The effect of geographical distance on community detection in flow networks

Sebastijan Sekulić
School of Geography and Sustainable Development,
University of St Andrews
Irvine Building, North Street
KY16 9AL, St Andrews, Fife
ss372@st-andrews.ac.uk

Jed Long
School of Geography and Sustainable Development,
University of St Andrews
jed.long@st-andrews.ac.uk

Urška Demšar
School of Geography and Sustainable Development,
University of St Andrews
urska.demsar@st-andrews.ac.uk

Abstract

A flow network is a directed graph in which each edge has an origin and a destination, and the edge represents movement from one point to another. With constantly increasing number of devices and methods for collecting location information, flow networks are becoming larger, denser and more complex. Many physical models have been developed for exploring the structure of flow networks, however in most cases interpretation considers only the number of flows and ignores relevant geographical context. In this paper we demonstrate how geographic distance can be used to improve the process of a standard physical model for flow network structure: community detection. We incorporate distance between two areas in the computation of the Louvain community detection algorithm. The results are compared to the result of the original method using a flow data set on commuting flows in Scotland. A preliminary evaluation of results shows that including geographical distance enables us to obtain a different insight from the network and provides more detailed information about local patterns of commuting in Scotland, which is lost in a non-spatial version of community detection.

Keywords: community detection, distance decay, flow networks, movement analytics

1 Introduction

Flows are commonly represented as directed mathematical graphs. That is, each location where the flow can come from or lead to is represented as a node. A connection between every two nodes is then weighted with the flow size. Flow networks are used to investigate a number of social science phenomena, for example flows inferred from mobile phone data (Expert et al. 2011), migration flows (Peter J. Taylor, 1975), commuting flows (Sila-Nowicka et al. 2016), or taxi flows between pick-up and set-down points (Demšar et al. 2018). In geography, the most common approach for analysing flow data is to fit spatial interaction models to describe the movement across space that results from a decision-making process (Kordi and Fotheringham 2016). In this paper, we take an alternative approach to studying spatial flows using a *community detection* algorithm, where communities are defined as subsets in a network in which connections between nodes are dense, but connections between two such subsets are sparse. (Girvan & Newman 2002)

Physical models have been used in community detection for flows of mobile phone, web or social network data (Girvan & Newman 2002). However, physical models typically do not consider the value of the geographical information contained inside nodes of the network (Expert P et al. 2011). They mainly use the number of connections between two nodes, completely ignoring geographical linkages between them, consequently giving the same value to the connections between two geographical neighbours and connections regardless of their geographical context. When dealing with geographical flow data, using the location of the nodes and the distance between them has a potential to lead to more insight into the phenomenon of human movement. For example, Kempinska et al. (2018) create the flow network from the street network, where street intersections are used as nodes and streets as links to detect communities from GPS tracks of the police cars.

To demonstrate the importance of distance in the analysis of geographical flows, we have taken an existing community detection algorithm (Louvain algorithm; Blondel et al. 2011) and modified it in a way that explicitly considers the distance between two nodes in the calculation of communities. We applied the new algorithm to the Scotland commuting data from UK Census (ONS 2011). We further compared the two methods, the original community detection and our spatialised version, by calculating their modularity scores, the number of generated communities and through a visual comparison.

As this is work in progress, we only present some preliminary results, which however do already show that using distance in community detection influences the result.

The rest of the paper is structured as follows: in section 2 we explain how the chosen community detection algorithm works and how we have integrated the distance into the algorithm, in section 3 we present some preliminary results, and in section 4 discuss the results and list some of our conclusions.

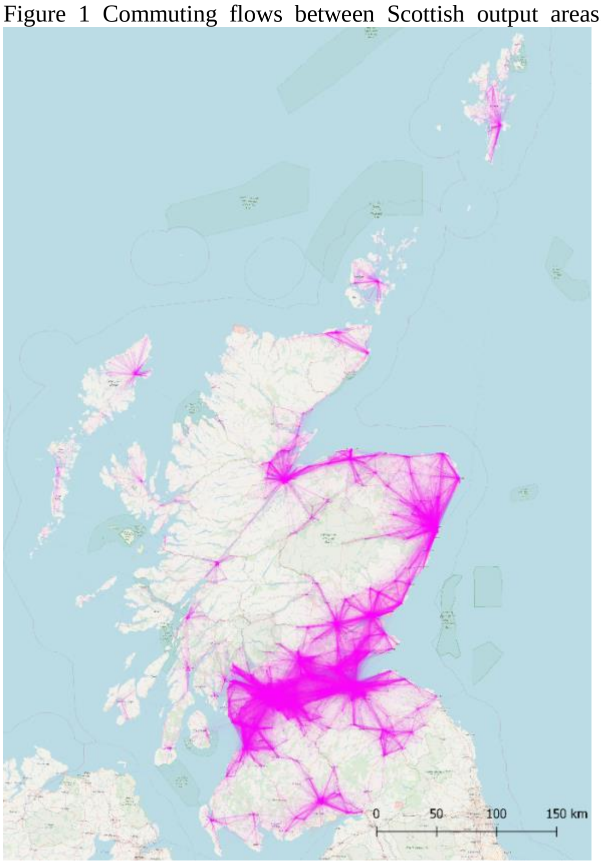
2 Materials and methods

To test the effect of geographical distance on community detection, we have constructed an undirected flow network using commuting data from UK Census 2011 (ONS 2011, Figure 1). The lowest level of the geographical level at which census data are provided is called the output area (OA). OAs were designed in a way that they have similar population size and to be as socially homogenous as possible (ONS 2018). In our network, every node represents one output area in Scotland and links between two output areas represent the number of commuter flows in-between two areas.

Table 1 Statistical information about dataset

Number of output areas	46 351
Number of flows	1 966 097
Number of flows (without intra-zonal trips)	1 944 545
Maximum length	647 km
Average length	13.4 km
Minimum number of flows from/to single OA	1 / 1
Maximum number of flows from/to single OA	100 / 4950

Figure 1 Commuting flows between Scottish output areas.



2.1 Defining the network

The network was constructed from commuting flows (work-home) using the total number of flows between each two output areas. That is, we constructed the undirected flow network by summing flows from place A to place B and flows from B to A into a single value which was then assigned to the edge between A and B. The reasoning behind this is, that in this experiment we do not evaluate orientation and direction of commuting, but rather how strong the connection between two output areas is (de Montis et al. 2013). Also, flows that are pointing to themselves (the so-called intra-zonal trips, i.e. trips of people who live and work in the same output area) are not considered in this analysis (Bhatta, 2011).

Next, we calculated the adjacency matrix of the flow network. Adjacency matrix for the undirected network is a symmetric matrix $\boldsymbol{A}$, whose element $\boldsymbol{A}_{ij}$ represents the weight of the connection between node i and node j. To incorporate distance in the adjacency matrix, we are using the so-called distance decay effect (Taylor, 1975), which says that spatial interaction is closely connected to distance, i.e., people are more likely to travel shorter distances to work to minimize transport costs.

We have incorporated the effect of the distance into the weights matrix $W_{ij} = nf(d)$ where n is number a of total flows between node i and node j, and $f(d)$ is a function of distance. In the continuation of this paper, we will be exploring how different choices for the distance function produce different results.

2.2 Community detection algorithm

To test the different effects of the distance, we have chosen to use an existing algorithm, the Louvain algorithm (Blondel et al. 2008), and included the distance effect in link weights. This community detection algorithm allows us to include link weights, find the community hierarchy structure and calculation speeds. Moreover, we can get insight into the quality of generated communities by comparing modularity values of the partitions.

The Louvain algorithm has two phases. In the first phase, it creates communities through modularity optimisation. Modularity is a measure of the quality of a particular division of a network. It ranges from -1 to 1 and represents how dense are the links inside communities in comparison to links between communities (Girvan & Newman 2003). For the weighted network as ours, modularity is calculated as follows (Newman 2004):

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta\left(c_i, c_j\right). \qquad (1)$$

Here $A_{ij}$is the weight of the edge between the nodes i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to node i, c_i is the community to which node i is assigned, the δ-function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and $m = \frac{1}{2}\sum_{ij} A_{ij}$.

We begin with a same number of communities as there are nodes in the network, then for each node, we check if moving it from its current community to another one gives a gain in the modularity. If the modularity gain is positive, the node is assigned to the community which has the highest increase in modularity. Otherwise, the node stays in its original community. This process is then repeated for each node until there is no more modularity gain (a single node can be checked and moved into a different community several times).

In the second phase, a new network is created by using communities from the first phase as nodes, where link weights between nodes are the sum of all weight between nodes of two connecting communities. Completion of these two phases creates one level of the hierarchy. Then the process is repeated by using the newly created communities as input data in next iteration, thus creating higher and higher level of community groupings. The algorithm usually finishes after two to four passes.

The efficiency of the algorithm is based on the ability to calculate the gain in the modularity ΔQ very fast. ΔQ for moving an isolated node i into community C is calculated as:

$$\Delta Q = \left[\frac{\Sigma_{in} + k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (2),$$

Where $\sum_{in}$ is sum of weights of the links inside C, $\sum_{tot}$ is the sum of the weights of the links incident to nodes in C, k_i is the sum of the links incident to node i, $k_{i,in}$ is the sum of the weights of the links from i to nodes in C and m is the sum of the weights of all the links in the network. (Blondel et al. 2008).

Note that this algorithm is general and has no consideration for geographic distance or geography (e.g. location) in any way.

2.3 Introducing the geography

To incorporate geographic distance into the community detection, we are proposing two ways of modelling the distance as part of the link weights: as an inverse power function (3) and as a negative exponential function (4). These both decrease in value with increasing distance but do so at different speeds. The two functions are:

$$W_{ij} = n * d^{-(k)} \quad (3),$$

$$W_{ij} = n * e^{-(dk)} \quad (4),$$

where n is the number of flows between nodes i and j, d is the geographic distance between the two nodes and k is a parameter that defines the importance of the distance measure and is subject to optimisation. We have chosen two approaches to setting the parameter; as a value dependant on how large we want our communities to be, and as an unbiased constant.

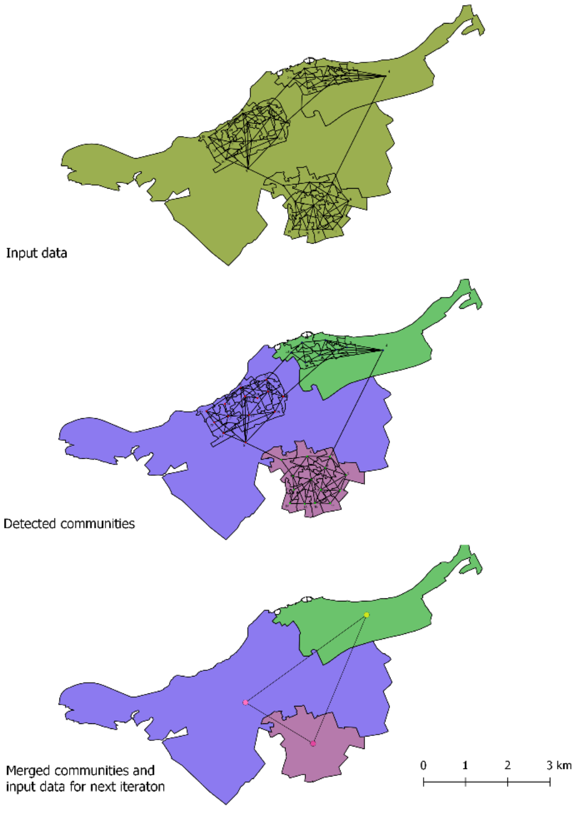
Choosing the value for parameter k is a computational problem of itself. We take a heuristics approach to choosing k, while in the future we will consider experiments to optimise this parameter properly, as the goal is to find a balance between the distance and the number of flows. To avoid punishing distance too hard, we have chosen for k to be a ratio of d_{ij} and average length of a flow d_{AVG} or maximum length of flow d_{MAX}. Changing the denominator allows us to present interaction at different scales. In addition to that, we will use fixed values of k, where $k \in [0,1,2]$. We are using value of $k = 0$ to get insight in how community detection works if the distance is not used (i.e. this is the original Louvain algorithm). As mentioned above, this is not the optimal way of choosing the value for k, as it depends on the type of network and planned usage of the results, but in this experiment, it will let us explore the general structure of the network and how including the distance changes the communities. Values for ***k*** are shown in Table 2.

Table 2 Distance function parameters

k for inverse power function (eq. 3)	**k for inverse exponential function (eq. 4)**
0	-
1	1
2	-
$\frac{d_{ij}}{d_{AVG}}$	$\frac{1}{d_{AVG}}$
$\frac{d_{ij}}{d_{MAX}}$	$\frac{1}{d_{MAX}}$

One of the advantages of the Louvain algorithm is its ability to identify the community structure on multiple hierarchy levels. As nodes in our network represent the lowest geographical level at which census data are provided, each pass of the algorithm creates a new geographical level with nodes from the previous level aggregated into new nodes based on the network structure, as explained above.

Figure 2 Visualisation of the Louvain model hierarchy structure found in each step of the algorithm.



3 Results

In this section, we show how the number of communities and modularity changes depending on used distance function and its parameters. Results are shown only for the top hierarchy level.

By using distance decay to the model importance of flows, we have a higher number of detected communities than by just using the number of flows, as visible in Figure 3. Using inverse power function has shown better and more controllable results than by using the negative exponential function. Figure 3 shows us the partition with optimal modularity for some of the functions. Partition where distance is not used (Figure 3A), shows a small number of large communities, while all the other partitions (Figures 3B-F) show increase in the number of communities and improvement in resolution. Especially interesting is increase in the number of the communities in rural areas (Scottish Highlands). Modularity of partition and number of generated communities for each of the functions are shown in Table 3 and Table 4.

As ***k*=*0*** represents results where distance is not included in community detection, we can see an increase in both the numbers of communities and modularity score when the distance is used. Values for power function show lager span in the number of communities and their size, while using exponential function does not show that difference in number between using average or maximum distance. Using k>2 would put too much importance on distance and ignore number of flows.

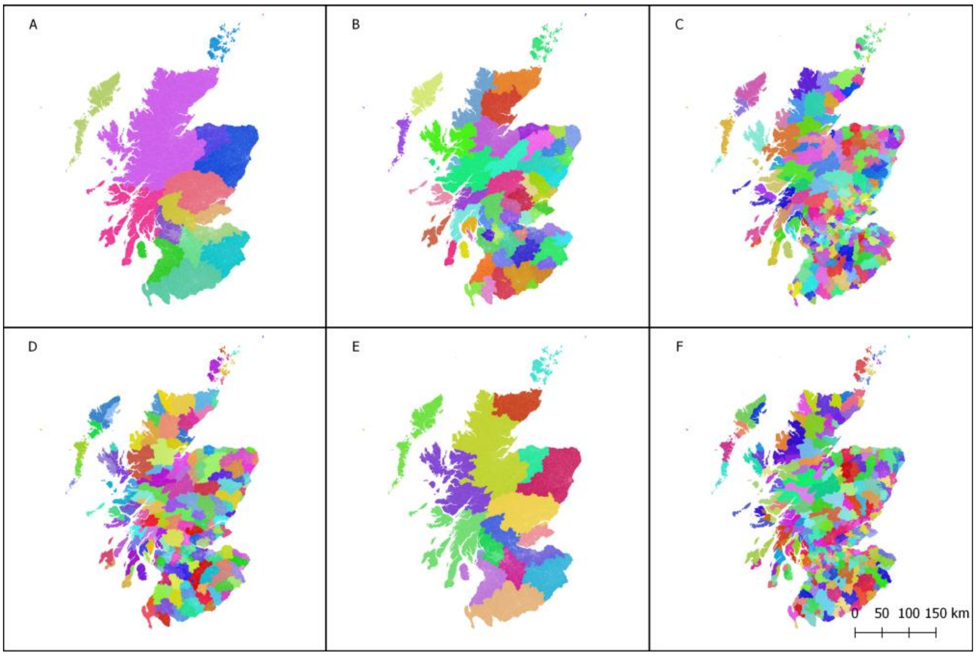
Table 3 Modularity scores and number of communities for inverse power function

$\boldsymbol{k}$ values for (3)	Optimal number of communities	Optimal modularity score
0	17	0.732
1	68	0.906
2	**287**	**0.975**
$\boldsymbol{\frac{d_{ij}}{d_{AVG}}}$	**183**	**0.945**
$\frac{d_{ij}}{d_{MAX}}$	20	0.770

Table 4 Modularity scores and number of communities for inverse exponential function

$\boldsymbol{k}$ values for (4)	Optimal number of communities	Optimal modularity score
1	**271**	**0.959**
$\frac{1}{d_{AVG}}$	33	0.838
$\frac{1}{d_{MAX}}$	18	0.737

Figure 3 Side by side comparison of six different community partitions: A – no geography, B – power function with **k = 1**, C - power function with **k = 2**, D – power function with $\mathbf{k} = \frac{d_{ij}}{d_{AVG}}$, E – power function with $\mathbf{k} = \frac{d_{ij}}{d_{MAX}}$, F – exponential function with **k = 1**



4 Conclusions and Discussion

In this paper, we incorporate geographic distance in a popular community detection algorithm to show how this can provide more detail on local patterns in mobility flows.

Our method enables us to extract information from migration flows that were previously not available. By changing the ***k*** parameter, we can manipulate the size of the communities and how far they reach, which would allow us to investigate how far the effect of distance decay on commuting reaches. This could be useful in urban planning, for example, to delineate new regional divisions or in market-based research (Farmer and Fotheringham 2011). Best results were achieved by using inverse power function (3) with ***k*** as a ratio between the distance of the flow and some proprietary distance (average distance in our case). While some results show higher modularity scores, we believe that having a balanced ratio between number of flows and distance is more important.

As the method is based on the Louvain algorithm, it inherits its scalability and speed (processing of the graph with more than 1M of flows takes seconds) but preparing the data can be a lengthy process (calculating the length of the flows in the network took just over 4 hours). Additionally, the right balance between the number of flows and the distance still needs to be explored and optimized. Another improvement would be using alternative measures of distance (e.g., the road network distance) instead of the Euclidian distance to take natural boundaries into account.

By using predefined parameters of community size and reach, we can get a new insight into flow structure and human movement. With the right data, we could also track how communities are changing over time and correlate those with outside factors. For example, using subsets of commuting data, we could track how commuting regions change depending on the weather, how the opening (or closure) of a new road changes movement patterns or study broader scale migration patterns.

5 Acknowledgements

This work is supported by the Economic and Social Science Research Council (ESRC) through an Industry Aligned Advanced Quantitative Methods / Data Science scholarship. Commuting data are licensed under the terms of the Open Government Licence available at [http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3].

6 References

Bhatta, B. &. L. O. I., (2011). Are intrazonal trips ignorable?. *Transport,* 18(1), pp. 13-22.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E., (2008). Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, P10008, DOI:10.1088/1742-5468/2008/10/P10008

de Montis, A., Caschili, S. & Chessa, A.,(2013). Commuter networks and community detection: A method for planning sub regional areas. *European Physical Journal: Special Topics*, 215(1), pp. 75-91.

Demšar U et al., (2018). Revisiting the past: Replicating fifty year old flow analysis using contemporary taxi flow data. *Annals of the Association of American Geographers,* Advanced Online Publication. https://dx.doi.org/10.1080/24694452.2017.1374164

Expert P et al., (2011). Uncovering space-independent communities in spatial networks. *PNAS,* 108(19), p. 7663-7668.

Farmer, C. J. & Stewart Fotheringham, A., (2011). Network-based functional regions. *Environment and Planning A,* 43(11), pp. 2723-2741.

Girvan, M. & Newman, M. E. J., (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), pp. 7821-7826.

Kordi M & Fotheringham AS, (2016), Spatially Weighted Interaction Models, *Annals of the American Association of Geographers*, 106(5): 990-1012.

Kira Kempinska, Paul Longley & John Shawe-Taylor (2017): Interactional regions in cities: making sense of flows across networked systems, *International Journal of Geographical Information Science*, DOI: 10.1080/13658816.2017.1418878

Newman, M. E. J., (2004). Analysis of weighted networks. *Phys. Rev. E,* Volume 056131, pp. 1-9.

Newman, M. & Girvan, M., (2003). Finding and evaluation community structure in networks.

Office for National Statistics, (2011) Census: Special Migration Statistics (United Kingdom) [computer file]. UK Data Service Census Support. Downloaded from: https://wicid.ukdataservice.ac.uk

Office for National Statistics, (2011), accessed 10 January 2018 https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

Peter J. Taylor, S. O., (1975). Distance decay in spatial interactions. *Concepts and Techniques in Modern Geography*

Siła-Nowicka K et al., (2016), Analysis of Human Mobility from Volunteered Movement Data and Contextual Information, *International Journal of Geographical Information Science*, 30(5): 881-906.