Extracting Backbone Structure of a Road Network from Raw Data

Hoai Nguyen Huynh[1][0000−0002−8432−7435] and Roshini Selvakumar[2]

[1] Institute of High Performance Computing
Agency for Science, Technology and Research, Singapore
huynhhn@ihpc.a-star.edu.sg
[2] Raffles Institution, Singapore

Abstract. The representation of roads as a networked system of nodes and edges has attracted significant interest in the network literature, generating a large number of studies over the years. Such representation requires a proper identification of what constitute an edge or a node. Intuitively, nodes represent the junctions where roads intersect, and edges are the road segments connecting these junctions. In practice, however, such simplified presentation is not trivial to achieve due to extra details of individual roads. In this paper, we present a set of novel and efficient computational techniques based on elementary geometry and graph theory that can be employed to obtain the essential structure of a road network, while also retaining the crucial geometry of roads, such as shape and length. This is done by dissecting the network into clusters of nodes of degree other than 2 and curves, which contain consecutive nodes of degree 2, connecting these clusters. These clusters of nodes and curves will be collapsed into cleaned nodes and paths, respectively in a simplified mathematical graph. We apply this method to obtain the simplification of road network in Punggol new town in the northeast region of Singapore, and show that application of network analyses such as centrality measures could be performed in a more meaningful and concise manner than on the original road network.

Keywords: Road network · Clustering · Geometrical graph.

1 Introduction

Modern computers with various architectures developed in recent years have enabled the storage and processing of a large amount of data, especially the spatial ones. Such development has led to a rapid growth in the availability of (open) urban data for almost every corner on the Earth's surface. The available geospatial data allow us to study different aspects of urban systems, or cities, including their physical elements of road networks. The representation of roads as a networked system of nodes and edges has attracted significant interest in the network literature, generating a large number of studies over the years (see *e.g.* [15, 16, 18, 5, 10, 7]).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_43

The available geospatial data on road networks, typically exist in the digital format of a shapefile [8] that contains points listed in a spatially sequential order to store the geometrical information of the road lines defined by the points. Although the data format is convenient in storing spatial features of point or lines for roads, the network information concerning the connections among points or intersections between roads is generally not available and requires further processing before studies such as network analysis could be performed [7]. On the other hand, the detailed geometrical information of roads unpremeditatedly makes the application of network analysis on the raw data challenging. The extra elements included in the data, such as multiple (clustered) points when two roads intersect (*e.g.* slip road or filter lane) or multiple lines representing the road segments (*e.g.* lanes) between a pair of junctions, create spatially redundant information that hinders network analyses such as centrality measures, since a single intersection may contain many unnecessary and noisy data points.

As a result, the presentation of roads as networks requires a proper identification of what constitute an edge or a node to ensure that accurate models can be created to study road networks in a meaningful way. In the literature, various methods and algorithms have been proposed to simplify or generalise road networks by removing redundant information and performing merging processes to properly represent nodes and edges in the network. At the simplest level, intermediary nodes along a road line are treated redundant and removed from the network as they are only intermediate to provide connections between a pair of junctions [6]. For more sophisticated simplification, a number of approaches have been proposed including the combination of nodes' degree with road length and road attributes [21, 17] or making use of the proximity of network nodes to facilities [14, 19]. Alternatively, with the additional availability of dynamic traffic data, main structure of urban road networks have been extracted based on their usage through the determination of importance (or grade) of roads using vehicle trajectories [23] or traffic flow pattern [22, 20].

The above-mentioned approaches are either too simplistic (only removing intermediary nodes) or require additional data which may not always be available. In this work, we propose a procedure to simplify a road network from its raw geospatial data using a novel set of techniques based on elementary geometry and graph theory. The simplification would yield the essential structure of the road network, which could be represented as a mathematical graph object. In this approach, we only use spatial information of points along road lines, omitting other road attributes such as name or grade as these may be incorrect or inconsistent at times. The problem we tackle, therefore, boils down to extracting the backbone structure of a network of pure geometric connections. This structure of the road network in an urban system would subsequently allow the studies of other aspects of the system, such as its morphology, beyond the road network itself. In the remainder of this paper, we first describe our proposed procedure for simplification of a road network. After that, we apply the procedure to the network of roads in an area in Singapore and illustrate the value of the simplified network, especially when applying network analyses such as centrality measures.

2 Data and methods

2.1 Data

OpenStreetMap (OSM) [3] is a collaborative source that is free and features a map, from which geospatial data for different regions of the world can be downloaded. The data come in a number of file formats, with `shapefile` being one of the most popular ones. For the analysis in this work, we utilise shapefiles of road data from Nextzen service [2], which provides up-to-date extracts of OSM data, mostly for major cities in the world. In this format, the data contain entries of road chunks with sequential lists of points along the roads, which could be converted to network format by creating according nodes and edges.

2.2 Simplification of road networks

After obtaining the data, we proceed with a network of roads, in which a node represents a point on a road line and an edge connects a pair of successive points along the line. Depending on the level of detail in the data, a road could be represented by a single or multiple road lines, corresponding to the number of lanes of the road. A node could belong to multiple road lines, for example when two roads intersect or merge. However, very often, two road lines cross one another without sharing any common node in the data. In this work, we will treat a road network as a planar graph, *i.e.* no two segments on the same plane can cross without having a common node.

General framework The general framework we employ to simplify a road network is depicted in Fig. 1 and involves four steps. Firstly, the graph is planarised by identifying the pairs of segments that cross one another and adding the points of intersection as new nodes in the graph. In the second step, the graph is cleaned by removing the dangling paths of length shorter than 30 m as these are considered dead ends, which don't lead to a place different from the other nodes. Next, the graph is dissected into curves, which are lines of consecutive nodes of degree 2, and clusters of nodes of degree other than 2. Finally, each cluster is collapsed into a single node and the curves connecting the same pair of clusters are merged into a clean path between the newly collapsed nodes in the simplified graph. These four steps are repeated until no further simplication could be made to the final graph.

Graph planarisation A road network provides connections among places in an area on the Earth's surface. Therefore, it is reasonable to consider a road network a 2-dimensional object. This property, however, cannot be taken for granted as very often the data doesn't contain information on the points of intersection when roads pass through each other. Without these points, it would not be possible to properly obtain all the useful information from the network regarding the connections it provides. Therefore, it is imperative that we planarise the

DOI: 10.1007/978-3-030-50371-0_43

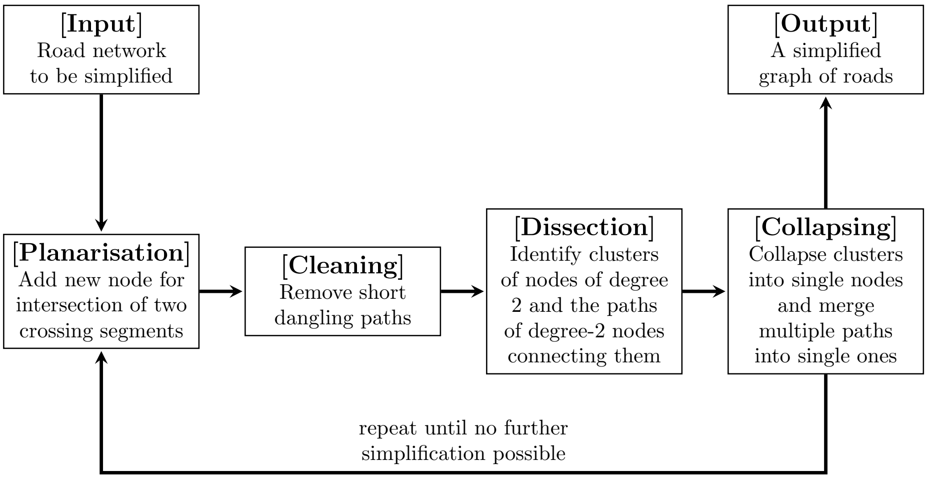

[Input]
Road network to be simplified
[Planarisation]
Add new node for intersection of two crossing segments
[Cleaning]
Remove short dangling paths
[Dissection]
Identify clusters of nodes of degree 2 and the paths of degree-2 nodes connecting them
[Collapsing]
Collapse clusters into single nodes and merge multiple paths into single ones
[Output]
A simplified graph of roads
repeat until no further simplification possible

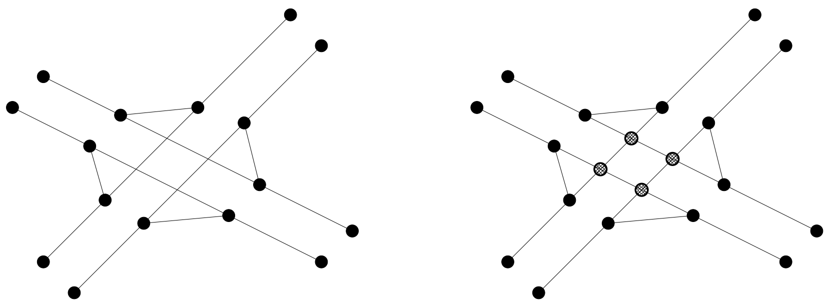
Fig. 1: Procedure for simplification of a road network.

graph such that no edges intersect one another, but rather meet only at their end points. In order to carry out the planarisation, intersection points were found and added as nodes (see Fig. 2). In reality, even real roads can lay on top of one another without crossing, it is still suitable and necessary to ensure a road network is planar (see Sec. 3.3 for an argument).

Graph cleaning An actual road network can contain short dead ends that branch out from a main road, for example, to provide access to a particular building which might have side entrance. While such access is necessary in reality, the branch does not make a significant difference in terms of connectivity between places. As such, dangling paths of length shorter than 30 m is trimmed off the graph, reducing extra unnecessary junctions in the road network (see Fig. 3).

Graph dissection In order to simplify the road network, it is vital that we took out any redundant information that would not contribute to a greater understanding of any one intersection and its connections to other intersections. For example, a junction or road intersection may contain at least one point for a simple intersection that has a 3- or 4-way approach. Most simple intersections tend to have at least four points since the roads have more than one lane. Another example of a junction that may contain redundant information is a junction that contains ramps in addition to the simple intersections. Moreover, roads that have multiple lanes and that intersection would result in one point for every lane intersection.

We notice that a network of roads could be dissected into two groups of elements. The first group contains the chunks of degree-2 nodes that connect from a non-degree-2 node to another non-degree-2 node. These chunks can be uniquely



(a) Raw roads with non-intersecting crossing road lines.

(b) Planarised roads with new nodes (crosshatched) added for intersections of crossing road lines.

Fig. 2: Planarisation of roads.

and effectively identified by picking a degree-2 node and start tracing in opposite directions (as the node has only 2 neighbours) until a non-degree-2 node is reached at both ends. The idea behind this approach is that if all connecting roads, characterised by paths of degree-2 nodes, were removed, it would leave clusters of (connected) nodes that can be considered as corresponding road junctions since there are no road chunks connecting nodes within them (see Fig. 4). These clusters would form the second group of elements in the road network.

Node and curve collapsing After dissecting the road network into two groups of non-degree-2-node clusters and degree-2-node chunks, we collapse them into simple nodes and curves, respectively, to construct the simplified network.

Nodes For each cluster, all the points are collapsed into a single node located at their (geometric) centroid and added as a new node to represent the corresponding junction in the simplified network (see Fig. 5). By replacing the points with their centroid, the representative node contains contribution from all member nodes.

Curves We apply the same concept to curves, *i.e.* to get the representative curve that could be considered "average" of all the pertaining curves. To do that, we notice that after the node collapsing above, all the curves share the same end points, which are also the end points of the averaged curve. It is reasonable to argue that the general direction of the curves is set by these two end points. Therefore, we use a line passing these two points as our reference line. We then find the projection of all points in the curves onto this reference line. The total number of projections on the reference line is the sum of number of points in all curves, unless some points in two different curves share the same projection.

DOI: 10.1007/978-3-030-50371-0_43

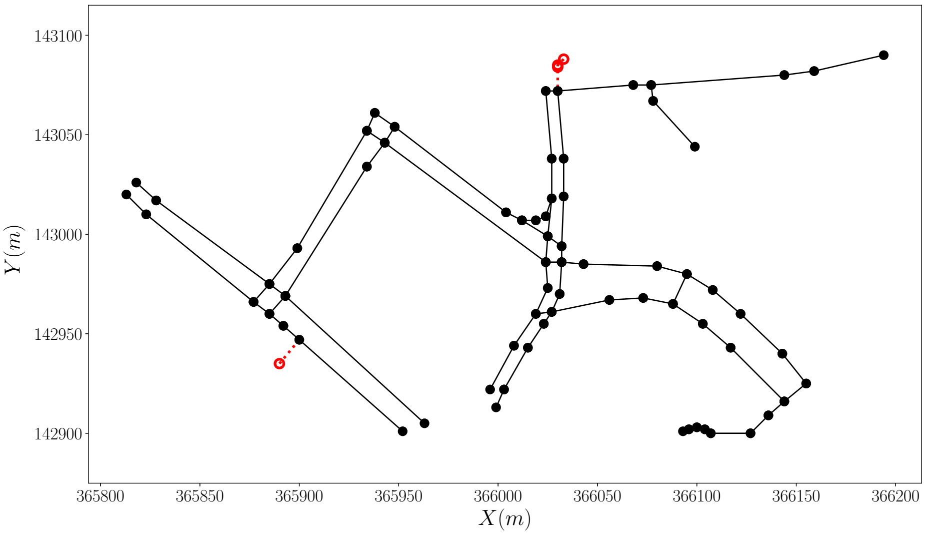


Fig. 3: Removal of short dangling paths. A dangling path starts from a degree-1 node and ends at the first non-degree-2 node encountered, upon tracing along the path. In this illustration, the dotted red paths have length shorter than 30 m and are to be trimmed off. After the operation, the nodes where those short dangling paths terminate (and get removed) will have 1 degree less.

For every projection identified, a line perpendicular to the reference line is plotted to find its intersections with all the curves. The centroid of these intersections will contribute as a point on the averaged curve, in the same order as the projections on the reference line (see Fig. 6).

3 Results and discussion

In order to apply the simplification procedure devised above, we report a case study of analysis of roads in the planning area of Punggol in the northeast region of Singapore. Punggol in the last 20 years has been developed into a new town with 11 districts, spanning a total area of 9.57 km^2 [1]. The simplified network after applying the procedure removes the extra details of roads by collapsing complex junctions into single nodes and merging multiple curves between the same pair of nodes into a thin simple path, to obtain the essential structure of the road network (see Fig. 7). This simplified network, however, still retains the overall geometrical properties of the original network such as the shape and (average) length of the paths between places.

3.1 Network attributes

In Table 1, we compare some basic attributes of the network of roads in Punggol before and after simplication. It can be seen that a large number of spatially

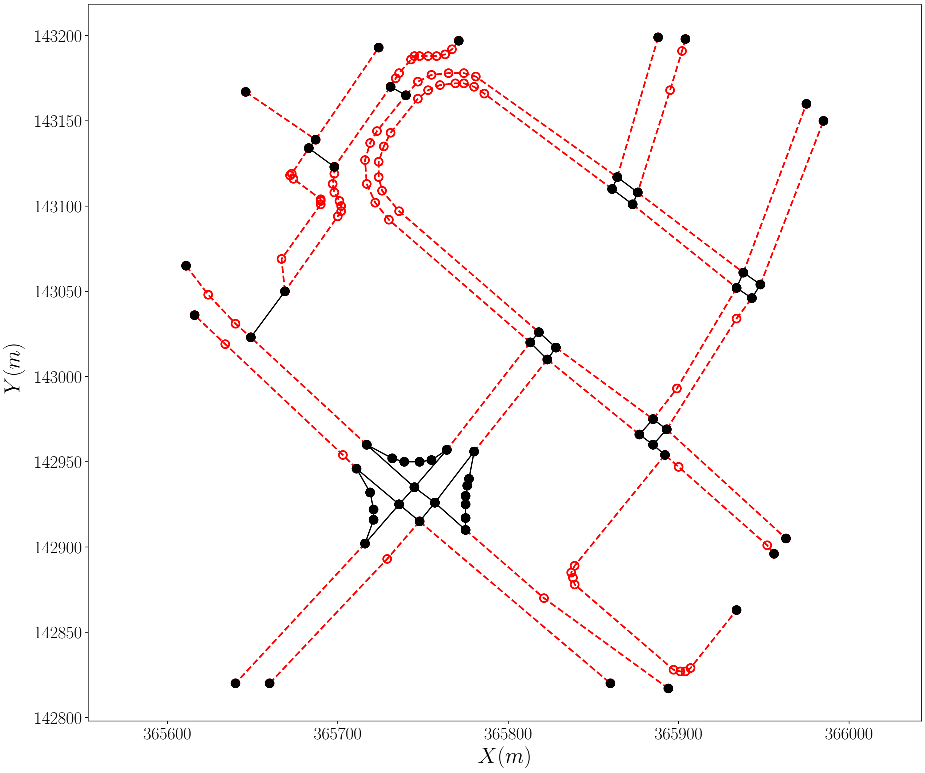
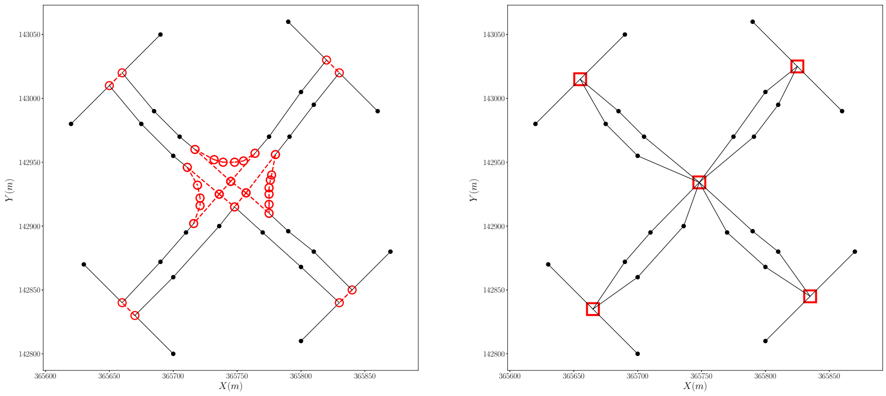


Fig. 4: Dissection of a road network into two types of elements, clusters of non-degree-2 nodes and chunks of degree-2 nodes connecting the clusters. The paths tracing the chunks are dashed in the plot with its nodes being hollow circles. The clusters are depicted with solid lines and filled circles.

redundant nodes have been removed (and/or replaced) in the simplification process, leaving only 3,265 necessary nodes compared to the original 6,168 or a gross 47% reduction. Similarity, a large number of edges have also been cleaned off in the operation, from 6,917 down to 3,401, which is more than two times slimming down. More importantly, it should be noted that the number of degree-2 nodes reduces significantly, due to merging of multiple paths between junctions.

The simplication also brings attention to high-degree nodes, of 5 or more, which indicate the emergence of proper junctions at which multiple roads from different directions meet. In the original presentation, such nodes are very rare as junctions tend to comprise a bulky set of degree-4 nodes, which arise from a pair of intersecting road segments, even when more than two roads cross each other.



(a) Original road network with clusters of nodes highlighted in red. Their nodes are plotted in hollow circle and the intra-cluster edges are dashed. Inter-cluster edges are solid.

(b) Simplified road network is obtained by replacing the clusers of nodes with their respective centroid. The newly added centroid nodes are marked with hollow square markers, while other nodes are intact and presented as black filled circles.

Fig. 5: Simplification of clusters of nodes at road junctions.

3.2 Centrality measures

With the simplified network, we can apply common network analysis techniques such as centrality measures. Four measures that are commonly used for spatial networks are degree, betweenness, closeness and eigenvectors [4, 7]. As shown in Table 1, degree centrality of a junction cannot be accurately understood in the original road network as most points at a junction have either degree 3 or 4 when two road segments intersect. In the simplified network, the degree centrality correctly reflect the number of connections that a junction has to others, which refers to the number of directions one can travel from a node.

Another benefit of having a simplified network structure compared to the original one is that spatial distribution of importance measures like betweenness

Table 1: Comparison of basic network properties before and after simplification.

Feature	Original	Simplified
Number of nodes	6,168	3,265
Number of edges	6,917	3,401
Number of nodes of degree 2	4,573 (74.14%)	2,881 (88.24%)
Number of nodes of degree 3	926 (15.01%)	137 (4.20%)
Number of nodes of degree 4	411 (6.66%)	59 (1.81%)
Number of nodes of degree 5	2 (0.03%)	25 (0.76%)
Number of nodes of degree more than 5	0 (0%)	19 (0.58%)

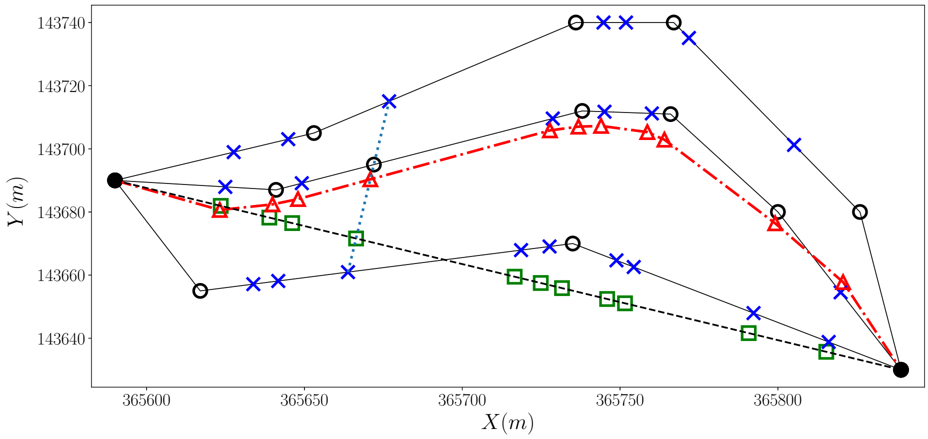
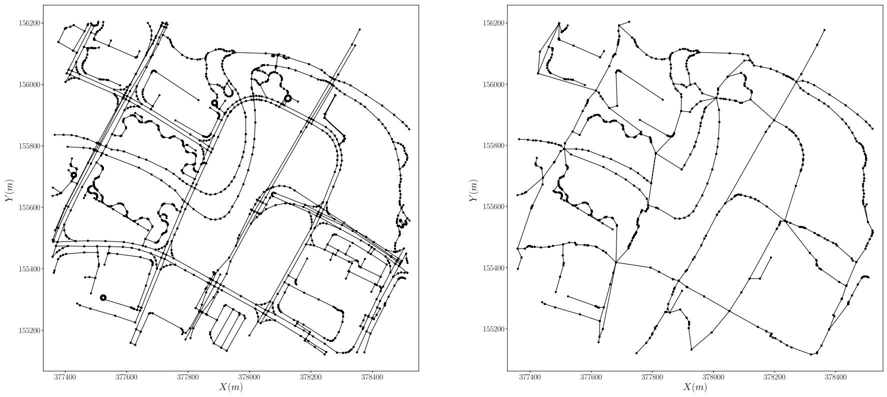


Fig. 6: Collapse of multiple curves into a single "averaged" curve. All the curves share the same end points, which are presented in two filled circle markers (•). Hollow circle markers (○) are other points in the curves, which are connected by solid black lines. The dashed line between two end points is the reference line. Every hollow circle marker has its projection on the reference line, which is shown as a hollow green square marker (□). For each of these projections, a perpendicular line is plotted to find intersections with all the curves, which are marked blue crosses (×). For each of the sets of intersections (including the curve point itself, in hollow circle marker), the centroid is found (red triangle marker (△)) and constitutes a point of the averaged curve.

centrality is more meaningful and properly represents significance of places on a map. As an illustration, the top 50 nodes in the network of roads in Punggol, Singapore, are shown in Fig. 8 for both original and simplified network. It could be seen that in the simplified network, the nodes with highest betweenness centrality measures locate at the major crossroads in the area. However, in the original network, such nodes are mostly located in clusters near the centre part of the network, which is due to redundancy of nodes at the same place and paths between places in the unsimplified network. This highlights the fact that the simplification is important before analysis of spatial networks could be meaningfully applied. This is to remove the spatial redundancy encoded in the original data.

3.3 Discussions

For simplicity of the procedure, we use a simple edge to represent each road segment in the network, with no specification of direction of the edge. In other words, we treat all roads as two-way, with flow possible in either direction, although this does not always hold true in the real world. Nevertheless, this would not affect the main objective in this study of modelling the geometric properties



(a) Actual roads

(b) Essential structure

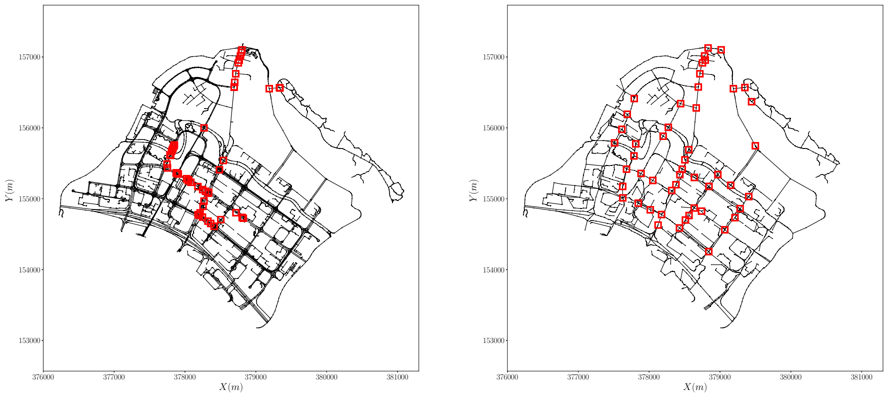
Fig. 7: Representation of network of roads in a zoom-in area in Punggol, Singapore, before and after applying the simplification.

of the network of roads, which essentially concerns the spatial pattern of roads that provide connections between places. In that context, roads are considered in more generic network sense, where junctions are taken to be representative nodes and (abstract) links are established to indicate the flow among the nodes, be it people, materials or information. This is particularly relevant in the context of urban morphology studies, which deal with the spatial organisation of an urban system [11, 12], whose structure is primarily laid out by the skeleton of the road network. Yet, for other, more detailed and conventional, purposes such as routing of vehicles, the directionality of roads remains important, which is, however, beyond the scope of this paper.

For the same reason that we're only concerned with the geometrical presentation of road networks that depicts the connections between places, it can also be argued that road network can be reasonably considered planar for the purpose of this study. In reality, a road network may not strictly be planar due to existence of flyover or tunnel. Yet, a flyover or tunnel is usually part of a system at a junction that come with the lanes leading traffic from the flyover of tunnel to the road below or above it. Hence, planarisation of such crossing (creating a point of intersection) does not introduce invalid information to the network. The new point of intersection would still be in the same junction and eventually get merged in the simplification process.

In the process of developing the simplification procedure reported in this study, we came across an approach which purely looks at the spatial location of nodes at road junctions. We find it beneficial to discuss this plausible approach as well as its limitation, to highlight the effectiveness and superiority of the reported method.

This point-based approach utilises the clustering algorithm DBSCAN (Density-Based Spatial Cluster of Applications with Noise) [9] (see also continuum perco-



(a) Original roads

(b) Simplified graph

Fig. 8: Spatial distribution of top 50 nodes (shown as hollow red square markers) with highest betweenness centrality values in the network of roads in Punggol, Singapore, before and after applying the simplification.

lation [13, 12]) in order to identify the points that belong to the same junction. This approach is based on the fact that points of a road junction are in close proximity to one another. The main concept of DBSCAN is to scan for cluster of points in different regions, taking the density of points into account. A distance parameter can be set for DBSCAN, which will determines the maximum distance between a point A and any other point B for B to be considered in the same cluster or neighbourhood as A. Using this algorithm, clusters of junction nodes in a road network can be easily identified, based on their proximity.

Despite its simplicity, this approach suffers multiple limitations that have to be overcome before any useful result can be obtained. Firstly, the main limitation of DBSCAN would be that the input parameter of the maximum distance between any 2 points to be considered in the same cluster is difficult to determine, but significantly influences the result. The problem with using DBSCAN in this context is that each intersection is unique and contains points distributed in different manners, which makes it harder or even not possible to determine one single parameter for an entire road network region. For instance, in one simple regular intersection, points are relatively close together since the roads are parallel and reasonably of close distance to one another. However, road lanes inevitably have different lengths and widths, and there is no one distance that could fit every single intersection in the real world road networks. This problem becomes even more evident when we consider road intersections that contain ramps, slip lanes, staggered junctions or other specialised types of junctions. It would then only mean that a different distance has to be determined for each intersection, which makes the problem counter-intuitive. DBSCAN is used for intersections, but the complexity of the different types of intersections makes it

DOI: 10.1007/978-3-030-50371-0_43

vital to identify the type of intersection and where they are before DBSCAN is even used.

4 Conclusions

In this paper, we present a procedure to extract the backbone structure of spatially embedded networks, such as road networks. The procedure involves simplification of multiple nodes present at the same junction due to multiple lines connecting the same pair of places in the raw data. This is achieved by dissecting the networks into clusters of nodes of degree other than 2 and chunks of nodes of degree 2 serving as paths between the clusters. Applying this simple yet effective technique, we obtain a simplifed structure of road network in the Punggol planning area in the northeast region of Singapore. We show that network analysis such as centrality measures on the simplified network can be performed in a more concise and meaningful manner than on the original network.

Acknowledgement

This work was done under the 2019 A*STAR Student Research Attachment Programme (RAP), for which RS was supported by the A*STAR Science Award (JC).

References

1. Data.gov.sg, `https://data.gov.sg/`
2. Nextzen metro extracts, `https://www.nextzen.org/metro-extracts/index.html`
3. OpenStreetMap, `https://www.openstreetmap.org`
4. Barthelemy, M., Bordin, P., Berestycki, H., Gribaudi, M.: Self-organization versus top-down planning in the evolution of a city. Scientific Reports **3**(1), 2153 (2013)
5. Barthlemy, M., Flammini, A.: Modeling urban street patterns. Physical Review Letters **100**(13), 138702 (Apr 2008)
6. Boeing, G.: OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, Environment and Urban Systems **65**, 126–139 (2017)
7. Brookes, S., Huynh, H.N.: Transport networks and towns in roman and early medieval england: An application of pagerank to archaeological questions. Journal of Archaeological Science: Reports **17**, 477–490 (2018)
8. ESRI: ESRI shapefile technical description (1998), `https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf`
9. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231. KDD-96 (1996)
10. Gudmundsson, A., Mohajeri, N.: Entropy and order in urban street networks. Scientific Reports **3**(1), 3324 (Dec 2013)

11. Huynh, H.N.: Continuum Percolation and Spatial Point Pattern in Application to Urban Morphology, pp. 411–429. Birkhäuser, Springer Nature, Cham (2019)
12. Huynh, H.N.: Spatial point pattern and urban morphology: Perspectives from entropy, complexity, and networks. Physical Review E **100**(2), 022320 (2019)
13. Huynh, H.N., Makarov, E., Legara, E.F., Monterola, C., Chew, L.Y.: Characterisation and comparison of spatial patterns in urban systems: A case study of U.S. cities. Journal of Computational Science **24**, 34–43 (Jan 2018)
14. Kim, Y., Fukuyasu, H., Yamamoto, D., Takahashi, N.: A road generalization method using layered stroke networks. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising - LocalRec '19. pp. 1–10. ACM Press, Chicago, Illinois (2019)
15. Marshall, S.: Streets and Patterns. Routledge (2004)
16. Rosvall, M., Trusina, A., Minnhagen, P., Sneppen, K.: Networks and cities: An information perspective. Physical Review Letters **94**(2), 028701 (Jan 2005)
17. Tian, J., Xiong, F., Lei, Y., Zhan, Y.: Revising Self-Best-Fit Strategy for Stroke Generating. In: Harvey, F., Leung, Y. (eds.) Advances in Spatial Data Handling and Analysis, pp. 183–192. Springer International Publishing, Cham (2015)
18. Xie, F., Levinson, D.: Measuring the structure of road networks. Geographical Analysis **39**(3), 336–356 (Jul 2007)
19. Yamamoto, D., Murase, M., Takahashi, N.: On-demand generalization of road networks based on facility search results. IEICE Transactions on Information and Systems **E102.D**(1), 93–103 (2019)
20. Yu, W., Zhang, Y., Ai, T., Guan, Q., Chen, Z., Li, H.: Road network generalization considering traffic flow patterns. International Journal of Geographical Information Science **34**(1), 119–149 (2020)
21. Zhang, Q.: Road network generalization based on connection analysis. In: Developments in Spatial Data Handling, pp. 343–353. Springer-Verlag, Berlin/Heidelberg (2005)
22. Zhang, W., Wang, S., Tian, X., Yu, D., Yang, Z.: The backbone of urban street networks: Degree distribution and connectivity characteristics. Advances in Mechanical Engineering **9**, 168781401774257 (2017)
23. Zhou, C., Li, W., Jia, H.: Rroad network generalization based on float car tracking. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLI-B4**, 71–77 (2016)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_43