Role discovery in node-attributed public transportation networks: The study of St Petersburg city open data

Yuri Lytkin
ITMO University
Petr Chunaev (chunaev@itmo.ru)
ITMO University
Timofey Gradov
ITMO University
Anton Boytsov
ITMO University
Irek Saitov
ITMO University

Research Article

Keywords: Node-attributed network, public transportation network, role discovery, network node classification, network topology, social infrastructure

Posted Date: April 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1532891/v1

Role discovery in node-attributed public transportation networks: The study of St Petersburg city open data

Yuri Lytkin[1], Petr Chunaev[1*], Timofey Gradov[1], Anton Boytsov[1] and Irek Saitov[1]

[1]National Center for Cognitive Research, ITMO University, 16 Birzhevaya Lane, St Petersburg, 199034, Russia.

*Corresponding author(s). E-mail(s): chunaev@itmo.ru;
Contributing authors: yuvlytkin@itmo.ru; timagradov@yahoo.com; aboytsov@itmo.ru; iasaitov@itmo.ru;

Abstract

In this paper, we propose a framework for solving the novel problem of role discovery in a public transportation network (PTN). We model a PTN as a weighted node-attributed network whose nodes are public transport stations (stops) grouped with respect to their geospatial position, node attributes store information about social infrastructure around the stations (stops), and weighted links integrate information about the travelling distance and the number of hops in the transportation routs between the stations (stops). Our framework discovers meaningful node roles in terms of both topological and infrastructural features of a PTN and is capable of extracting useful insights about the overall PTN's efficiency. We apply the framework to the newly collected open data of St Petersburg, Russia, and point out some transportation and infrastructural weaknesses that should be taken into consideration by the city administration to improve the PTN in the future.

Keywords: Node-attributed network, public transportation network, role discovery, network node classification, network topology, social infrastructure

MSC Classification: 62H30 , 05C90 , 68R10 , 93A30

1 Introduction

In recent years network theory has found its way into a variety of fields of science and technology. A network is a collection of *nodes* some of which are connected together by *links*. Being so simply constructed and versatile simultaneously, networks become very useful in analyzing, modelling, and studying all sorts of complex systems such as online and offline social networks, computer and technological networks, biological and brain networks, transportation networks, etc.

The study of public transportation systems from a network theory perspective started rather recently [1, 2]. Most works on this topic are aimed at analyzing the topological structure of *public transportation networks* (or *PTNs*) of different cities (e.g. in Poland [2], Hungary [3], China [4]) with regard to various modes of transportation like bus [2] or subway [5]. Usually in these cases the underlying network is defined with bus stops or subway stations as nodes and some rule to assign links between these stops and stations. The links are mainly unweighted although there are studies considering PTNs as *weighted*, too, see e.g.

[6] (and references therein), where weighted bus PTNs are analyzed by means of common network characteristics.

In addition to the PTN topology, it is also usual to consider the geospacial information about the nodes in the network. A popular approach that utilizes the geography of nodes is combining sets of closely situated nodes into groups called *supernodes* [4, 7]. Such approach is motivated by the fact that people usually take walks between closely positioned stops to make a connection, instead of sticking to a strict path through the network. Therefore, such supernode networks are more precise at modelling how people use public transport. From another perspective, some studies (see e.g. [8]) consider PTNs as geospatial ones so that the spatial configuration and topology of the network are used for the identification of macroscopic and mesoscopic statistical network characteristics.

Furthermore, another notable source of information that can be used in the public transportation system analysis is social infrastructure surrounding stations and stops that, in a sense, may provide "semantics" to a PTN. For instance, it can be used to analyze and model transport accessibility [9] or as an additional component for measuring PTN transportation efficiency [7].

As far as we know, the union of weighted geospartial networks (supernodes and weighted links) and node semantics (social infrastructure in our case) have not been considered in the PTN studies (as the so-called *node-attributed networks*), although it may certainly enrich our knowledge about processes of PTN formation. This is confirmed by the case of node-attributed networks modelling online social networks where not only connections between social actors (network topology) but also actors' content (profile information, posts, etc.) are taken into account within different tasks such as community detection, link prediction, outlier identification, etc., see e.g. [10–12]).

To get closer to the objective of our study, let us also mention that in the recent times *role discovery* (especially topological feature-based [13]) has become a popular topic, most notably in the domain of *non-attributed social* network analysis [13–19]. In the network context, roles refer to clusters, or classes, of nodes, where the nodes from the same cluster are structurally similar to each other in some way. The problem of role discovery is related to another network clustering problem called *community detection* in non-attributed [20–22] and node-attributed [10–12] social networks, where the clustering mainly aims to separate densely interconnected parts (called communities) of the network by means of network topology or both network topology and attributes (semantics). By contrast, role discovery aims to distinguish between various structural and other characteristics of different nodes. For instance, in a social network there can be multiple communities of people, and in each community there are people of various roles, i.e. leaders, influencers, etc., with possible transitions between roles and interaction preferences (see the recent studies on the topic e.g. in [23–25]). Let us specifically mention the study [26] here as it seems the first attempt to enrich role discovery methodology in social online networks by the content generated by social actors ("semantics"). Although the authors do not explicitly model online social networks as *node-attributed* networks, the experimental results in [26] show that the semantics helps to identify social network roles more effectively.

In this study we take into account the experience of studies in social network analysis connected with role discovery in non- and node-attributed social networks to model and analyse PTNs. Furthermore, we are motivated by the survey [27] where PTNs are considered from the network perspective of complexity, static and dynamic resilience, and it is emphasized that the study of PTN node roles (in particular, based on topological features — besides the well-known *hubs*, for example) is still limited although may offer useful insights into identifying the most critical nodes of PTNs.

Thus we propose a role discovery framework for weighted node-attributed PTNs that is able to discover roles both in terms of network topology (i.e. transition hubs, outskirts, etc.) and node infrastructural attributes — semantics (i.e. tourist, residential, industrial areas, etc.). In short, the main contributions of this paper are the following:

1. We model a PTN as a weighted node-attributed network where nodes are *supernodes*, i.e. groups of public transport stops and stations grouped with respect to their geospatial position, and node attributes are numerical vectors storing information about social infrastructure around

the supernodes. The weighted links in the network integrate information about the travelling distance and the number of hops in the transportation routs between the supernodes.
2. We propose a new framework for role discovery in weighted node-attributed networks. This framework uses semantics (i.e. node attributes) as well as structure (i.e. network topology). In the context of PTNs, this framework allows to discover meaningful roles in terms of both topological structure of stops and stations and social infrastructure around them. At the same time, the framework is not topic-specific and can be applied in other domains like social network analysis.
3. We test the framework on the newly collected open public transportation data of St Petersburg, Russia. It is shown to be capable of discovering different roles of public transport stops in terms of both structure and social infrastructure and extracting useful information about the overall PTN's transport and social infrastructure efficiency.

Let us additionally mention that with respect to previous studies, we

- define the supernodes formally as equivalent classes to avoid ambiguity, with the choice of reasonable thresholds;
- choose a trade-off between hop-based and distance-based routs to balance between the travelling distance and the number of hops corresponding to a given rout between two nodes in a PTN;
- define the problem of social infrastructure role discovery and propose a procedure for constructing social infrastructure attributes in our model;
- scrupulously select and analyse commonly-used topological features of network nodes in the context of PTN models;
- analyse correlations between topological and infrastructural node features.

We also point out several common misconceptions and errors in some previous studies of PTNs which we believe stem from the misunderstanding of some interpretations of different PTN models.

The paper is organized as follows. In Section 2 we survey the previous research and methods on PTN analysis (Section 2.1) and role discovery (Section 2.2). In Section 3 we describe the construction of our model and some issues that arise in the process. We present the results of our experiments and analysis in Section 4, in particular, the data overview is given in Section 4.1, the process of building the supernode network is described in Section 4.2, the different kinds of node features are discussed and constructed in Section 4.3 and the final results are presented in Section 4.4. Finally, we conclude our paper and express our thoughts on possible future research in Section 5.

2 Related work

2.1 Modelling public transportation networks

The study of *public transportation networks* (or *PTNs*) using network (graph[1]) theory began in [1, 2]. The main aim of such studies is usually to analyze the topology of the given city's PTN in order to extract useful information about the state and structure of that city's public transportation system.

The two most popular ways of constructing a PTN (both were introduced in [1]) are *L-space* and *P-space* models. In both cases the nodes of the network represent various public transportation stops and stations. What these models differ in is the way of assigning the links between the nodes. According to the *L*-space model, a link is assigned between two nodes that correspond to two consecutive stops on some route. Thus, the topology of an *L*-space model is visually similar to a normal scheme of a public transportation system that one can find on an information stand near a bus stop. By contrast, in the *P*-space model a link is put between all stops that are connected by some route (not just the consecutive ones). Therefore, in the *P*-space model and link is interpreted as a possibility of travelling directly between two nodes. (Note that as a result, the *P*-space model is normally much more dense, than the corresponding *L*-space model.) The difference between *L*-space and *P*-space is explained in Figure 2.1.

These models have been used in virtually all the papers dealing with PTNs and were applied to analyze various cities in Poland [2], Hungary [3],

[1]Here and throughout the paper we use the terms *network* and *graph* interchangeably.

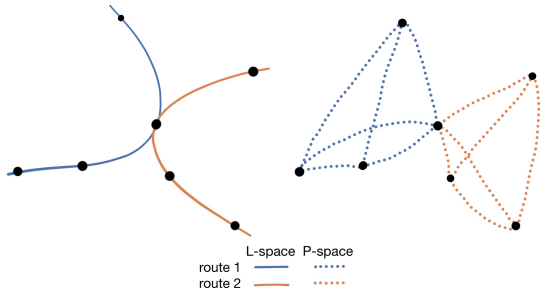

Fig. 1 Difference between L-space (left) and P-space (right).

China [4], among others. Such analysis is especially easy to conduct since the data needed to build a basic PTN is nowadays available publicly for most big cities around the world (see Figure 2). Usually authors aim to check some graph-theoretic and network-theoretic properties of the constructed graphs, i.e. degree distribution, clustering coefficient, scale-free property and so on. A comprehensive comparison of such properties between different cities around the world can be found in [28] along with interpretations of these properties in a sense of public transportation quality.

Another natural source of information for constructing a PTN is the geospacial component, i.e. the coordinates of the stops. As we mentioned previously, a conventional PTN (with separate stops as nodes) does not account for passengers' possibility to make walking connections between closely situated stops while moving around a city. Additionally, such approaches are not capable of combining different modes of transportation (like bus, trolleybus, tramway, and subway) in a single network. To overcome these issues, one can consider groups of nearby stops and stations as *supernodes* (see Figure 3), thus transforming the conventional node structure into the supernode structure (note that the node links are naturally transformed into the supernode links, given the defined node to supernode mapping). Such approach was used in [4, 7].

To further improve a public transportation model, one can also assign link weights, see e.g. [4, 6, 7]. In [7] the authors propose to assign weights to the links of the L-space network by counting the number of routes operating of each given link. Such weights can therefore represent the amount of passenger flow via each link. By contrast, the authors of [4] propose to assign link weights (both in L-space and P-space) as the minimal travel distance between the nodes along the corresponding route. Such approach is more suitable in terms of determining the optimal routes and connections while travelling around a city.

It should be noted that the choice of the network model as well as the method of assigning link weights greatly influences what one can then do with the resulting network model. For instance, when using the L-space model (as it was done in [7]), one should be careful in interpreting the shortest paths through the network, as these generally do not correspond to how passengers choose to travel in practice, since, for example, the number of connections is not minimized when using such paths, while normally a passenger would want to make as little connections as possible (see Figure 4). Such misinterpretation of shortest paths may lead to subsequent misinterpretation of various centrality measures, such as betweenness centrality and closeness centrality.

The P-space seems to be better suited for such shortest path analysis, although choosing the method of link weight assignment is still very important here. Assigning equal weights to links resolves the issue of minimizing the number of connections, since in this case a shortest path through the P-space network is precisely the path requiring the minimal number of connections. At the same time, such shortest paths can be excessively long in terms of travelling distance. However, setting travelling distances as link weights (as in [4]) brings back the issue of the number of connections, since a shortest path in terms of travelling distance can involve a suboptimal number of connections. Therefore, an intermediate approach is needed, taking into account both the number of hops in a shortest path, and the travelling distance corresponding to it. Such approach is used in our paper (see Figure 5).

There also exist methods of assigning link weights based on the flow of passengers during a certain part of the day (see [29, 30]), resulting in a dynamic structure of the PTN. It should be mentioned, however, that such data is usually quite hard to obtain, while in this paper we aim to construct the model using only the openly available data.

Finally, social infrastructure is also an available and important source of information when

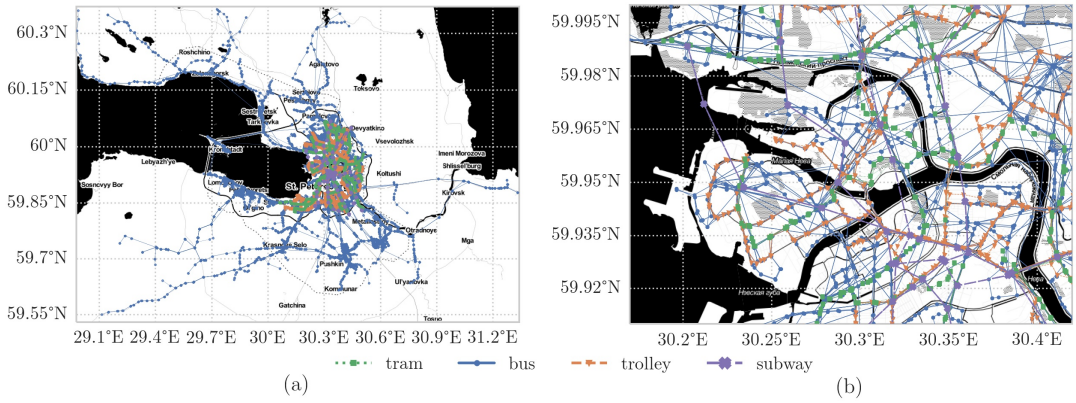

Fig. 2 The map of area surrounding St Petersburg (a) and the city center (b), indicating stops and routes of different modes of transportation.

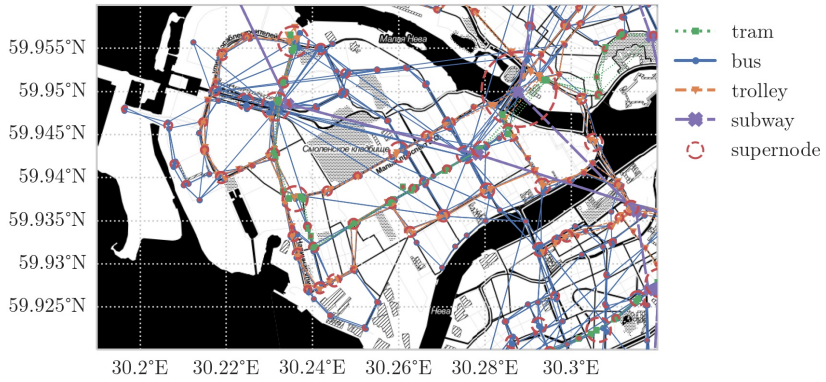

Fig. 3 The map of Vasileostrovsky District in St Petersburg, indicating stops and routes for different modes of transportation, as well as the supernodes (groups of nearby stops).

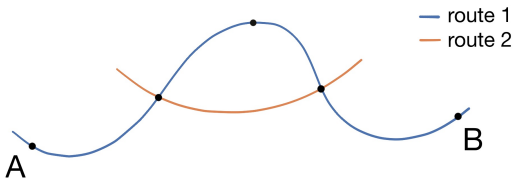

Fig. 4 In the *L*-space model, all consequent stops in each route are connected with a link. As a result, a shortest path in the *L*-space graph generally does not indicate an optimal route for a passenger. For instance, while travelling from point A to point B, the optimal travelling route is route 1 (blue), while the shortest path through the graph involves changing to route 2 midway.

constructing a PTN since it sheds light on why people actually travel to a given destination (there can be, for instance, a school, a hospital, or a sightseeing spot nearby). The infrastructural component was used in [7], where the authors assigned node weights depending on a number of factors such as the number of social infrastructure objects of certain types (recreation, emergency, education, and transportation), the total number of passengers accessing the node, etc. All these factors were then weighted, producing a single value with was chosen as the node weight.

This method is useful when trying to access importance (as a unidimensional characteristic) of each node from the infrastructural standpoint. At the same time it does not capture any information about the role of the node, i.e. its unique infrastructural characteristics. Therefore, in this paper we adopt a more general multidimensional approach, assigning not weights but attribute vectors to nodes.

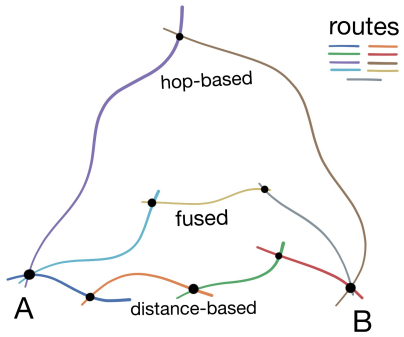

Fig. 5 In the *P*-space model, both hop-based and distance-based link weights result in shortest paths that are not indicative of optimal routes for passengers. When using hop-based link weights (i.e. each link having weight 1), a shortest path is the one with the least number of connections, but it can be arbitrarily long in distance. The contrary holds for distance-based weights: a shortest path is indeed shortest in distance, but can involve arbitrarily many connections in the process. A fused approach (considering both distance and hops) mitigates such problems.

2.2 Role discovery in public transportation and other networks

The main idea behind the role discovery is to group nodes by their connectivity patterns, where each group represents some topological role such as hub, bridge, near-clique, etc. Topological roles indicate which functions nodes serve in the network [13].

Initially role discovery was the point of interest in sociology, used to study the interactions between social actors and assign roles to actors, but networks in these studies were very small [31, 32]. In general, role discovery can be applied to any network, and the main difference across networks will be in the interpretation of roles. Lately, this concept was studied and implemented for biological networks [33], web graphs [34], and many others [35].

The process of role discovery usually consists of several steps. Firstly, centrality measures (or other chosen features) are chosen and calculated for every node in the network. Following this, nodes are clustered by using vectors of centrality measures. As a result, nodes are grouped by similarity among centrality measures, which shows how similar nodes are in terms of topology.

To the best of our knowledge, no purposeful attempts have been made to state and solve the problem of role discovery in the above-mentioned sense for PTNs. Indeed, the survey [27] (where PTNs are considered from the network perspective of complexity, static and dynamic resilience) emphasizes that the study of PTN node roles (in particular, based on topological features — besides the well-known *hubs*, for example) is still limited although may offer useful insights into identifying the most critical nodes of PTNs.

Nevertheless, we can mention e.g. the study [8], where the geospatial configuration of a PTN is analysed and some conclusions about the roles of the PTN nodes (by means of importance) are made. Furthermore, the topic-related work [36] aims to detect and analyse node clusters in the intercity transportation networks. The authors propose using a distance measure based on the K shortest paths between a pair of nodes to measure the proximity between all node pairs, and then use the hierarchical clustering method in order to obtain the clusters. The resulting clusters correspond to the groups of nodes that are in close proximity of each other. However, this work is more in line with the problem of community detection than role discovery, since these clusters do not reflect different roles of these nodes in the network.

Another notable attempt at geospatial PTN clustering is the work [37], the authors of which introduce a problem of node-attributed spatial graph partitioning. This problem aims at obtaining clusters of nodes that are densely interconnected, homogeneous with respect to their attributes and also meet a certain size constraint in terms of the geographical coordinates of the nodes. Even though this problem can indeed be formulated in terms of PTNs and also accommodate the presence of node-attributed social infrastructure vectors, it is however more in line with community detection in node-attributed networks [10–12] rather than role discovery [13], since in general the nodes of a certain role (like transition hubs, for instance) do not need to be in close proximity of each other.

One should note that the richest experience on the role discovery task is nevertheless in the field of social network analysis, where non- and node-attributed networks are deeply studied within the task [13–19]. One can find a comprehensive overview of role discovery approaches in [13], where graph-based, feature-based, and hybrid definitions of roles and methods for their discovery

from social network data are discussed. Let us also mention several further studies on the topic.

In [16], a novel role discovery approach is proposed for extracting *soft* roles of social actors with similar behavioural and functional characteristics in online social networks. The study [24] is focused on the problem of research role identification (i.e., principal investigator, sub-investigator or research staff) for large research institutes in which similar yet separated teams coexist. Furthermore, [25] states and proposes a framework for solving the multiple-role discovery task and conduct an experimental study of their framework on several real-world online document/social networks. Finally, let us mention the study [26] that seems the first attempt to enrich role discovery methodology in social online networks by the content generated by social actors e.g. posts. In the paper, a novel method which integrates both user behavior and his/her content to identify roles is proposed. Although the authors do not explicitly model online social networks as *node-attributed* networks, the experimental results in [26] show that the semantics helps to identify various roles more effectively and to get more insights on how the network is functioning.

As we have already mentioned, in our study we take into account the experience of studies in social network analysis connected with role discovery in non- and node-attributed social networks to model and analyse PTNs.

3 Description of the model and the role discovery task

3.1 The model of a node-attributed public transportation network

We now proceed to describing the node-attributed PTN model that we are going to use for role discovery later. The data needed to construct such model will be described in detail in Section 4.1, but for now we note that only the general public transportation and social infrastructure data, which is available for the majority of cities around the world, is needed here. Below, we illustrate our model with the PTN data for St Petersburg, Russia (see Section 4.1) in order the make it clearer for the reader.

Formally, the model can be defined as a tuple

$$G = (V, E, A),$$

where V is the set of nodes, $E \subseteq V \times V \times \mathbb{R}$ is the set of undirected weighted links, and $A : V \to \mathbb{R}^n$ is a mapping that defines the set of node-attributed vectors. In what follows we will define each component of this graph.

3.1.1 Supernodes (nodes of the node-attributed network)

The first step is combining the public transportation stops and stations into supernodes, i.e. groups of nodes that are located close to each other, thus making it possible to make a transition between them on foot. Suppose that $S = \{s_1, \dots, s_N\}$ is the set of public transportation stops (N in total). To combine them into supernodes, we first need to calculate the pairwise distances between each pair $s_i, s_j \in S$. This can be done using their geographical coordinates. The distances are calculated using the well-known Haversine formula:

$$d(s_i, s_j) = 2r_0 \arcsin\sqrt{\Theta(\varphi, \lambda)}, \tag{1}$$

$$\Theta(\varphi, \lambda) = \sin^2\frac{\varphi_j - \varphi_i}{2} + \cos\varphi_i \cos\varphi_j \sin^2\frac{\lambda_j - \lambda_i}{2},$$

where $d(s_i, s_j)$ is the distance between stops s_i and s_j, r_0 is the radius of Earth, φ_l, λ_l, $l \in \{i, j\}$, are latitudes and longitudes of the two points, respectively.

The most common way of grouping the closely situated stops is by using a *distance threshold* [4, 7]: all stops that are closer to each other than some constant d_0 are added to a common supernode. Since this construction is not an equivalence relation, in order to define the supernodes correctly, we also close this relation *transitively*. When this is done, the supernodes are defined as *equivalence classes* with respect to this closed relation, i.e. two stops $s_i, s_j \in S$ belong to the same supernode $\hat{s}$ *if and only if*

$$\begin{aligned} &\exists n_1 = s_i, n_2, \dots, n_K = s_j \in S : \\ &\quad \forall k < K \; d(n_k, n_{k+1}) \le d_0. \end{aligned} \tag{2}$$

We denote the set of all supernodes as $\hat{S}$ and use it as the set of nodes V of the graph G. In some practical cases we will also need coordinates of

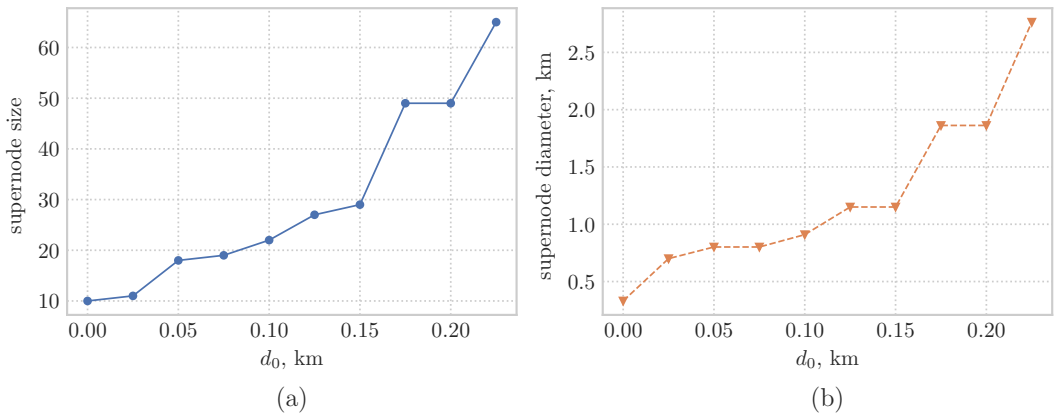

Fig. 6 Maximal supernode size (a) and diameter (b) for different values of d_0. Even for relatively small values ($d_0 > 0.15$) these characteristics grow quite rapidly, resulting in some supernodes having diameter as large as 2 km and more.

supernodes. For these cases we define coordinates of a supernode as simply the mean of latitude and longitude over all stops belonging to the given supernode.

Note that in general there can be nodes inside a single supernode with distance greater than d_0, provided there is a sequence of nodes

$$n_1 = s_i, n_2, \ldots, n_K = s_j \in S,$$

such that each pair n_k, n_{k+1} is closer than d_0. This can potentially result in some supernodes being arbitrarily large. This issue cannot be resolved in a symmetrical way, and we have no choice but to allow it (even though it has not been discussed in any of the previous papers, we assume that the authors of those papers also faced this issue), but we stress that an appropriate value of d_0 should therefore be chosen carefully, taking into account the sizes of the resulting supernodes (see Figure 6). Some of the characteristics of supernodes that can be considered here is the supernode size (i.e. the number of nodes inside it) or the supernode diameter (i.e. the maximal distance between two nodes inside it).

For instance, in Figure 6 we see that when $d_0 > 0.1$ (i.e. the distance of 100 meters), the maximal supernode diameter gets beyond 1 km, which is not really acceptable as a walking distance between the stops. Therefore, for our study we take $d_0 = 0.1$.

3.1.2 Weighted links of the node-attributed network

The second step is defining the set of links E. This is done traditionally using the information about different routes that comprise the public transportation system. Suppose that R is the set of all public transportation routes, where each route is defined as a sequence of stops from S:

$$r = (s_{i_1}, \ldots, s_{i_k}). \tag{3}$$

Here k is the route length, and each s_{i_j} is a stop from S. Since each stop $s \in S$ is mapped uniquely to a supernode $\hat{s} \in \hat{S}$, these routes can be easily converted into the sequences of supernodes:

$$\hat{r} = (\hat{s}_{i_1}, \ldots, \hat{s}_{i_l}), \tag{4}$$

where $l \leq k$ and $\hat{s}_{i_j} \in \hat{S}$.

Recall that in the P-space model, links are defined as all pairs of stops (not necessarily consecutive) on all the routes, i.e.

$$\left\{(s_i, s_j) \in S^2 \mid \exists r \in R : s_i, s_j \in r\right\}.$$

A P-space link, therefore, means that there exists a route connecting the given pair of stops.

In order to assign weights to these links, consider an arbitrary route $r = (s_{i_1}, \ldots, s_{i_k})$ and take two arbitrary stops $s_{i_j}, s_{i_l} \in r$, $i_j < i_l$. Since there exists a sub-route $(s_{i_j}, s_{i_{j+1}} \ldots, s_{i_l}) \subseteq r$, we can define a *route distance* between s_{i_j} and s_{i_l} with

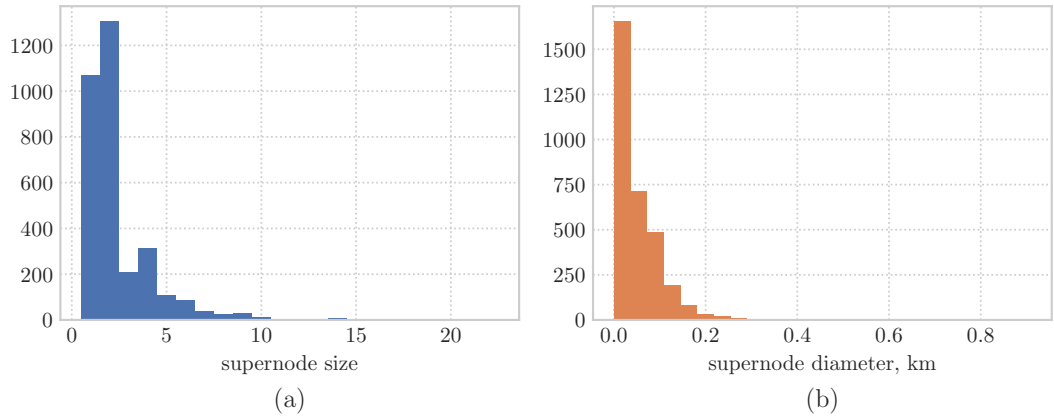

Fig. 7 Histograms of supernode sizes and diameters for $d_0 = 0.1$. Most supernode diameters are within 0.2 km, but some are as large as 0.9 km.

respect to the route r as follows:

$$rd_r(s_{i_j}, s_{i_l}) = \sum_{k=j}^{l-1} d(s_{i_k}, s_{i_{k+1}}), \tag{5}$$

where d is the distance defined in Eq. 1. Notice that there can be several routes connecting the same pair of stops s_i, s_j, and the corresponding route distances $rd_r(s_i, s_j)$ can vary. We thus define the route distance between two nodes s_i, s_j as the minimal route distance between them across all the available routes:

$$rd(s_i, s_j) = \min_{r \in R} rd_r(s_i, s_j). \tag{6}$$

Route distances were used as link weights in [4], but, as it was discussed in Section 2, such approach to assigning link weights brings up an issue in that a shortest path between two nodes with respect to route distances, while being optimal in terms of travel distance, can be suboptimal in terms of the number of connections made while travelling via this path. Using unweighted links solves the problem of minimizing the number of connections, but can result in shortest paths that are inadequate in terms of travelling distance.

This issue is illustrated in Figure 8. In both cases we have two routes between the same pair of stops, and route A is obtained by minimising the travel distance, while route B is obtained by minimising the number of hops. In the first case we see that route B, while having less transfers than route A, is about 10 times longer than the latter, therefore it is much less convenient for a passenger. The second case is the opposite: route A is shorter (albeit marginally) than route B, but has 10 times more transfers, and it is very unlikely that a passenger will decide to take route A over route B.

Therefore, an intermediate approach should be adopted. Here we propose the following weighing scheme:

$$w(s_i, s_j) = \alpha \cdot rd(s_i, s_j) + 1 - \alpha. \tag{7}$$

(The term $1 - \alpha$ can be thought of as multiplied by a 'hop-weight' of a link, which is always equal to 1.) This approach makes it possible to balance between the travelling distance and the number of hops corresponding to a given path between two nodes. We use these values as link weights in our model:

$$E = \left\{ (\hat{s}_1, \hat{s}_2, w(\hat{s}_1, \hat{s}_2)) \mid \hat{s}_1, \hat{s}_2 \in \hat{S} \right\}. \tag{8}$$

In order to choose an appropriate value of α, consider the two borderline cases, namely $\alpha = 0$ and $\alpha = 1$. In the first case we get an unweighted graph (each link having weight 1), thus the shortest paths have the minimal possible number of hops. For an arbitrary pair of nodes $s_i, s_j \in S$ denote such minimal number of hops as $H_{min}(s_i, s_j)$. In the latter case (i.e. $\alpha = 1$) we get a graph weighted with geographical distances along the links, thus the shortest paths

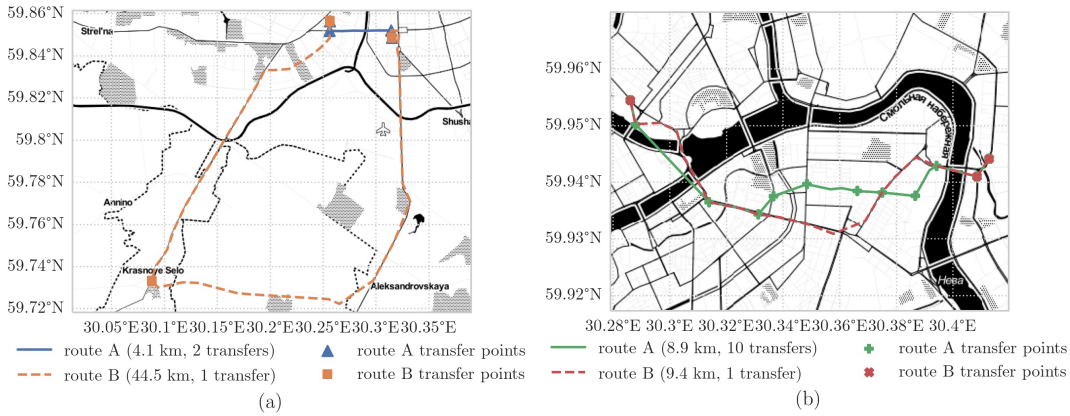

Fig. 8 Minimising the number of hops can lead to excessively long routes (a), while minimising the travel distance can lead to routes requiring an excessive number of transfers (b).

in this case are minimal in terms of travel distance. Denote these minimal travel distances as $D_{min}(s_i, s_j)$, $s_i, s_j \in S$.

Now, for an arbitrary $\alpha \in (0, 1)$ notice the shortest paths are sub-optimal in terms of both the number of hops (denote these as $H_\alpha(s_i, s_j)$) and travel distance (denote these as $H_\alpha(s_i, s_j)$). Therefore, we can consider *mean percentage difference* between these values and their corresponding minima, i.e.

$$MPD_H(\alpha) = \frac{100\%}{\mathsf{V}(\mathsf{V}-1)} \sum_{u,v \in V} \frac{H_\alpha(u,v) - H_{min}(u,v)}{H_{min}(u,v)} \quad (9)$$

for hops, and

$$MPD_D(\alpha) = \frac{100\%}{\mathsf{V}(\mathsf{V}-1)} \sum_{u,v \in V} \frac{D_\alpha(u,v) - D_{min}(u,v)}{D_{min}(u,v)} \quad (10)$$

for distances.

These values can be used to determine the optimal value of α. For instance, in Figure 9 we see that for $\alpha = 0.2$ both MPD_H and MPD_D are less than 10%, which means that on average both the number of hops and travel distance are no more than 10% greater than their corresponding minima.

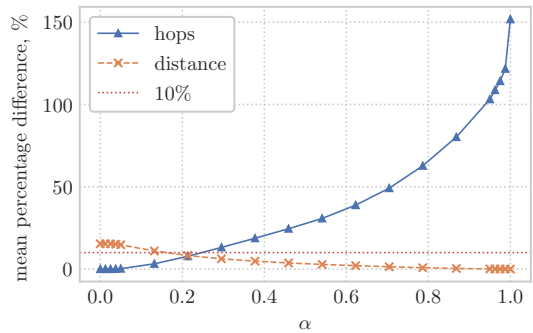

Fig. 9 Mean percentage difference of hops (see Eq. 9) and distance (see Eq. 10) for different values of α. When $\alpha \approx 0.2$, MPD is less than 10% for both hops and distance.

3.1.3 Attribute vectors of the node-attributed network

Finally, we want to assign each node $\hat{s} \in \hat{S}$ a vector $A(\hat{s}) \in \mathbb{R}^n$ describing it in terms of social infrastructure surrounding it. This can be done using the information about various infrastructural objects $I = \{i_1, \ldots, i_m\}$ around the city. Each object i_j is a tuple (φ, λ, t), where φ, λ are latitude and longitude of the object, and t is a categorical marker of the type of this object (i.e. be it a shop, a hospital, a sightseeing place, etc.). The set of different infrastructural object types $T = \{t_1, \ldots, t_n\}$ is usually pre-defined.

To construct node attributes, we first assign each infrastructural object to some stop. The most natural way of doing this is by assigning each infrastructural object to a stop that is closest to it. We note however that such approach is not

the most accurate, since there are generally multiple ways of getting to a given destination (for instance, one can take multiple routes to work or school), and these can involve getting off a bus at different stops. To account for this, we propose using a *distance window* d_1 when assigning infrastructural objects to stops. To do so, take an infrastructural object i and suppose that d_{min} is the minimal distance from i to a stop. We then assign the object i to all stops s such that

$$d(i,s) \leq d_{min} + d_1,$$

where $d(a,b)$ is the distance between geographical points (see Eq. 1). In this study we take $d_1 = 0.2$, i.e. the distance of 200 meters (see Figure 9).

Denote $I_s \subseteq I$ as the set of all infrastructural objects assigned to a stop s. When this is done, we construct a vector v_s corresponding to the given stop s by counting the infrastructural objects of different types assigned to this stop, i.e. $v_s \in \mathbb{N}^n$ and

$$(v_s)_j = \#\left\{i \in I_s \mid i = (\varphi, \lambda, t), t = t_j\right\}. \quad (11)$$

These vectors are used as note attributes in our network model, i.e. $A : \hat{s} \mapsto v_{\hat{s}}$. Such attributes reflect the characteristics of each node in terms of what kind of social infrastructure this node is surrounded by (see Figure 10). The definition of our public transportation model is thus complete.

3.2 Role discovery task for the node-attributed public transportation network

The task of role discovery originated in the field of social network analysis, but found its way into a variety of different domains of science (see Section 2.2). This task usually involves clustering of network nodes, but not in a sense of connectivity structure (the so-called *community detection*), but rather in terms of topological features of nodes (for instance, various centrality measures, more on that below). Thus, the goal is to obtain clusters not of densely connected nodes, but rather of nodes having similar structural characteristics.

The basic approach to this task is therefore to extract some features of the network nodes and then use machine learning algorithms (i.e. KMeans [38]) to extract clusters based on these features. Even though originally only topological features were used in this approach, the basic framework can naturally be extended to include also node-attributed vectors (that too can be used as a separate set of node features). One can then combine these two sets of features in some way and perform clustering simultaneously, or alternatively obtain two separate clustering (with respect to topological features and node attributes) and then analyse their relationship, for instance, using a contingency table.

In this study we adopt the latter approach, i.e. we perform separate clustering with respect to topological features (derived from the network structure) and infrastructure features (using the supernode attributes, see Section 3.1) and then compare the two. The reason for this is that these two feature sets have their own interpretations, thus interpreting clusters with respect to only one of the feature sets is much more intuitive than if one uses, for instance, concatenated features.

4 Experimental study: role discovery in the public transportation network of St Petersburg

We apply our model for the task of role discovery on the PTN of St Petersburg, Russia. The task is to analyze and separate in an *unsupervised* manner various public transportation stops into homogeneous clusters, i.e. groups of nodes with similar characteristics, including infrastructural attributes and topological features. This task is useful in the analysis of public transportation systems since these clusters can then be further analyzed in terms of their role within the public transportation system and reachability between them.

4.1 Dataset description

We first describe the data[2] needed to build the PTN model proposed above. As we mentioned before, we only use the general public transportation and infrastructural data, that is available for

[2]The data along with all preprocessing and analysis procedures is available in the Github repository at https://github.com/AlgoMathITMO/public-transport-network.

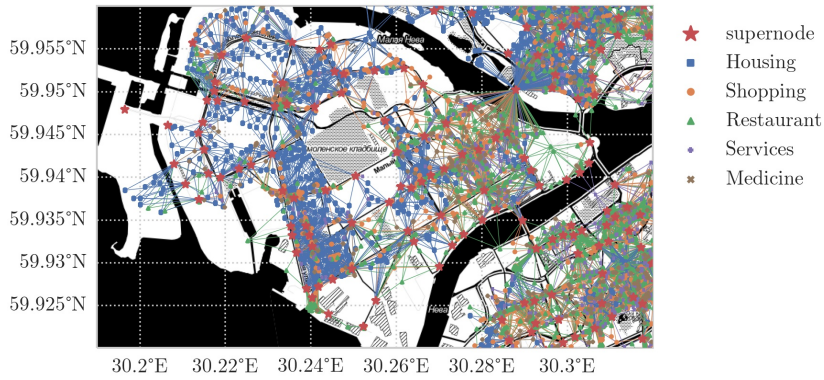

Fig. 10 The map of Vasileostrovsky District in St Petersburg, indicating supernodes and various infrastructural objects attached to them.

most of the cities in the world, thus making it possible to use the proposed techniques for analyzing public transportation systems of virtually any city.

The two sources of original data are:

- St Petersburg city Open Data[3], containing information about different public transportation stops and routes operating in St Petersburg, Russia.
- OpenStreetMap (OSM)[4], containing information about various infrastructural objects around St Petersburg.

The public transportation data is presented in form of a table with each row containing information about a stop as part of some transportation route. Each such row consists of information about the current route (its ID and mode of transportation), the current stop (its ID and coordinates), and the next stop corresponding to the route. The three modes of transportation presented here are: bus, tramway, and trolleybus. Note that there is no subway data available here, thus it will be extracted from the second source of data.

The data from OSM is presented in form of a JSON list containing information on various objects in and around St Petersburg. These objects can be either *nodes* or *ways*. A node usually denotes a single point on the map, and these are used in cases where the size of an object does not matter too much (for instance, bus stops, historical monuments, etc.). By contrast, ways (sequences of nodes) are usually used to represent larger objects (big buildings, industrial areas, etc.). Each object is represented as a dictionary that contains information on this object (namely, its ID, coordinates and some attributes corresponding to its type). These attributes are usually quite precise and can be used to extract more topic-specific information about each object.

This data is used in two ways. Firstly, we extract information on subway routes and stations to add another mode of transportation to those present in the first data source. The basic statistics on the completed public transportation data can be seen in Table 1.

Secondly, we group various infrastructural objects into 20 groups related to different types of social infrastructure (i.e. housing, shopping, restaurants, medicine, etc.). All of this is done using the OSM attributes of these objects, and the specific correspondings between these attributes and the resulting infrastructure types can be found in the Github repository. The number of infrastructural objects of each type can be seen in Figure 11.

4.2 Supernode network

This data is then used to build the model described in Section 3.1. First of all, supernodes are produced by combining the closely situated stops and stations. We use the distance threshold $d_0 = 0.1$, i.e. all stops that are closer than d_0 are combined into a single supernode (see Fig 3). In practice, this can be achieved using the following algorithm:

[3] https://classif.gov.spb.ru/irsi/7830001067-marshruty-dvizheniya-gorodskogo-transporta/structure_version/186/

[4] https://www.openstreetmap.org/

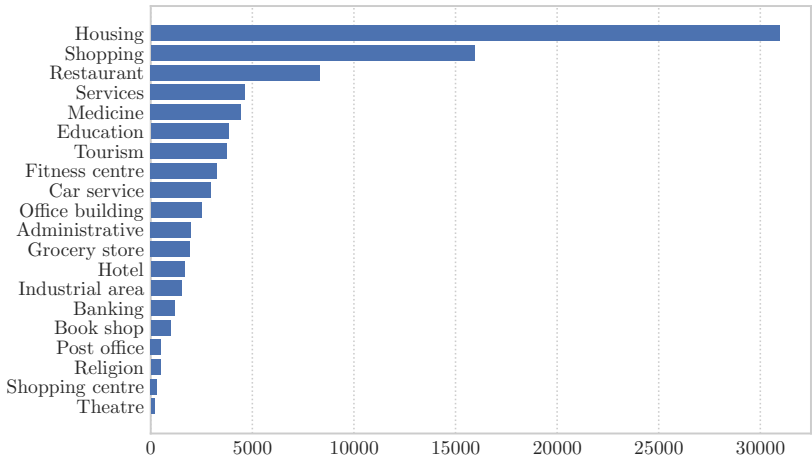

Fig. 11 Number of infrastructural objects of various types in St Petersburg

Table 1 Statistics on public transportation of St Petersburg

	Bus	Subway	Tram	Trolley
Number of stops	5511	71	887	1192
Number of routes	1070	10	83	90
Average route length, km	14.15	24.02	10.38	11.32

1. Calculate the distances between all pairs of stops.
2. Build a graph, in which a link between two stops s_1 and s_2 exists if $d(s_1, s_2) \leq d_0$.
3. Each connected component of this graph is a separate supernode.

Secondly, we need to construct links for our network. As we mentioned previously, we adopt the P-space model for constructing links, i.e. a link between nodes s_i and s_j means that there exists a route connecting these nodes (not necessarily consecutively). We also use link weights defined in Eq. 7 with $\alpha = 0.2$. In practice, this can be done using the following algorithm:

1. Iterate through all routes. For each route r do the following.
2. Iterate through all pairs of nodes in r. For each pair of nodes s_i, s_j calculate and store the route distance $rd_r(s_i, s_j)$ (Eq. 5).
3. When this is done, for each pair of nodes s_i, $s_2 j$ take the minimal route distance $rd(s_i, s_j)$ (Eq. 6) and use it to calculate the link weight $w(s_i, s_j)$ (Eq. 7).

Note that the built graph is not always connected. In Figure 12 we can see two small portions of nodes disconnected from the main body of the graph. This can happen if each of the route stops is too far away from the rest of the nodes in the graph, thus making it impossible to make a short walk to reach it.

Finally, we need to construct the supernode attributes, based on the social infrastructure around them. As it was described in Section 3.1, we use a distance window $d_1 = 0.2$ here. This value controls the additional distance (with respect to the minimal distance to any supernode) allowed in order to assign a given infrastructure object to a supernode (see Figure 10). In general, assigning infrastructure objects to supernodes can be done using the following algorithm:

1. Iterate through all infrastructure objects. For each object i do the following.
2. Calculate distances from i to each supernode.
3. Take the minimal distance d_{min}. Assign object i to each stop s such that $d(i, s) \leq d_{min} + d_1$.

Note that in general not all nodes in the graph will be assigned infrastructure objects, and there

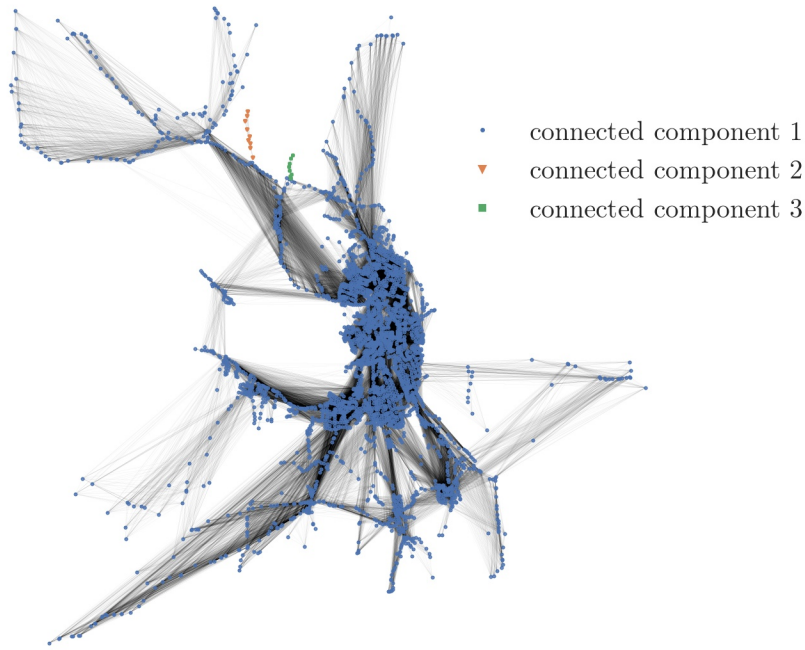

Fig. 12 The P-space supernode graph built using the St Petersburg data. The node positions correspond to the geographical locations of the corresponding nodes. There is a small portion of nodes that are disconnected from the main body of the graph.

can be some nodes with no infrastructure around them. Figure 13 shows all the supernodes of the St Petersburg PTN with colours indicating the number of infrastructure objects assigned to them.

4.3 Supernode features

Recall that in this study we perform separate clustering with respect to topological features (derived from the network structure) and infrastructure features (using the supernode attributes, see Section 3.1) and then compare the two.

Firstly, we consider the supernode attributes defined in Section 3.1 and built in Section 4.1. To construct *infrastructure features* from these attributes, we additionally divide each vector $v = A(s)$ by its sum $\sum v$ (obviously, excluding the cases, where $\sum v = 0$). Therefore, each such infrastructure feature vector shows the orientation of a given supernode towards one or multiple infrastructure types (for example, some nodes can be mainly housing-oriented, having a large value corresponding to the housing infrastructure type and smaller values on other positions), regardless of the total number of infrastructure objects around the node.

The second set of supernode *topological features* is constructed based on the topology, i.e. the connectivity structure of the network. Here we use various well-known centrality measures, namely, *degree centrality*, *betweenness centrality* [39] (considering weights w when calculating shortest paths), and *closeness centrality* [40] (using both weights w and the number of hops), as well as other topological features like the *local clustering coefficient* [41] and *PageRank* [42].

As we already mentioned in Section 2, the choice of a network model (L-space or P-space) is of paramount importance when using and interpreting such topological features. For instance, we argue against using centrality measures based on shortest paths with the L-space model, since such shortest paths are not indicative of the optimal

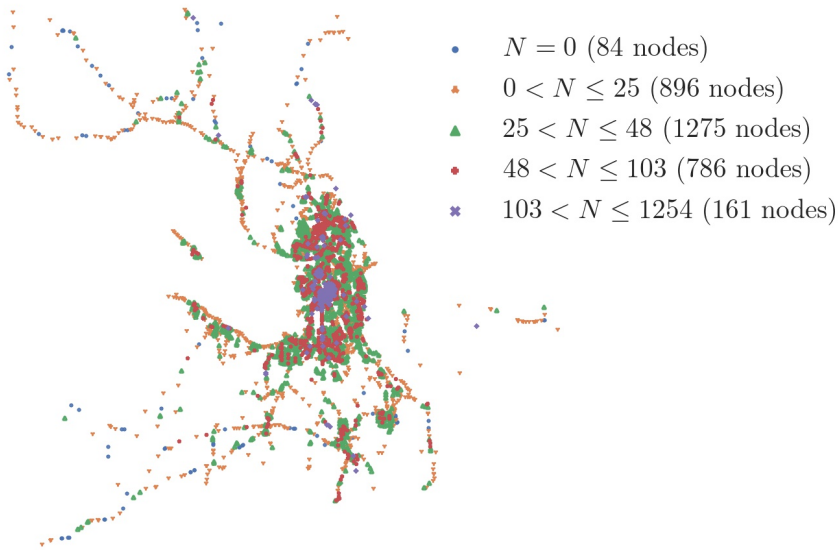

Fig. 13 The nodes of the St Petersburg graph with colours indicating the number of infrastructure objects assigned to them.

travelling routes of passengers (see Figure 4), thus making any analysis of these centralities (as in [7]) rather questionable.

Contrarily to this, the mentioned centrality measures offer a natural interpretation when using them with the P-space model. For instance, degree centrality emphasizes the so-called accessibility hubs, i e. nodes from which a lot of other nodes are accessible without need to make a connection. Betweenness centrality emphasizes the transportation hubs, i.e. the nodes at which a lot of connections happen. Closeness centrality emphasizes the nodes that on average require the least travelling distance (in sense of either weighted distance w, or the number of connections) to reach. PageRank is similar to degree centrality, but it also promotes the nodes that are connected to many important nodes of the graph. All these centrality measures therefore highlight different aspects of centrality that can occur in a PTN.

The least intuitive feature here is the local clustering coefficient, which is the fraction of closed triangles that exist in the neighbourhood of a given node. Since in the P-space model all the node pairs inside each route are connected with a link (and therefore all possible triangles exist around these nodes in these cases), local clustering coefficient does the contrary to the measures described above and actually emphasizes the nodes that are a part of the least number of different routes.

Before turning to the clustering task, we examine these features a bit more. Figure 14 shows the heatmap of Spearman correlations between the features. Note that some of the higher correlations (in absolute terms) are actually expected, such as the strong positive correlation between the centrality measures and the strong negative correlation between the latter and the local clustering coefficient. The other correlations are more interesting though. For instance, we can see a significant positive correlation between the centrality measures and some infrastructure features such as Restaurant, Service, Office building and Banking. These (expectedly) indicate that there are on average more of such infrastructure objects towards the city center. Such correlation is less noticeable for Shopping centres and Post offices, indicating that these infrastructure objects can generally be found everywhere throughout the city, not just in the center. On the other hand, we can note a slight negative correlation between Housing and many of the other features (notably, except Grocery stores and Education), which indicates a tendency

towards higher isolation of the residential districts in St Petersburg.

4.4 Supernode clustering

Finally, in this section we perform the cluster analysis of the supernodes with respect to both infrastructure and graph features (see Section 4.3). The framework of this analysis is as follows. For a given set of features, we first obtain their clusters using the KMeans algorithm [38]. The number of clusters is determined based on the inertia plot and interpretability of the clusters. The clusters are then plotted in different views and interpreted, based on their features.

We first perform cluster analysis of the infrastructure features. Figure 15 shows the t-SNE projection [43] and geographical positions of the supernodes, coloured based on the obtained clusters. To analyse the difference between these clusters and interpret them, we also plot the aggregated features over each cluster (Figure 16). In the upper part of the figure the mean feature values are plotted for each cluster, as well as the global mean. Also, to emphasize the difference between the feature values in each cluster, in the lower part we plot the values of the 2-sample Welch's t-test statistic [44] for comparing the mean of each feature over the given cluster, compared to the mean of this feature over the rest of the clusters (this is the so-called one-vs-rest strategy).

The obtained clusters can be summarised as follows:

1. Tourism area. Nodes with mostly tourist attractions around them, not much else.
2. Residential area — unimproved. Nodes with mostly housing around them and no other common urban amenities such as grocery stores, hospitals, etc.
3. Center. Nodes located around shops, restaurants, office buildings, banks, etc.
4. Residential area — improved. Nodes with housing as well as various amenities like schools, hospitals, grocery stores, fitness centres, etc.
5. Industrial area. Nodes surrounded by industrial areas and not much else.
6. No infrastructure. Nodes that have no social infrastructure around.

It should be noted that the proposed interpretation is not exactly strict, since, as it can be seen in the t-SNE projection in Figure 15, there are no clear boundaries between these clusters (except for the cluster with no social infrastructure). Thus by means of social infrastructure, one can see smoothly changing and highly various set of supernode types in the PTN under consideration.

The cluster analysis of graph features is done in a similar fashion: the clustering is performed using the KMeans algorithm, the clusters are plotted using their t-SNE projection as well as the geographical positions in Figure 17, and their aggregated features are shown in Figure 18. Based on the presented data, these clusters can be summarised as follows:

1. Hubs. Nodes that serve as points of transition between different routes when travelling through a city. These nodes have higher degree and betweenness centrality, compared to the other nodes.
2. Center. Nodes that represent the well-accessible part of the city center. These nodes have high values of centralities.
3. Inaccessible center. Nodes that represent the less accessible part of the city center. These nodes have high closeness centrality based on the link weights, which indicate their close proximity to the center, but at the same time these nodes have low closeness centrality based on hops (which indicates that these nodes on average require much more connections to reach), as well as low betweenness and degree centrality.
4. Towns. Nodes located outside the main city in separate towns (low betweenness, closeness and degree centrality).
5. Suburbs. Nodes that are located moderately far away from the city center (lower closeness centrality), but are still well-connected to the transportation network (moderate betweenness and degree centrality).
6. Disconnected nodes. A few nodes that are not connected to the transportation network and form separate connected components. (Note that this cluster is not presented in Figure 18 since its 2-test statistic values are extremely low and render all the other plots impossible to read.)

As with the infrastructure-based clustering, it should be noted that there are no clear boundaries

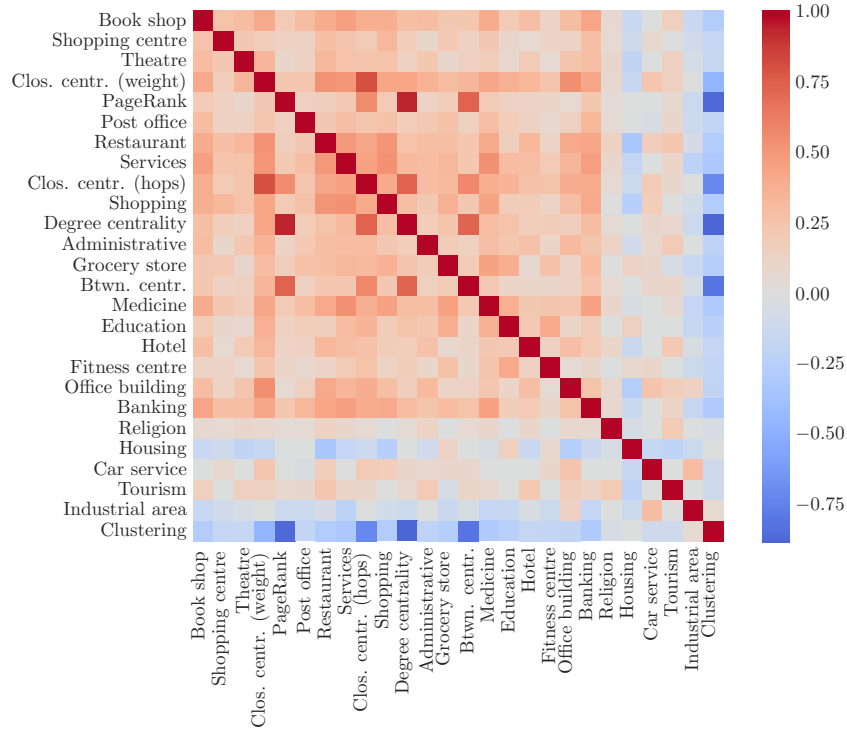

Fig. 14 The heatmap of Spearman correlations between various supernode graph features. (For interpretation of the references to colour in this heatmap, the reader is referred to the web version of this article.)

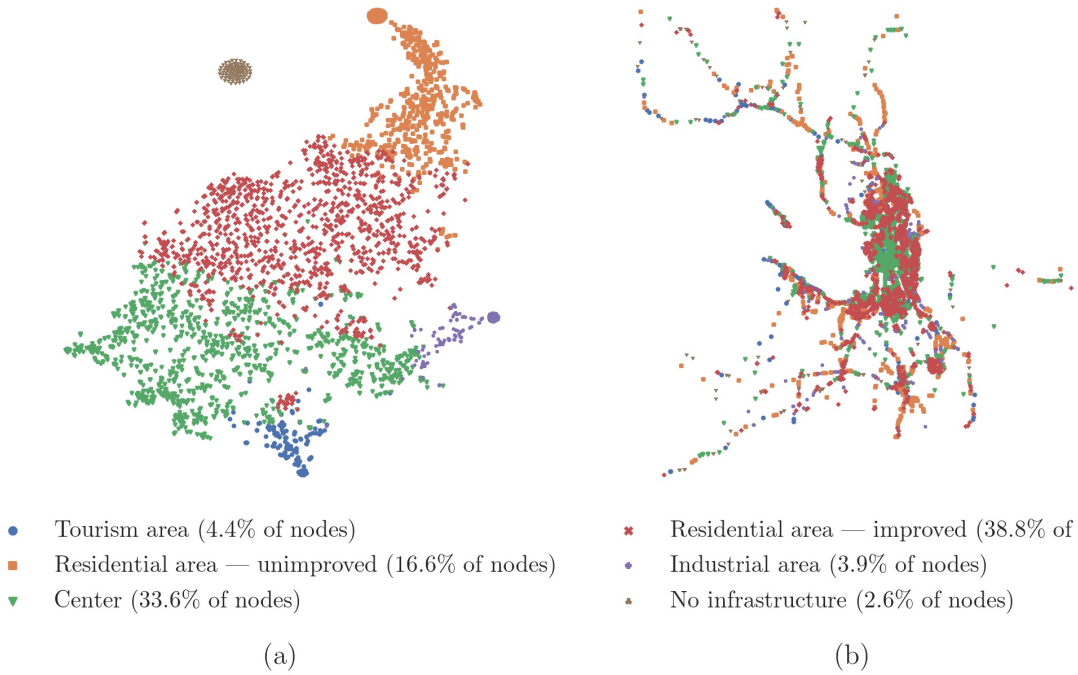

Fig. 15 t-SNE projection (a) and geographical positions (b) of supernodes, coloured based on the clusters obtained using their infrastructure features.

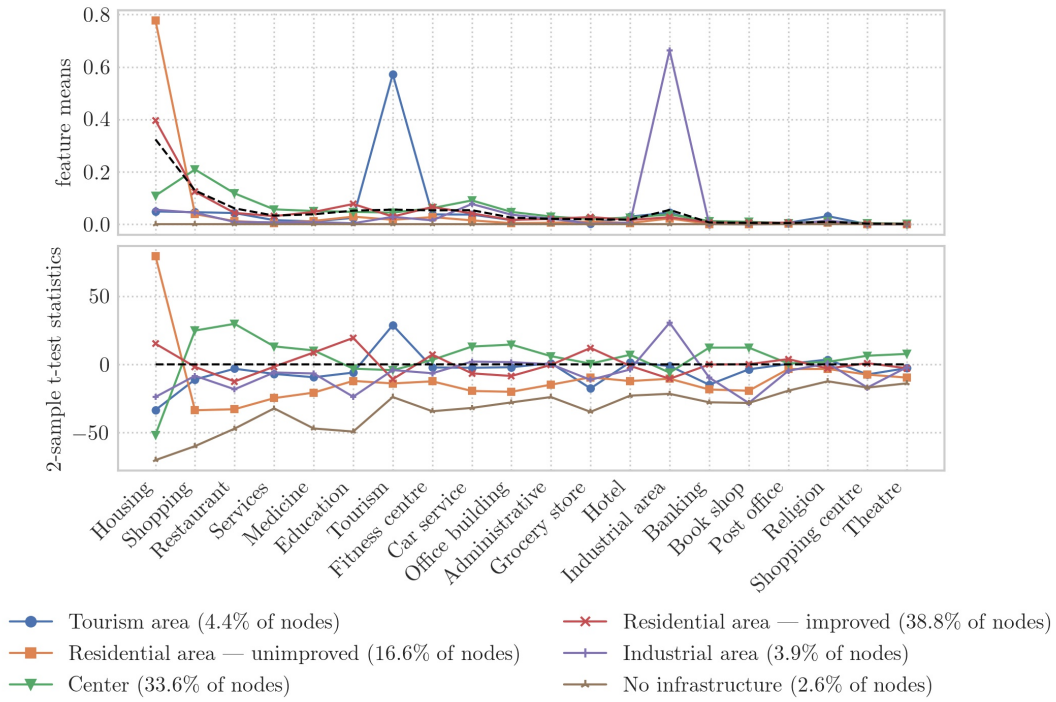

Fig. 16 Aggregated features of supernodes from different infrastructure clusters. The upper plot shows a mean value of each feature across each cluster, as well as the global mean. The lower plot shows the values of the 2-sample Welch's t-test statistic [44] for comparing the mean of each feature over the given cluster, compared to the mean of this feature over the rest of the clusters.

between the topology-based clusters. Nevertheless, this clustering shows the different high-level roles of the network nodes and provides insight into the relations between these clusters.

Finally, in order to assess the relations between the infrastructure and topology feature clusters, we build a contingency table by counting the number of nodes in different intersections of these clusters. These values are presented in Table 2. The rows of the table represent the infrastructure clusters and the graph feature clusters are represented by the columns.

From this table some interesting interconnections between the two clusterings arise. We can see that most of the infrastructure clusters are well-represented in all of the graph-feature clusters (and vice versa), which means that these two clusterings both carry important and unique information about the roles of each node. For instance, we can see that the nodes corresponding to improved residential areas (with better-developed urban amenities) have more members in graph-based clusters 'Center', 'Hub' and 'Inaccessible center', as well as 'Suburbs', at the same time there are more undeveloped residential areas in the 'Towns' cluster.

Another important cluster to consider from the urban development point of view is the graph-feature cluster 'Inaccessible center', which contains nodes that are fairly central in terms of closeness centrality (i.e. average distance from the rest of the nodes), but are low on betweenness and degree centrality, which means that public transportation is under-developed in these areas. We can see that this cluster contains many members of the infrastructure clusters 'Center' and 'Residential area — improved', which means that these areas are well-developed in terms of urban amenities, but are quite separated from the rest of the PTN, which means less convenience of daily commuting, for instance.

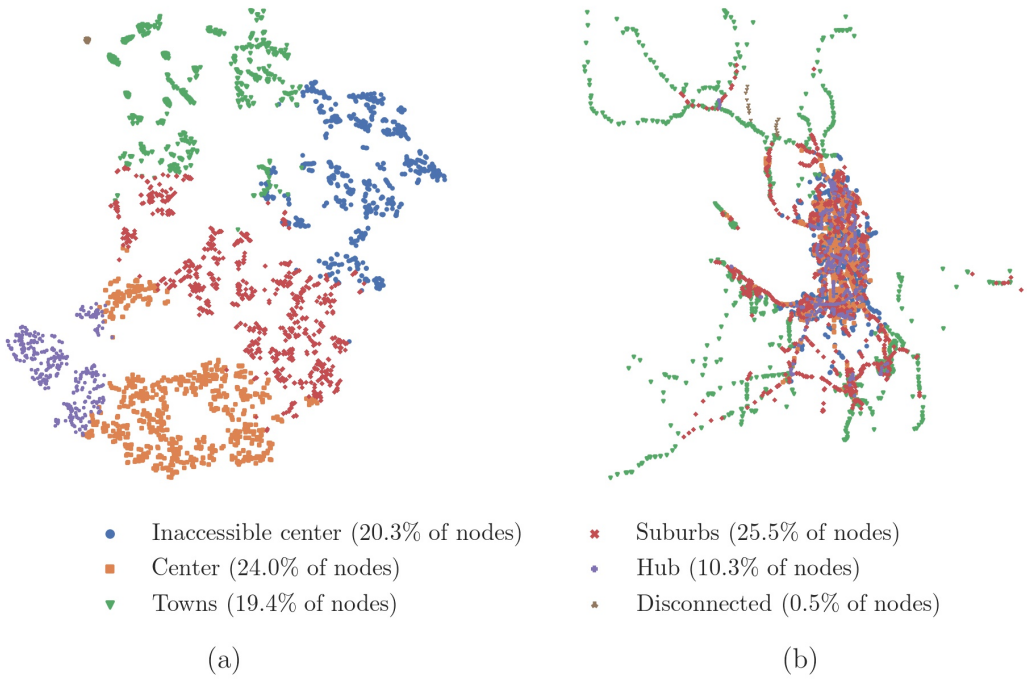

Fig. 17 t-SNE projection (a) and geographical positions (b) of supernodes, coloured based on the clusters obtained using their graph features.

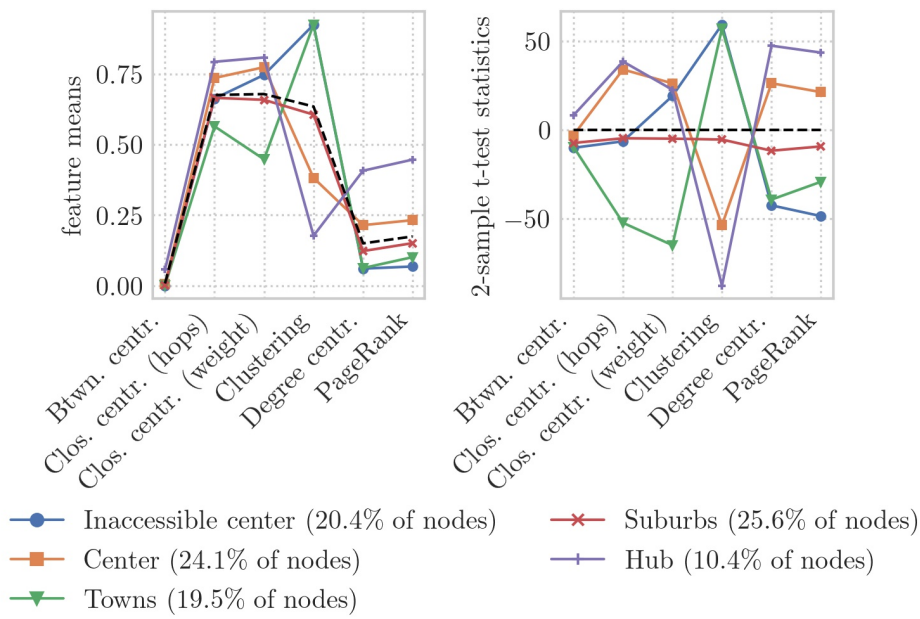

Fig. 18 Aggregated features of supernodes from different graph feature clusters. The left plot shows a mean value of each feature across each cluster, as well as the global mean. The right plot shows the values of the 2-sample Welch's t-test statistic [44] for comparing the mean of each feature over the given cluster, compared to the mean of this feature over the rest of the clusters.

Table 2 Contingency table showing the relations between infrastructure clusters and graph feature clusters. The rows of the table represent the infrastructure clusters and the graph feature clusters are represented by the columns.

	Center	Disconnected	Hub	Inaccessible center	Suburbs	Towns
Center	300	3	169	265	225	115
Industrial area	3	0	0	33	43	47
No infrastructure	0	3	0	2	11	68
Residential area (improved)	371	2	144	249	344	131
Residential area (unimproved)	74	9	15	92	139	203
Tourism area	20	0	2	10	54	56

5 Conclusions and future work

In this work we develop a novel weighted node-attributed PTN model (using information about a city's social infrastructure to construct the node attributes) and apply it to discover roles of public transport stops and stations of St Petersburg, Russia. We also point out some of the common misconceptions and errors in some previous analyses of the PTNs which we believe stem from the misunderstanding of some of the interpretations of different PTN models.

A novel role discovery framework is introduced which uses both structural (i.e. network topology) and semantic (i.e. social infrastructure around the nodes) aspects of a node-attributed PTN. This framework is shown to be capable of extracting useful information about the properties and overall efficiency of a city's public transportation system from both the structural and infrastructure standpoints. For instance, in case of St Petersburg, it is able to point out some under-developed areas of the city, e.g. less accessible parts of the city center or residential areas that are low on urban amenities. These weaknesses can lead to better development of the city in the future, if taken into consideration by the city administration.

The performed analysis uses only the generally available data, which means that similar analysis can be performed on any large city's public transportation system. In general, the proposed approach to role discovery in node-attributed networks can be applied beyond the scope of PTNs and to any other kind of network (e.g. social, biological, technical, etc.), given the appropriate set of node attributes.

It is noted in Section 3.1 that the most common method of constructing supernodes (i.e. just grouping together all the closely located stops) is not without its drawbacks. Additional research should be conducted regarding this problem. Another potential direction of future research is developing more interpretable graph-based node metrics that would highlight even more peculiarities in the different roles of the nodes in a PTN. For instance, as it is mentioned in Section 4.3, the metric of betweenness centrality over a P-space model graph highlights the nodes at which a lot of transfers happen. At the same time the actual stops that these shortest routes though a P-space graph go through are not highlighted by any of the existing metrics (and are not actually even considered in a P-space model). A metric like this could bring up very important information about the actual workload of different PTN nodes without the need for any dynamic data like transportation of passenger flows.

Declarations

Data and code availability

The data that support the findings of this study are openly available in the GitHub repository at https://github.com/AlgoMathITMO/public-transport-network. These data were derived from the following resources available in the public domain:

- St Petersburg city Open Data[5];

[5] https://classif.gov.spb.ru/irsi/7830001067-marshruty-dvizheniya-gorodskogo-transporta/structure_version/186/

- OpenStreetMap (OSM)[6].

Research involving Human Participants and/or Animals

No.

Informed consent

Not applicable.

Conflicts of interests

The authors declare no competing interests.

Funding

This study is financially supported by the Russian Science Foundation, Agreement 17-71-30029, with co-financing of Bank Saint Petersburg, Russia.

Author contributions

The contributor's impact on this paper is as follows. Conceived and designed the experiments: Y.L., P.C., T.G., A.B. Performed the experiments: Y.L., T.G., A.B., I.S. Collected and pre-processed the data: Y.L., T.G., A.B., I.S. Analyzed the data: Y.L., P.C., T.G., A.B., I.S. Wrote and reviewed the the main manuscript text: P.C., Y.L., T.G.

References

[1] Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P.A., Mukherjee, G., Manna, S.S.: Small-world properties of the indian railway network. Phys. Rev. E **67**, 036106 (2003). https://doi.org/10.1103/PhysRevE.67.036106

[2] Sienkiewicz, J., Hołyst, J.: Statistical analysis of 22 public transport networks in poland. Physical review. E, Statistical, nonlinear, and soft matter physics **72**, 046127 (2005). https://doi.org/10.1103/PhysRevE.72.046127

[3] Háznagy, A., Fi, I., London, A., Nemeth, T.: Complex network analysis of public transportation networks: A comprehensive study. In: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 371–378 (2015). https://doi.org/10.1109/MTITS.2015.7223282

[4] Yang, X.-H., Chen, G., Chen, S.-Y., Wang, W.-L., Wang, L.: Study on some bus transport networks in china with considering spatial characteristics. Transportation Research Part A: Policy and Practice **69**, 1–10 (2014). https://doi.org/10.1016/j.tra.2014.08.004

[5] Zhang, J., Zhao, M., Liu, H., Xu, X.: Networked characteristics of the urban rail transit networks. Physica A: Statistical Mechanics and its Applications **392**, 1538–1546 (2013). https://doi.org/10.1016/j.physa.2012.11.036

[6] Wang, L.-N., Wang, K., Shen, J.-L.: Weighted complex networks in urban public transportation: Modeling and testing. Physica A: Statistical Mechanics and its Applications **545**, 123498 (2020). https://doi.org/10.1016/j.physa.2019.123498

[7] Shanmukhappa, T., Ho, I.W.-H., Tse, C.K.: Spatial analysis of bus transport networks using network theory. Physica A: Statistical Mechanics and its Applications **502**, 295–314 (2018). https://doi.org/10.1016/j.physa.2018.02.111

[8] Wang, Y., Deng, Y., Ren, F., Zhu, R., Wang, P., Du, T., Du, Q.: Analysing the spatial configuration of urban bus networks based on the geospatial network analysis method. Cities **96**, 102406 (2020). https://doi.org/10.1016/j.cities.2019.102406

[9] Lantseva, A., Ivanov, S.: Modeling transport accessibility with open data: Case study of st. petersburg. Procedia Computer Science **101**, 197–206 (2016). https://doi.org/10.1016/j.procs.2016.11.024

[10] Bothorel, C., Cruz, J., Magnani, M., Micenková, B.: Clustering attributed graphs: Models, measures and methods. Network Science **3**(3), 408–444 (2015). https://doi.org/10.1017/nws.2015.9

[6] https://www.openstreetmap.org/

[11] Chunaev, P.: Community detection in node-attributed social networks: A survey. Computer Science Review **37**, 100286 (2020). https://doi.org/10.1016/j.cosrev.2020.100286

[12] Atzmueller, M., Günnemann, S., Zimmermann, A.: Mining communities and their descriptions on attributed graphs: a survey. Data Mining and Knowledge Discovery **35**(3), 661–687 (2021). https://doi.org/10.1007/s10618-021-00741-z

[13] Rossi, R.A., Ahmed, N.K.: Role discovery in networks. IEEE Transactions on Knowledge and Data Engineering **27**(4), 1112–1131 (2015). https://doi.org/10.1109/tkde.2014.2349913

[14] Ahmed, N., Rossi, R.A., Willke, T.L., Zhou, R.: Revisiting role discovery in networks: From node to edge roles. ArXiv 1610.00844 (2016)

[15] Martínez, V., Berzal, F., Cubero, J.-C., Bueno, A.: An automorphic distance metric and its application to node embedding for role mining. Complex. **2021** (2021). https://doi.org/10.1155/2021/5571006

[16] Gupte, P.V., Ravindran, B., Parthasarathy, S.: Role discovery in graphs using global features: Algorithms, applications and a novel evaluation strategy. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp. 771–782 (2017). https://doi.org/10.1109/ICDE.2017.128

[17] Revelle, M., Domeniconi, C., Johri, A.: Persistent roles in online social networks. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds.) Machine Learning and Knowledge Discovery in Databases, pp. 47–62. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46227-1_4

[18] Rossi, R.A., Gallagher, B., Neville, J., Henderson, K.: Modeling dynamic behavior in large evolving graphs. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. WSDM '13, pp. 667–676. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2433396.2433479

[19] Vega, D., Meseguer, R., Freitag, F., Magnani, M.: Role and position detection in networks: Reloaded. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 320–325 (2015). https://doi.org/10.1145/2808797.2809412

[20] Yang, Z., Algesheimer, R., Tessone, C.J.: A comparative analysis of community detection algorithms on artificial networks. Scientific Reports **6**(1) (2016). https://doi.org/10.1038/srep30750

[21] Fortunato, S.: Community detection in graphs. Physics Reports **486**(3), 75–174 (2010). https://doi.org/10.1016/j.physrep.2009.11.002

[22] Souravlas, S., Sifaleras, A., Tsintogianni, M., Katsavounis, S.: A classification of community detection methods in social networks: a survey. International Journal of General Systems **50**(1), 63–91 (2021). https://doi.org/10.1080/03081079.2020.1863394

[23] Bartal, A., Ravid, G.: Member behavior in dynamic online communities: Role affiliation frequency model. IEEE Transactions on Knowledge and Data Engineering **32**(9), 1773–1784 (2020). https://doi.org/10.1109/tkde.2019.2911067

[24] Ni, W., Guo, H., Liu, T., Zeng, Q.: Automatic role identification for research teams with ranking multi-view machines. Knowledge and Information Systems **62**(12), 4681–4716 (2020). https://doi.org/10.1007/s10115-020-01504-w

[25] Liu, S., Toriumi, F., Nishiguchi, M., Usui, S.: Multiple role discovery in complex networks. In: Benito, R.M., Cherifi, C., Cherifi, H., Moro, E., Rocha, L.M., Sales-Pardo, M. (eds.) Complex Networks & Their Applications X, pp. 415–427. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-93413-2_35

[26] Liu, Y., Du, F., Sun, J., Silva, T., Jiang, Y., Zhu, T.: Identifying social roles using heterogeneous features in online social networks. Journal of the Association for Information Science and Technology **70**(7), 660–674 (2019). https://doi.org/10.1002/asi.24160

[27] Zhang, L., Lu, J., Fu, B.-b., Li, S.-b.: A review and prospect for the complexity and resilience of urban public transit network based on complex network theory. Complexity **2018**, 1–36 (2018). https://doi.org/10.1155/2018/2156309

[28] Shanmukhappa, T., Wang-Hei Ho, I., Tse, C.K., Leung, K.K.: Recent development in public transport network analysis from the complex network perspective. IEEE Circuits and Systems Magazine **19**(4), 39–65 (2019). https://doi.org/10.1109/MCAS.2019.2945211

[29] Xu, Q., Mao, B., Bai, Y.: Network structure of subway passenger flows. Journal of Statistical Mechanics: Theory and Experiment **2016**, 033404 (2016). https://doi.org/10.1088/1742-5468/2016/03/033404

[30] Feng, J., Li, X., Mao, B., Xu, Q., Bai, Y.: Weighted complex network analysis of the beijing subway system: Train and passenger flows. Physica A: Statistical Mechanics and its Applications **474**, 213–223 (2017). https://doi.org/10.1016/j.physa.2017.01.085

[31] Faust, K., Wasserman, S.: Blockmodels: Interpretation and evaluation. Social Networks **14**(1), 5–61 (1992). https://doi.org/10.1016/0378-8733(92)90013-W. Special Issue on Blockmodels

[32] Batagelj, V., Mrvar, A., Ferligoj, A., Doreian, P.: Generalized blockmodeling with pajek. Metodoloski Zvezki **1**, 455–467 (2004)

[33] Luczkovich, J., Borgatti, S., Johnson, J.C., Everett, M.: Defining and measuring trophic role similarity in food webs using regular equivalence. Journal of theoretical biology **220 3**, 303–21 (2003)

[34] Ma, H., Zhou, D., Liu, C., Lyu, M.R., King, I.: Recommender systems with social regularization. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM '11, pp. 287–296. Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1935826.1935877

[35] Golder, S., Donath, J.: Social roles in electronic communities. Internet Research **5** (2004)

[36] Yue, H., Guan, Q., Pan, Y., Chen, L., Lv, J., Yao, Y.: Detecting clusters over intercity transportation networks using k-shortest paths and hierarchical clustering: a case study of mainland china. International Journal of Geographical Information Science **33**(5), 1082–1105 (2019). https://doi.org/10.1080/13658816.2019.1566551

[37] Bereznyi, D., Qutbuddin, A., Her, Y., Yang, K.: Node-attributed spatial graph partitioning. In: Proceedings of the 28th International Conference on Advances in Geographic Information Systems. SIGSPATIAL '20, pp. 58–67. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397536.3422198

[38] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (1967)

[39] Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977). https://doi.org/10.2307/3033543

[40] Freeman, L.C.: Centrality in social networks conceptual clarification. Social Networks **1**(3), 215–239 (1978). https://doi.org/10.1016/0378-8733(78)90021-7

[41] Onnela, J.-P., Saramäki, J., Kertész, J., Kaski, K.: Intensity and coherence of motifs in weighted complex networks. Phys. Rev. E **71**, 065103 (2005). https://doi.org/10.1103/PhysRevE.71.065103

[42] Page, L., Brin, S., Motwani, R., Winograd,

T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (November 1999). http://ilpubs.stanford.edu:8090/422/

[43] van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

[44] Welch, B.L.: The generalisation of student's problems when several different population variances are involved. Biometrika **34**(1-2), 28–35 (1947). https://doi.org/10.1093/biomet/34.1-2.28