Applied Network Science

RESEARCH Open Access

Neighborhood discovery via augmented network community structure

Aaron Bramson[1,2,3*]

*Correspondence:
a_bramson@ga-tech.co.jp

[1] GA Technologies Inc., Roppongi 3-2-1, Minato-ku, Tokyo 106-6290, Japan
[2] Laboratory for Symbolic Cognitive Development, RIKEN Center for Biosystems Dynamics Research, Minatojima-Minamimachi 6-7-3, Chuo-ku, Kobe 650-0047, Japan
[3] Department of General Economics, Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium

Abstract

The geospatial characteristics of transportation networks structurally constrain their features, and as a result, analysis methods designed for social networks typically fail to capture useful characteristics or make informative comparisons. In the case of road networks, natural constraints on the edge distribution weaken the ability of standard community detection algorithms to find clusters of nodes that align with natural neighborhood extents. We show that by adding edge weights based on the similarity of localized subgraph features, we can apply modularity-based community detection algorithms to uncover improved neighborhood shapes and extents. The use of local network characteristics allows the feature analysis to be completed in linear time, thus making the approach expandable to very large networks. We demonstrate this technique with an application to central Tokyo.

Keywords: Transportation networks, Spatial networks, Community structure, Classification, Machine learning

Introduction

When analyzing geospatial data at the mesoscopic level (e.g. regions, counties, metropolitan areas), we may struggle to find a breakdown of a wide area into intuitive and useful regions of analysis. Official administrative boundaries may not exist at the appropriate level and are unlikely to divide the area into natural clusters. Our goal is to use micro level geospatial and network data to divide a large area into its organic neighborhoods. The identification of such neighborhoods is valuable for city planning, pricing models, real estate recommendations, and geospatial visualizations among other applications.

We define neighborhoods as "collections of localities with similar characteristics separated by localities with dissimilar characteristics." Such a description draws an obvious parallel with network community structure. However, due to physical constraints and their transportation purposes, road networks rarely exhibit sufficient density and connectivity variation to allow community structure algorithms alone to identify coherent mesoscopic structures. Specifically, community structure typically succeeds in distinguishing areas separated by rivers, highways, or railways, but fails to consistently separate areas with more nuanced differences in road patterns. Our approach is to first

generate edge weights for the road network based on the similarity of nodes' local network features, and then use these edge weights to assist the modularity maximization algorithm's ability to cut the network at natural boundaries

Using this technique, neighborhoods are emergent properties of the road network structure. Clearly the incorporation of population, employment, store, building height, greenery, water, etc. data would help identify perceptually similar and dissimilar areas. However, such data comes in the shape of grids, polygons, or administrative areas that are large with respect to the size of natural neighborhoods and would impose unnatural data gradations at the shape boundaries. Our assumption is that characteristics such as building height, floor area, neighborhood age, and zoning (residential, commercial, or industrial) are sufficiently correlated with features of the road network (length and straightness of edges, proportions of intersection types, etc.) that measures of the road network can indirectly distinguish these perceptual characteristics. For these reasons we focus on road network features that exist at the level of nodes and edges for discovering neighborhoods, with the later integration of rich geospatial data for describing and classifying them.

This treatment extends the work presented at the Complex Network 2021 conference (Bramson 2021) by expanding the list of subgraph attributes, analyzing larger and only distance-based subgraphs, changing the measure of feature-space similarity, generating neighborhood polygons from node sets, and refining the criteria for evaluating community detection results. The current approach is superior both in its methods and its outcomes. Further extensions, such as integrating additional demographic and environmental data and fuzzy community detection algorithms, are discussed in "Conclusions and future work" section.

Data sets

We limit our analysis to the central 23 wards of Tokyo; a region covering 614 km^2 (Ministry of Land, Infrastructure, Transport and Tourism 2020) of predominantly urban land with a population of 9,172,273 (Official Statistics of Japan 2015) and 7,153,658 jobs (Official Statistics of Japan 2014). (By comparison, New York City has a population of 8,804,190 distributed over 778 km^2.) Our current analysis focuses on exploring the ability to discover neighborhoods using only information from the driving network (as a proxy for other features); and as such, the data needed is limited to the network itself and additional data necessary for determining the width of roads and presence of sidewalks.

Network data

Our base network data is the road network of selected types for central Tokyo from Open Street Map (OSM) (OpenStreetMap 2022). The OSM road network includes nodes for all intersections as well as nodes to capture the curvature of the roads with straight segments. We simplify the network from OSM by merging edges across nodes with degree 2 so that most nodes in the simplified graph correspond to intersections. However, our network data is segmented into 1500 m $\times$ 1500 m tiles to make it manageable in computer memory, and a road segment is kept unmerged if it crosses a tile boundary to facilitate fusing tiles. Nodes of degree 2 are also kept when they occur at a road structure change (e.g., surface to tunnel or bridge), but the tile boundary condition

Table 1 Degree distribution of nodes in Central Tokyo

Degree	Count
1	15,629
2	16,213
3	110,741
4	34,923
5 +	723

Fig. 1 Steps needed to process the raw "Kiban" data (Geospatial Information Authority of Japan 2022) into usable geospatial information such as road width and sidewalk coverage. Part of the West Shinjuku area mapped using keplergl (2020)

accounts for most of the remaining degree-2 nodes seen in the degree distribution in Table 1.

Because of how the road network is encoded in OSM, the degree k of a node may not match the perceptual k-way intersection. For example, large roads are often separated for the different directions, so a 3-way "T" intersection of two large roads actually consists of 2 degree-4 nodes and 2 degree-3 nodes (e.g., see the center of Fig. 1c). There exists a single intersection where 9 roads (including overhead expressway exits) intersect, but in the highly detailed OSM representation the highest degree node at that intersection is 4.

Our driving network for the 23 wards region has 177,924 nodes and 261,254 edges in the simplified undirected graph. We buffer the 23 wards region by 400 m to avoid boundary effects in our analysis of node-based subgraph attributes. The nodes/edges in this buffered region are used to collect attributes for the subgraphs, but are not included in the network during community detection.

Cartographic data

In the case of Japan, road widths are rarely input into OSM. Roads are tagged by type in the OSM data (e.g. motorway, primary, residential); however, the road attributes (such as speed limit and road widths) vary within each category and overlap across categories.

In order to obtain accurate road width values we process cartographic "Kiban" data from Geospatial Information Authority of Japan (2022) that provides outlines of road edges, road components, and waterlines as well as building polygons (Fig. 1a).

This data is not immediately usable for measuring road widths or sidewalk coverage. It comes as just a table of LineString geometry objects and needs to be converted into polygons of blocks and sidewalks. Furthermore, the lines are often not well-formed; e.g., lines marking the outer edges of sidewalks often fail to connect to the block at the endpoints and/or are left open at the inner edge when they span multiple blocks. We perform a multi-step image processing analysis to generate polygons for blocks and sidewalks, the result of which can be seen in Fig. 1b.

Although the results are imperfect, we are able to achieve an overall high degree of accuracy with respect to road and sidewalk widths in most areas [as confirmed by comparison with measurements on Google Maps (2022)]. The generated blocks and sidewalks, along with the included building polygons, are then used to measure road widths and the percentage of road lengths that are serviced by sidewalks (as well as building spans and mean sidewalk widths not used here; Fig. 1c) (more details in "Road widths and sidewalk coverage" section).

Analysis methods

As stated in "Introduction" section, our method weighs edges of the road network by the similarity of the nodes' local network features. We perform our analysis separately for four different values of subgraph extent: 100, 200, 300, and 400 meters. For each node in the road network, we collect subgraphs containing the nodes and edges within that range. Edges are included if (1) both end nodes are within range or (2) the edge directly connects to the focal node regardless of its length.

Measures

There is a vast literature on analyzing road networks to estimate/predict movement activity using features of the network structure (Omer et al. 2017; Serra and Hillier 2019), optimize logistics (Goczyła and Cielatkowski 1995; Gai et al. 2019), and perform structural comparisons (Barthélemy 2011; Austwick et al. 2013). Naturally, these studies use network measures appropriate for the task at hand (e.g. angular closeness and betweeness centralities). Because our goal here is instead capturing features that can distinguish neighborhoods, a different suite of measures and a novel method to integrate them are necessary.

For each ego-centric subgraph we collect the twelve variables listed in Table 2. These measures were chosen both because they capture perceivable features of a road network that may contribute to a neighborhood's identity and because they are relatively fast to compute (i.e., compared to centrality measures and angularity). Specifically, for each subgraph, all measures can be computed in linear time with a single pass through that subgraph's edge list.

We can separate the measures into two categories: (1) measures of the network structure and (2) measures of the roads. The main distinction being made here is that network measures are purely topological and require no geospatial characteristics (measures 1, 2, 8, 9, 10, and 11 from Table 2), while road measurements require the network be

Table 2 Measures collected for each ego-centric subgraph

1	Number of nodes
2	Number of edges
3	Total edge lengths (measure of density)
4	Mean road segment length
5	Average edge straightness weighted by length
6	Average road width weighted by length
7	Average percent sidewalk coverage weighted by length
8	Proportion of degree 1 nodes
9	Proportion of degree 3 nodes
10	Proportion of degree 4 nodes
11	Square clustering (meshedness)
12	Triangle clustering

embedded in an a spatial context (measures 3, 4, 5, 6 and 7 from Table 2). We measure similarity using different combinations of measures in order to assess which kinds of measures generate more natural neighborhoods: (1) all measures, (2) uncorrelated measures (see "Measure analysis" section), (3) network features only, (4) selected network features, and (5) road features only. Measure set (4) takes the network-only measures and ignores the proportions of degree 3 and 4 nodes because, given the OSM representation, they do not correspond to the perceived number of roads at an intersection.

Straightness

Because we have simplified the network, the edge geometries are no longer straight lines; they are 'linestrings' capturing the concatenated segment lines. In this way, we can generate an edge straightness attribute from the ratio of Euclidean endpoint distance over the linestring length. This measure is a less rich description of the perceived road straightness than a circuity measure across all the subgraph leaf nodes because it measures the curvature of the roads themselves rather than of the paths across the network, but it has the benefit of being computationally simple.

Road widths and sidewalk coverage

The Kiban data, once processed into block and sidewalk polygons, can be used to determine the widths of roads and presence of sidewalks for an embedded road network. Road widths are determined as the sum of the distances on each side of an edge to the closest point on a sidewalk, block, or building polygon (minimum road width along that edge). Using buffers on each side of each edge, we determine the proportion of the edge length that contains sidewalk polygons, and take the greater value of each side.

For small connecting road segments inside intersections, there are no sidewalks or blocks on either side. Rather than assigning unrealistic values (e.g., road widths of 400 m), these edges inherit the values from the roads they connect to and align with. The full details of our processing from OSM network and Kiban data into road widths and sidewalk coverages is particular to these datasets and the Japanese context, so it is omitted here. However, those interested in this process are invited to contact the author for further details.

Measure analysis

Increasing the range of the subgraph makes neighboring nodes more similar because there is greater overlap in their subgraphs. This creates a smoothing effect that can be seen in Fig. 2 as more coherent and larger patches of similar values.

More than just smoothing the attribute values across the nodes, because edge weights are calculated as the similarity of their end nodes' subgraphs, these coherent patches are also expected to induce patterns in the edge weights. Specifically, areas that are similar across multiple attributes internally, but distinct from surrounding areas with respect to those attributes, will form neighborhoods using community detection on the network. More smoothing, and hence more coherent patches, should make it easier to identify those neighborhoods.

As seen in Fig. 3 for the 300 m subgraph case, the level of correlation among these variables is surprising low overall. The number of nodes, number of edges, and the sum of edge lengths are very highly correlated, as one should expect, because they are all measures of road network density. Square clustering (meshedness) is only moderately (0.48) correlated with the proportion of degree-four nodes, and even less correlated (0.11) with road straightness—both lower than expected. This can be partially explained by the OSM representation of larger roads as split in the two directions, so that (as described above) a square is created within a 'T' intersection of large roads, and many intuitively grid-like areas would not be captured as 'squares' in the network. Given the OSM representation, a more sophisticated measure is necessary to capture perceptual meshedness among large roads.

The overall low level of correlation among these attributes means that, even when the individual variables exhibit clear spatial clusters through subgraph size smoothing (as seen in Fig. 2), each variable generates a different pattern of spatial clusters. Because all variables are aggregated through subgraphs, they are all smoothed, but neighboring nodes can be highly similar in some attributes and very dissimilar in others. This is

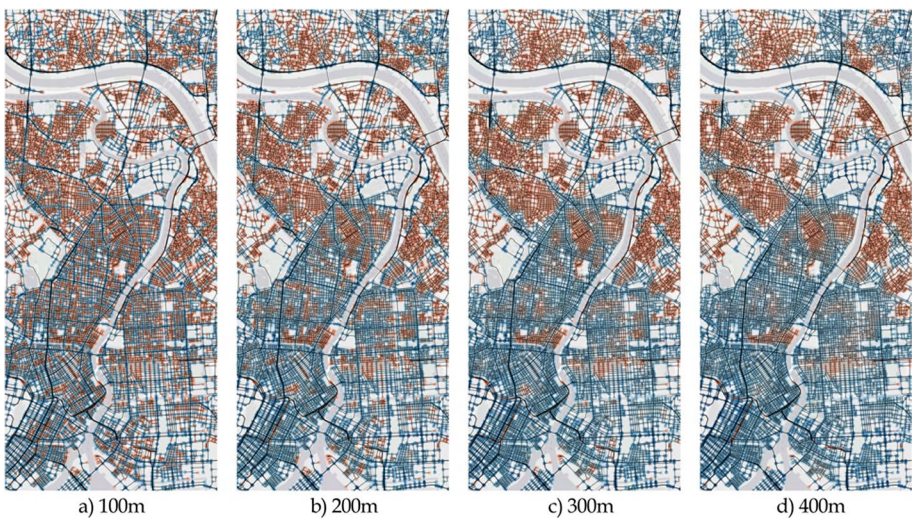

Fig. 2 The normalized percents of sidewalk coverage for each node's subgraph at varying subgraph sizes (blue is more coverage). Larger subgraphs tend to increase the similarity of neighboring nodes, thus smoothing the spatial distribution of attribute values

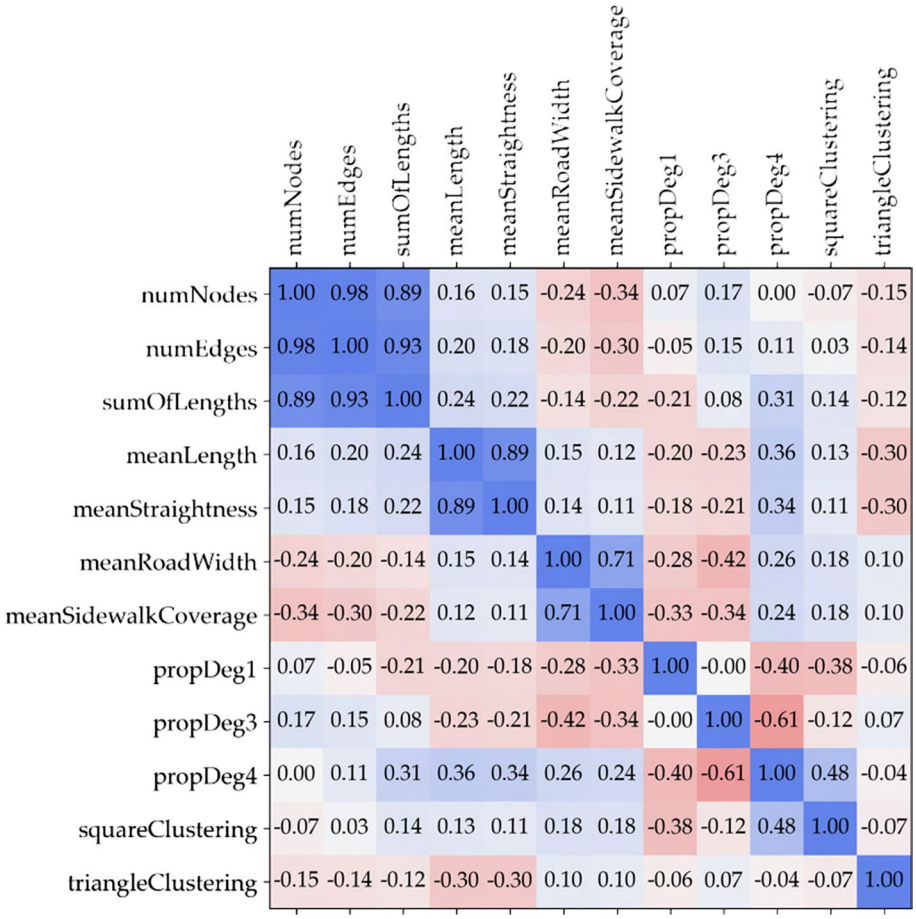

Fig. 3 Pearson correlation matrix for the local network measures using 300 m subgraphs. Other subgraph extents exhibit the same qualitative pattern with greater correlation for larger distances

especially true for areas with long street segments. End nodes for roads having lengths of 200 m or more can have little overlap among their subgraphs, and hence may vary significantly in their attributes.

Based on these patterns in measure correlations, we use the following measures for our 'uncorrelated' measures set: sumOfLengths, meanLength, meanRoadWidth, meanSidewalkCoverage, propDeg1, propDeg3, propDeg4, squareClustering, and triangleClustering.

Algorithms

The computational methods for this work include the network community detection algorithms, the methods to evaluate the performance of the discovered communities, and an additional method to convert spatially overlapping collections of nodes in communities into partitioning polygons for neighborhoods.

Network communities

Recall our definition of neighborhoods as "collections of localities with similar characteristics separated by localities with dissimilar characteristics." The unweighted version of the

greedy modularity communities algorithm, as implemented in NetworkX (Hagberg et al. 2008), finds communities wherein there are more internal connections than external connections. As already noted, the degree of nodes in a road network is highly constrained by physical limits, and methods appropriate for social networks are rarely useful for such spatially embedded networks. That said, some areas are densely packed with roads and separated from other areas by sparse connections imposed by parks, rivers, railways, and large expressways; intuitively, modularity does play some role in describing and individuating neighborhoods.

The modularity maximization algorithm has a resolution parameter (γ) that adjusts the tradeoff between internal and external edges so that values less than 1 favor larger communities while values greater than 1 favor smaller communities. After exploring a range of values, we find that using a resolution parameter of $\gamma = 4$ produces community sizes that approximate many intuitive community extents, so all experiments presented here use $\gamma = 4$.

Edge weights

In order to strengthen the ability of modularity maximization to identify coherent neighborhoods, our idea is to weight edges by the similarity of their end nodes. Modularity would therefore reflect *stronger similarity within a community and less similarity outside*. First, for each subgraph extent, we separately standardize the values x_i of each measure m using the typical method:

$$\hat{x}_i = \frac{x_i - \mu_m}{\sigma_m}. \tag{1}$$

where μ_m is the mean value and σ_m is the standard deviation for that measure m across all nodes for that subgraph size. Now, all the measures are in the same scale space where mean values are at zero and the unit is standard deviations. For each measure set $\mathcal{M}$, we then create the feature vectors of the standardized variables. From here we experimented with calculating the distance d_{ij} between the lists of features of two nodes i and j for variable set $\mathcal{M}$ using a few different distance metrics (log of Euclidean distance, truncated distance, geometric distance, etc.), but achieved the best results with ordinary Euclidean distance:

$$d_{ij} = \sqrt{\sum_{m}^{\mathcal{M}} \left(i_m - j_m\right)^2} \qquad \text{for measure } m \text{ in set } \mathcal{M}. \tag{2}$$

Although these distances are all in terms of standard deviations across multiple dimensions, they can be arbitrarily large, and we need weights based on similarity rather than distance. In order to convert pairwise node feature vector distances into similarity weights, we use the following two functions:

$$\omega_1(ij) = \frac{1}{1 + d_{ij}} \tag{3}$$

$$\omega_2(ij) = \frac{1}{1 + d_{ij}^2} \tag{4}$$

Both functions yield weights between 0 and 1 (which we multiply by 10,000 so we can store them as integers to speed up processing while retaining sufficient resolution), but generate different edge weight distributions as can be seen in Fig. 4.

Evaluating detection performance

There is no ground truth for what the "correct" neighborhoods are in an objective sense because they are intrinsically perceptual and rely on an implicit consensus and shared understanding among a large number of people. However, insofar as there are generally accepted cohesive neighborhoods, we can evaluate how well each of the parameter and dataset combinations reveals those neighborhoods. Using domain knowledge of the Tokyo area we have manually created a set of polygons representing 26 known neighborhood cores as shown in Fig. 5.

By "core" we mean a portion of a neighborhood that any reasonable person or algorithm would identify as cohesive. While both intuitive and detected neighborhoods are likely to extend beyond their cores, the degree and directions in which they extend may depend on one's perspective and the particular attributes considered. In some cases it is ambiguous whether to consider an area as one cohesive neighborhood or two nearby distinct but similar neighborhoods; we may accept both clusters in which those two are together or separate neighborhoods. In such cases we chose only one of them to act as a core so as to keep the evaluation criteria from becoming too complicated. By defining neighborhood cores in this way, we can evaluate the performance of community detection using standard machine learning-style tests.

We use the polygons from Fig. 5 to identify the set of nodes in each core after removing the nodes for motorways and motorway links (which we consider as not parts of communities, even though they are important for identifying communities). This subset of 1381 nodes in the systems constitutes our ground truth by placing each relevant node in exactly one mutually exclusive community. A successful neighborhood discovery will

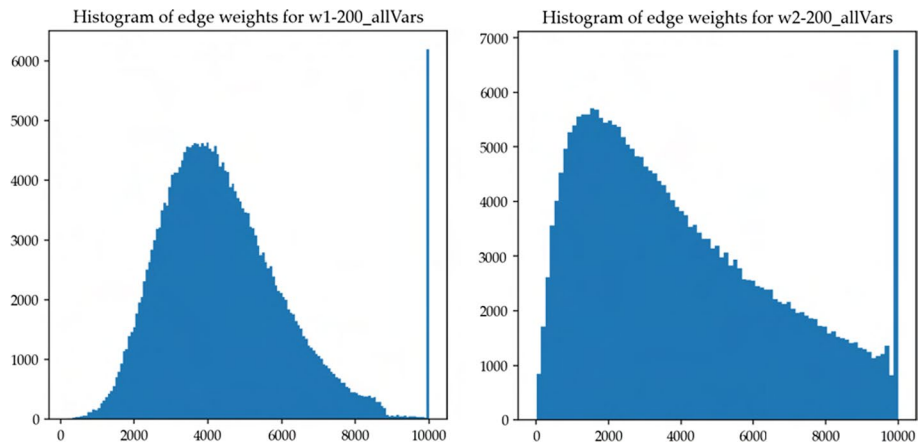

Fig. 4 Using the same extents and variable set, the ω_1 weights are nearly normally distributed with a spike at 1. ω_2 weights, on the other hand, are more even and skewed lower

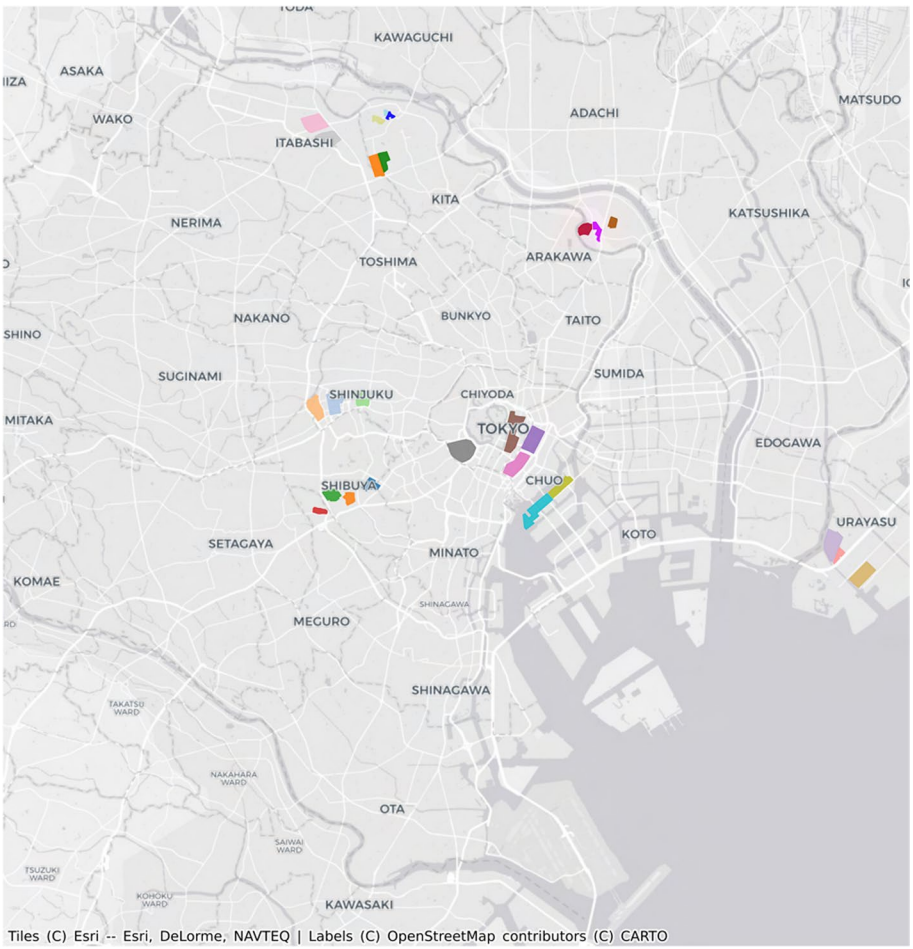

Fig. 5 The set of neighborhood cores used to evaluate the performance of detected communities. An algorithm performs well if all the nodes in each core are in the same community, and nodes in different cores are in different communities

(at least) keep all the nodes in the same core together in the same community while separating the nodes in different cores into different communities. In order for this method to effectively evaluate neighborhood discovery performance, it is necessary to identify sets of distinct cores that are spatially near each other; however, they needn't share edges because the gaps can capture ambiguous in-between and/or transition zones.

Our approach here parallels the way machine learning algorithms are typically evaluated. We manually define a ground truth (annotated training and test data), and test the ability of different attributes and parameter combinations to match that pattern. After identifying the combination that performs the best on the test areas, we then explore the particular neighborhood boundaries in generates as well as the discovered neighborhoods in areas where we don't have clear a priori knowledge.

Generating neighborhood polygons

The modularity maximization community detection algorithm will create a 'crisp' partition of the network nodes, but because they are spatially embedded, nodes in different communities may overlap in physical space. In future work we will address fuzzy

neighborhood boundaries with a non-partitioning community detection algorithm (see "Other community detection algorithms" section for more details), but for this work we wish to maintain the partitioning feature so that the neighborhoods can be intuitively plotted on a map and used for quick visual reference. For these reasons we need to convert collections of nodes assigned to communities into polygons for the neighborhoods.

In order to generate neighborhood polygons from the collections of nodes, we first assign a community membership to an edge if both end nodes are in that community, then remove edges that do not belong to any community (as well as motorway and motorway link edges). We then buffer the community edges by 50 m (so they are now 100 m wide bars) and merge them into a single object. If the result is a multipolygon, then we only keep the largest component polygon. We close any holes in these polygons, erode it by 90 meters, and again remove isolated areas by keeping the largest polygon for each community. At this stage, the community polygons are not a partition because they may overlap and have gaps, so we cut out any overlapping regions from all polygons that overlap and then fill in gaps using the *tessellation* function from the Momepy library (Fleischmann 2019). Figure 6 shows the relationship between nodes from community detection and the generated polygons for one case. Although we use the node sets resulting from community detection in the quantitative analysis in "Clustering accuracy comparison" section, we use this technique to create the maps seen in "Discovered neighborhoods details" section.

Neighborhood discovery results

First we present the accuracy results of our discovered communities based on standard clustering performance measures as well as aggregated error rates. We then explore the discovered communities beyond the neighborhood cores. Finally, we examine some

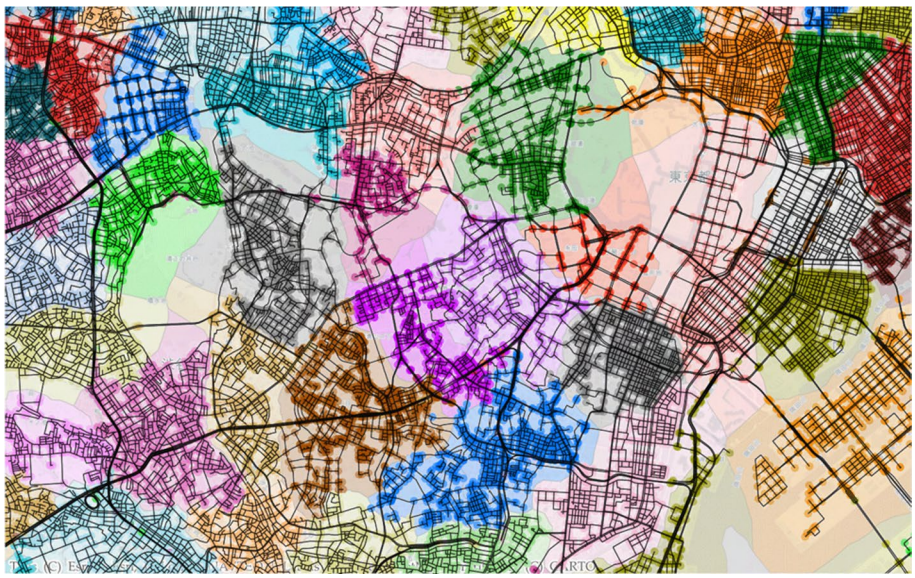

Fig. 6 The nodes belonging to each detected community and the neighborhood polygon generated from those nodes using ω_1, 300 m subgraphs, and all variables. Some nodes at the peripheries of communities fall into a different neighborhood through the smoothing and tessellation process

selected areas in order to assess the strengths and weaknesses of the methods for future improvements.

Clustering accuracy comparison

As described in "Evaluating detection performance" section, there are no "correct" organic neighborhoods to use as ground truth, so we evaluate the performance of our models using manually annotated neighborhood cores. For each weighting equation (ω_1 and ω_2), each subgraph extent (200, 300, and 400 m), and each of the five variable sets, we compare the nodes in clusters found by greedy modularity maximization with our defined neighborhood cores. The results are presented in Table 3.

The adjusted mutual information (AMI) and Rand index are common methods for comparing the similarity of two partitions, and as such are often applied to clustering results. Although we present both values in Table 3, they are 95.7% correlated across

Table 3 Community detection accuracy results for each parameter combination

	Weight	Subgraph extent	Variable set	AMI	Rand Index	Proportion missed	Extra communities
0			unweighted	0.864	0.662	0.124	**0.273****
1	ω_1	200	allVars	**0.942****	**0.885****	**0.025*****	0.318
2	ω_1	200	unCorr	0.933	0.835	**0.027****	**0.182*****
3	ω_1	200	netOnly	0.919	0.785	**0.043***	0.318
4	ω_1	200	networkSome	**0.936***	**0.871***	0.059	0.318
5	ω_1	200	roadFeatures	**0.943*****	**0.896*****	0.044	**0.273****
6	ω_1	300	allVars	0.908	0.805	0.068	0.500
7	ω_1	300	unCorr	0.854	0.638	0.079	0.364
8	ω_1	300	netOnly	0.862	0.658	0.075	0.409
9	ω_1	300	networkSome	0.872	0.748	0.101	0.545
10	ω_1	300	roadFeatures	0.927	0.839	0.071	0.318
11	ω_1	400	allVars	0.889	0.771	0.088	0.545
12	ω_1	400	unCorr	0.898	0.787	0.101	0.636
13	ω_1	400	netOnly	0.902	0.802	0.086	0.500
14	ω_1	400	networkSome	0.894	0.771	0.106	0.455
15	ω_1	400	roadFeatures	0.896	0.761	0.071	0.364
16	ω_2	200	allVars	0.927	0.852	0.085	0.545
17	ω_2	200	unCorr	0.919	0.838	0.093	0.591
18	ω_2	200	netOnly	0.920	0.848	0.098	0.636
19	ω_2	200	networkSome	0.929	0.838	0.072	0.455
20	ω_2	200	roadFeatures	0.924	0.840	0.078	0.364
21	ω_2	300	allVars	0.860	0.671	0.152	0.682
22	ω_2	300	unCorr	0.874	0.703	0.186	0.682
23	ω_2	300	netOnly	0.865	0.718	0.172	0.682
24	ω_2	300	networkSome	0.869	0.736	0.143	0.727
25	ω_2	300	roadFeatures	0.883	0.671	0.112	0.545
26	ω_2	400	allVars	0.881	0.713	0.167	0.727
27	ω_2	400	unCorr	0.886	0.728	0.122	0.727
28	ω_2	400	netOnly	0.886	0.749	0.129	0.591
29	ω_2	400	networkSome	0.888	0.770	0.120	0.636
30	ω_2	400	roadFeatures	0.906	0.811	0.096	0.545

*** indicates the best results, ** the second best, and * the third best result

our 31 clusterings (using Spearman rank correlation), so we focus our discussion on AMI. AMI takes the value of 1 when two partitions are identical, and 0 when the overlap can be explained fully by chance. This adjustment for chance means that the AMI does not exactly track the number of misclassified nodes, it matters how they are distributed across the communities. The four highest AMI scores were all achieved using ω_1 and 200 m subgraphs. The highest uses only road features, the second uses all measures, and the third uses selected network measures; however the ω_1|200 m results range from 0.919 to 0.943 across variable sets, meaning they are all quite high.

Although the AMI and Rand score are common measures of clustering quality, we can dig a bit deeper and look at two additional aspects of misclassification. First, the *proportion missed* column of Table 3 reports the proportion of nodes covered by each ground truth polygons that are not in the same community as the community with the most nodes in that polygon. The best performing clustering in this respect (Table 3 row 1) missed only 2.5% (i.e., 34) of the 1381 nodes covered by neighborhood cores. This row achieves the second highest AMI (again, because the AMI depends on the distribution of errors in its adjusting for randomness). By contrast, the row with the best AMI and Rand scores (Table 3 row 5) missed 4.4% of the nodes across all neighborhood cores. Overall, the proportion missed values are − 72.9% correlated with the AMI values using Spearman rank correlation.

The *extra communities* column of Table 3 tells us the degree to which the incorrectly assigned nodes fall into few or many discovered neighborhoods. It is calculated as the number of discovered communities that intersect the ground truth polygons, minus the number of ground truth polygons, then divided by the number of ground truth polygons. A perfect score is 0, and the values for our results range from 0.182 to 0.727. The ground truth communities are covered by between 1 and 4 discovered communities; some cores were perfectly identified by every clustering, while others were problematic (to varying degrees) for most clusterings. The best performing clustering in this respect (Table 3 row 2) split just 4 of the cores into just 1 additional community. Overall, the extra communities values are − 49.4% correlated with the AMI values using Spearman rank correlation.

Unweighted results comparison

The unweighted results (Table 3 row 0) achieve the second best results with respect to the number of extra communities, but otherwise the results were among the worst. An AMI of 0.864 places it at the 5th worst, while its proportion missed is 7th worst. Its AMI is 3rd worst and its proportion missed is worse than all ω_1 clusterings. These results tell us that while unweighted modularity is indeed capable of distinguishing reasonable organic neighborhoods in many cases (an AMI of 0.864 is not terrible), adding weights based on similarity generally improves the performance of neighborhood discovery.

Weighting function results comparison

In general, the AMI results are similar but slightly worse for ω_2 compared to ω_1 across subgraph sizes and measure sets. On average, ω_1 is only 0.01 better, but being better or worse depends on the particular parameters. The Spearman correlation of the AMIs between ω_1 and ω_2 is 0.742 (for the Rand index it is 0.421). The proportions of missed

nodes are also similar with a Spearman correlation of 0.63, but ω_1 is consistently better for every parameter combo.

Subgraph extent comparison

Like with the weight equation differences, it is not the case that one subgraph extent is always better than others. The 200 m subgraphs are always better than 300 m and 400 m results for the same ω and variable set, but there is no consistent pattern when comparing 300 m and 400 m results. Larger subgraph extents means more smoothing, and although the 100 m results were insufficiently interesting to even be worth presenting, using less smoothing fosters a slightly better ability to differentiate neighborhood boundaries.

Quantitative analysis summary

In the final analysis, the best results came from using ω_1 and a 200 m subgraph extent. All of the variable sets performed well, and which ones are considered best depends on which metric is used. If minimizing the number of misclassified nodes is the benchmark, then using all the variables yields the best results; however, using just the road features is better when the overall preservation of neighborhoods is the goal (as measured by AMI and the Rand index).

Discovered neighborhoods details

Although the quantitative analysis above can tell us which clusterings capture the clear and distinct neighborhood cores we identified, evaluating the naturalness and usefulness of the discovered neighborhoods requires examining them on a map. Below we present some details of the discovered neighborhoods in order to gain a better understanding of where and why they succeed and fail.

Unweighted results

As a baseline, and an example of a lower-accuracy clustering, we first present the results of using *greedy modularity* without edge weights in Fig. 7. As expected, unweighted modularity suffices to separate neighborhoods when rivers, large motorways, or train tracks act as barriers, but fails to separate areas with very different features. As an example, in the upper-left part of Fig. 7 we can see the large blocks west of Shinjuku station are split between a small red neighborhood and a large light blue neighborhood around Shinjuku station. This comparatively recently built-up area has long, straight roads, sparse but large buildings, and exceptionally wide sidewalks, and it really should be its own neighborhood. Considering the small sizes of the red neighborhoods to the west and south of this area, this is not a problem with the resolution (γ) parameter. Because the unweighted algorithm has no information on the road widths, sidewalks, etc., it cannot distinguish this area from its surroundings. In fact, because of the relatively sparse roads in this sky scrapper dominated area, it serves as a natural cutting point between neighborhoods similar to parks, large stations, the Emperor's Palace, and bodies of water.

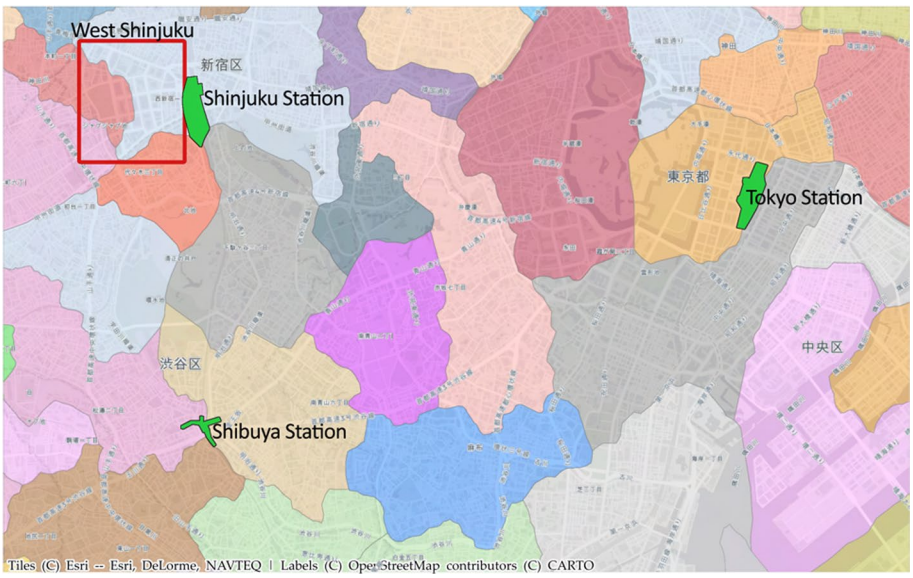

Fig. 7 Close-up map of the communities discovered using the greedy modularity algorithm without edge weights. Pictured area is 9.33 km wide by 5.6 km tall and covers the areas around Shinjuku, Shibuya, and Tokyo stations

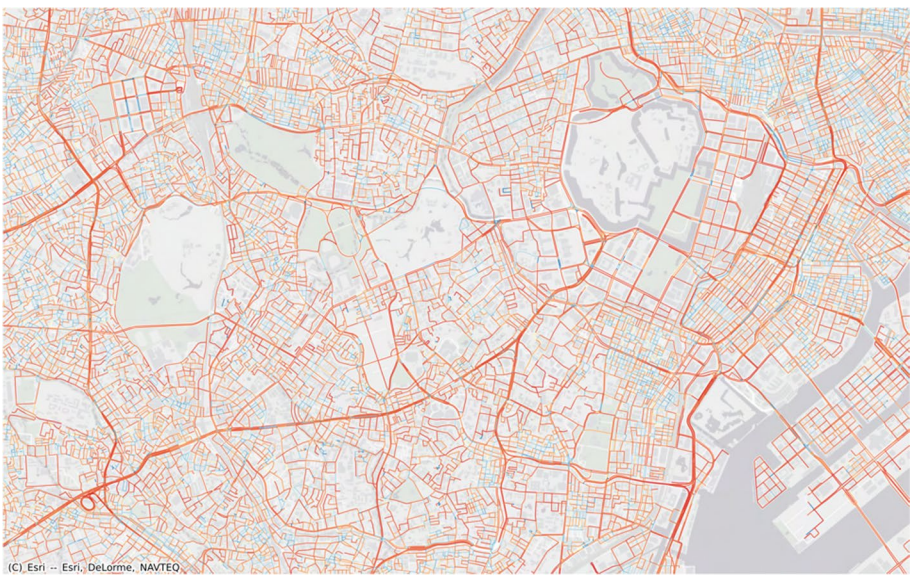

Fig. 8 Close-up map of the edge weights capturing node similarity using ω_1, 200 m subgraphs, and all variables (lowest error combination). Dark red indicates minimal similarity, dark blue indicates the maaximum similarity, colors diverge at the mean similarity value. Coherent patterns are few and small within this central area

Generated edge weights

Augmenting the network with similarity weights is expected to allow the greedy modularity algorithm to find communities with similar features. In order for the edge weights to actually help the modularity algorithm, they need to form coherent patterns useful for

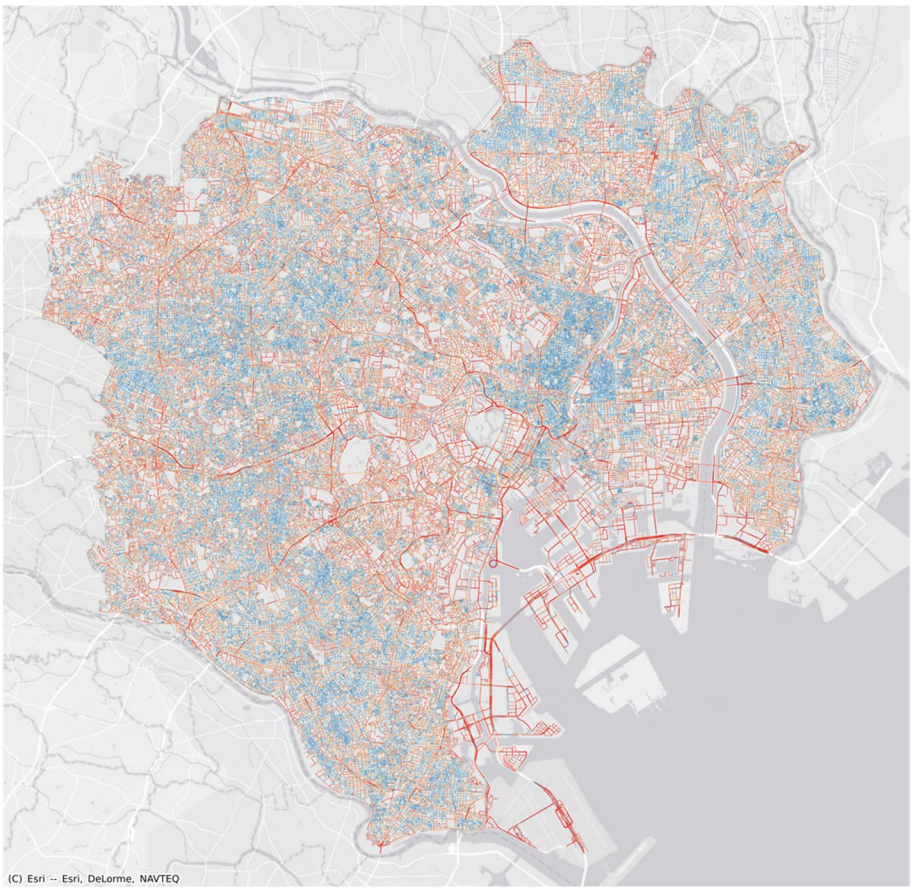

Fig. 9 Map of the edge weights capturing node similarity among all variables using ω_2, 400 m subgraphs, and only road features. Dark red indicates minimal similarity, dark blue indicates the max similarity, colors diverge at the mean similarity value. Fairly large collections of similar nodes appear throughout the area

identifying organic neighborhoods. However, as shown in Figs. 8 and 9, the edge weights do not in general reveal coherent clusters of similar nodes for much the study area.

Because each node's features aggregate across (in the case of Fig. 8 200 m) subgraphs, we expected a stronger smoothing effect, but in the dense and heterogeneous road network of central Tokyo, swapping a few roads can (and does) make a significant difference in the aggregated subgraphs features. However, the degree of edge weight clumpiness depends on the weighting function and measures used, and also the area. Figure 9 shows the edge weights for the case using ω_2 and 400 m subgraphs on the road feature variables. Here we can clearly see coherent patterns in the high vs low weights that establish the expected sets of nodes that are similar within and dissimilar between communities. There are still many areas lacking such a pattern, including many of the areas identified as neighborhood cores, that result in the relatively low, but not especially poor, performance of this parameter combination.

Even without a clear pattern of strong inner and weak outer edge weights, because we are using these similarity-based edge weights to augment the modularity maximization, we can still rely on modularity to do the heavy lifting in discovering neighborhood boundaries, while utilizing these weights to modulate exactly where the separations occur. This is because modularity maximization works across all the edges within and

between proposed communities, and it is relative strengths that matter for where to split the nodes. However, these edge weight results imply that additional/different features and/or a different method for identifying communities based on these weights may be necessary to clearly identify neighborhoods.

Weighted modularity results

As described in Section "Generating neighborhood polygons" we take the nodes within each network community found by modularity maximization and generate polygons for the neighborhoods so we can more easily assess the shape and process geospatial data for those areas. The resulting neighborhoods for row 1 of Table 3 (the least proportion missed and second best AMI by 0.001 point), which come from the edge weights in Fig. 8, are shown in Fig. 10 for the downtown region. This is a clear improvement over the unweighted results in Fig. 7: the area west of Shinjuku station gets its own (medium blue) neighborhood, Shibuya is split between primarily retail and primarily residential areas, Roppongi get its own clearly defined area, the two sides of Tokyo Station are separate, etc. Only some of these were defined as neighborhood cores, so the results beyond that also produce very natural neighborhoods in many areas. On the other hand, the Marunouchi area is combined with the Ginza area, part of the government building zone is in the Akasaka neighborhood, and the distinct parts of Tsukishima island are grouped together. So, although this parameter combination performed well on our tests, and has many desirable features, it also leaves room for considerable improvement.

We can compare the results for all variables in Fig. 10 with the results for road features in Fig. 11. This is row 5 of Table 3 that achieved the highest AMI and second fewest extra communities. This parameter combination achieved similar accuracy on our cores, and some neighborhoods are extremely similar in size and shape, but there are also important differences. Using only the road features correctly distinguished the Marunouchi

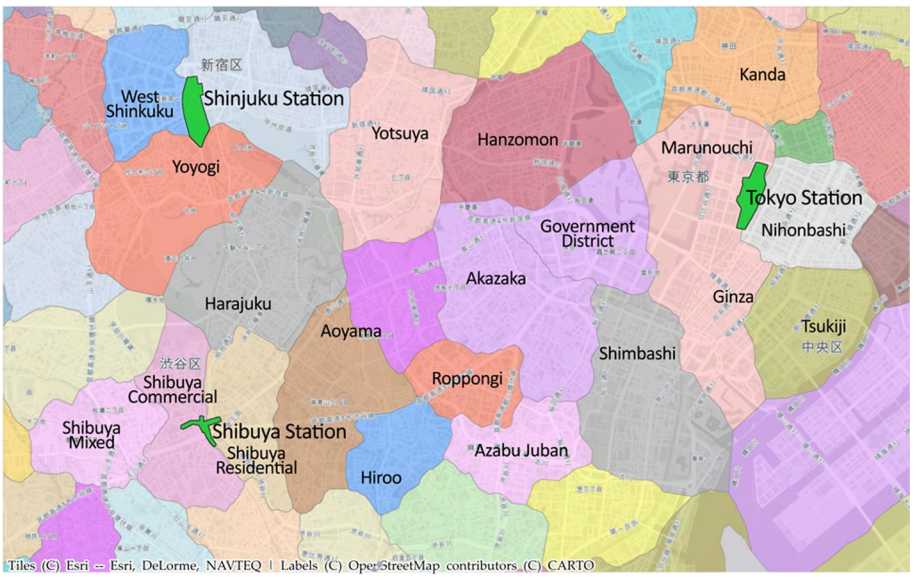

Fig. 10 Close-up map of the communities discovered using the greedy modularity using ω_1 edge weights based on all variables for 200 m subgraphs

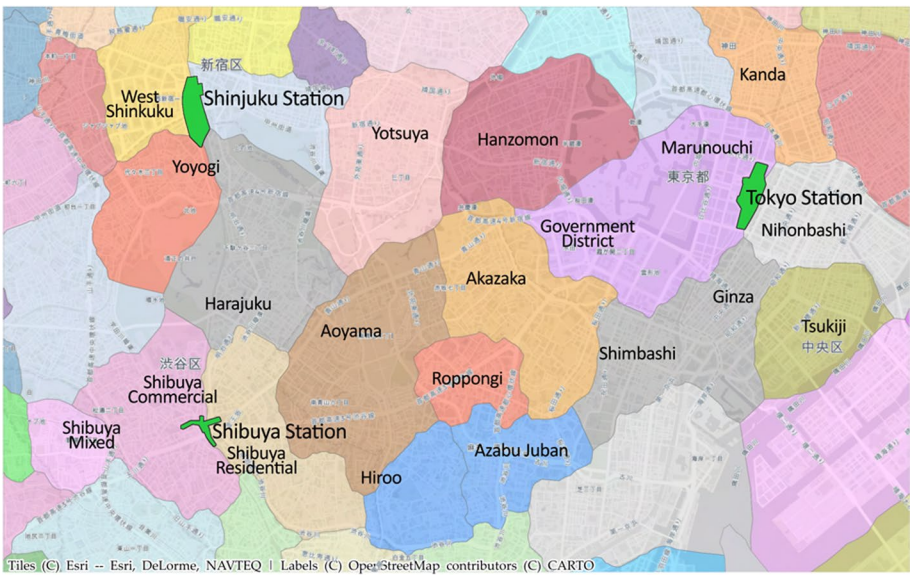

Fig. 11 Close-up map of the communities discovered using the greedy modularity using ω_1 edge weights based on road features for 200 m subgraphs

area from the nearby Ginza area, and joined Marunouchi with the government zone, which is much more natural, although it would be even better if these too were separated. They both mistakenly group two of the four Shibuya area cores together, but they group together different ones. Subjectively, the neighborhoods discovered using only the road features seem slightly more natural than using all variables, but they both have strengths and weaknesses.

The revealed communities also conform to other intuitive characteristic neighborhoods in the detailed area as well as across the whole 23 wards (Fig. 12). Figure 12 In some of the regions (especially in the suburbs) the neighborhood separations seem arbitrary because the community size prescribed by the resolution parameter is smaller than the natural neighborhood. However, it is easier to join such communities in post-processing than to modify the modularity algorithm to accommodate larger community size disparities. The shapes of discovered neighborhoods conform well to many unofficial identified regions such as the Roppongi Area, Ginza Area, and Shimbashi Area. There are stations and/or administrative areas with these names, but the intuitive neighborhood bearing that name is better captured by these discovered neighborhoods than the official boundaries. In this sense, the results are already useful; they offer advantages in characterizing, scoring, and associating localities compared to the current next best alternative (official administrative areas).

Numbers of communities

For ω_1, all the variable sets and subgraph spans generate a similar, but varying, number of communities in the 331 to 402 range, as shown in Table 4. In contrast to that, ω_2 generates a wider range of values (from 408 to 634) that are across-the-board larger than the ω_1 results. However, the pattern of *rankings* of the number of communities across subgraph span and variable sets is essentially identical between ω_1 and ω_2. Considering

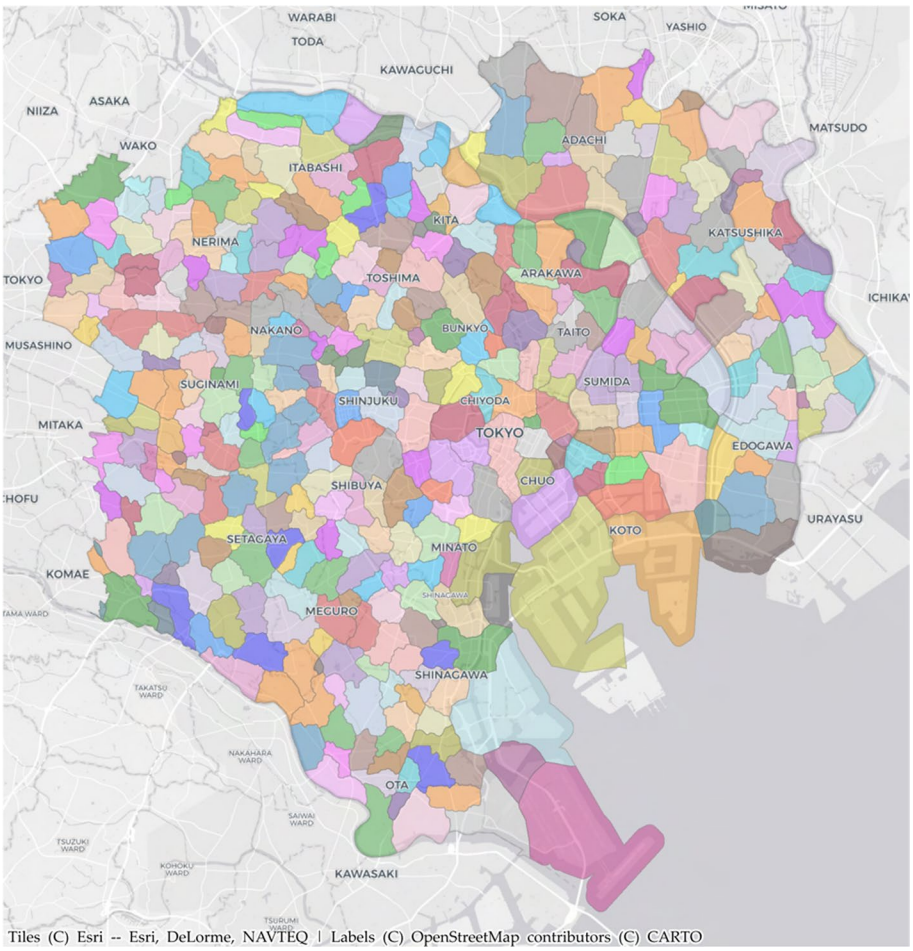

Fig. 12 Discovered neighborhoods for the 23 ward region using all variables on 200 m subgraphs

Table 4 Number of communities found for each parameter set

Variable set	Number of communities ω_1			Number of communities ω_2		
	200 m	300 m	400 m	200 m	300 m	400 m
All variables	402	378	370	634	561	503
Uncorrelated	397	378	366	602	527	461
Network only	393	373	359	582	511	461
Select network	385	375	359	563	512	438
Road features	376	365	331	507	459	408

how greedy modularity maximization works, and that the difference in the weighting schemes alters the values but not the rankings of the edge weights themselves, this preserved pattern is unsurprising. However, the result that stretching the distribution of weights towards extremes has the effect of generating consistently, and significantly, more groups is interesting. Although intuitively we could make more, smaller communities and merge similar ones in post-processing to generate more heterogeneous neighborhood sizes, this may not be a viable approach because the ω_2 results generated more

communities but performed less well in separating according to our annotated ground truth.

Conclusions and future work

We have shown that the standard *greedy modularity maximization* algorithm can already find reasonable neighborhoods in parts of Central Tokyo due to typical dividing/connecting features such as large roads, rivers, and railways. However, that alone is insufficient to distinguish dissimilar areas that are nonetheless well-connected. By augmenting the network with edge weights based on the similarity of local network characteristics, we successfully separate dissimilar areas to discover improved intuitive neighborhoods.

Compared to previous results (Bramson 2021), adding accurate road widths and sidewalk coverages based on the Kiban data, as well as changing the function used to convert the Euclidean distances of normalized values in feature-space to similarity weights, improved the consistency of edge weights into coherent patterns. This edge-weight refinement improved the results enough that it required additional and more stringent tests for the quality of the detected communities to evaluate them. In order to facilitate the quantitative comparison of results, we manually generated a ground-truth of neighborhood cores: areas in which nodes must be considered as the same neighborhood but distinct from other cores. This method allows us to utilize this approach to simultaneously ensure that known neighborhoods are distinguished and explore natural delineations where domain knowledge and intuitions are lacking. Although the discovered neighborhoods are already useful improvements over using official administrative areas, there is still much work to be done to uncover the generally recognized organic neighborhoods.

Some aspects of real neighborhoods complicate this research. Although some neighboring neighborhoods exhibit starkly different ambiances, many more blend into each other smoothly. This was one motivation for using only neighborhood cores to differentiate them for evaluation: it is actually ambiguous/vague where to draw the line between them. There are also gradations in the relative strengths of neighborhood differences: Ginza and Shimbashi may feel like distinct neighborhoods when walking from one to the other, but they are very similar to each compared to the rest of Tokyo. So, it doesn't feel wrong to lump them together or keep them as separate (but similar) neighborhoods. In some cases there are large swaths of land for which there are no differentiating features, while in other cases a single street can have its own ambiance that sets it apart from the rest of the area, a kind of micro neighborhood. The current methods are a good start at identifying neighborhoods with results that are useful for many practical applications in urban planning and real-estate, but they lack the ability to make some of these more nuanced discriminations.

Evaluation criteria

The quantitative results here are clearly dependent on the particular neighborhood cores chosen. If one had a preferred method, one could pick and choose neighborhood cores to make that particular method yield the best outcome. Furthermore, the 26

neighborhood cores relied upon here represent only a few easy-to-discriminate types of neighborhoods; therefore, adding more cores of different types could change the ranking of the best-performing methods. Despite these potential issues, annotating clearly distinct neighborhoods and using these to train/test neighborhood discovery methods is a powerful approach. It is compatible with all kinds of clustering methods (even beyond network community detection), facilitates standard comparison measures, and provides concrete goals that simultaneously allows for fuzzy communities, gradual transitions, and ambiguous boundaries. Because none of the discovery methods we used here could perfectly identify all the neighborhood cores already identified, it suffices for our needs. However, expanding the coverage of areas and types of cores, even including smaller regions such as single streets, can improve the usefulness of this evaluation approach.

Other community detection algorithms

There are only a few community detection methods available in NetworkX that can handle weighted edges (Hagberg et al. 2008), and we applied all available methods. However, both the *label propagation* algorithm and the *Girvan-Newman* method (using the maximum edge weight as the iterative removal function) yield tens of thousands of communities. This is due to the largely homogeneous and low degree distribution of road networks and the fragmented nature of the edge weights noted above (Fig. 8). As such, these methods are ill-suited to the problem of detecting communities using similarity weights based on our data. However, with changes to the edge weights (e.g., using different variables) these methods may become useful in the future.

The *greedy modularity communities* algorithm we use creates partitions of contiguous nodes into communities. This is a desirable feature for some applications, but we acknowledge that some nodes are interstitial and should not be a member of any community. On the other end, communities may overlap and blend into each other. Using a community detection method based on edge-weighted probabilistic walks or other fuzzy approaches could achieve both the community gaps and overlaps that intuitively exist in neighborhood identifications. We attempted to implement such a fuzzy community detection algorithm (e.g. Kundu and Pal (2015)), but were unable to find or adapt Python code that could handle large networks with weighted edges. Furthermore, it is unlikely that existing methods that are targeted at community detection in social networks would perform especially well on road networks for reasons already stated. We have plans to develop our own network community detection algorithm specifically designed for geospatial transportation networks, but developing and testing such a method is left for future work.

Integrate demographic and environmental data

As already noted, the current method does not consider demographic or environmental information when discovering neighborhoods. One can imagine a predominantly residential area of large apartment buildings contiguous to an area of large office buildings both with similar road network structures. In this case, the two areas may exhibit distinct neighborhood feels that are indiscernible from the road network alone. Incorporating population, employment, greenery, zoning, and other data into the analysis is targeted for future work, but because this data is typically only available at much larger

scales than the micro-subgraphs used here, we expects it to be used in post-processing (i.e., identifying similar neighborhoods and measuring the heterogeneity/cohesiveness of neighborhoods).

Bottom-up neighborhood discovery

Although neighborhoods can be described as regions with a coherent ambiance, there are naturally variations in the ambiance within them. This leads to the question of how much variation to tolerate and still consider the neighborhood coherent. Label propagation and Girvan-Newman style community detection methods, as well as hierarchical clustering from machine learning, can be implemented to fuse micro communities that are similar enough to each other. One plan going forward is to evaluate the ambiance using features around individual edges, and then merge edges with similar qualities. By applying different thresholds, we can generate neighborhoods of different scale to adapt to different purposes.

Recall again our definition of neighborhoods as "collections of localities with similar characteristics separated by localities with dissimilar characteristics." Because we are looking for areas that are internally similar and externally distinct from neighboring areas, here we tried using modularity with similarity-based edge weights to approximate maximizing internal similarity, but we can also develop or adapt an algorithm that specifically aims to maximize internal similarity and external dissimilarity of sets of nodes.

Summary

This approach has been largely successful in its task: the unsupervised learning of organic neighborhoods from local road network features. Successful enough to be useful for the purposes of identifying and visualizing similar organic neighborhoods in Tokyo. Further improvement will likely be achieved through additional data, refinement of the edge weights, and/or different community detection algorithms. Different desiderata, different locations, and robustness considerations will drive our efforts towards better methods along these lines. We hope to also apply these techniques to other major cities with unofficial, but well known and clearly defined, neighborhoods such as New York, Boston, and Barcelona through collaborations to both assess its generality and expand its usefulness.

Abbreviations

AMI	Adjusted mutual information
OSM	Open Street Maps

Acknowledgements
Not applicable.

Author contributions
Single author work. The author read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The road network and Kiban data used are openly and freely available from the references cited. Additional data and/or plots of data generated (correlation matrices, maps of other experiments, scatter plots of variables, etc.) are available upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not Applicable.

Competing interests
This research was performed by employees of GA Technologies and may lead to the development of products or information services which may be used by GA Technologies for business operations.

Received: 10 March 2022 Accepted: 13 June 2022
Published online: 15 July 2022

References

Austwick MZ, O'Brien O, Strano E, Viana M (2013) The structure of spatial networks and communities in bicycle sharing systems. PLoS One 8(9):74685
Barthélemy M (2011) Spatial networks. Phys Rep 499(1–3):1–101
Bramson A (2021) Neighborhood discovery via network community structure. In: International conference on complex networks and their applications. Springer, pp 769–779
Fleischmann M (2019) momepy: urban morphology measuring toolkit. J Open Sour Softw 4(43):1807. https://doi.org/10.21105/joss.01807
Gai W, Du Y, Deng Y (2019) Multi-objective route planning model and algorithm for emergency management. Decision-making analysis and optimization modeling of emergency warnings for major accidents. Springer, Singapore, pp 113–150
Geospatial Information Authority of Japan (2022) Basic map information, basic items download site. https://fgd.gsi.go.jp/download/menu.php
Goczyłla K, Cielatkowski J (1995) Optimal routing in a transportation network. Eur J Oper Res 87(2):214–222
Google (2022) Tokyo Japan. https://www.google.com/maps/place/Tokyo,+Japan/@35.6922069,139.6936926,20.25z/data=!4m5!3m4!1s0x60188b857628235d:0xcdd8aef709a2b520!8m2!3d35.6803997!4d139.7690174?hl=en
Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using networkx. Technical report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States)
keplergl (2020) kepler.gl. https://github.com/keplergl/kepler.gl
Kundu S, Pal SK (2015) Fuzzy-rough community in social networks. Pattern Recognit Lett 67:145–152
Ministry of Land, Infrastructure, Transport and Tourism (2020) Administrative area data. https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N03-v2_4.html. Accessed 26 Jan 2021
Official Statistics of Japan (2014) Economic census for business frame, tabulation of establishments, results for Japan. www.e-stat.go.jp. Accessed 12 Jan 2020
Official Statistics of Japan (2015) Subregional population by age and sex from the 2015 census. www.e-stat.go.jp Accessed 11 Nov 2020
Omer I, Kaplan N, Jiang B (2017) Why angular centralities are more suitable for space syntax modeling? In: Proceedings of the 11th international space syntax symposium, Lisbon, Portugal, pp 3–7
OpenStreetMap contributors (2022) Planet dump. Retrieved from https://planet.osm.org. www.openstreetmap.org
Serra M, Hillier B (2019) Angular and metric distance in road network analysis: a nationwide correlation study. Comput Environ Urban Syst 74:194–207

Publisher's Note