaries, the description length appeared to get larger when longer displacements were used, which corresponded to less and larger geographically cohesive, non-overlapping urban areas. We believed that different movement ranges of users changed the weights of this graph and affected the interaction strength between two fishnet cells. As we have illustrated that the geographically cohesive, non-overlapping urban areas were instances of spatial proximity effects, where the employed gravity model suggested the interaction strength between two fishnet cells decreases as the geographical distance increases, the longer displacements were used, the larger were produced from the map equation algorithm.

# 6. Conclusion

In this study, we described a method that connected human mobility research with the delineation of the non-administrative anthropographic urban boundaries in Great Britain using a mobility network of Twitter user spatial interactions. In contrast to administrative urban boundaries, our “bottom-up” approach constructed a virtual fishnet over the islands of Great Britain to partition the space. By studying the probability distributions of the radius of gyrations of individual Twitters users, we selected a cell-size of 10 km at the national level and 1 km for the greater London region. Twitter user movements were used to establish a connectivity network of the fishnet cells. We applied the map equation algorithm to partition the network and associated geographic regions. The strongly connected communities within the network space yielded geographically cohesive, non- overlapping urban areas that provided a clear delineation of the urban boundaries in Great Britain. By performing a statistical analysis of Twitter user mobility patterns in Great Britain, in particular the distribution of collective Twitter user displacements, we found multi-scale and multi-modal urban movements that were divided into several distance ranges starting from short intra-city to inter-city movements with clear destination points. Identifying the connected regions at each of these distance ranges revealed hierarchical boundaries of the urban space in Great Britain.

The power of using Twitter user mobility to delineate non-administrative anthropographic boundaries enables urban representation at different mobility ranges inferred objectively from the collective mobility patterns. Urban boundaries redrawn based on Twitter user movement represent physical commutes rather than social ties or phone call initiation to reflect human interactions with the geographical space. More importantly, as the depicted urban boundaries exhibited a strong instance of spatial proximity, we further employed a gravity model to connect human mobility research to understand and justify the distance decay effects in shaping the delineated urban boundaries. This well fitted gravity model explains how geographical distances found in the mobility patterns affect the interaction strength among different non-administrative anthropographic urban areas, gaining new insights into the interactions between human activities and urban space. A final remark from this study for future work is that geographical distance plays an important role in affecting human mobility patterns and the strength of human spatial interactions across space. Our approach can be considered to assist in understanding human spatial interactions from the mobility perspective, which is applicable based on detailed geo-located Twitter data in many countries, as well as future mobility datasets with detailed location information of individuals and large spatial coverage.

### Acknowledgements

xx

### References