Geographical context in community detection: a comparison of a node-based and a link-based approach

Sebastijan Sekulić
Bell Edwards Geographic Data Institute
School of Geography and Sustainable Development,
University of St Andrews,
Scotland, UK
ss372@st-andrews.ac.uk

Jed Long
Department of Geography,
Western University,
London, ON, Canada
jed.long@uwo.ca

Urška Demšar
Bell Edwards Geographic Data Institute
School of Geography and Sustainable Development,
University of St Andrews,
Scotland, UK
urska.demsar@st-andrews.ac.uk

Abstract

With the ability to easily record spatial information, spatial flow networks are increasingly common. There are many existing methods for analysing spatial flows, but they typically do not consider the spatial component of the flow network. To analyse flow networks, methods which originate from computer science and physics, such as community detection (CD), have become very popular. In this paper we modify CD methods to make them geographically sensitive. We explore two different approaches to CD, a node-based one and a link-based one and add a spatial element to each. Our hypothesis is that by using link-based CD we will be able to find different overlapping communities in the same area and that classifying links will enable us to classify movement instead of location. We take the Louvain and Hierarchical Link Clustering (HLC) algorithms and test how the results change if we add spatial weighting to the input flow network. The results show that link-based CD provides additional information, which would not be available to us by using the node-based approach, but at the cost of additional post analysis and computing time.

Keywords: community detection, flow networks, network analysis, movement networks, node-based CD, link-based CD

1 Introduction

One of the most common ways to represent movement in geographical databases are flows, where only an origin and a destination of a specific route are recorded. Commonly, flows are recorded alongside a value that represents the weight of a flow (i.e., the number of people travelling, or the number of packages sent), but not the full and accurate geometry of the route. Using these data we can generate flow networks, which are collections of interconnected flows between points (nodes) of the network. Every node can have one or more flows assigned to it, but a flow can have only two nodes.

One of the ways to analyse a flow network is by utilising community detection (CD) methods. Detecting communities in the network means identifying a partition of the network into subsets or clusters that have more connections internally than connections to other clusters (Girvan and Newman, 2002). Communities also represent a group of nodes/links which share common properties and/or play similar roles within the network (Fortunato, 2010). CD algorithms have been used in biological networks (Guimerà and Nunes Amaral, 2005), social networks and mobile phone networks (Ahn et al., 2010). However, typical CD algorithms only take the number of connections between two nodes into account and ignore the spatial characteristics of the flows (Expert et al., 2011). As a result, these methods consider spatially the closest node in the same way as the furthest node. This is especially disadvantageous in the analysis of movement networks (e.g., commuter flows, migration flows), where location and distance are two of the most important factors to consider.

In a social network, the importance of a social subject is based on how many connections someone has and not on the distance between the object and its neighbours. In travel networks however, the distance is vitally important, as it is related to both the cost and time of travel.

In this paper we show how a spatial component of the network can be utilised to generate new knowledge about the structure of the given network and corresponding movement patterns. We also compare two conceptually different approaches to community detection. One approach is detecting which node belongs to each community and the others is detecting which links belong to the community. There have been attempts to use spatial information in CD, but most have focused on detecting node communities (regions) (Adam et al., 2018; Expert et al., 2011; Farmer and Fotheringham, 2011; Guo et al., 2018). Our hypothesis is that detecting link communities instead of node communities will reveal a different geographical pattern in movement flow networks.

The rest of the paper is structured as follows. In Section 2 we explain the difference between the two types of community detection, then in Section 3 we explain how we modified the existing methods to take space into account. In Section 4 we present our results and algorithm outputs. These are discussed in Section 5, where we compare geographic and non-geographic results and explore similarities and differences between the different types of community detection. We conclude with some ideas for future work.

2 Community detection methods

The standard approach to community detection is node-based (Newman, 2004). This approach partitions the nodes of the network into groups that do not intersect each other, that is, each node is assigned to exactly one group. It can therefore be used to define regions or groups of similar subjects as communities do not overlap (Figure 1a). Many real-world networks have a highly overlapping structure, and there are many cases were one node should belong to multiple communities. This is especially the case in social networks, where an individual can belong to many different social groups (Lancichinetti et al., 2009).

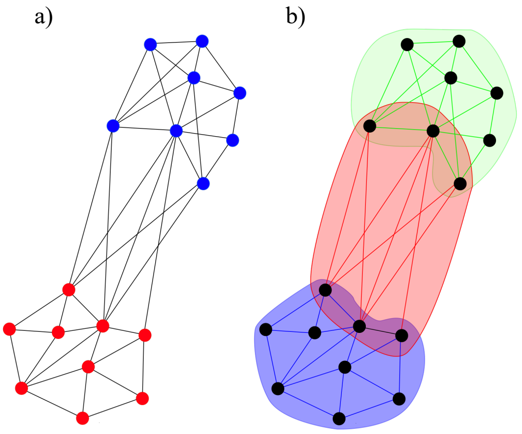
In a link based approach, a community is defined as a group of closely released links (Ahn et al., 2010). In this case, each node can belong to one or more communities (Figure 1b) which means that communities can overlap. The idea of classifying links was primarily envisioned for detecting overlapping communities, however there is no evidence that it is better than classifying nodes (Fortunato, 2010). Previous studies that have incorporated geographical context into community detection (Kempinska et al., 2018; Sekulić et al., 2018) or to remove spatial interference from the network (Expert et al., 2011) are all based on classifying nodes.

In a flow network which models movement, a node represents a location (city, town, checkpoint) and a link represents the flow (i.e., the actual movement) of an individual between the two places. If we want to create communities that represent similar movement flows, we need to classify links using a link-based method, which is what we propose in this paper.

Figure 1:

a) An example of node-based classification. We get two communities and every node belongs to a single community.

b) An example of link-based classification. We get three link communities and every link belongs to a single community, while nodes can be shared between them.



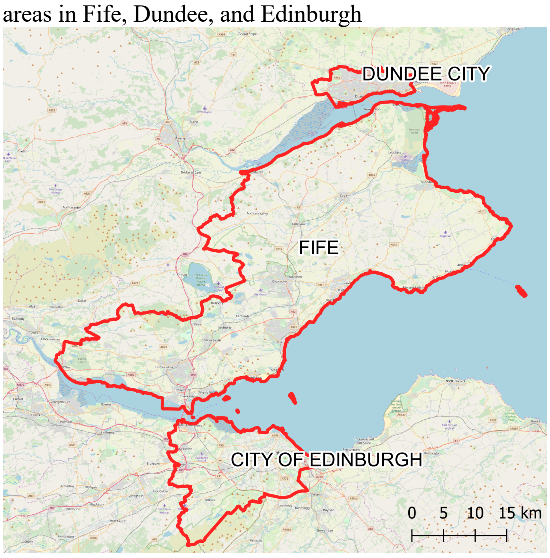
3 Methodology

We calculated the partition into communities by using two different methods and the same data. We used the number of flows and the spatial information to calculate flow weights, which we then used as input for a node-based and a link-based CD method.

3.1 Data

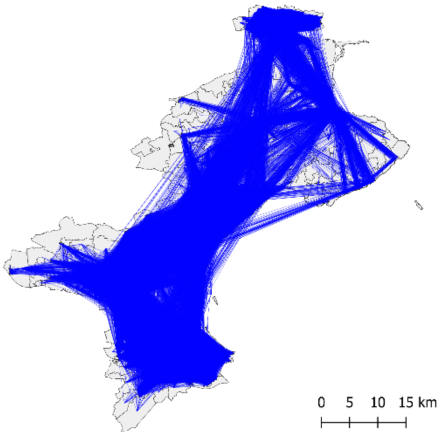
We used a subset of commuting flow data from the Scottish Census 2011 (ONS 2011). These data describe commuting flows between administrative units, that is, how many inhabitants live in one area and work in another. The highest resolution census data are provided at the output area (OA) level. OAs are designed such that they are as socially homogenous as possible and that they have a similar population size. One OA represents around 80 people, and has a varying extent, depending on population density (smaller in cities, larger in rural areas) (ONS 2018). The subset used for this analysis consists of the region of Fife, the city of Dundee and the city of Edinburgh. to cover the entire area where inhabitants of Fife commute to.

Figure 2: Map of the study area; consisting of output areas in Fife, Dundee, and Edinburgh



To construct our network, for each OA we calculated a centroid, and then assigned that centroid as a node in the network. To create our flows, we constructed straight lines that connect a node where people live with a node where they work and summed the number of total commuters on the single route of travel to get the flow weight. Occurrences of flows that start and end in the same OA were ignored (Bhatta and Larsen, 2011). Our final dataset consisted of 8842 nodes and 200028 flows.

Figure 3: Our flow network is overlayed over the study area



3.2 Adding geography to the network

Commuting is heavily impacted by distance, thus we desire to include geographical weights in our network to account for the effect of distance on commuting. To get a geographically weighted flow w_{ij}, we multiplied the existing weight (the number of commuters) n, with a function of distance $f(d_{ij})$.

$$w_{ij} = n * f(d_{ij}), \quad (1)$$

where $w_{ij}$is the weight of the flow between nodes *i* and *j* and $d_{ij}$the distance between nodes *i* and *j*.

.

There are several different ways we can calculate the geographic weight for our network. We can use the fixed distance approach where every flow whose length is larger than specified is not considered (i.e., distance threshold). That means that if the length of a flow d_{ij} is smaller than a predefined distance d, we multiply our number of people traveling on that flow with 1, otherwise we multiply by 0. A different approach would be to consider only m nearest nodes and flows that connect those nodes (i.e., nearest neighbour approach). The third approach, and the one we are using here, is to continuously scale our multiplication factor with the distance from the origin (i.e., distance decay effect). This means that shorter flows contribute more than longer flows, but they are all considered. The metric we use to scale flows in this is an inverse power function, where we raise distance of a flow to a power of the ratio of the flow and a predefined distance as shown in Equation 3 (Sekulić et al., 2018). However, other forms are equally viable for modelling the distance decay effect.

$$s_{ij} = d_{ij}^{-\frac{d_{ij}}{d}} \quad (3),$$

where s_{ij} is the spatial weight, d_{ij} the distance between the nodes *i* and *j,* and d is a predefined distance.

3.3 Running the algorithms

To test the effect of geography and to compare the two types of community detection we run both types of CD with both the standard network and network with geographical weights.

For the node-based approach we used the Louvain algorithm (Blondel et al., 2008). This algorithm uses the modularity optimisation to create the communities, allows for a weighted network and unfolds the complete hierarchy of the community structure. The algorithm first generates small communities by optimising modularity locally, then aggregates nodes belonging to the same community and generates a new network where communities from the previous step are converted into nodes. These steps are then repeated iteratively, until the maximum modularity is achieved. While the complexity of the algorithm cannot be explicitly calculated, it is estimated that it runs in time O(n log n), where n is the number of nodes. The results are travel to work regions.

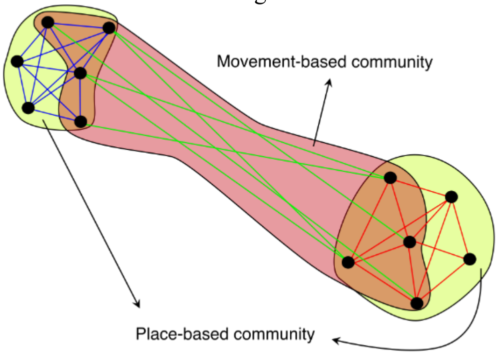
To detect link communities, we use the Hierarchical Link Clustering (HLC) CD algorithm (Ahn et al., 2010). This algorithm uses a set of nodes and connected neighbours, then for each link it calculates the similarity between links with the complexity of O(nK^2), where n is the number of nodes and K is the average degree of the network. Using the single link similarities, the algorithm builds a dendrogram. Cutting the dendrogram at some threshold yields link communities. The expected result are groups of commuters that move in a similar way and direction.

3.4 Identifying different types of link communities

For link communities to be meaningful, they should contain three or more links (Ahn, 2010) and we first filter out all communities that do not fulfil this requirement. Then we calculate the total number of people moving inside every community to quantify the magnitude of the community.

We can categorise movement based on the prevalent direction of the flows to get two different types of link communities as displayed on Figure 4.

Figure 4: Difference in between two types of link communities shown as a diagram



When the direction of the flows varies greatly, the link community represents place-based movement (i.e., within a town area). If most of the flows have a similar direction, the community represents movement between two locations (i.e., a regional commuting pattern). To identify each type, we calculate the standard deviation of bearings of all the flows in each link community. Low values of standard deviation will indicate which community is movement-based, and high values a place-based one. For each flow we calculate its bearing using Equation 3, where (x_1, y_1) and (x_2, y_2) represent the coordinates of the origin and the destination of the flow.

$$\rho = \operatorname{atan2}\left(\frac{y_1-y_2}{x_1-x_2}\right) \qquad (3)$$

From this, we then calculate the standard deviation of bearing for all the flows for each community.

4 Results

4.1 CD without considering geography

In the first step we used the original CD algorithms, without accounting for geographical location. Using the node-based approach, our study area was classified into three different node communities: one covering most of Fife, one for Dundee and the surrounding area and one for Edinburgh and the surrounding area.

When using the link-based approach, all of the flows get classified into a single link community. This means that we don’t find any new information about how and where people commute at all.

Figure 5: Place based communities filtered by the total number of commuters. Gray areas represent the results from node-based CD.
A – between 1000 and 10 000 commuters, B– more than 10 000 commuters.

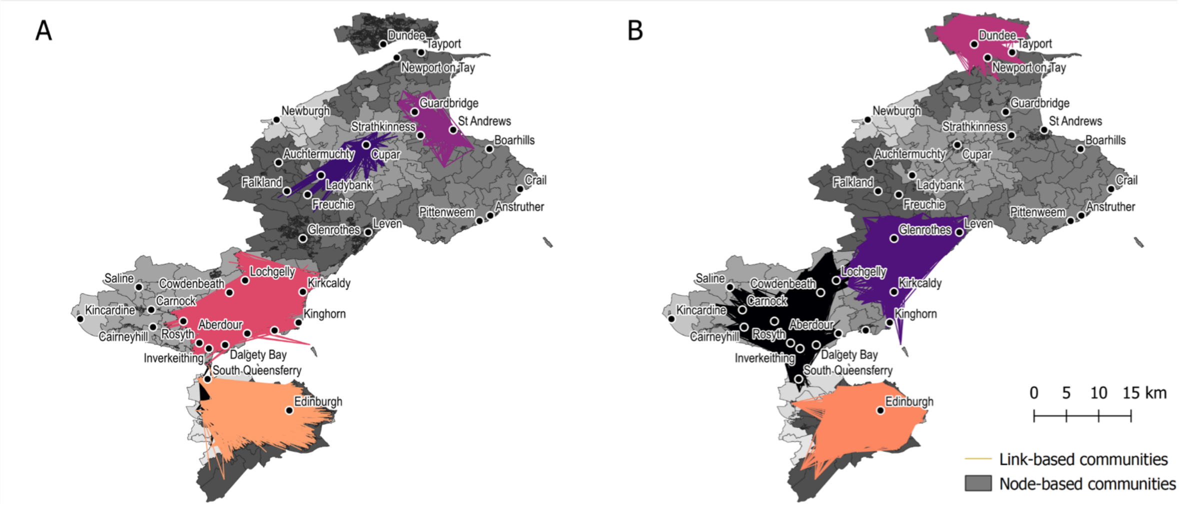
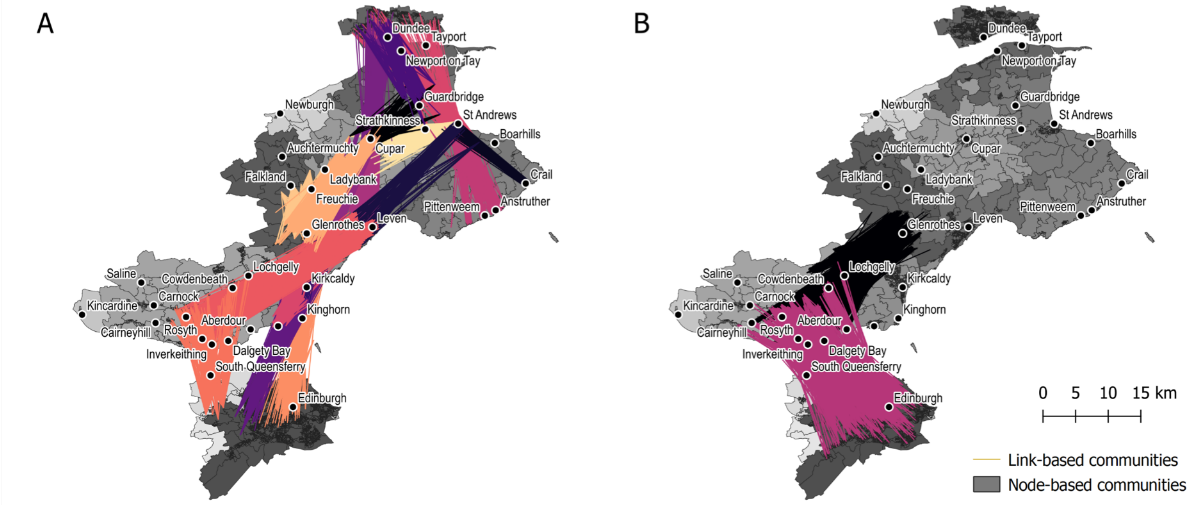


Figure 6: Movement based communities filtered by the total number of commuters. Gray areas represent the results from node-based CD.
A – between 200 and 1000 commuters, B – more than 1000 commuters.



4.2 Adding the geographical component

In the second step we added geographical weighting into both types of CD.

The geographical node-based community detection resulted in 14 different node communities. These areas are shown as grayscale regions on Figures 5 and 6.

By using the geographical link-based community detection on the same flow network we found 99 distinct link communities. Out of those, 69 were movement-based communities, and 30 were place-based. The decision if a given link community was place- or movement-based was done by selecting a threshold for standard deviation of the bearing of the flows, in this case we used 0.6 (radians).

Table 1 Comparison of the number of communities per different method and type of input network

Method	Non-spatial	Spatially weighted
Node-based (Louvain)	3	14
Link-based (HLC)	1	99

Table 2 shows the properties of link-based communities. The same information is unavailable for node-based CD as in that case flows cannot unambiguously be assigned to a community.

Table 2 Details about the minimum and maximum values for results of link-based CD

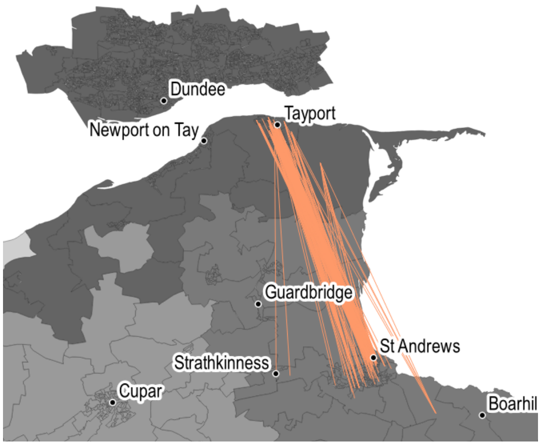
	Total	Place based	Movement based
Number of communities	99	69	30
Total population	294532	22967	271565
Total number of flows	200028	182327	17701
Minimum number of flows in a community	-	41	31
Maximum number of flows in a community	-	5631	105547
Minimum number of commuters in a community	-	51	53
Maximum number of commuters in a community	-	7114	146927

5 Conclusion

In this paper we have shown how using the spatial characteristics of the flow network influences the results of two different approaches to community detection. When geographical information about the locations is ignored, we get very little new knowledge about the movement patterns captured by the flow network.

We further compared the node- and link-based approaches to CD to investigate the hypothesis that classifying links is going to produce new knowledge on the movement inside the network. Node-based CD is able to detect regions of movement but is unable to detect the relationship between the regions. In addition to detecting regions inside the network using node-based CD, we have discovered that by using link-based CD it is possible to distinguish between two different types of movement (commuting movement, and within-city movement in our case) by looking at the variation in bearings of flows inside each community. The results show that it is possible to distinguish movement inside an area (i.e., a town area) and traveling between two locations (two towns) as shown on Figure 7. The place-based link communities roughly correspond to the regions generated by node-based CD, while the movement-based link communities contribute completely new information about movement patterns in our study region.

Figure 7: An example of a movement-based link community, which shows a group of commuters that travel from and to St Andrews.



This information can be used in many ways. For example, for targeted advertising, road management and planning, the locations of movement-based flows will be highly valuable. Further, movement-based link communities may be useful in identifying regions which are not well connected into existing workforce patterns. This information can be used to inform transportation policy and guide the development of infrastructure.

The disadvantage of the link-based approach is that it requires more computing resources compared to node-based approaches. In the continuation of our work we will optimize the algorithm to enable it to process larger datasets in less time and find better ways to visualise the results, for example using methods for visualising spatial interaction (Guo, 2009). We will also explore other forms of distance (e.g., road network) in developing these models together with adding additional contextual information in the networks (temperature change, pollution, weather). We also plan to analyse flows temporally to observe how communities change over time (day of the week, time of the day).

6 Acknowledgments

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].

7 Bibliography

Adam, A., Delvenne, J.-C., Thomas, I., 2018. Detecting communities with the multi-scale Louvain method: robustness test on the metropolitan area of Brussels. J. Geogr. Syst. 20, 363–386. https://doi.org/10.1007/s10109-018-0279-0

Ahn, Y.-Y., Bagrow, J.P., Lehmann, S., 2010. Link communities reveal multiscale complexity in networks. Nature 466, 761–764. https://doi.org/10.1038/nature09182

Bhatta, B.P., Larsen, O.I., 2011. Are intrazonal trips ignorable? Transp. Policy 18, 13–22. https://doi.org/10.1016/j.tranpol.2010.04.004

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Expert, P., Evans, T.S., Blondel, V.D., Lambiotte, R., 2011. Uncovering space-independent communities in spatial networks. Proc. Natl. Acad. Sci. 108, 7663–7668. https://doi.org/10.1073/pnas.1018962108

Farmer, C.J.Q., Fotheringham, A.S., 2011. Network-Based Functional Regions. Environ. Plan. A 43, 2723–2741. https://doi.org/10.1068/a44136

Fortunato, S., 2010. Community detection in graphs. Phys. Rep. 486, 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Guimerà, R., Nunes Amaral, L.A., 2005. Functional cartography of complex metabolic networks. Nature 433, 895–900. https://doi.org/10.1038/nature03288

Guo, D., 2009. Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data. IEEE Trans. Vis. Comput. Graph. 15, 1041–1048. https://doi.org/10.1109/TVCG.2009.143

Guo, D., Jin, H., Gao, P., Zhu, X., 2018. Detecting spatial community structure in movements. Int. J. Geogr. Inf. Sci. 32, 1326–1347. https://doi.org/10.1080/13658816.2018.1434889

Kempinska, K., Longley, P., Shawe-Taylor, J., 2018. Interactional regions in cities: making sense of flows across networked systems. Int. J. Geogr. Inf. Sci. 32, 1348–1367. https://doi.org/10.1080/13658816.2017.1418878

Lancichinetti, A., Fortunato, S., Kertész, J., 2009. Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. 11, 033015. https://doi.org/10.1088/1367-2630/11/3/033015

Newman, M.E.J., 2004. Detecting community structure in networks. Eur. Phys. J. B - Condens. Matter 38, 321–330. https://doi.org/10.1140/epjb/e2004-00124-y

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

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.

Sekulić, S., Long, J., Demšar, U., 2018. The effect of geographical distance on community detection in flow networks 5.