Edge-based communities for identification of functional regions in a taxi flow network

U. Demšar[1], J. Reades[2], E. Manley[3], M. Batty[3]

[1]Centre for GeoInformatics, University of St Andrews, UK
Email: urska.demsar@st-andrews.ac.uk

[2]Geography Department, King's College London, UK
Email: jonathan.reades@kcl.ac.uk

[3]Centre for Advanced Spatial Analysis, University College London, UK
Email: {ed.manley, m.batty}@ucl.ac.uk

1. Introduction

Recent technological advances in spatial data collection have caused an explosion of new data volumes and their availability. One of these data types are flow networks, sometimes also called origin-destination (OD) networks which are now being increasingly captured using various forms of sensor technology from bespoke system which track vehicles and passengers to smart phone locations can that can be associated with individual travellers. These networks consist of vertices representing locations where flows start and end. Edges of the network bear information on the flow size and direction, thus forming a directed weighted network on spatial vertices. Examples of flow networks are transportation networks (e.g. flows of passengers between subway stations), migration/commuting networks and mobile phone communication networks.

In the geographic tradition which can be traced back forty years at least, one of the uses for flow networks has been in context of regionalisation: flow information was used to derive regions of functional interaction between origin and destination locations. Studies used a number of different flow data for this purpose: transportation flows (Black 1973), phone calls (Clark 1973, Goddard 1973), and taxi journeys (Goddard 1970). However, these studies were limited due to the limits on computer power and data available and perhaps due to this, network-based regionalisation seems to have been temporarily forgotten. Only recently has this topic received renewed attention: for example, regionalisation studies use commuting networks (Farmer and Fotheringham 2011, Landré and Håkansson 2013) and mobile phone communication networks (Expert et al. 2011, Thomas et al. 2012). This renewal is largely based on an interdisciplinary transfer of methods from network science research in physics, in particular various community detection methods that have been used to partition and summarise clusters that comprise such networks (Newman 2006).

In network science, a community is defined as a set of vertices in the network which are more densely interconnected with each other than with the rest of the network (Newman 2004). In geographic regionalisation, this corresponds to the concept of functional regions, which are spatially contiguous, internally well connected and relatively cohesive in terms of flows (Farmer and Fotheringham 2011). A number of algorithms have been developed for community detection in physics, where most approaches partition the vertex set to obtain communities. This means that each vertex can belong to at most one community and it is not possible for communities to share vertices with each other, which may be problematic in real-world networks (Palla et al. 2005).

In spatial flow networks the non-intersection criterion is in contradiction with the idea of polycentricity in movement, which is the concept that there exist several central locations which generate and receive large numbers of flows across a wide area and where trips from

each of the centres are not exclusively delimited from trips from all other centres. The polycentric movement process has been observed both at the level of mega cities as well as smaller areas (Hall and Pain 2006; Zhong et al., 2014). Indeed, the observation of the necessity of overlapping regions in movement-based regionalisation can again be traced back forty years to work by Goddard (1970) and his work on partitioning the centre of London into travel regions based on taxi flows. However, none of the recent network-based regionalisation approaches takes this overlapping necessity into consideration.

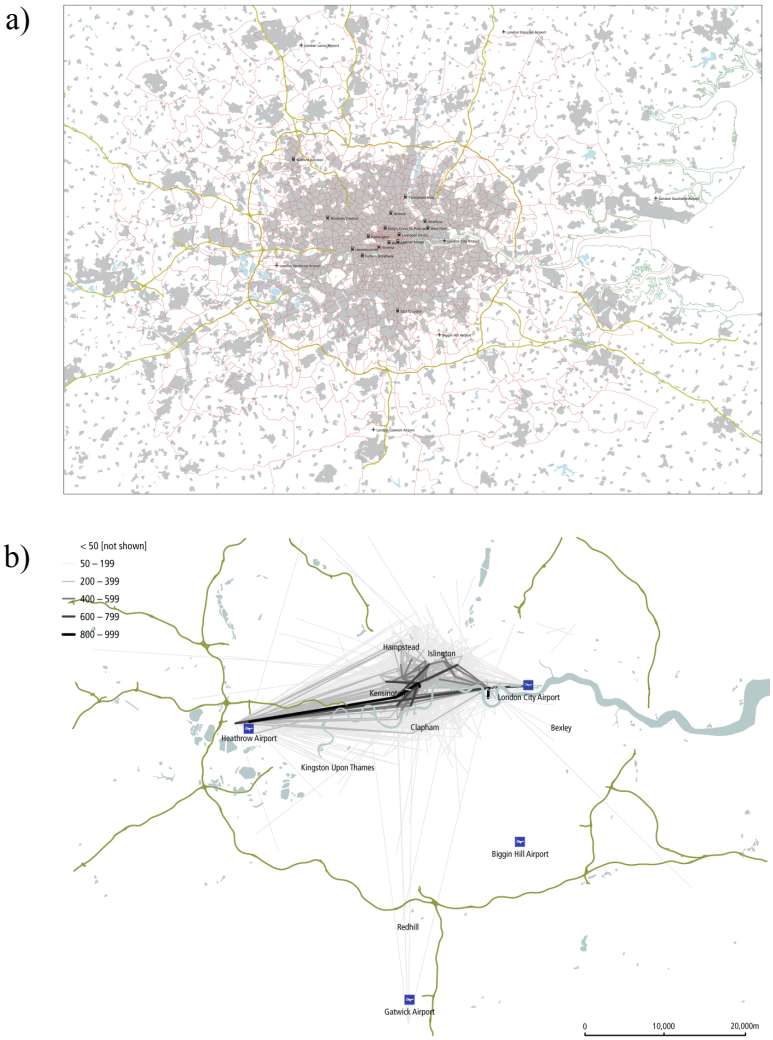

Figure 1: a) Traffic Area Zones (TAZes) in our study area. b) Taxi flows among TAZes.

In this paper we take inspiration from Goddard (1970) and investigate the possibility of using an edge-based community detection algorithm (Ahn et al. 2010) for identification of overlapping functional regions defining taxi flows in Greater London Area. This is work in progress: we demonstrate some initial results and discuss further directions for edge-based community detection in the context of spatial flow networks.

2. Data: taxi flows

For this study we were given access to three months of work day taxi flow data by Addison Lee minicabs (Dec 2010 – Feb 2011). Data consisted of GPS trajectories of taxis, and we aggregated origins and destinations of each trajectory into a set of Traffic Analysis Zones

(TAZ) to obtain a flow network. Figure 1 shows the TAZes covering the Greater London Area and the taxi flows.

We performed our analysis at two spatial scales: for Central and Inner London and for the Greater London Area. Table 1 presents the sizes of the two flow networks at these two spatial scales.

Table 1. Network sizes.

	Number of vertices (TAZes)	Number of edges (non-zero flows between TAZes)
Central and Inner Area	391	50,786
Greater London Area	1,165	104,587

3. Edge-based community detection

A typical community detection algorithm operating on vertices (Girvan & Newman 2002, Newman 2006) starts by calculating the similarity between each pair of vertices. Vertices are then aggregated using hierarchical clustering. This procedure starts with each vertex as representing one cluster. Vertices/clusters are then joined iteratively so that at each step the two clusters are joined that contribute the least to the increase in overall dissimilarity. This builds a dendrogram representing the temporal sequence as to how vertices/clusters are joined. A partition of the vertex set is obtained by cutting this dendrogram at some level. The best level is defined through optimisation of a modularity function, which reaches the maximum value when intra-cluster similarity is maximised and inter-cluster similarity is minimised. The resulting optimal partition splits the set of vertices into non-overlapping groups (communities), i.e. each vertex can only be a member of one of the groups.

Edge-based community detection (Ahn et al. 2010) operates in the same way as vertex-based community detection with one difference: it is the set of edges that is being partitioned rather than the set of vertices. Partitioning edges rather than vertices makes sense in cases where the network under consideration is a social network, since each vertex (a person) can belong to several not necessarily overlapping communities (social groups), e.g. colleagues, friends, family, etc. (Palla et al. 2005). This also makes sense for our taxi flow data, since it is reasonable to expect that each vertex (each TAZ) could feed flows of taxi traffic into several other TAZes, which do not necessarily feed flows among each other.

We start with a directed network of taxi flows between TAZs. Note that the Ahn et al. (2010) algorithm is suitable for undirected networks, whereas the flow networks are directed and weighted. We consider possible adjustments to directed networks as future work, while here we transform our flow network into an undirected one. This can be done in several ways, but we use the simplest and the most frequently used approach which is to sum the bidirectional flows and then discard direction (Leicht and Newman, 2008), which produces an undirected weighted network.

In the next step we calculate the similarity of each pair of connected edges (edges that share a common vertex) using a function that compares the topological structure of the neighbourhoods of both edges (i.e. edges that share a common vertex with the two original edges), the Tanimoto coefficient. This coefficient also uses edge weights (i.e. flow sizes) in the similarity calculation (see Ahn et al. 2010 for details)..

Once pair-wise similarity is calculated for all pairs of connected edges, we produce a dendrogram based on edge-similarity. Further we calculate the modularity function - partition density. Density of one community is a topological measure, defined by the number of links in the community, normalised with the maximum possible number of edges between the nodes in the community. Partition density is the average community density over all communities in one particular partition. This value has one global maximum between the top and the bottom of the dendrogram (Ahn et al. 2010) – this is at level k where the inter-

community density is maximal and intra-community density is minimal, thus defining the best possible partition of the edges into edge-communities.

In the final step we cut the dendrogram at level k, to obtain the best possible partitioning of edges into edge-communities.

a) b)

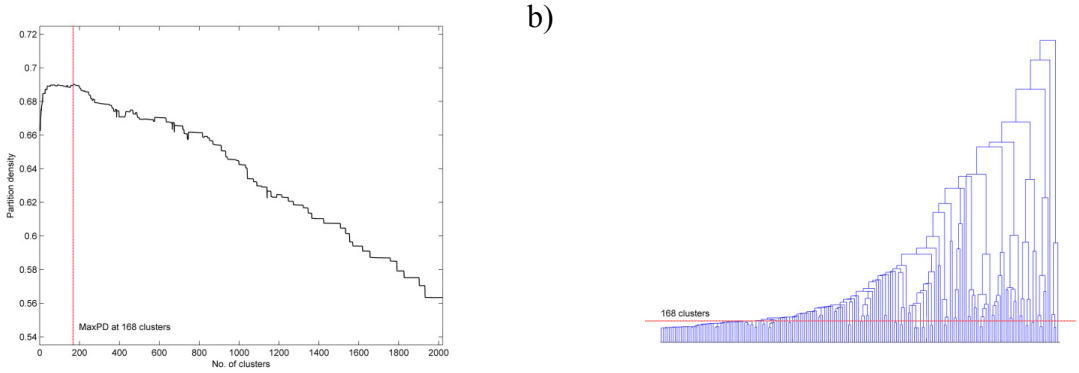

Figure 2: a) Partition density and b) dendrogram for the Central and Inner Area, cut at the optimal partition density level.

4. Results

The edge-based community detection algorithm splits our two taxi flow networks as follows: the optimal number of clusters is 168 for the Central and Inner Area, and 6,276 clusters for entire Greater London Area. Such large numbers of clusters in the optimal partition pose a particular challenge for interpretation of results, which we comment on in the discussion.

For illustration, figure 2 shows the maximisation of partition density and the resulting dendrogram cut for the 168 clusters in the Central and Inner Area. Further sorting the 168 clusters according to their size (number of edges included in each cluster), and taking into consideration only clusters containing more than 10 edges leaves 25 clusters as potential functional regions in taxi traffic in Central and Inner Area. These edge clusters and our tentative interpretation of the type of traffic they represent are shown in figure 3.

Addison Lee minicabs have a strong business bias: in contrast with London's more familiar black cabs they cannot be hailed on the street and need to be pre-booked. Further, the company prioritises customers with accounts, the majority of which are large businesses with extensive mobility requirements. As a consequence, we expect to see most of the flows to be for either business travel purposes or for trips to events. The resulting edge-communities (figure 3) appear at first exploration to be consistent with the expected business bias in the taxi traffic, but this bears further detailed investigation of results..

One of the surprises in our edge-based regionalisations is that the first cluster is vastly larger than all other clusters combined and that flows in this cluster do not seem to have any particular spatial distribution (note that clustering is purely data-driven and done on flow information only and therefore location does not play any role in cluster assignment). We speculate that this cluster may represents the `usual state of affairs' or, to put it in another way, the everyday traffic that encapsulates a dominant configuration that is both fairly uniform and geographically extended, but its size masks interesting secondary functional groupings in taxi flows, which we can identify from the rest of the communities.

The results however bear further investigation and perhaps algorithm validation on a synthetically generated flow network, where we would be able to control for patterns that we would expect to be able to identify with our method.

We further encounter a "big data" problem when attempting to partition the whole Greater London Area. This partition is optimal when 104,587 edges are split into 6,278

clusters (fig. 4) and we are contemplating possibilities as to how to visualise let alone interpret such a large partition.

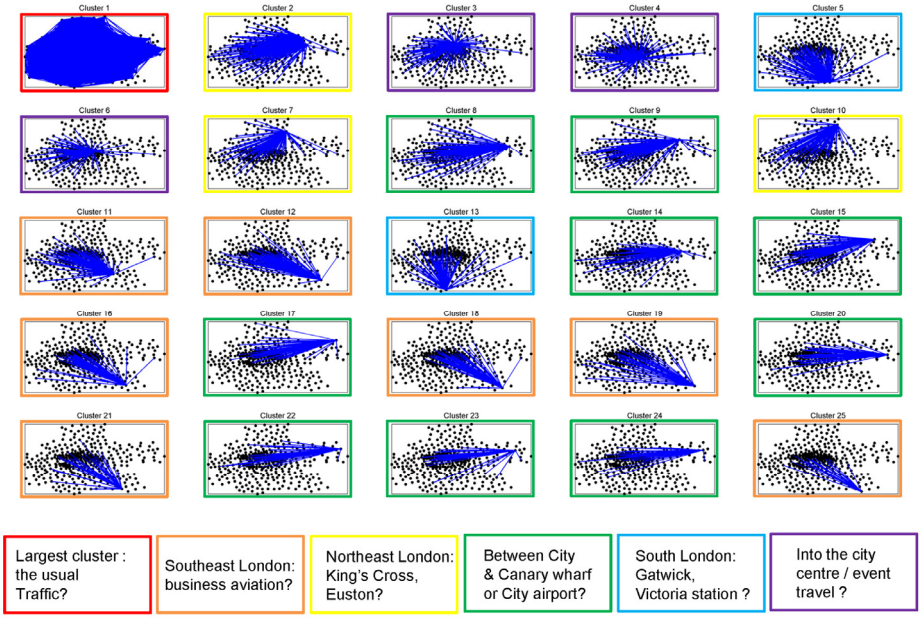

Figure 3: Tentative categorisation of the largest edge-based taxi flow clusters for Central and Inner Area. Interpretations shown are only possibilities and need to be further investigated.

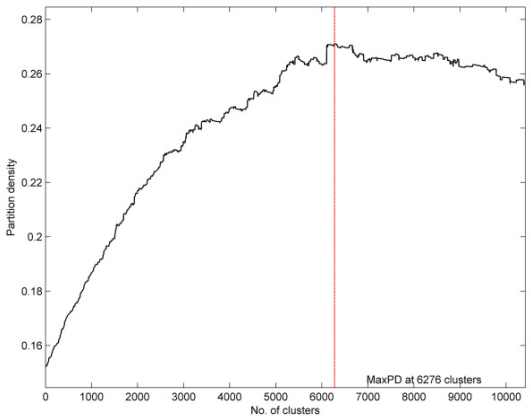

Figure 4: Partition density max for the Greater London Area. is reached at 6278 clusters.

5. Conclusions and outlook

This paper presents an attempt at edge-based regionalisation of taxi traffic in London and is a work in progress. As mentioned above, one of the major problems we encounter is that while we demonstrate that the algorithms taken from complex networks research can be potentially applied on real-life spatial flow networks with a tentatively plausible geographical results, the size of results may be prohibitory towards proper understanding and interpretation of results (e.g. 6278 clusters in taxi traffic in Greater London Area). The task of interpreting thousands of clusters resulting from the mathematically-derived optimal partition is demanding and further consideration will be needed to address it. The problem is not limited to our case: most of flow data sets are very large and can produce a large number of optimised regions in the best partition. For example, UK migration flow data are collected between 223060 census

output areas, which generates a network with more than 49 million edges to be classified into regions. GIScience can not solve problems of this type on its own – we believe that interdisciplinary knowledge exchange of spatial sciences with other disciplines that deal with network-style big data (such as physics and computer science) would be necessary to approach these problems.

Further, in this attempt we implemented an algorithm originally developed for non-spatial undirected networks. How to include direction in this process is a topic for future research. In flow networks, space also matters and this needs to be recognised in the regionalisation procedure. Space can be taken into account in different ways, either by incorporating spatial autocorrelation into community-detection (Cerina et al. 2012) or by using a geo-aware modularity function (Hanningan et al. 2013). The second point is particularly relevant, since currently the best partition obtained through optimisation of partition density is not linked to any spatial properties of the flows, but only to topology of the network. There is an implicit inclusion of the sizes of the flows in the procedure through Tanimoto coefficient, which uses flow sizes as weights in the calculation of edge similarity, but locations of edges/vertices and other spatial properties are currently not considered. We plan to investigate these possibilities further and explore how these or similar novel space- and/or direction-aware methods can be used for improved regionalisation from flow networks.

Acknowledgements

The authors would like to acknowledge support from Addison Lee and Transport for London (TfL) who contributed data for this study.

References

Ahn YY, Bagrow JP and Lehmann S, 2010, Link communities reveal multi-scale complexity in networks. *Nature*, 466:761-765.

Black WR, 1973, Toward a Factorial Ecology of Flows. *Economic Geography*, 49(1):59-67.

Cerina F, De Leo V, Barthelemy M and Chessa A, 2012, Spatial Correlations in Attribute Communities. *PLoS One*, 7(5): e37507.

Clark D, 1973, Normality, transformation and the principal components solution. *Area*, 5:110-113.

Expert P, Evans TS, Blondel VD and Lambiotte R, 2011, Uncovering space-independent communities in spatial networks. *PNAS*, 108(19):7663-7668.

Farmer CJQ and Fotheringham AS, 2011, Network-based functional regions. *Environment and Planning A*, 43(11):2723-2741.

Girvan M and Newman MEJ, 2002, Community structure in social and biological networks. *PNAS*, 99(12):7821-7826.

Goddard JB, 1970, Functional Regions within the City Centre: A Study by Factor Analysis of Taxi Flows in Central London. *Transactions of the Institute of British Geographers*, 49:161-182.

Goddard JB, 1973, Office linkages and location: A study of communications and spatial patterns in Central London. *Progress in Planning*, 1:109-232.

Hall P and Pain K, 2006, *The polycentric metropolis: learning from mega-city regions in Europe*. Earthscan.

Hannigan J, Hernandez G, Medina RM, Roos P and Shakarian P, 2013, Mining for Spatially-Near Communities in Geo-Located Social Networks. *Social Networks and Social Contagion, AAAI Technical Report* FS-13-05.

Landré M and Håkansson J, 2013, Rule versus Interaction Function: Evaluating Regional Aggregations of Commuting Flows in Sweden. *European journal of transport and infrastructure research* 13(1):1-19.

Leicht EA and Newman MEJ, 2008, Community structure in directed networks. *Physical Review Letters*, 100:118703-1-118703-4.

Newman MEJ, 2006, Modularity and community structure in networks. PNAS, 103(23):8577-8582.

Palla G, Derenyi I, Farkas I and Viscek T, 2005, Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814-818.

Thomas I, Cotteels C, Jones J and Peeters D, 2012, Revisiting the extension of the Brussels urban agglomeration: new methods, new data… new results? *Belgeo* 1-2:1-12.

Zhong C, Arisona SM, Huang X, Batty M and Schmitt G, 2014. Detecting the dynamics of urban structure through spatial network analysis, *International Journal of Geographic Information Science*, forthcoming.