Percolation-based congestion analysis and its applications in transportation networks

MONASH University

Homayoun Hamedmoghadam Rafati

Institute of Transport Studies

Department of Civil Engineering

Faculty of Engineering

A thesis submitted for the degree of Doctor of Philosophy at Monash University in 2021

To my Mom and Dad and in loving memory of Aziz for their untiring support and faith in me.

Copyright notice

©Homayoun Hamedmoghadam-Rafati (2021)

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Summary

Transportation networks are engines of economic development and growth, moving goods and providing human activities with their pivotal basic requirement, i.e., mobility. Considering the rapid urbanization and unprecedented increase in population of cities, the efficiency of urban transportation networks has never been a more important issue. The constant conflict with different forms of congestion is the most important factor compromising the efficiency of transportation systems. In urban road networks, traffic jams impose immense additional costs to the society every day, by wasting time, exhausting energy, and deteriorating citizen's health.

Fundamental laws of traffic flow theory have been studied intensively for decades. There is now a deep understanding of how congestion changes in road networks at different scales. Especially, the relation between the traffic density and congestion on single road segments and over the whole network is well studied. In comparison, however, the organization of different traffic congestion levels in cities is rarely studied, and only recently the attention to this is heightened in the area of complex networks. The criticality of the issues with traffic congestion and increasing availability of pertinent real-world transportation data, warrant further investigations into organization of congestion in urban transportation networks.

In the present dissertation, we focus on applications of percolation theory, a popular tool in network science and statistical physics, in transportation network analysis. At its core, percolation provides a framework to characterize the complex topological properties of networks. Built upon the most recent advancement in network percolation analysis, we tackle two different problems related to traffic congestion using percolation framework to unpack the organization of different levels of congestion in transportation networks.

The first problem is the conflict between passenger flows in on-road (bus and tram) public transportation network and congestion on road infrastructure. To tackle this, we use data including over 120 million passenger smartcard records collected in real public transportation networks. From the available data we model the transportation network, augmented with the temporal congestion and passenger flow information. We propose a percolation analysis to measure network reliability, in terms of its ability to provide congestion-free routes for traveling flows. A major finding of this study is a theoretical relationship between link-level congestion and network-level reliability, which allows for identifying the most critical bottleneck links in the network. We demonstrate the effectiveness of our reli-

ability measurement, and prove that the identified bottlenecks have guaranteed improving effect on network reliability.

The second study is concerned with the propagation of congestion in road networks. We utilize percolation analysis to characterize the phase transition from small isolated pockets of congestion into a large congested cluster in the network. Based on this new knowledge, we propose a percolation-based strategy to modify the timing of intersection traffic signals in urban road networks. In particular, we control the signals on the time-varying boundary of the identified congested cluster, to mitigate the congestion. Simulations demonstrate that the proposed strategy can effectively reduce congestion and boost the traffic-carrying capacity of the network, by dynamically balancing congested queues that form around a hotspot region.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes ***two*** original papers published in peer reviewed journals. The core theme of the thesis is ***percolation-based congestion analysis and its applications in transportation networks***. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the ***Department of Civil Engineering*** under the supervision of ***Prof. Hai Vu***.

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.)

In the case of *Chapter 3: Section 3.1* my contribution to the work involved the following:

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
3	Automated extraction of origin-destination demand for public transportation from smartcard data with pattern recognition	in press	60%. Concept, methodology, data analysis, and writing	1) Hai L. Vu, methodology and input into manuscript (15%). 2) Mahdi Jalili, input into manuscript (7.5%). 3) Meead Saberi, input into manuscript (7.5%). 4) Lewi Stone, input into manuscript (5%). *5)* Serge Hoogendoorn, manuscript review (5%).	N

In the case of *Chapter 4: Section 4.1* my contribution to the work involved the following:

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
4	Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks	published	60%. Concept, methodology, data analysis, and writing	1) Mahdi Jalili, design of the study and manuscript input (10%). 2) Hai L. Vu, design of the study and manuscript input (10%). 3) Lewi Stone, methodology and writing (20%).	N

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature: ..... **Date:**

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature: ..... **Date:**

Acknowledgements

First, many thanks to my supervisor Prof. Hai Vu for the exceptional support and mentorship that he provided me with, in every step of the work. Also, I would like to thank Prof. Shlomo Havlin for his extremely helpful discussions of the ideas and for his suggestions regarding a work presented in a chapter of this dissertation. Many thanks to Dr. Meead Saberi for his advice in early stages of my PhD and that he kept helping me through collaboration and with the occasional fruitful discussions. I am grateful to Prof. Mark Hickman and Dr. Zhenliang Ma for their generous help with a critical part of my work by running my codes on their data. I am especially indebted to A/Prof. Mahdi Jalili and Prof. Lewi Stone for their incredible mentorship and the significant contributions they made to my research work. And last but not least, many thanks to Dr. Nora Estgfäller for providing me with valuable help and feedback in every step of the research.

Table of Contents

List of Figures

Chapter 1

Introduction

It has been more than a decade, since the world reached the point of having the majority of its population living in cities [1]. Yet, worldwide rapid urbanization has not been slowed down since and continues to increase the population of cities. Major cities, once viewed as chaotic disordered systems, are far better understood over the past 40 years. Cities are now viewed as highly ordered complex systems showing clear patterns in their evolution and growth [1, 2]. Functioning, growth, and prospect of cities are highly dependent on efficiency of flow-carrying infrastructures such as power grids, communication networks, and transportation systems [3].

Urban transportation systems are critical components of cities, bringing access to goods and providing the society with opportunities. The interplay between the demand for mobility and the limited transportation infrastructure and supply produces different forms of congestion in urban transportation networks [4], e.g., train delays due to excessive number of passengers and road traffic jams due to excessive number of vehicles and pedestrian crowding. On the other hand, the unprecedented urbanization, highly correlated with the availability of transportation supply [5], often adversely affects the congestion, especially in mega cities around the world. Alleviating the effect of congestion is essential to the efficiency of the transportation system and is the main focus of the present thesis.

Almost any user of urban road systems can recognize 'congestion' as a major problem, wasting time and money while diminishing the convenience of most trips at least to some degree. Vehicles fuel consumption increases by an order of 80% when riding along congested roads [6], which leads to an increase in harmful emissions including carbon dioxide and threatens public health [7]. TomTom, a leading multinational developer of location technologies, reported that during 2019 in the most traffic congested cities of the world, namely, Bengaluru, Manila, Bogota, Mumbai, Pune, Moscow, Istanbul, Kyiv, and Bucharest, drivers spent an average of over 50% extra travel time in traffic [8]. Due to the delay caused by rush-hour traffic congestion, each commuter had lost an average of 128 hours during the

year 2019 in Melbourne, Australia [9], while pedestrians are able to walk faster than a car moves on the street in central Manhattan, New York [10].

Effective approaches towards resolving or mitigating the congestion problem can save an immense amount of time and resources while contributing to the comfort and health of countless number of users. To achieve the ultimate goal of freeing traveling flows from congestion, a great deal of work is required on various aspects of transportation systems including policy making, development, and operation.

Traffic congestion is a complex phenomenon and it is an inevitable product of economic growth and increase in social activities [11]. It is very difficult to systematically study all factors contributing to the onset of traffic congestion and determining its propagation or dissipation dynamics. However, the complex interplay between the travel demand and transportation supply is known to be the major determinant in traffic congestion dynamics [4, 12]. Travel demand can be well explained by the number of passengers moving between places. Transportation supply can be described by its quantitative properties such as the structure and capacity of the transportation routes and the size of individual and Public Transportation (PT) fleet. The pivotal factor driving the interplay between the travel demand and transportation supply is travel behavior, which refers to how travelers choose to move between places using the available transportation option.

Many congestion alleviation strategies have been developed and successfully applied in urban transportation literature [4, 10, 11, 12, 13]. In one category of strategies to mitigate the traffic congestion on roads, the base capacity of the system is modified, for example, through increasing the size and number of the roads or providing more PT services [14, 15]. Another category of strategies attempts at reducing the congestion production by modifying the demand through planning to encourage better travel and land use patterns [11, 13, 16, 17]. The aim of the work presented in this dissertation is not entirely detached from the previous efforts in the above two categories. However, the general strategy to the problem of congestion acquired here, belongs to a third category that aims at operating the existing capacity more efficiently for the existing demand.

Tools in the area of transportation management and operation are widely investigated as means to mitigate traffic congestion. However, the existing congestion alleviation measures in management and operation can be substantially improved if we better understand the physics of congestion and its formation in cities. In the literature, more attention is paid to investigating the local (in a road segment or a highway) formation of congestion [18], or investigating the factors affecting global congestion (vehicle accumulation in the whole system) [19]. The dynamics of congestion organization, i.e., distribution of different congestion levels over a transportation system and its variation over time, have rarely been studied. Thus, there is a gap in the understanding of congestion propagation dynamics, the organization of different congestion levels over the system, and the interaction between

traveling flows and the congestion. Here, we study transportation systems and their vehicular (or passenger) flows as complex 'network' systems, which offers the opportunity to use advanced techniques to study congestion organization and system's performance under congestion.

In particular, the present work pays attention to less studied fundamental questions with regards to congestion in transportation networks: "how does the traffic congestion propagate and organize over the network?" and "how to characterize the conflict between traveling flows and the formation of traffic congestion?", given an existing travel demand, travel behavior, and transportation supply. When tackling the above questions, we also pay specific attention to further steps towards improving the transportation system. On that account, the aim of our analyses is to pinpoint the problems in congested transportation networks and accordingly, propose possible measures and solutions that can effectively optimize traveling over the network.

1.1 Studying congested transportation networks

Arguably, around two decades ago, the field of network science emerged as a result of the most recent advancements in graph theory coinciding with a surge in the possibility of computer-aided analysis of real-world data [20, 21, 22]. Network systems are ubiquitous in nature and their structure determines how they function. Studying the structure of complex network systems proved to be an effective path to advance the understanding of mechanisms that our life depends on. Early empirical network analysis in different areas of science and engineering shed light on a variety of systems, from biological and ecological networks to critical infrastructure networks [23, 24, 25]. Ever since, network science has been a rapidly developing field providing new tools to unravel the obscure properties of a variety of complex network systems [26, 27, 28, 29, 30, 31, 32, 33, 34, 35].

Here, we focus on studying transportation systems as complex networks. Transportation systems can be well represented as network structures, where network nodes represent spatially distributed places of concern, and links connecting node pairs represent the transportation infrastructure which enable movement in specific directions between places [22, 36, 37]. Link dynamics such as traffic load or congestion and demand for movement of vehicles and/or passengers can be embedded quantitatively on transportation networks. In Fig. 1.1 a schematic framework is shown, where a complex transportation system is first represented by its network structure, where links connecting nodes represent the direction of transport between different locations. Next, the enriched network representation of the system can be generated by augmenting the travel demand between different places and representing the congestion levels as link attributes. Enriched transportation network representation allows for studying physics of congestion and at the same time, its conflict

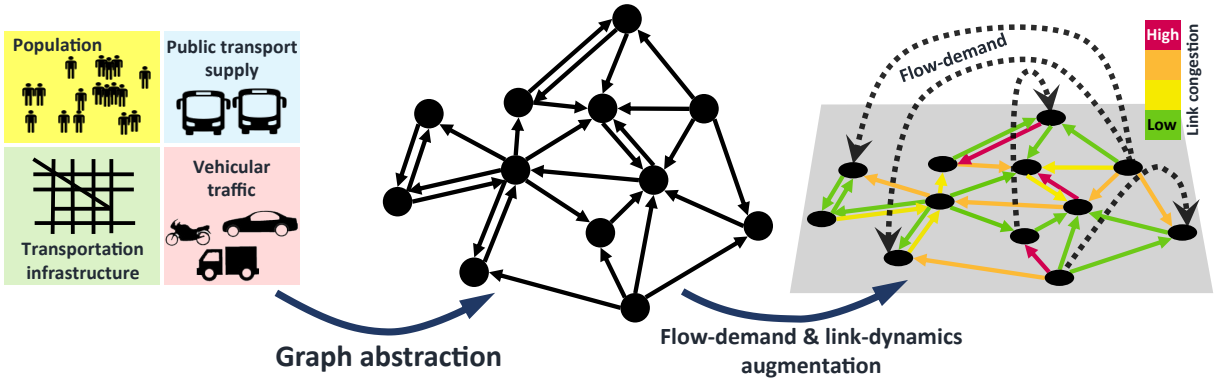


Figure 1.1: Enriched network modeling of complex transportation systems. A complex transportation system can be represented well by its network structure, including components and directed connections between them. An enriched version of this network structure can be constructed by augmentation of flow-demand and link-dynamics (e.g., temporal road congestion levels).

with traveling flows. This can lead to better understanding of transportation networks dynamics, which in effect, will aid addressing the congestion problem in different areas of transportation system management.

Transportation networks are inherently difficult to understand due to their structural and dynamical complexity. Additionally, these networks involve what is known as 'meta-complications' referring to the fact that various components of these complex systems can influence each other [38]. As an example, the relationship between road network geometry and both travel demand [39] and route choice [29] is known, meaning that the structure of road network influences the traveling flow dynamics. Mathematical network modeling of transportation systems allows for utilizing the well-established theories in network science and statistical physics. These tools enable us to effectively characterize the congestion- and traveling flow-dynamics, analyze the conflict between congestion and traveling flows, pinpoint the critical parts of the network, and provide mitigating solutions accordingly.

A set of tools heavily used in the network science area is developed within the realm of percolation theory [40, 41, 42, 43, 44]. Percolation theory provides a mathematical framework to characterize the structure and dynamics of networked systems and to study spreading phenomena in such networks. Both road traffic congestion and traveling flows behave similar to spreading phenomena, with the former spreading from a congested road to its upstream roads (e.g., as a result of queue spillback [45]) and the latter spreading from origin points on different available paths until reaching the desired destination points. This makes a great opportunity to analyze their dynamics and interaction on transportation networks using percolation and network theories.

The research work for this dissertation is carried out around two distinct problems that are commonly concerned with congestion and traveling flows on transportation networks.

We tackle both of these problems using ideas from percolation theory applied together with well-established techniques and concepts from network science and transportation engineering. The first problem is analyzing the reliability of on-road PT networks (bus-tram networks) under the congestion imposed by the road network condition. The second problem is traffic signal control to mitigate the propagation of congestion in road networks. In the following two sections, these problems are outlined and the main knowledge gaps related to each problem, which will be addressed via percolation approaches, are briefly discussed.

1.2 Reliability analysis of transportation networks

On-road PT systems are in constant conflict with traffic congestion on roads as trams and buses share the road space with vehicular traffic and pedestrian crowds. With regards to time-varying conditions on road networks, the reliability of on-road PT networks can be viewed as their ability to provide passengers with less congested passages between origin-destination locations. At a particular time of the day, congestion is distributed at different levels of intensity over the PT network links. Percolation theory allows for unpacking the hierarchical organization of congestion on the network. In other words, percolation analysis involves dissecting the network into parts exposed to different road conditions, and examining how these network parts come together to connect different locations. We mainly seek to achieve three objectives in tackling this problem:

- To formulate an analysis based on percolation theory that can account for different important factors governing the dynamics of on-road PT networks.
- The analysis should allow for quantifying the reliability of on-road PT networks under different travel demands and road conditions.
- Identifying the most critical links in the network, where the level of congestion has the most impact on the overall reliability of the network.

In transportation networks the uneven travel flow demand between different places is a major determinant of the global dynamics [46]. The signature of an urban transportation network is a particular flow demand and its daily or day-to-day evolution patterns. However, the demand is ignored by the existing percolation models which motivated us to improve the existing paradigm. Here, a new theoretical framework is developed, which involves the heterogeneity of the demand in percolation analysis of the network. By characterizing the organization of congestion on the network via percolation analysis, the relation between the demand and different levels of congestion can be monitored on the network. This allows us to measure the macroscopic reliability of the network based on the adversarial effect of

congestion on flow-movement with respect to the travel demand. In addition, we theoretically tie the microscopic congestion (i.e., congestion level on each link) to the macroscopic reliability of the network, which is rarely done in the existing relevant literature. Thereby, it is possible to find the most critical links (bottlenecks), most responsible for hindering the traveling flows all over the network.

A reason behind the scarcity of attempts in the literature to comprehensively analyze real-world networks for the conflict between passenger flows and congestion, is the unavailability of appropriate data sources. Here, we use large-scale passenger smartcard data collected in the on-road PT network of Melbourne and Brisbane, Australia. The detailed transportation data is used to digitally reconstruct the PT network, estimate the level of congestion on each link, and extract the passenger travel demand over time. To extract passenger travel demand from PT smartcard data, a parameter-free procedure is developed which unlike most existing approaches does not require extensive parameter tuning and is applicable to various smartcard data settings. The proposed percolation-based reliability analysis is applied to PT networks of Melbourne and Brisbane. The performed experiments suggest the effectiveness of the proposed reliability measurement and bottleneck identification approaches.

1.3 Traffic signal control in road networks

Perimeter signal control is a well-studied area of traffic engineering [47]. It refers to strategizing the action of traffic signals on the boundary between particular regions of the network. The goal of perimeter control is to regulate the traffic operation in the entire network. A simple and effective example, is controlling signals of intersections placed on the boundary of an important region of the network, say the city center, aiming to protect the region from being overflown by delaying and balancing the unwanted traffic outside the protected area.

A common issue with perimeter signal control is the development of congestion queues outside an active perimeter that is trying to reduce the inflow of the protected region inside it. A possible solution to this is implementing multiple concentric perimeters, where control at the innermost boundary aims at protecting the encompassed region but control at each larger boundary mitigates the development of congested queues outside the next boundary inside it. The issue with applying this solution is that congestion propagates in a complicated and heterogeneous manner which makes it a difficult task to determine the positioning of these extra perimeters to effectively combat the propagation. Working on this problem, We mainly pursue the below two objectives:

- To apply percolation theory in order to effectively characterize the propagation of con-

gestion resulted from gating at a perimeter signal control. In other words, the objective is to tailor the classic percolation-based analyses for this specific purpose.

- The second objective is to develop a control scheme that incorporates the percolation-based analysis of congestion when determining the timing of signals, so that the new control strategy effectively prevents the congestion propagation.

Most traffic signal control methods are developed to modify the timing of a spatially fixed set of intersections [47]. The control methods often function based on the condition on individual roads or measuring the traffic flux between regions neighboring the fixed set of controlled intersections. Even the state-of-the-art methods that account for congestion propagation use simple measurements such as the length of individual congested queues attached to the predetermined controlled intersections. Here, we improve this through characterizing the propagation of congestion using percolation analysis. A region of the road network can be protected by a fixed perimeter and the propagation of congestion outside this perimeter can be analyzed by percolation analysis. We propose a strategy to control signals at a time-varying second perimeter (outside the fixed perimeter). The boundary of this second perimeter is determined at each point in time to effectively mitigate the congestion propagation triggered by gating at the fixed perimeter inside it. The performance of the proposed strategy is demonstrated using micro-simulations on a grid traffic network.

1.4 Dissertation content overview

In this chapter, we provided an overview of the problem of congestion propagation in transportation networks and highlighted the research questions regarding congestion that require more attention. In particular we use percolation analysis on transportation networks to tackle two problems outlined in sections 1.2 and 1.3 of this chapter (see the leftmost panel in Fig. 1.2). In the next chapter, first, a brief background is provided on studying traffic congestion and network percolation analysis. It is then followed by a detailed review of the literature related to the defined two main problems. The first problem is studied on real-world transportation networks, built upon raw transportation data. Hence, the next chapter also provides a review of the literature on processing smartcard data and challenges to extract networks' travel demand from such data.

The first major research problem is tackled in Chapters 3 and 4. Chapter 3 is dedicated to the detailed methodology used to extract information from raw smartcard records, mainly, the passenger travel demand over the PT network. The chapter includes a brief introductory section followed by the below journal publication [48], which covers our work on PT smartcard data:

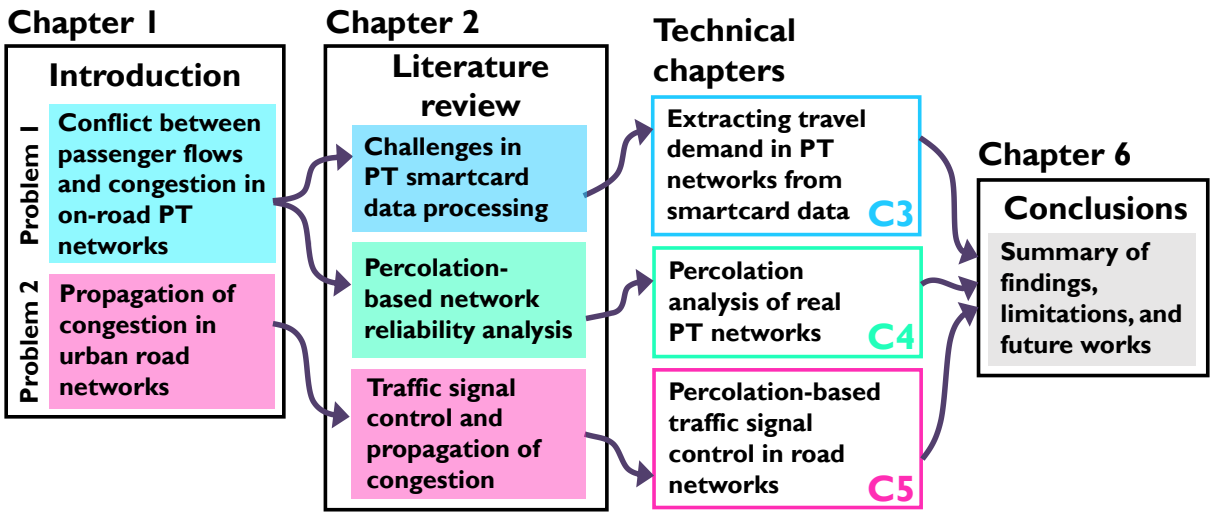


Figure 1.2: Dissertation content flowchart. The flowchart shows the content of the present dissertation, divided into 6 chapters (including three technical chapters). The important components of each chapter are highlighted above, and arrows indicate the logical flow in the material presented in different chapters.

> H. Hamedmoghadam, H. L. Vu, M. Jalili, M. Saberi, L. Stone, and S. Hoogendoorn, "Automated extraction of origin-destination demand for public transportation from smartcard data with pattern recognition," *Transportation Research Part C: Emerging Technologies*, vol. 129, p. 103210, 2021.

This work is complemented by the following Chapter 4, which presents our proposed percolation-based framework to assess the reliability and identify the bottlenecks of transportation networks with respect to the conflict between organization of congestion and passenger flows. This work is published in a journal paper accompanied by detailed supplementary information, and these two published documents form the chapter together. Below is the complete reference to this article[49]:

> H. Hamedmoghadam, M. Jalili, H. L. Vu, and L. Stone, "Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks," *Nature Communications*, vol. 12, no. 1, pp. 1–10, 2021.

The last technical chapter (Chapter 5) is written in standard form (no publication included) and presents our proposed methodology and results regarding the second main research problem, defined here. There, we optimize traffic in road networks using a signal control scheme that is based on percolation analysis of congestion propagation.

Below is the list of the following chapters in this dissertation. A brief overview of the content of each chapter is also provided below.

- **Chapter 2: Literature review.** Provides background to the theories and concepts

used in this project, and reviews the existing literature related to the problems addressed here.

- **Chapter 3: Smartcard data processing.** This chapter reports a part of the work on the first problem. It is mainly concerned with the method that we propose to extract the enriched network representation of PT systems from their raw smartcard data. The result is then used in the analysis presented in the following chapter.

- **Chapter 4: Percolation-based reliability analysis.** The chapter lays out our developed theoretical framework for percolation analysis of congested PT networks with respect to a certain heterogeneous travel demand between different points. The theories are formulated, proved, and their application and effectiveness are demonstrated on theoretical graph models and real-world systems.

- **Chapter 5: Percolation-based traffic signal control.** This chapter presents the developed strategy for signal control in road networks, which is built upon a percolation analysis of congestion propagation. The proposed methods are evaluated using simulations on synthetic road networks.

- **Chapter 6: Conclusions.** This chapter summarizes the studies presented in the previous chapter, mainly by encapsulating the main findings. It also describes the limitations of the technical approaches used and depicts a roadmap for possible future works to be built upon the work presented here.

Chapter 2

Literature Review

Physics of traffic have been studied for decades since the early observational measurements and analogies with fluid dynamics [50, 51, 52]. The theory is especially concerned with how vehicles, drivers' or travelers' behavior, and the environment affect the traffic flow [53, 54]. Perhaps the grand problem related to traffic phenomena is that of congestion. The existing literature has intensively studied the physics of congestion in a street or over a whole city, and also factors influencing congestion outside the realm of physics, such as commuters' and drivers' psychological behavior, are investigated [55, 56]. To date, however, only a few studies have been carried out on spatial dynamics of intra-urban traffic congestion, i.e., how different levels of congestion organize and evolve with respect to the topology of the city road networks [57, 58, 59, 60, 61]. Our efforts in this dissertation are focused on this area that was fairly neglected until receiving the recent attention in network science studies.

The structural evolution of Paris' road network over time as a result of infrastructure refinements and its effect on redistributing congestion is studied in [57]. A recent study [58] suggests that the interconnections between arterial and local roads (the way they are entangled) in the network explain the spatial transitions in road congestion (i.e., how congestion level varies). The same team of authors investigates the impact of the structural properties of city road networks on spatial transition of congestion levels [59]. They discuss the structural features that seem to be able to control congestion displacement in the network, suggesting that the findings have the potential to be applied in planning and developing optimal city road networks.

Percolation analysis is shown to be an effective tool in analyzing the organization of congestion against the network's geometrical properties. In [62], the organization of traffic congestion in a central area of Beijing is studied using percolation approaches, leading to identification of congestion bottlenecks and their temporal evolution in the network. Studying percolation properties of road networks has been able to successfully reveal the pivotal role of congestion level on long-range connections (e.g., highways) in determining the be-

havior of traffic flow circulation [60]. By applying percolation approaches to two real urban road networks (enriched with dynamic traffic data), it is shown in [61] that there exists a daily pattern in regime shifts between metastable states of the system, from which the operators can benecritical transition to low-performance states can be identified and treated by operators.

Just as the body of traffic flow theory, the recently growing literature on spatio-temporal organization of network congestion applies to the whole spectrum of transportation system analysis. They especially allow for profound understanding of the rules governing urban transportation dynamics which is the prerequisite to undertaking the impressive tasks of alleviating congestion and optimizing transportation via planning and operation in cities.

On the one hand, understanding the organization of congestion may be used to seek congestion alleviation in the area of transportation planning. This often involves one or more of the following conventional approaches: modifying the travel demand (e.g., based on the impact of land use or pricing strategies), modifying the capacity of the network (e.g., adding/removing roads or adjusting their widths), and modifying travelers' route or mode choice (e.g., through public or active transportation encouragement strategies) [4, 10, 63, 64, 65, 66, 67]. On the other hand, here our percolation-based analyses allow for acting against network congestion through practical solutions that do not require significant modifications of the transportation network and may be classified as operational improvements.

In particular, our first analysis leads to identification of bottlenecks links in on-road public transportation (PT) networks which can be treated by allocation of separated lanes for bus and tram vehicles. In the second study, percolation of congestion allows for identifying a time-varying congested cluster which can be treated well by adjusting the timing of traffic signals at its boundary. We will discuss both of these treatments in more detail in Chapters 4 and 5.

The remaining of this chapter is structured as follows. First, a brief historical background on percolation theory, and a preliminary technical introduction to percolation-based analysis of congested transportation networks is provided in Section 2.1. This is followed by a review of the existing literature on percolation-based network analysis (Section 2.2), where we also identify the gaps that we attempt to address in our work. Our first study is focused on analyzing PT networks based on real-world data, so in two subsections we cover the background related to percolation analysis for these networks (Section 2.2.1) and the challenges in processing PT smartcard data (Section 2.2.2). Our work on this study is then laid out by presenting the methodology and results of passenger smartcard data processing in Chapter 3 and the proposed network analysis in Chapter 4. The final section of this chapter (Section 2.3), provides the background and reviews the literature related to our second study on tackling the propagation of congestion by traffic signal control in road networks, with the study itself presented in Chapter 5.

2.1 Background on percolation analysis

The initial development of the classic percolation theory is credited to the works of Flory and Stockmayer [68, 69] on modeling polymerization, i.e. formation of a network of chemical bonds between basic molecular units of polymers (monomers) [70]. Close to two decades later, the theory was first named and framed more mathematically in a 1957 publication [71] by Broadbent and Hammersley [40]. Due to its simplicity and applicability to diverse problems, percolation had soon become a popular theory in the physics community.

Since the emergence of modern network science [72], percolation theory has found a great deal of interest and a large number of applications, especially when it comes to studying real-world networks. Network percolation analysis is often based on a simple percolation process simulating a gradual addition or removal (inverse percolation) of network nodes (site percolation) or links (bond percolation) [73], while the theory helps achieving a deeper understanding of the network system by interpreting the network's behavior during the percolation process. In particular, the statistical properties of the network under percolation process, reveals the geometrical and functional properties of the network system [44, 74]. Percolation models are heavily applied in the area of complex networks to address a variety of problems in different contexts [44], with examples including virus transmission in sexual contact networks [75], spread of information in social networks [76], prosperity of cooperation (and resolution of dilemmas) in human populations [77, 78], and wiring in human brain architecture [79, 80].

Although there are various approaches and the possibility of numerous objectives to percolation analysis, for our purposes, we study the network representation of transportation systems via percolation processes on network links guided by links' congestion dynamics. Let us imagine a transportation network, where nodes represent physical locations of interest and each link e_{ij} represents directed transportation between two points (node i to node j). Nodes (links) represent intersections (road segments) for road networks and represent stops (service between consecutive stops) for PT networks. This approach to model transportation systems is sometimes referred to as 'primal' network representation [81]. To augment the congestion dynamics on such a network representation, one approach is to have multiple network structures corresponding to different snapshots of the system in time, then for a particular time t, the level of congestion on each link e_{ij} can be simply represented by the link 'quality' attribute $q_{ij}(t)$. An example of such a network-snapshot in time is shown in Fig. 2.1.a, where values of quality attribute are color-coded on links.

Let us define the link quality attribute within the range $q \in [0, 1]$, inversely indicating the level of congestion. The link quality can be also interpreted as the quality (or level) of service for transportation between two nodes. The quality of transportation on a link is high (q close to unity) when the traffic is moving freely, and the link quality is low (q close to

zero) if the congestion level is high and slowing down the traffic. The quality attribute can be calculated as the ratio between the instantaneous transportation speed and the free-flow speed (or speed limit) on each link. Alternatively, the link quality may be calculated from relative link density (instantaneous density divided by the jam density of the link), relative link travel time, or more advanced measures of congestion level [82, 83].

The aim of our analysis is to characterize the organization of congestion over the congestion of the network. Broadly speaking, this is to understand how congested links are placed on the pathways between nodes or how pockets of congestion are forming, propagating, and dissipating on the network. To do so, a possible starting point is to divide network links into congested and free-flow (non-congested) classes, which requires defining a particular threshold for quality attribute of links, separating the congested regime (low q) from the free-flow regime (high q) [84]. Instead of studying congested links determined by a certain fixed quality threshold, say, links with quality below 0.3, we choose to study the congested links at all possible thresholds. Thus, starting with a very small quality-threshold, we can examine only extremely congested links, and then by increasing the threshold, less congested links and their relation to those extremely congested ones can be investigated. This will be a more comprehensive way of analyzing the network's congestion, and it can be naturally mapped to a percolation process which is based on the congestion level of network links. See two examples of such percolation processes illustrated in Fig. 2.1.b and Fig. 2.1.d.

We provide more details regarding the above explanations, especially in Chapters 4 and 5, but with this brief background, we move forward to review the literature related to the main problems tackled here, i.e., conflict with congestion in on-road PT networks and congestion propagation in road networks.

2.2 Percolation-based analysis of network reliability

Percolation theory has been frequently employed in modern studies of complex network systems to investigate the properties of natural [23, 85], technological [86], and social networks [87], especially in terms of their robustness and resilience to perturbations. In percolation-based approach to network robustness (or resilience) analysis, a percolation model is used to simulate a series of link/node failures or dysfunctionalities, by progressively removing links[1] (or nodes) from the network [41, 72]. The network's behavior during the percolation process indicates the system's response to link removals (simulating relationship terminations, connection failures, etc.), and this response can be monitored and quantified to

[1]Here, we only deal with percolation on network links, which is sometimes referred to as bond percolation.

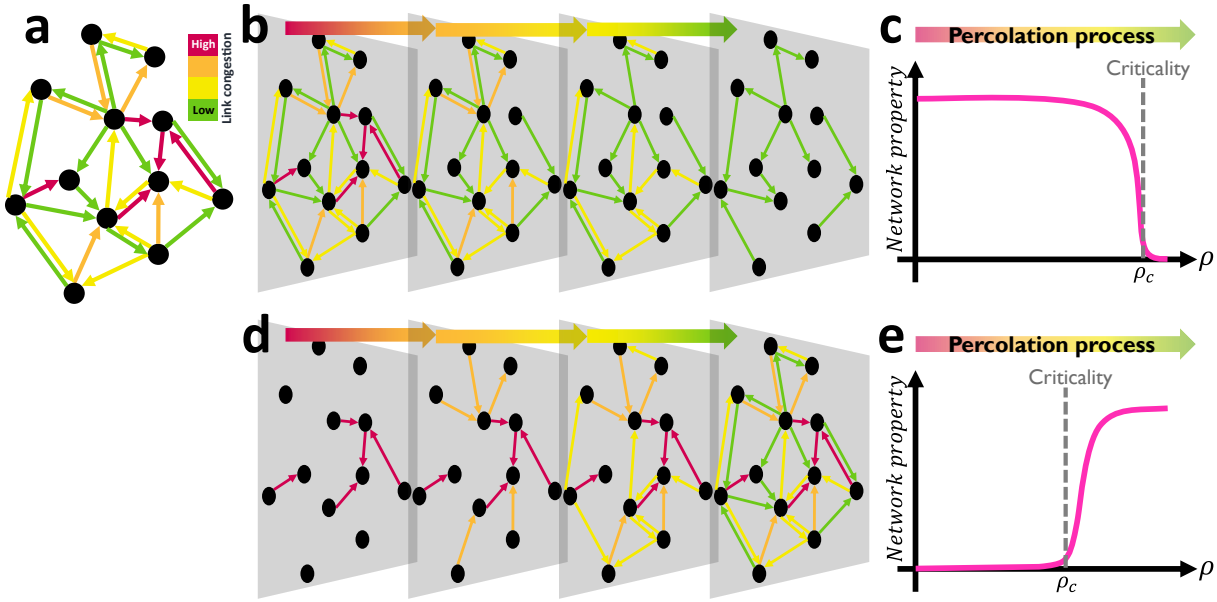


Figure 2.1: Percolation on an example network with link dynamics. **a** An example network representation of a transportation network, where link congestion level (color-coded in the figure) is modeled as a link attribute. **b,d** Example percolation models are applied to the network (shown in **a**) and the process is demonstrated at different points. A shell of most congested links is removed from (added to) the network in **b** (**d**) progressively. **c,e** Network's percolation behavior during the percolation process, shown via the evolution of one of its properties as a function of the threshold ρ. Different behaviors seen in **c** and **e** correspond to different percolation models depicted in **b** and **d**, respectively.

characterize the network's robustness [88, 89, 90].

A common approach to capture the percolation properties (reflecting the impact of link removals) is measuring change in the size of the network's largest connected component[2], or Giant Component (GC), during the percolation process [91, 92]. Different strategies for simulating link failures, make it possible to study a range of different network characteristics. Two widely-applied percolation-based analyses are based on random (simulating errors) and targeted (simulating attacks) link removals, respectively, to assess the vulnerability of the networks to unplanned faults in its parts and targeted attacks on its important parts of the system [25, 89, 93].

Let us formalize percolation on a network G (or G_0) as a process controlled by the threshold parameter ρ, gradually increasing from 0 to 1. We use G_ρ to denote the network under percolation at a particular threshold ρ. Typically, error and attack tolerance of networks are measured by monitoring G_ρ as ρ is gradually increased and simultaneously the fraction

[2]The largest connected component (or giant component) of a network is its largest possible subgraph (i.e., a graph, comprised of a subset of the network's nodes and all of the network links between those nodes), for which there exists a path between every pair of nodes. The size of the giant component is the cardinality of the set of its nodes.

ρ of all links are being removed from the actual network G to obtain G_ρ. Of special interest is often the critical threshold $\rho = \rho_c$ during the process at which the GC suddenly disintegrates into components of smaller size [94]. The percolation threshold ρ_c is a widely-used informative measure of the network's robustness, indicating that the network fails to provide global connectivity after critical fraction of its links are removed [41, 43]. Note that when the percolation process is both governed and monitored by the network's topological features, the percolation behavior of the network can only be used to characterize the structural properties of the network.

By modifying the mechanism or rules guiding the percolation simulation (e.g., [79, 95, 96]), more complicated network properties can be examined. For example, percolation process can account for that failure of some parts may trigger failure in new parts of the system. This is a common phenomenon in many systems ranging from infrastructure to human body. For example, in ecological networks extinction of a species endangers some others or in financial networks, failure of a bank puts some others at risk [85, 97]. Analyses based on such percolation models have been carried out widely to study cascading failure on networks, especially in the case of critical infrastructure networks [98]. In these networks, coupling between components of a certain network or interdependence between components of coupled networks can magnify the damage caused by an initial problem [26, 92, 99, 100]. As examples, failure of one connection can lead to overload in other links in a power grid, or failure(s) in a power grid can fail parts of its dependent transportation network.

In real infrastructure networks, pervasive phenomena such as various forms of congestion (e.g., packet congestion in communication or traffic jams in transportation) reduce the quality of flow movement on links in a continuous manner rather than necessarily causing complete link failure. Despite numerous valuable studies on the impact of failures on transportation networks [101, 102, 103], the hindering effect of congestion at different levels over these networks is rarely explored in the area of complex network analysis. By involving link-level dynamics (such as congestion) in the percolation process, the network's behavior under percolation will reflect its dynamical properties as well. As mentioned in Section 2.1, to consider this we follow the state-of-the-art percolation analyses, which model link-level flow dynamics on a network G by associating each link e_{ij} (connecting node i to node j) with its own 'quality' attribute $q_{ij} \in (0, 1]$ at each time [104, 105, 106].

The link quality indicates the temporal link performance relative to an observed or predetermined maximum level of performance. For example, in a communication network, link quality can be the instantaneous delivery rate of packets on a link [107]. In a transportation network, where susceptibility to congestion causes the speed on each link to change temporally, link quality q_{ij} can be defined as the ratio of instantaneous traffic speed to the speed limit of link e_{ij} [60, 105]. To involve the link-level dynamics in the analysis, percolation is simulated on such networks by increasing the threshold ρ from 0 to 1 and simultaneously

removing all links having $q_{ij} \leq \rho$ from the actual network G to obtain G_ρ [79, 108, 109]. See this process illustrated on a small network in Fig. 2.1.b. The network's behavior during this percolation process is symbolically depicted in 2.1.b. which in practice can be examined to characterize the network's topological and dynamical features.

2.2.1 Reliability of on-road public transportation networks

Consider the percolation process during which the links are gradually removed in an order determined by their quality attribute, i.e., inverse congestion level. The 'percolation threshold' (or percolation criticality) during this process can be defined as the threshold $\rho = \rho_c$ at which the GC disintegrates into components of smaller size. (See the percolation criticality marked in the example processes depicted in Fig. 2.1.c.) The percolation threshold ρ_c can be an informative measure of network's global quality, indicating that the network fails to provide global connectivity with links only having quality above ρ_c [41, 88, 107]. Note that rupture of network paths (or separation of network components) due to removal of congested links, reveals how the congestion is separating different places in the actual network. While the generic critical phenomenon is of vital importance for characterizing networks, we will show that limiting attention exclusively to the GC and its sudden disintegration reveals only a part of the full picture when studying real-world problems such as the conflict between passengers and congestion in transportation networks.

The primary goal in many critical infrastructures such as communication, power distribution, water supply systems, and transportation networks is to serve the demand for a certain amount of flow; we refer to such systems as *demand-serving* networks. In reality, the flow demand between Origin-Destination (O-D) node pairs is often distributed heterogeneously over the network. This is especially the case in transportation networks, where, for example, the travel demand is much larger between O-D points when one or both of them are hotspot locations [110]. The larger the passenger travel demand between two nodes, the more crucial are the paths connecting the two nodes.

When studying percolation in demand-serving networks, despite the loss of global connectivity at percolation criticality, there might be a substantial volume of flow (a large number of passengers) inside isolated components in subcritical phase. This highlights a problem with interpreting ρ_c as a reliability index (as per [62, 88, 111]) if the main interest is on heterogeneous passenger flow demand. For example, at percolation criticality the bulk of the passenger travel demand may be contained within small and medium-sized isolated clusters[3] (resulting from the disintegration of the GC). This reveals that most passengers

[3]Disjoint subgraphs of the actual network G, for which there remains no connecting path on G_{ρ_c}, i.e., the network under percolation at criticality.

are traveling between places that are not separated by higher levels of congestion ($q \leq \rho_c$), which means that the network is highly functional with only its less congested links ($q > \rho_c$) that remain unremoved after the GC collapse. In other words, the global dynamics in transportation networks (and in general in demand-serving networks), is not only controlled by the structure and organization of link congestion, but also by the distribution of the flow demand. This motivated us to develop a new approach to capture the reliability of transportation systems as heterogeneous demand-serving networks.

Our goal is to add further realism to percolation-based transportation network analysis by inclusion of heterogeneous passenger travel demand. The concept of travel demand distribution is fundamental to transportation theory [112], but only in our work it has been involved in percolation-based analysis of dynamical transportation networks (see the first objective stated in 1.2). In Chapter 4, we provide detailed description of a percolation-based framework to analyze the conflict between passenger flows and congestion on on-road PT networks. Thereby, we measure the reliability of these networks (second objective in 1.2) and identify the most critical links to enable optimal targeted improvements (third objective in 1.2).

Application of the proposed framework is demonstrated on the bus and tram PT networks in two major Australian cities, Melbourne and Brisbane, modeled using smartcard transaction data collected during September and October 2017 in Melbourne and over March 2013 in Brisbane. We use the large-scale real smartcard data, first, to digitally reconstruct the temporal network representation of on-road PT systems enriched with link-level congestion information. Also, we are especially interested in extracting the node-to-node passenger flow demand from the smartcard data, which is often represented as a matrix referred to as O-D travel demand matrix or simply 'O-D matrix.' The next section reviews challenges and the existing literature related to extraction of O-D demand from smartcard data. We report the methodology and results regarding the smartcard data processing in Chapter 3, before we lay out the details on the proposed percolation-based reliability analysis and demonstrate its application to the data-driven transportation networks in Chapter 4.

2.2.2 Demand extraction from passenger smartcard data

Passenger smartcard data are collected by Automated Fare Collection (AFC) systems implemented in PT networks. AFC systems rely on the accuracy of their equipment and passengers' interaction with them, both of which are liable to perform erroneously. As a result, PT smartcard data are often contaminated with inaccurate or missing information. Missing information can substantially lower the quality of smartcard data and the passenger travel demand extracted from it. Missing transactions, especially missing 'alighting' transactions, is the major problem in our available smartcard data from Melbourne's PT network.

Statistical inference based on passengers' transaction history has been proved to be effective in estimating missing alighting transactions [113, 114, 115]. Also, there is consensus in the literature that a missing transaction from a particular passenger can be estimated with high accuracy if that passenger's immediate previous and following transactions are available [116, 117, 118, 119].

Often, estimation of missing alighting transactions has to be performed by relying on a set of preliminary assumptions on passengers' travel behavior. The common assumptions used in majority of the existing methods are: i) at the end of each day, passengers return to the first boarding location on the same or the next day, known as the 'day's symmetry trip assumption,' ii) a missing alighting point is most likely in a convenient walking distance from the next boarding stop, iii) passengers alight at the closest possible point (*closest* option) to the location of their next boarding, or alternatively, at the point which leads to the earliest arrival (*fastest* option) to their next boarding location [114, 120, 121, 122].

These assumptions are not consistently applicable to data from different cities. Regarding the first assumption, for example, different PT networks have different schedules and time-varying service supply, which can lead to different passenger travel behaviors at the start and the end of a day. A convenient walking distance to access PT stops (the second assumption), depends on factors such as availability of alternative transportation, PT network design, walking environment, and passengers' active travel behavior [123, 124]. Different values chosen by the existing studies as the maximum distance that passengers walk to transfer between two stops, attests to the variability of this parameter with respect to different environments; e.g., 750 m for London [125] and 400 m for Brisbane [126].

Another major challenge in generating an OD demand matrix from individual PT trips is understanding the passengers' 'trip chaining' behavior and aggregating rides belonging to the same passenger journey [127, 128, 129]. More specifically, the problem is to identify the sequence of single-leg PT trips that a passenger had taken to move from an initial origin point to undertake an activity at the final destination point [130, 131]. A chain of PT rides, linked to one another by transfers for the purpose of an activity at the last destination, is often called an O-D trip or a 'journey.' The time between two consecutive PT trips may be associated with either a 'transfer' within the PT system (passenger walking between the stops and waiting for the next service), or an 'activity' outside the PT system. The challenge is to process alighting-then-boarding (interchange) incidents appearing in the data and to categorize them into transfers and activities. Then, consecutive single-leg trips of each particular passenger connected via interchanges identified as transfers will be aggregated into a single journey. The passenger journeys (as opposed to passenger trips) are the appropriate input to the process of generating the OD matrix, as they accurately describe the demand for travel between origin and destination locations.

Duration of an interchange event, between two consecutive trips taken by a single pas-

senger, is sometimes referred to as Inter-Transaction Time (ITT). As transfers typically have shorter ITTs compared to activities, the common practice is to determine a fixed time-threshold and identify each interchange event as a transfer (an activity) if its corresponding ITT is smaller (larger) than the threshold. Some existing studies have chosen a fixed time-threshold within a wide range, from 30 to 90 min, based on the expert knowledge of the particular PT system under study [116, 132], while others have explored the outcome of different thresholds and used the one leading to more favorable results [117, 133]. Another group of studies derive a set of quantitative rules and constraints governing the passengers' trip-chaining behavior, and then test each interchange event against those rules to decide if it should be identified as a transfer or an activity [125, 126, 129].

In our proposed methodology to extract the travel demand from smartcard data, our contribution is toward elimination of system-specific assumptions. We develop a procedure that minimizes the need for expert knowledge and manual parameter setting for estimating the missing alighting transactions and identifying transfers/activities. In order to estimate missing alighting transactions, we develop a procedure that uses the information available in the smartcard data to model passengers' choice between possible alighting options that have different advantages over one another. Thereby, for missing alighting information from a trip, the model is used to predict the passenger's alighting time and stop.

To identify the transfers/activities between consecutive passenger rides, we follow the common approach of determining an ITT threshold. However, we propose a simple method that systematically derives an appropriate ITT threshold for a PT network from the patterns in the smartcard data. The idea is mapping this problem to a binary classification problem, where interchange events described by their ITTs should be classified into two classes: transfers and activities. Our approach is to first derive the statistical properties of transfers and activities from the data, and then, to build an optimal classifier which is able to effectively identify the transfers and activities between each passenger's consecutive rides. The proposed procedure processes and enhances the smartcard data and derives the passenger O-D demand, and is applicable to various PT smartcard data settings.

2.3 Percolation-based signal control in urban road networks

Traffic signal control refers to adjusting the timing of road traffic signals to optimize the traffic operation of the entire network [47, 134, 135]. The classic perimeter control regulates the operation of traffic flows via traffic signals at the fixed boundaries between selected partitions of the network to maximize the efficiency of the entire network. Traffic flow dynamics in a region of the network can be parameterized through the relationship between the vehicular density and the vehicular flow within the region, explained by the so called Macroscopic Fundamental Diagram (MFD) which is also known as the 'network fundamen-

tal diagram' and 'network exit function' in the literature [136, 137, 138]. The MFD describes the network's traffic flow dynamics within two regimes marked by a critical density, below (above) which the addition of vehicles increases (decreases) the vehicular flow in the network (see Fig. 2.2). Perimeter signal control aims at restraining the density of the network below the critical point of the MFD to prevent the overall flow from declining. With the examples including [139, 140, 141, 142, 143, 144], such control strategies are abundant in the literature.

Recent studies have made extensive efforts in this research direction, and the state-of-the-art methods are able to determine the optimal perimeter entry flow for single- and multiple-regions, while recognizing and treating the heterogeneity of congestion and the delayed effect of control on local queues at the perimeter. Representative studies include the application of perimeter flow control in multi-region cities [143, 144, 145] and multi-modal networks [146, 147], and also those analytically incorporating the local traffic performance into the network-level control algorithms [148, 149, 150].

While perimeter control can be very effective in keeping a target region free of congestion, its gating principle imposes a spillback effect on the approaching roads to the boundary of the region [151]. The resulting dynamic queue spillbacks can cause congestion outside

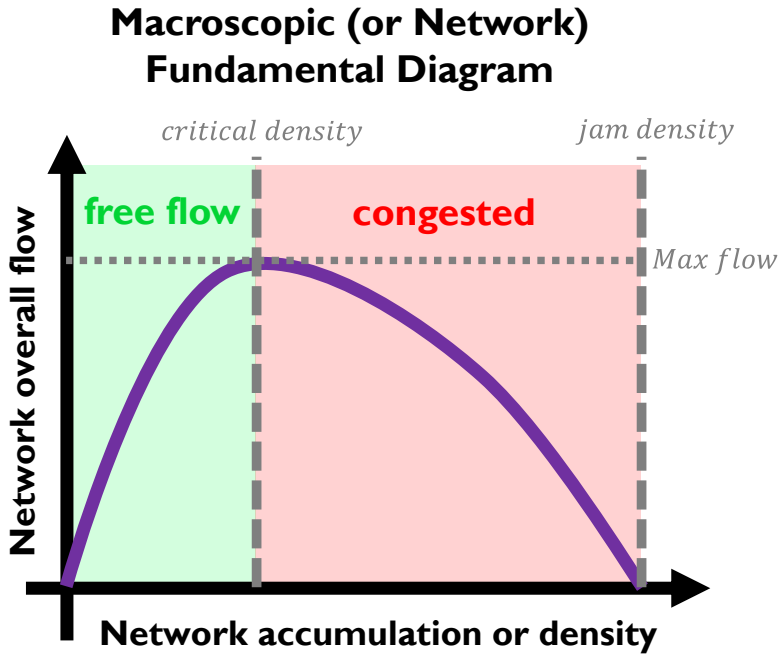


Figure 2.2: Macroscopic fundamental diagram for urban traffic. A symbolic depiction of macroscopic fundamental diagram, which relates vehicular flow [veh/h] to the vehicular accumulation [veh] (or density [veh/km]) in urban road networks (or a region of the network). The critical density marks the separation of the free-flow from the congested regime; the critical value is characteristic of the network's topological and infrastructural properties.

the region and cancel out the benefit of the perimeter control at a global level. We take a new step toward resolving the spillback effect of perimeter control, by integrating the spatio-temporal variations of congestion into the control scheme. Some initial efforts have been made towards this direction. For example, [148] accounted for the total length of congested queues at the perimeter when adjusting the metering (limiting the flow entry) rate, [149] developed a multi-scale control scheme where the metering rate given to local signals at the perimeter is queue-dependent. In [150], it is argued that since local queues affect congestion dynamics and thus the target of control, it would be more efficient to adapt the size and location of perimeter control in real time. Existing control schemes in the literature have rarely touched on this type of control scheme. This is mainly because i) such control requires detailed modeling of the local queue dynamics, which can impede the merits of a parsimonious modeling approach such as those based on the MFD, ii) a size- and control-changing network may have MFDs with different shapes that affect system dynamics and control performances, and iii) traffic control based on dynamic network partitioning is yet an open question in macroscopic traffic flow community.

Motivated by this challenging problem, we propose a multi-perimeter control scheme that uses percolation-based analysis to characterize the propagation of congestion and respond to it accordingly. In the proposed scheme, we assume a classic fixed perimeter control, implemented at the boundary of a hotspot region attracting substantial traffic; this can be the central business district of a city or an activity/shopping center. Although this classic approach is known to be effective in optimizing the traffic within the protected region and even improving the network's overall flow [140], we are interested in mitigating the spillback effect of the perimeter and regulating the propagation of congestion at its upstream.

By representing static (e.g., road capacity or node connection) and dynamic (e.g., shockwaves or traffic state) properties of the traffic system in its network model, percolation analysis is able to characterize the evolution of small pockets of congestion growing into a congested cluster of substantial size. (See the first objective stated in Section 1.3.) Figure 2.1.d illustrates a percolation model based on link congestion-levels which can be used for such an analysis; this is actually the simple representation of the model we use in our proposed control scheme.

The spatio-temporally evolving traffic congestion around the fixed perimeter is an essential input to the proposed controller. Thus in our control scheme, a second perimeter on top of the single-region perimeter control will be triggered with the aim of preventing the phase transition from small pockets of congestion to a large congested cluster as a result of queue spillback from the fixed perimeter. The formation and boundaries of this congested cluster are determined through percolating analysis of congestion (see the symbolic example in Fig. 2.1d,e) at different points in time, leading to a time-varying second perimeter. (See the second objective stated in Section 1.3.)

In Chapter 5, we show that controlling the travelling flows in a hierarchical manner both at the fixed perimeter around the primary hotspot region and then at the perimeter of the time-varying buffer space, leads to improvement of overall traffic flow over the network. Building upon the pioneering works on percolation-based analysis of transportation systems [60, 62, 101, 106, 107], our proposed method is the first attempt to examine congestion dynamics using percolation approaches for the purpose of road traffic signal control. The application of the percolation analysis in signal control for urban road networks is demonstrated in detail in Chapter 5 of this manuscript.

Chapter 3

Smartcard Data Processing

The conflict between congestion and passenger flows is a significant issue in transportation networks which is worthy of attention from various aspects. In this dissertation, our aim is to achieve a better understanding of this conflict and devise mitigating solutions to the problem through application of percolation theory. In particular, in the first study, the analysis is performed on real on-road (bus and tram) Public Transportation (PT) networks, where sharing the road space involves the operation of the PT system with road congestion. This chapter is the first of the two chapters covering this study. Real passenger smartcard data, collected in public transportation networks of two Australian cities, are used to demonstrate the real-world application of the proposed analysis. Processing smartcard data involves certain challenges concerned with enhancing the quality of the data and extracting useful information from raw transaction records. In this chapter, the methodology used to estimate missing transactions and extract the passenger travel demand from PT smartcard data is presented by a journal publication [48]. The enhanced data and its products, especially the travel demand, are important ingredients of the analysis proposed and performed in the next chapter.

3.1 Publication

The rest of this chapter is covered by the following article:

> H. Hamedmoghadam, H. L. Vu, M. Jalili, M. Saberi, L. Stone, and S. Hoogendoorn, "Automated extraction of origin-destination demand for public transportation from smartcard data with pattern recognition," *Transportation Research Part C: Emerging Technologies*, vol. 129, p. 103210, 2021.

Transportation Research Part C 129 (2021) 103210

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Automated extraction of origin-destination demand for public transportation from smartcard data with pattern recognition

Homayoun Hamedmoghadam [a,*], Hai L. Vu [a,*], Mahdi Jalili [b], Meead Saberi [c], Lewi Stone [d], Serge Hoogendoorn [e,a]

[a] Department of Civil Engineering, Faculty of Engineering, Monash University, VIC 3800, Melbourne, Australia
[b] School of Engineering, RMIT University, VIC 3000, Melbourne, Australia
[c] School of Civil and Environmental Engineering, UNSW Sydney, NSW 2032, Sydney, Australia
[d] Mathematical Sciences, School of Science, RMIT University, VIC 3000, Melbourne, Australia
[e] Department of Transport and Planning, Delft University of Technology, 2628 CN, Delft, the Netherlands

ARTICLE INFO

Keywords:
Smartcard data
Public transportation network
Origin–destination matrix
Passenger travel demand
Pattern recognition
Destination inference
Transfer identification
Trip chaining

ABSTRACT

Origin-destination travel demand matrix is the signature of travel dynamics in transportation networks. Many fundamental analyses of transportation systems rely on the origin–destination demand matrix of the network. Although extraction of origin–destination travel demand for public transportation networks from ticketing data is not a new problem, yet it entails challenges, such as 'alighting transaction inference' and 'transfer identification' which are worthy of further attention. This is mainly because the state-of-the-art solutions to these challenges, are often heavily reliant on network-specific expert knowledge and extensive parameter setting, or multiple data sources. In this paper, we propose a procedure that effectively applies statistical pattern recognition techniques to address the main challenges in extracting the origin–destination demand from passenger smartcard records. Learning from patterns in the available data allows the procedure to perform well under minimum case-specific assumptions, thus it becomes applicable to smartcard data from various public transportation systems. The performance of the proposed framework is tested on a dataset of over 100 million smartcard transaction records from Melbourne's multi-modal public transportation network. Evaluations on different aspects of the proposed procedure, suggest that the identified tasks are well addressed, and the framework is able to extract an accurate estimation of the origin–destination demand matrix for the system.

1. Introduction

Timestamped location is now constantly recorded on land, sea, and in the air, with hundreds of millions of GPS-equipped devices from personal smartphones to hardware serving public systems. Primarily, this can be attributed to the popularity of location-aware services, and that the value of longitudinal data applications has been increasingly materialized (Ferraro and Aktihanoglu, 2011; Hazas et al., 2004). Among countless applications of spatio-temporal data, those regarding urban human mobility are of special significance. Human mobility management is one of the most important aspects of smart cities, enabling efficient public services and optimal use of resources and assets (Batty et al., 2012). With applications of automated data collection being recognized and the recent advancements

* Corresponding authors.
E-mail addresses: homayoun.hamed@monash.edu (H. Hamedmoghadam), hai.vu@monash.edu (H.L. Vu).

https://doi.org/10.1016/j.trc.2021.103210
Received 13 September 2020; Received in revised form 2 April 2021; Accepted 3 May 2021
Available online 3 June 2021
0968-090X/© 2021 Elsevier Ltd. All rights reserved.

in information technology solutions, Public Transportation (PT) systems are progressively shifting from paper ticketing to contactless smartcards ticketing. Pervasive PT smartcard data are a rich source of information which can substitute data collected through time- and resource-consuming traditional processes (Pelletier et al., 2011; Utsunomiya et al., 2006; Zhu et al., 2018). Yet, when it comes to effective processing of smartcard data, there are still some gaps to be filled.

A primary product of mining PT smartcard data is the Origin-Destination (OD) travel demand, which in the past could be only generated through expensive surveys (Ickowicz and Sparks, 2015; Wong et al., 2005). Often described by a matrix, the OD demand is the volume of passengers traveling from each origin to each destination location via the transportation system. The OD demand matrix is an essential input for many transportation network analyses, such as studying passengers' route and mode choice behavior, modeling passenger loads and flows in transportation supply systems, and measuring the efficiency of transportation networks (Hamedmoghadam et al., 2021; Hamedmoghadam et al., 2019; Mohamed et al., 2016; Shafiei et al., 2020; Zhu et al., 2018). The procedure of estimating the OD matrix from smartcard data can vary depending on the PT network design, its fare collection system design, and the employed method of data collection. However, OD estimation from almost any smartcard data involves two major challenges, namely, i) missing and inconsistent records in the data as a result of either ticketing system fault or human error, and ii) the absence of any explicit indicator in the data on whether a sequence of rides by a passenger are aimed at visiting multiple destinations (unlinked trips) or only reaching a single destination (linked trips). Respectively, the above problems are often dealt with through performing two tasks known as: i) 'inference of missing alighting transactions' and ii) 'identification of transfers/activities.'

Our approach in this study, is to merely rely on information available from the raw smartcard data to derive the OD passenger travel demand of the PT system. The state-of-the-art approaches, not only sometimes rely on additional data sources such as the time-table of the PT services, but also often take a set of assumptions and parameters which require the expert knowledge of the particular PT network under study (Hussain et al., 2021). Relying on system-specific assumptions and parameters, reduces the generalizability of the OD estimation process. Our proposed approach, however, uses statistics to learn from the available data and infer what used to be manually set or assumed based on expert knowledge. In particular, our main focus is eliminating two commonly predetermined assumptions (or parameters), one about the passengers' choice of alighting stop and the other about the duration of transfers between linked trips. Each of these assumptions play a pivotal role in addressing the two major OD estimation tasks identified above. The idea here is to map each task to a classification problem. Then, instead of taking advantage of any predetermined assumption or parameter, we find the optimal solution to each classification problem, through an automated process of pattern inference from the observations provided by the available data. The effectiveness of the proposed framework is demonstrated by applying it to passenger trip data from the large-scale PT system of Melbourne, Australia. Melbourne's PT system functions in three modes, namely, train, tram, and bus, and it is equipped with an Automated Fare Collection (AFC) subsystem where passengers' payments are made by a contact-less smartcard, named 'Myki.' The available smartcard data used in this study are collected during a two-month period, containing over 100 million smartcard transaction records.

The rest of the paper is organized as follows. First, the background on the two well-known challenges in OD matrix extraction from PT smartcard data is provided through a review of the existing relevant literature. These two problems are 'inference of missing alighting transactions' to enhance the PT smartcard data and 'identification of transfers/activities' to perform trip chaining. Then, the *Methodology* section lays out the proposed procedure to perform the above tasks and extract the OD matrix of a PT network from its smartcard transaction records. Next, a section is devoted to description of the smartcard dataset used in this paper which also covers the data preparation stage in detail. Finally, the proposed methodology is applied to the pre-processed PT smartcard data and the results are presented, validated, and discussed.

2. Literature review

2.1. Inference of missing alighting transactions

Smartcard usage in different PT systems is often following one of the two common schemes. One scheme requires passengers to validate their smartcards only when entering the system; such a system is sometimes called a 'tap-in system' (He and Trépanier, 2015). In the other scheme, passengers are required to tap their smartcard also when exiting from the system (tap-in and -out). Also, the fare policies of a particular PT system impact the way passengers interact with the AFC system. For example, in systems where tapping-out is optional, a distance-based pricing encourages passengers to tap-out while flat-rate pricing makes it unnecessary, and thus, reduces the passenger alighting information in the collected data. Regardless of the validation scheme in use, the smartcard data is often contaminated with inaccurate and/or missing information. This is mainly due to the nature of the AFC subsystem which relies on the accuracy of both its equipment and passengers' interaction with them. Furthermore, the quality of the data collected by AFC systems is largely dependent on the manner in which the AFC system is implemented in the PT system. We discuss the latter in more details in the *Data Preparation* section.

Missing information in PT smartcard data lowers the quality of the resulting data-driven products. In datasets containing only the boarding information, statistical inference based on the history of transactions in the network have been shown effective in estimating passengers' alighting transactions (He and Trépanier, 2015; Ma et al., 2012; Trépanier et al., 2007). Also, there is consensus in the majority of the existing studies that a missing transaction from a passenger can be estimated with high accuracy using the passenger's immediate previous and following transactions (Alsger et al., 2016; Ma et al., 2013; Munizaga et al., 2014; Wang et al., 2011). The evaluation process in many existing methods for estimating missing transactions is limited to reporting the proportion of missing transactions that they can handle, as there is often no ground-truth data to compare the estimation results against (He and Trépanier, 2015; Munizaga and Palma, 2012). Alternatively, the OD matrix extracted using the estimated transactions, can be validated by

checking its consistency with another available source of travel information (Barry et al., 2009). In this paper, the available data contain the boarding and alighting information of a sufficient number of trips. This enables us to properly evaluate the accuracy of the proposed method by masking different portions of the data and then comparing the estimation results against the available ground-truth.

Estimation of missing alighting transactions is usually performed based on a set of preliminary assumptions on passengers' travel behavior. The assumptions commonly used in most of the existing methods include: i) at the end of each day, passengers return to the location of their first boarding of the day (or the next day), ii) a missing alighting point is most likely in a convenient walking distance from the next boarding stop (a maximum convenient walking distance is determined according to a rule of thumb), iii) passengers alight at the closest possible point to their next boarding location, or in an alternative approach, at the point which leads to the earliest possible arrival to their next boarding location (Kurauchi and Schmöcker, 2017; Munizaga and Palma, 2012; Trépanier et al., 2007).

The first assumption may result in different outcomes when applied to data from different cities. For example, different PT networks have different schedules and time-varying service supply, which can lead to different passenger travel behaviors at the start and the end of a day. A convenient walking distance to access PT stops (the second assumption), heavily depends on features such as availability of alternative transportation, PT network design, walking environment, and passengers' active travel behavior (Estgfaeller et al., 2017; Guo, 2009). This makes this parameter specific to each particular urban environment, and it can even vary between different local areas in a single urban environment. As a result, different values have been chosen by the existing studies as the maximum distance that passengers walk to transfer between two stops; e.g., 750 m maximum walking distance for London (Gordon et al., 2013) and 400 m maximum walking distance for Brisbane (Nassir et al., 2015). Our proposed methodology eliminates these conventional assumptions, yet, as it will be demonstrated via detailed evaluations, it performs effectively. However, if a behavioral trait is well-established for passengers in a specific PT network, it can be easily considered in our framework.

In order to estimate missing alighting transactions, we pay particular attention to the third assumption. For a passenger on-board of a PT vehicle and intending to take a subsequent PT trip at a particular boarding point, two alighting options are considered to be of particular importance. The first option is alighting at the closest stop to the next boarding point, i.e. the alighting point offering the least amount of walking. The second one is to alight at an earlier-visited stop, which requires more walking but leads to the passenger's earliest possible arrival to the next boarding point. For simplicity, let us refer to the former as the *closest* alighting point, and the latter as the *fastest* alighting point. As mentioned, different studies choose one over the other according to their own reasoning, but as our results will show, the effect of this naïve assumption about passengers' behavior should not be undermined. Here, we develop a procedure that uses the information available in the smartcard data to model passengers' choice between a pair of possible alighting time-stop points with different advantages over one another. Thereby, when alighting information is missing from a trip, the model is able to predict the passenger's choice in every pair of points from the set of all possible alighting points, which allows us to predict passenger's final choice of the alighting point. This choice behavior prediction is the main component of our proposed method for estimating the missing tap-off transactions in PT smartcard data. What we discussed here, makes an important stage of our framework applicable to any PT smartcard data with no adaptation or manual parameter-setting required.

2.2. Identification of transfers and activities at interchanges

Another major challenge in generating an OD demand matrix using the information from individual PT trips is understanding the passengers' 'trip chaining' behavior and aggregating rides belonging to a single passenger journey (Adler and Ben-Akiva, 1979; Primerano et al., 2008; Robinson et al., 2014). More specifically, the problem is to identify a sequence of PT trip-legs taken by a single passenger to move between two anchor locations for the purpose of undertaking an activity at the anchor destination point (Li et al., 2018). The anchor locations are the boarding point of the first trip and alighting point of the last trip in the chain of trip-legs taken by that single passenger (Alsger et al., 2015). A chain of single-leg trips, linked to one another by transfers between different PT services or modes for the purpose of an activity at the last destination, is often called an OD trip or a journey. The time between two consecutive PT trips may be associated with either i) passenger walking between the stops and waiting for the next service which is considered as a 'transfer' within the PT system, or ii) passenger undertaking an 'activity' outside the PT system. So, the challenge is to process alighting-then-boarding (interchange) incidents recorded in the data and then, to categorize them into transfers and activities. Any sequence of single-leg trips taken by a particular passenger should be aggregated into a journey (OD trip) if the interchange events between those trip-legs are identified as transfers. The passenger journeys (as opposed to passenger trips or rides) are the appropriate input to the process of generating the OD matrix, as journeys accurately describe the passengers' demand for movement between origin and destination locations.

Duration of an interchange event, between two consecutive trips taken by a single passenger, is sometimes referred to as Inter-Transaction Time (ITT). ITT is one of the most important descriptive features of an interchange event and is often used to determine whether an interchange event is associated with a transfer or an activity. As transfers typically have shorter ITTs compared to activities, the common practice is to determine a fixed time-threshold and identify each interchange event as a transfer (an activity) if its ITT is smaller (larger) than the threshold. Some studies have chosen a fixed time-threshold often within a wide range from 30 to 90 min based on the expert knowledge of the particular PT system under study (Munizaga et al., 2014; Nassir et al., 2011), while others have explored the outcome of different threshold choices and finally used the threshold leading to more favorable results (Alsger et al., 2016; Alsger et al., 2018). A group of studies, has focused more on improving the accuracy of transfer and activity identification (Gordon et al., 2013; Nassir et al., 2015; Robinson et al., 2014). They derive a set of quantitative rules and constraints governing the passengers' trip-chaining behavior, and then test each interchange event against those rules to decide if it should be categorized as a transfer or an activity.

Here, we follow the common approach of determining an ITT threshold to identify the transfers/activities between consecutive passenger rides. However, we propose a method that for the first time systematically derives the appropriate transfer/activity ITT-threshold for a PT network from its smartcard transaction data. The idea is mapping this problem to a binary classification problem, where interchange events described by their ITTs should be classified into two classes: transfers and activities. Conforming with the broad aim of this study, our generalized approach first derives the statistical properties of transfers and activities from the data. This allows for building an optimal classifier which is able to effectively identify the transfers and activities between each passenger's consecutive rides.

3. Methodology

Here, we present our general approach to address the two main tasks in estimating OD travel demand of any PT networks from its smartcard data. An overview of our framework is illustrated in Fig. 1. The first main task is estimating the missing alighting information in bus and tram modes, and the second task is identifying transfers to connect the linked passenger trips (marked as Task I and Task II in Fig. 1). The framework follows a standard process, starting with a data preparation stage (the top panel in Fig. 1) which is detailed in the *Data Preparation* section, where the smartcard dataset used in this study is also introduced. The preparation stage produces a cleaned smartcard record dataset and trajectories of PT vehicles, which become the inputs to the next stages. Next, each of the two major tasks is solved through an exploratory data analysis and modeling procedure, followed by a model deployment stage (see Task I and Task II in Fig. 1). In this section, the pivotal second stage is detailed, which is aimed at characterizing the patterns in the data and mathematically modeling them, in order to address each of the major tasks.

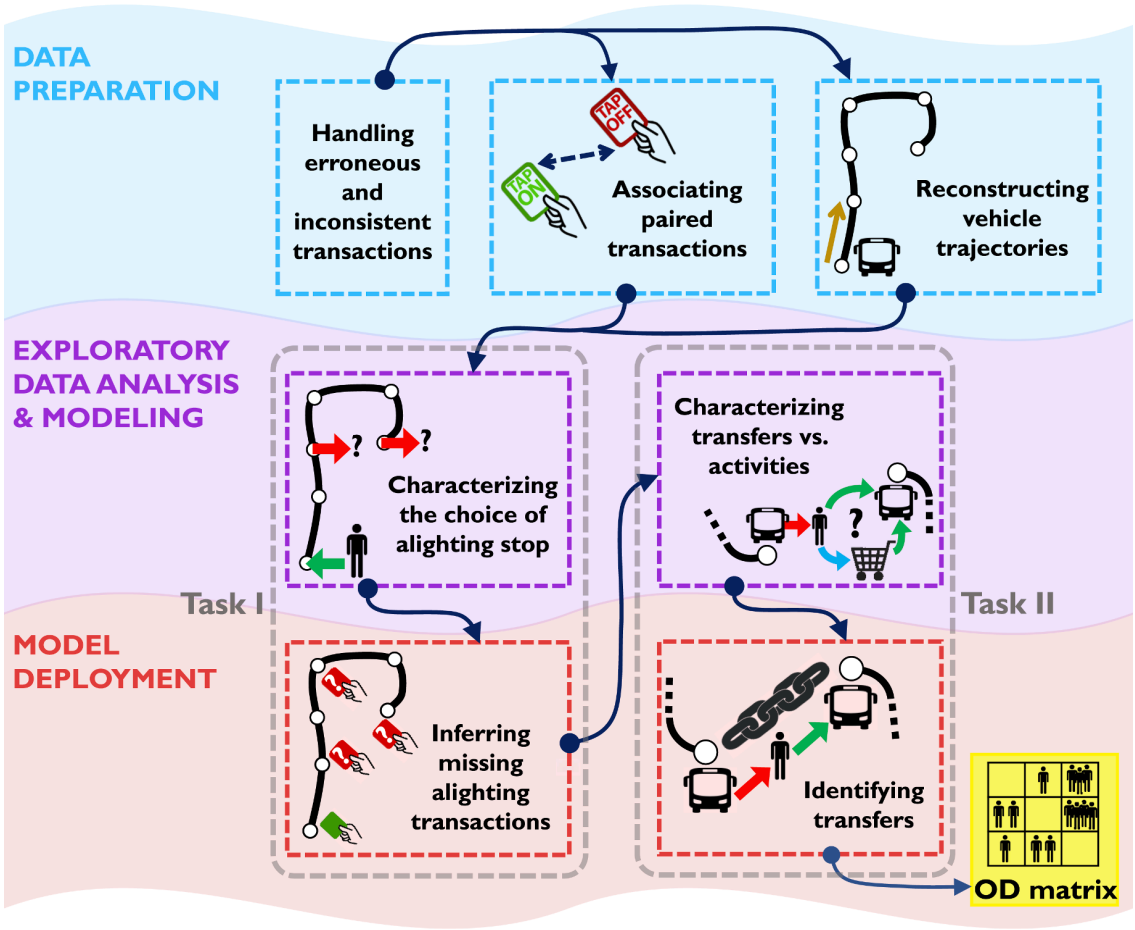


Fig. 1. Overview of the proposed framework. Three main stages in the framework are seen in different colors. Dark blue arrows show the workflow, starting with smartcard data pre-processing and ending with the origin–destination matrix of the public transportation network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Characterizing the choice of alighting stop

We propose a procedure to estimate the missing alighting transactions (Task I in Fig. 1) on bus and tram modes. Inputs to the procedure are the cleaned smartcard data and trajectories of bus and tram vehicles. The estimation method is applicable where a passenger tap-on record has a missing tap-off pair, but it is immediately followed by another tap-on record from that passenger during the same or the very next day on any PT mode. As pointed out in the previous section, we do not uphold the commonly-used assumption that a missing tap-off has to be within a predefined radius from the next boarding point. More importantly, unlike most existing procedures, we do not assume that all passengers choose to alight at the stop with minimum walking distance to their next boarding point (closest choice), or alternatively, at the stop leading to the earliest arrival to their next boarding point (fastest choice). The proposed procedure considers all possible choices for the passenger to alight at, and predicts the passenger's choice for each missing alighting transaction using a model that is built from the available trip data.

3.1.1. Finding plausible alighting points

Suppose that for a particular passenger a valid tap-on transaction is recorded on a bus or tram vehicle associated with the vehicle identifier *vid*, but the passenger's paired tap-off transaction is missing from the records. Also, the next recorded transaction for the same passenger is a tap-on during that very day or the next day on any PT mode. Let us denote the two consecutive boarding transactions, respectively as R_{b0} and R_{b1}, where $R_{bi} = (s_{bi}, t_{bi})$ is an ordered pair indicating the stop identifier and the timestamp associated with the transaction. Figure 2a illustrates an example of this scenario from real data, where R_{b0} and R_{b1} are recorded at green and red stops, respectively.

Our method begins by following the trajectory of the embarked vehicle *vid*, which can be represented by a sequence of stop-visits $TR_{vid} = (TR_1, TR_2, \ldots)$, where each stop-visit is a tuple $TR_i = (s_i, t_i^a, t_i^d)$ with its elements indicating the stop identifier, arrival time, and

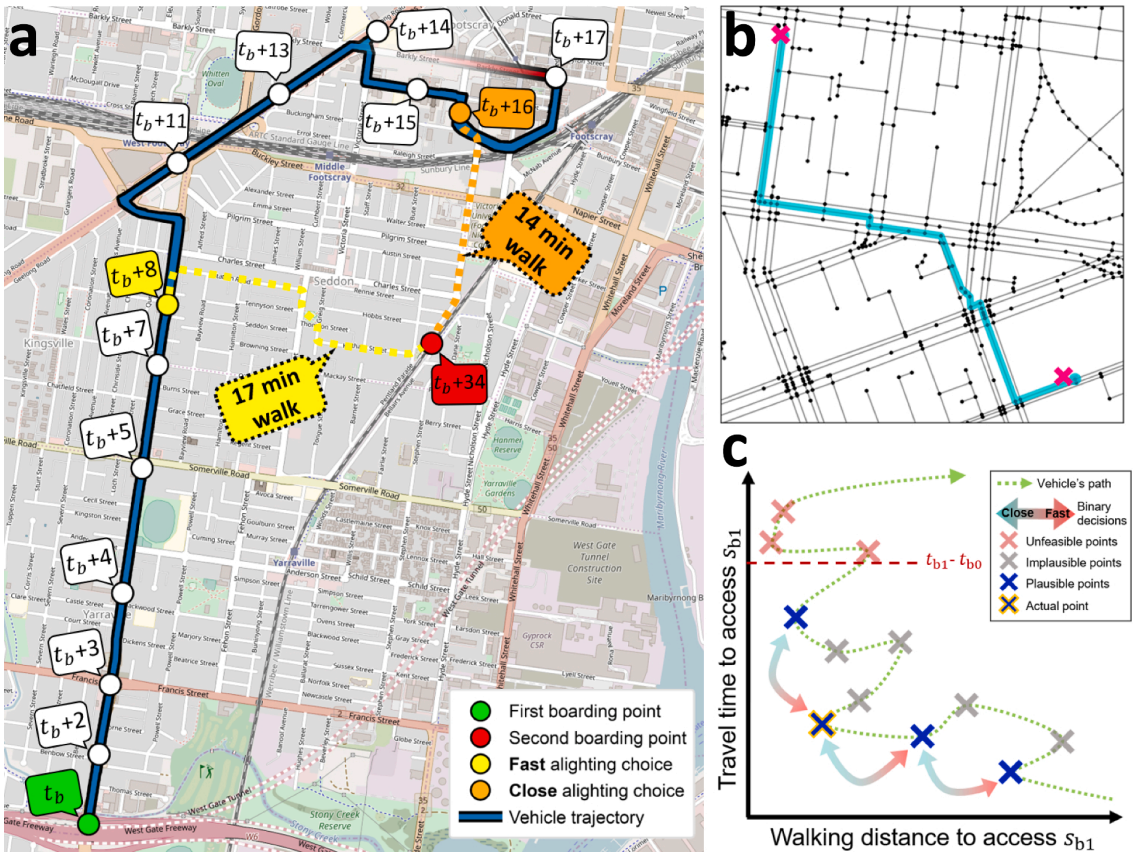


Fig. 2. An example scenario of a missing tap-off transaction. (a) The map illustrates a scenario from the actual data where a tap-on record for a passenger (green) is followed immediately by another tap-on (red), and the tap-off pair for the first boarding is missing. Figure is created using a map tile from OpenStreetMap (OpenStreetMap Copyright and License, 2020). (b) Shortest walking path (cyan) found between two PT stops (magenta) in Melbourne's footpath network. (c) Representation of a PT vehicle's stop-visits after a particular passenger embarks the vehicle, where position of crosses on the plane indicates the time and walking distance from every stop-visit (potential alighting point) to the passenger's next boarding. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

departure time from that stop, respectively. (Here, for simplicity we use the mid-dwell time of a stop-visit instead of using both arrival and departure time of the vehicle, and represent each stop-visit as $TR_i = (s_i, t_i = (t_i^a + t_i^d)/2)$.) In the example of Fig. 2a, the trajectory of the embarked PT vehicle is shown in blue, and the mid-dwell time at each stop is shown (in minutes) relative to the timestamp of passenger's first tap-on t_{b0}. An initial set of candidates CS for the missing alighting transaction can be made from the vehicle's stop-visits after the passenger's first boarding (at time t_{b0}) which allow for the passenger to walk to the next boarding point s_{b1} in time. This can be formulated as

$$CS = \{(s_i, t_i) \in TR_{vid} | t_{b0} < t_i < t_{b1} \wedge d(s_i, s_{b1})/(t_{b1} - t_i) < v_{walk}\}, \quad (1)$$

where v_{walk} is the average walking speed for passengers set as 4.5 km/h in this study (Hof et al., 2002; Montufar et al., 2007), and function $d(.,.)$ returns the walking distance between the locations of its input stop identifiers. The walking distance between two locations is calculated using the detailed footpath network with the link weights expressing the length of each footpath segment. The returned distance between the two input locations, is the shortest weighted path between the two network nodes, which are the closest nodes to the two given locations (see Fig. 2b for an illustrated example).

It is unlikely that passengers hold onto a ride after it is turning (e.g., to service another direction of the route), only to alight slightly closer to their intended destinations. Therefore, we update the candidate set CS by eliminating those stops that the passenger is able to reach sooner than the embarked PT vehicle, by alighting and walking from any previously-visited stop. In other words, we keep only those points $(s_i, t_i) \in CS$ for which there exists no other point $(s_j, t_j) \in CS$ that satisfies both following conditions: i) it is visited by the vehicle earlier in time (i.e., $t_j < t_i$), and ii) its walking distance to the stop s_i can be traversed faster than it takes the vehicle to move between the stops s_j and s_i (i.e., $d(s_i, s_j)/v_{walk} < (t_i - t_j)$). So, updating the candidate set can be formalized as

$$\dot{CS} = \{(s_i, t_i) \in CS | \nexists (s_j, t_j) \in CS : t_j < t_i \wedge d(s_i, s_j)/v_{walk} < (t_i - t_j)\}. \quad (2)$$

In Fig. 2, the part of the vehicle's trajectory removed via Eq. (2) is shown with a fading red line.

We use the illustrated example in Fig. 2c to explain the rest of the process of modeling and predicting passengers' alighting behavior. Consider the aforementioned two consecutive boarding transactions (R_{b0} and R_{b1}) from a single smartcard. Each cross ("x") in Fig. 2c corresponds to a stop-visit by the vehicle *vid* after the stop s_{b0} (or time t_{b0}) at which the passenger has embarked the vehicle. The position of each cross in Fig. 2c indicates the walking distance $d(s_i, s_{b1})$ and arrival time (or travel time from the first boarding) to the passenger's next boarding stop s_{b0}. For simplicity, we denote the arrival time at the next boarding stop, as a function $\tau(s_i, t_i)$ of candidate alighting points (s_i, t_i), calculated as the sum of on-board time (from s_{b0} to s_i) and the walking time (from s_i to s_{b1}), i.e.,

$$\tau(s_i, t_i) = (t_i - t_{b0}) + (d(s_i, s_{b1})/v_{walk}). \quad (3)$$

The dashed green arrow shows the trajectory of the vehicle with its stop-visits. Red crosses at the top (in Fig. 2c) are examples of unfeasible alighting points (removed by Eq. (1)) as they are visited by the vehicle after the passenger's next boarding (t_{b1}). We divide all the candidate alighting points in $\dot{CS}$, into *implausible* (gray crosses) and *plausible* (blue crosses) alighting points. For an alighting point (s_i, t_i) to be plausible, there should be no other point $(s_j, t_j) \in \dot{CS}$ with both smaller walking distance ($d(s_j, s_{b1}) < d(s_i, s_{b1})$) and arrival time ($\tau(s_j, t_j) < \tau(s_i, t_i)$) to the next boarding stop s_{b1}. This can be formally written as below to define the set of plausible alighting points:

$$\ddot{CS} = \left\{ (s_i, t_i) \in \dot{CS} | \nexists (s_j, t_j) \in \dot{CS} : \tau(s_j, t_j) < \tau(s_i, t_i) \wedge d(s_j, s_{b1}) < d(s_i, s_{b1}) \right\}. \quad (4)$$

The set $\ddot{CS}$ is the Pareto front of all alighting points with respect to i) walking distance and ii) arrival time at the next boarding point. Thus, an important feature is that for every implausible alighting point (Pareto dominated), there is a plausible alighting point (Pareto optimal) in the set $\ddot{CS}$ that has both a shorter walking distance and earlier arrival time to the next boarding stop. Thus, it is intuitive that under normal circumstances a passenger chooses an alighting point from those within $\ddot{CS}$. Note that the set $\ddot{CS}$ always includes the closest and fastest alighting points (respectively, orange and yellow stops in the example illustrated in Fig. 2a), one of which is taken as the passenger's final choice in most of the existing studies (as discussed earlier).

3.1.2. *Characterizing the choice between each pair of plausible alighting points*

To model the passengers' choice of alighting stop, we perform the below procedure (involving steps 'i' to 'v'):

i) First, we collect all instances of consecutive boarding transactions in the data, i.e., R_{b0} and R_{b1}, where R_{b0} is paired with an available alighting transaction ($R_{a0} = (s_{a0}, t_{a0})$).
ii) Then, for each of those instances the set of plausible alighting points $\ddot{CS}$ is found.
iii) We filter these instances to keep only those with $|\ddot{CS}| > 1$ and also $\ddot{CS}$ including the stop-visit $(s_i, t_i) \in \ddot{CS}$ associated with the passenger's actual chosen alighting point (i.e., $s_i = s_{a0}$).

Moving along the trajectory of the PT vehicle, any plausible alighting point in $\ddot{CS}$ compared to the next one, has necessarily a larger walking distance but a smaller arrival time to the passenger's next boarding point (see the order in which the plausible alighting points

(blue crosses) are visited by the PT vehicle in the example of Fig. 2c). The idea behind our modeling is that as the PT vehicle approaches a plausible alighting point, the passenger makes a 'binary choice' between alighting at that stop (fast stop) versus continuing the ride to alight at any other plausible alighting point (close stop). In this sense, for any set $\ddot{CS}$ from step 'iii', the passenger has chosen the close stop in the binary choice between s_{a0} and any plausible alighting point visited before that, but has chosen the fast stop between s_{a0} and any alighting point visited later along the PT vehicle's trajectory.

iv) From each set of plausible alighting points $\ddot{CS}$ associated with transactions R_{b0}, R_{a0}, and R_{b1}, we extract $\left|\ddot{CS}\right| - 1$ 'binary choices' made by the passenger, where each choice is between the passenger's actual choice and another plausible alighting points in $\ddot{CS}$. Each binary choice between a pair of fast/close alighting points is then treated as a data sample for which we calculate three attributes to describe the relation between the two options.

v) Finally, all available samples are used to train a model. The trained model will be able to predict a passenger's choice between any pair of fast/close alighting options, given a pair of consecutive boarding transactions (R_{b0} and R_{b1}) and plausible alighting points ($\ddot{CS}$) between boardings. Thereby, the model can be used to predict the final choice made by the passenger.

Consider a binary choice made by a passenger between two plausible alighting points, e.g., alighting points marked as s_{p2} and s_{p3} in Fig. 2c, where s_{p2} leads to a relatively early arrival and s_{p3} having a smaller walking distance to the passenger's next boarding point. Let us generally denote these two options participating in a binary choice with stop-time pairs (s_f, t_f) and (s_c, t_c), respectively representing the fast and close alighting points relative to one another. The binary choice between these two options can be denoted as $(s_f, t_f) \leftrightarrow (s_c, t_c)$. We generate three attributes to describe $(s_f, t_f) \leftrightarrow (s_c, t_c)$ with respect to transactions R_{b0} and R_{b1}. Let us formally show this with

$$X\big((s_f, t_f) \leftrightarrow (s_c, t_c), R_{b0}, R_{b1}\big) = (x_1, x_2, x_3), \tag{5}$$

where X is a mapping function of the binary choice between two plausible alighting points with respect to the passenger's immediate boardings before and after, and (x_1, x_2, x_3) is an attribute vector describing the binary choice as a data point in 3-dimensional space.

The first attribute is the ratio between in-vehicle times to reach the two alighting points (i.e., $x_1 = (t_f - t_{b0})/(t_c - t_{b0})$), and the second attribute is the ratio between walking distances from the two alighting points (i.e., $x_2 = d(s_c, s_{b1})/d(s_f - s_{b1})$), both always in the range $[0,1)$. The fast alighting option becomes more attractive when the in-vehicle time ratio is closer to zero (significantly less in-vehicle time) or when walking distance ratio is closer to unity (insignificant extra walking).

We also found another attribute for a binary alighting choice (comparing the two plausible alighting options), which fairly characterizes the decision between the two options. This third attribute is the difference in journey times divided by the difference in riding times. The difference between journey times, or in other words, time saved by choosing the fast option over the close option is $t_c - t_f - \frac{d(s_f, s_{b1}) - d(s_c, s_{b1})}{v_{walk}}$. This, divided by the extra riding time to alight at the closest stop (i.e., $t_c - t_f$), gives the third attribute describing the binary alighting choice. The third attribute (x_3), can be formally written as

$$x_3 = 1 - \frac{d(s_f, s_{b1}) - d(s_c, s_{b1})}{v_{walk} \cdot (t_c - t_f)}. \tag{6}$$

It is not difficult to show that always $x_3 \in (0,1)$. For a pair of plausible alighting options, the journey time to the next boarding point via a fast alighting point is lower than the journey time via a close alighting point, thus, $t_c + d(s_c, s_{b1})/v_{walk} > t_f + d(s_f, s_{b1})/v_{walk}$ which follows $t_c - t_f > (d(s_f, s_{b1}) - d(s_c, s_{b1}))/v_{walk}$. As the walking distance from the fast alighting point is larger than that of the close point, we have $t_c - t_f > \frac{d(s_f, s_{b1}) - d(s_c, s_{b1})}{v_{walk}} > 0$. Furthermore, as the in-vehicle time is larger for the close alighting point, we get $1 > \frac{d(s_f, s_{b1}) - d(s_c, s_{b1})}{v_{walk} \cdot (t_c - t_f)} > 0$ from which it is straightforward to see that $0 < x_3 < 1$. When the extra walk required from the fast option is not much larger than that of the close option (i.e., $d(s_f, s_b) - d(s_c, s_b)$ is small), but relatively large riding time is required to reach the close option (i.e., $t_c - t_f$ is large), x_3 will be close to unity. This is when a short extra walk saves a relatively significant journey time. Thus, for larger values of x_3 passengers are expected to choose the fast alighting option easier over the close option. When extra walking distance from the fast option is large relative to the extra riding time to the close option, x_3 will be close to zero. This implies that taking the shortcut is not effective in saving time, and thus, passengers are expected to choose the close option over the fast one for low values. This third attribute (x_3) indicates the effectiveness of extra walking in lowering the journey time when taking the shortcut (choosing the fast option) to the next boarding point; therefore, we refer to it as *shortcut effectiveness*.

Without loss of generality, we can use the example of Fig. 2c to explain how to construct a model that captures the passengers' alighting behavior (i.e., steps 'iv' and 'v' of the procedure described in the beginning of the section). In the example of Fig. 2c there are four plausible alighting points for the passenger, i.e., $\ddot{CS} = \{(s_{p1}, t_{p1}), (s_{p2}, t_{p2}), (s_{p3}, t_{p3}), (s_{p3}, t_{p3})\}$. We extract three binary choices, each between the known actual chosen point (s_{p3}, t_{p3}) and one other plausible point, and calculate the 3-dimensional attribute vector that compares the two alighting points. As previously explained, the attribute vector contains the ratio between the in-vehicle times to arrive at each point (x_1), the ratio between walking distances from each point (x_2), and the shortcut effectiveness (x_3) associated with the pair of plausible alighting points. Thus, each choice will be a data point in a 3-dimensional space. When the final chosen point is known (here s_{p3}) we can use two labels (say, '+' and '-') to categorize the samples into those with the close or the fast options chosen. In the example, the close option (s_{p3}) is chosen in the binary choices $(s_{p1}, t_{p1}) \leftrightarrow (s_{p3}, t_{p3})$ and $(s_{p2}, t_{p2}) \leftrightarrow (s_{p3}, t_{p3})$, while in the choice $(s_{p3},$

$t_{p3}) \leftrightarrow (s_{p4}, t_{p4})$ the passenger has taken the fast option.

3.1.3. Modeling passengers' alighting behavior with random forest classifier

To model the passengers' choice of alighting stop, we train a *random forest* model (Ho, 1995) using all labeled samples made from passengers' choice between pairs of fast-close alighting options. Random forest is an ensemble of decision trees (Mitchell, 1997) and classifies a new data point by feeding it as input to each tree and combining the prediction of all trees according to the majority vote. Decision tree models start with a root with branches leading to nodes in the next level and possibly each of those nodes branching into new nodes. Each node splits an attribute at a threshold into different ranges represented by branches, unless the node is a leaf which determines the label for the data point. For a new data point, at each node a question is asked about one attribute and depending on that attribute's value in the data point one branch is taken to the next level in tree, until a leaf node is visited which provides the label predicted for the data point. Decision trees are trained based on the information entropy concept. Given a set S of samples, each labeled as '+' or '-', the *entropy* of the Boolean classification is defined as

$$E(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-, \tag{7}$$

where p_+ and p_- are the proportions of positive and negative samples in S, respectively. Let samples in S be split into subsets $S_{\underset{<\tilde{x}_i}{\sim}}$ and $S_{\underset{>x_i}{\sim}}$ with respect to the attribute x_i and the threshold $\tilde{x}_i$. The *conditional entropy* for this split is defined as

$$E\left(S, \tilde{x}_i\right) = \frac{\left|S_{\underset{<x_i}{\sim}}\right|}{|S|} E\left(S_{\underset{<x_i}{\sim}}\right) + \frac{\left|S_{\underset{>x_i}{\sim}}\right|}{|S|} E\left(S_{\underset{>x_i}{\sim}}\right). \tag{8}$$

To train a decision tree using a set of training samples, at each node (where a set of samples S end up) an attribute x_i with a threshold $\tilde{x}_i$ is selected to yield the maximum *information gain* defined as

$$I(S, x_i) = E(S) - E(S, x_i). \tag{9}$$

Random forests use bootstrap aggregating (or bagging) technique to train decision trees. From all available data samples in the training set, random forest takes a random subset of samples with replacement to train each decision tree and classifies a new sample by taking the majority vote of all trained trees. As a result, although the decision trees are prone to overfitting and the prediction of one decision tree can be sensitive to noise in its training set, the random forest model made of uncorrelated trees is not prone to overfitting and has better performance than decision trees. The trained random forest model separates the 3-dimensional attribute-space into two regions associated with two labels '+' and '-' that best describes the choice of passengers to alight at the close versus the fast option between two plausible alighting points.

In a scenario where two consecutive passenger boarding transactions (R_{b0} andR_{b1}) with missing alighting transaction in-between appears in the travel data, we find the set of plausible alighting options $\ddot{CS}$ and sort them temporally so that for any (s_i, t_i) and (s_{i+1}, t_{i+1}) from the set we have $t_{i+1} > t_i$. Then, for the first plausible alighting point (s_i, t_i) in the set $\ddot{CS}$ we generate all possible choices $(s_i, t_i) \leftrightarrow (s_j, t_j)$ with $j > i$, and construct the corresponding attribute vector with $X\big((s_i, t_i) \leftrightarrow (s_j, t_j), R_{b0}, R_{b1}\big)$. The attribute vectors are then fed to the trained random forest model.

If the model predicts the passenger to take the fast alighting point (s_i, t_i) in all these choices, the final prediction for the missing alighting transaction will be made using the time-stop pair (s_i, t_i). However, if the model predicts at least one close option (s_j, t_j) to be more attractive than (s_i, t_i), we rule out the alighting point (s_i, t_i) as the passenger's final choice, and investigate the choice between the next plausible point (s_{i+1}, t_{i+1}) and all (s_j, t_j) in the set $\ddot{CS}$ with $j > i + 1$. This is repeated until the fast option is chosen over all close options or there is only one option left, and the final chosen point is then used to reconstruct the missing transaction.

As discussed, some PT systems around the world function only with passengers' boarding transactions. Smartcard dataset from such systems does not provide information directly showing how passengers choose between a pair of fast-close alighting points. However, i) trajectory of vehicles can be reconstructed using time-location of tap-on transactions, and then, from each pair of consecutive boarding transactions recorded for a passenger, ii) the set of plausible alighting points can be derived using Eq. (4), and iii) the choice between each pair of plausible alighting points can be mapped to a 3-dimensional vector of attributes (Eq. (5)). There is no known label for these data points showing the actual decision made, however, 'unsupervised learning' is theoretically able to separate data points into classes with the goal of maximizing intra-class similarity and inter-class dissimilarity. In the *Results* section and further in the *Appendix A*, we explain the application of clustering on the set of binary alighting choices mapped to the 3-dimensional attribute space. In case of a smartcard dataset including only tap-on transactions, where the passengers' decision on binary alighting choices is not known, clustering can replace the random forest model used in our proposed procedure. Our results, show that the proposed clustering-based procedure (see *Appendix A*), can lead to an acceptable accuracy in estimating missing alighting transactions for tap-on-only smartcard data.

3.2. Characterizing transfer and activity duration

The building blocks of an accurate OD demand matrix are passenger *OD trips* (or *journeys*) rather than their single-leg trips, as OD trips reflect the actual passengers' demand to travel from different origin locations to their intended destinations to undertake their

desired activities. Therefore, after enhancing the smartcard data information by estimating the missing alighting transactions, one needs to identify OD trips (or journeys) made up of a single or multiple trip-legs in order to construct the OD demand of the PT system. Let us refer to an alighting-then-boarding event between two consecutive PT rides (recorded for a passenger), simply as *interchange* and its duration as Inter-Transaction Time (ITT). The goal is first, to identify whether each interchange is associated with an activity or a transfer, and then, to integrate consecutive passenger trips linked together by transfers, into passenger journeys.

In order to develop a simple and yet effective approach, we focus on determining whether an interchange is associated with a transfer or activity based on its duration. Following the common practice in the existing literature, we make use of a maximum allowable ITT (ITT threshold). If the ITT between two consecutive passenger trips is larger than a threshold, it is more likely that the passenger is undertaking an activity and not merely transferring between PT services. Unlike the common practice though, we propose a solution to find the optimum maximum allowable ITT in any PT system from its smartcard data.

Similar to characterizing the choice of alighting point tackled previously, we solve the transfer identification problem as a binary classification problem. Each interchange with a particular ITT value belongs to either the transfer class or the activity class. The goal thus becomes finding the optimum threshold which divides the domain of ITT values into two regions, each being more representative of one of the target classes, i.e., transfers or activities. To find this optimum threshold, we first need to know the actual distribution of transfers and activities with respect to their ITT. This information is not directly accessible from smartcard data. Thus, we propose a method (Alg. 1) to estimate these two distributions from any given PT travel data where both end-points of trips are available.

3.2.1. *Extracting the distribution of transfer and activity duration*

To estimate the ground-truth distribution of activities and transfers over their ITTs, we detect a set of *return trips* for which it can be intuitively concluded if the intermediate interchanges are associated with activities or transfers. Here, we use the aid of the visualized example in Fig. 3 to explain the procedure. Imagine two anchor locations, the return and the target point, far from one another and possibly a number of other locations in between the two. We define a return trip as a sequence of trips starting from and ending to the first anchor location (the return point) through one or more interchange events. In Fig. 3, the sequence of trips 1, 2, and 3 provide an example of a return trip in Melbourne's PT network, starting and ending at the return point marked by a pink cross. It is intuitive that such return trips involve at least one activity, most likely at the target point located furthest away from the return point (the target point is colored red in Fig. 3).

In order to eliminate the PT trips possibly connected via a motorized non-PT trip, we require each interchange between two rides to be between stop clusters with a walking distance that can be feasibly traversed in time with a walking speed of 4.5 km/h. Also, to exclude the return trips likely to entail multiple activities we limit the return trips to those in which ITTs between all trip legs are shorter than the ITT at the target point.

The high-level pseudo-code in Alg. 1 describes the proposed procedure for detecting a set of credible return trips and estimating the distribution of transfer and activity durations between PT trip-legs from those return trips. Algorithm 1 processes a temporally sorted sequence of trips carried out by a single passenger, and runs through the smartcard data one passenger at a time. Lines 4–14 in Alg. 1 pick the first boarding transaction of a passenger and search through the rest of that passenger's transactions to find a possible return trip to that location. If the algorithm fails to detect a return trip to/from the picked transaction, that transaction will be removed from

Fig. 3. Example of a detected return trip. The map visualizes a return trip from Melbourne's smartcard data, comprised of three public transportation trips, numbered 1, 2, and 3 according to their order in time. The pink cross marks the location of the return point. The red and purple crosses mark the midpoint between alighting-then-boarding locations. Dashed circles indicate interchange events between trips and the length of each interchange event is annotated in minutes next to the associated circle. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the transactions sequence and the process is repeated until all return trips are detected or there is no transaction left in the sequence.

High level pseudo-code for estimating the distribution of transfer and activity durations

Input: R: The set of paired smartcard transaction records from all public transportation modes during a day, each as a tuple with timestamp, location, and transaction-type.

Output: h_a, h_τ: histogram of transfer and activity durations.

(1) Initialize empty sets of observations: $\alpha_{obs} \leftarrow \{\ \}$,$\tau_{obs} \leftarrow \{\ \}$
(2) **For** each smartcard ID c_{ID} in R:
(3) $T \leftarrow$ A sequence of temporally sorted trips $(T^1, T^2, \cdots)$ for c_{ID}, with the i-th trip T^i explaining the location $T^i_{B_s}$ ($T^i_{A_s}$) and timestamp $T^i_{B_t}$ ($T^i_{A_t}$) of boarding (alighting) stops.
(4) **While** T contains more than two trips:
(5) $j \leftarrow$ Find the earliest trip T^j ($j > 1$) in T where $T^j_{A_s} = T^1_{B_s}$.
(6) **If** j was not found **then:** remove the first trip in T and go to Line (4); **else:** $p_{ret} = T^1_{B_s}$ and $I = Interchanges(T)$.
(7) $d_{target} \leftarrow -\infty$; $\alpha_0 \leftarrow \{\ \}$;$\tau_0 \leftarrow \{\ \}$
(8) **For** each interchange event I^i in I:
(9) **If** $I^i_d / I^i_t > v_{walk}$ **then:** remove first $i+1$ trips from T and go to Line (4).
(10) **If** $d(I^i_p, p_{ret}) > d_{target}$ **then:** $d_{target} = d(I^i_p, p_{ret})$; $\tau_0 \leftarrow \tau_0 \cup \alpha_0$; $\alpha_0 \leftarrow I^i_t$; **else:** $\tau_0 \leftarrow \tau_0 \cup I^i_t$.
(11) **End For**
(12) **If** $\forall x \in \tau_0 : x \leq \alpha_0$ **then:** $\alpha_{obs} \leftarrow \alpha_{obs} \cup \alpha_0$ and $\tau_{obs} \leftarrow \tau_{obs} \cup \tau_0$
(13) Remove the first j trips from T.
(14) **End While**
(15) **End For**
(16) **Return** $histogram(\alpha_{obs})$,$histogram(\tau_{obs})$
(17) **Function** $Interchanges(T)$:
(18) **For** each pair of consecutive trips T^i and T^{i+1} in T:
(19) $I^i_d \leftarrow d\left(T^i_{A_s}, T^{i+1}_{B_s}\right)$
(20) $I^i_p \leftarrow (T^i_{A_s} + T^{i+1}_{B_s})/2$
(21) $I^i_t \leftarrow T^{i+1}_{B_t} - T^i_{A_t}$
(22) $I^i \leftarrow (I^i_d, I^i_p, I^i_t)$; // i.e., (walking distance, location, duration) of the interchange
(23) **End For**
(24) **Return** the ordered sequence of interchange events$I = (I^1, I^2, \cdots)$
(25) **End Function***Interchanges*

Algorithm 1. ***Estimating the distribution of activity and transfer durations.*** *Pseudo-code details the process of detecting a set of return trips to extract activity/transfer durations. Function $d(.,.)$ returns the shortest walking distance on the walking pathway network between the two input locations (see Fig. 2b). All locations ($T^i_{B_s}$, $T^i_{A_s}$, I^i_p, and p_{ret}) are points in a 2-dimensional space and any operation involving them should be regarded as a vector operation.*

In Alg. 1, for each detected return trip, the location of the return point is stored as p_{ret}. Interchange events between consecutive trip legs of the return trip are then extracted as tuples describing the interchanges' location (I^i_p), duration (I^i_t), and alighting-to-boarding walking distance (I^i_d); see lines 17–25 in Alg. 1. The location of each interchange is calculated as the midpoint of the line connecting the location of the alighting point from the first trip to the boarding point of the next trip. The information of the detected return trip is used, only if the interchange furthest away from the returning point (i.e., the target point) has also the longest duration among all interchange events (line 12). Finally, for an acceptable return trip, such as the example shown in Fig. 3, duration of the interchange event at the target point (79 min at the red cross) is added to the set of activity durations, while duration of all other interchange events (14 min at the purple cross) is appended to the set of transfer durations.

Alg. 1 returns two histograms for the duration of transfers and activities observed in the data, which provide an estimation of the actual distribution of transfer/activity durations. Evidently, transfers are generally shorter than the activities, yet the remaining question to be answered is: what is the ITT threshold that optimally separates all ITTs into those associated with transfers and activities. Next (in *section 3.2.2*), we propose a method to analytically determine a decision threshold, which guarantees achieving the highest possible classification performance when ITTs shorter than the decision threshold are labeled as transfers and those longer than the threshold are labeled as activities.

3.2.2. Finding the optimal threshold for binary classification

Here, the feature ITT associated with consecutive trip-legs will be used to classify each interchange event into either a transfer or an activity (Task II in Fig. 1). Here, we describe a method to solve the above binary classification (thresholding) problem. The method finds the optimum threshold value in the characteristic feature (ITT) domain, to build a model which optimally performs the transfer/activity identification task.

To formulate the method, let X_t and X_a be the two random variables respectively denoting the values associated with samples from 'transfers' and 'activities' target classes. Each sample is associated with a value of the characteristic feature ITT denoted by θ. The distribution of samples from A and B over θ, corresponds to the distribution of activity and transfer events for ITT values. It is evident that under normal circumstances, transfer between PT services takes less time than undertaking activities between PT trip legs. Thus, we can assume that samples from the transfer class have lower θ values on average compared to those from the activity class. An

example is illustrated in Fig. 4a, where samples from the two classes have Gaussian distributions with different means and standard deviations over the domain of θ. (Note that the important feature in the provided example, is the different mean θ between the samples from each of the two distributions, and the methodology presented here is independent of the shape of these two distributions.) Using a decision threshold θ_T, samples are labeled as transfers if $\theta \leq \theta_T$ or otherwise as activities. The problem is to find the optimal threshold θ_T^{opt} for which the classifier achieves the best classification performance.

We find the optimum threshold θ_T^{opt} to perform the classification, according to Youden's J statistic, also called *informedness* (Youden, 1950), which evaluates a dichotomous (binary) classifier. Informedness $J \in [0, 1]$ is the probability that a decision made by the classifier is an informed decision as opposed to a random guess, taking into account all predictions made by the classifier (Powers, 2011). Informedness (J) can be defined as below:

$$J = TPR + TNR - 1, \tag{10}$$

where TPR (TNR) stands for True Positive Rate (True Negative Rate) which is better known as *recall* (*inverse recall*) in information retrieval and pattern recognition literature. Assume that the actual labels of samples belonging to classes 'transfers' and 'activities' are positive and negative, respectively. Then, recall (TPR) and inverse recall (TNR) can be defined as:

$$TPR = \frac{TP}{TP + FN}, \tag{11}$$

$$TNR = \frac{TN}{TN + FP}, \tag{12}$$

where TP (FN) is the number of positive samples labeled correctly (incorrectly), and FP (FN) is the number of negative samples labeled correctly (incorrectly). So, given our definition of the problem, TP is the number of samples in class 'transfers' with a θ value below the selected threshold and TN is the number of samples in class 'activities' with a θ value above that threshold.

Let $P_t(\theta)$ and $P_a(\theta)$ be the distribution of samples over the characteristic feature θ, respectively for transfer and activity classes. Then, for a binary classifier which uses the decision threshold θ_T to classify samples into our two target classes, we can rewrite the definitions in Eqs. (11) and (12) with respect to θ_T:

$$TPR(\theta_T) = \int_{-\infty}^{\theta_T} P_t(\theta) d\theta = P(X_t \leq \theta_T), \tag{13}$$

$$TPR(\theta_T) = \int_{\theta_T}^{+\infty} P_a(\theta) d\theta = P(X_a > \theta_T). \tag{14}$$

$P(X_t \leq \theta_T)$ is the Cumulative Distribution Function (CDF) of the random variable X_t and $P(X_a > \theta_T)$ is the Complementary CDF (CCDF) of the random variable X_a. Equations (13) and (14) can be used to rewrite the definition of informedness J in Eq. (10), this time

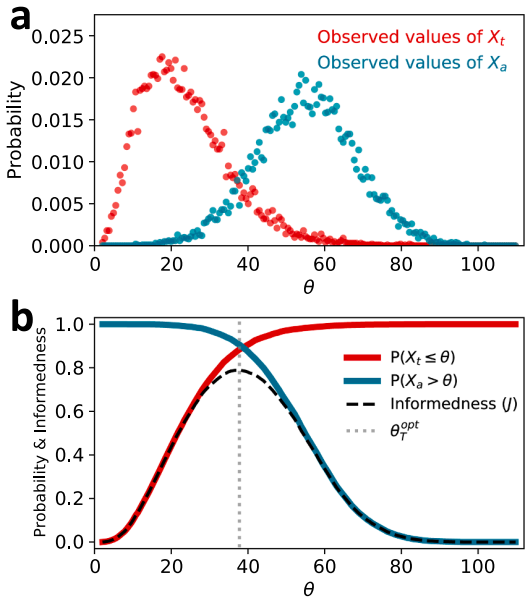


Fig. 4. Example procedure to build a binary classifier to separate samples from different classes. (a) Histogram of θ values for samples from X_t (transfer class) and X_a (activity class) random variables. (b) Finding the optimal decision threshold based on the histograms in (a).

allowing to calculate J as a function of the decision threshold θ_T (see Fig. 4b):

$$J(\theta_T) = P(X_t \leq \theta_T) + P(X_a > \theta_T) - 1. \tag{15}$$

Accordingly, the optimal decision threshold that maximizes the informedness J of the resulting classifier can be calculated as:

$$\theta_T^{opt} = \underset{\theta_T}{\text{argmax}} J(\theta_T). \tag{16}$$

We can now build an optimal classifier which uses the decision threshold θ_T^{opt}, and labels each sample represented with a θ value, as 'transfer' if $\theta \leq \theta_T^{opt}$ or as 'activity' if $\theta > \theta_T^{opt}$. According to the definition of our performance measure J, a decision made by this classifier has the maximum probability of being an informed decision as opposed to a random guess (Powers, 2011).

The example in Fig. 4 can aid summarizing the proposed procedure. First, from the histogram of observed ITT values (θ) associated with samples from X_t (transfer class) and X_a (activity class) random variables (Fig. 4a), CDF of X_t (the red curve in Fig. 4b) and CCDF of X_a (the cyan curve in Fig. 4b) over θ are calculated. Then, using the CDF of X_t ($P(X_t \leq \theta_T)$) and CCDF of X_a ($P(X_a > \theta_T)$), informedness of the classifier is calculated as a function of the ITT threshold θ_T according to Eq. (15) (dashed black curve in Fig. 4b). This allows for finding the optimal decision threshold θ_T^{opt} for the classifier (Eq. (16)) which maximizes its Informedness $J(\theta_T)$ (dotted gray line in Fig. 4b). The resulting model trained from the passenger behavior data optimally separates the ITTs into those associated with transfers and activities.

4. Data preparation

In order to show the effectiveness of the proposed methodology, we apply it to real smartcard data, recorded in the PT network of Melbourne, Australia. In this section, we provide a detailed description and necessary statistics of this smartcard dataset. Also, the application of the components in the data preparation stage of our proposed framework (see Fig. 1) to the available data is discussed in this section.

4.1. Data description and pre-processing

The data includes all passenger transactions made on train, tram, or bus recorded during 61 days of September and October 2017. The available data from the three PT modes include an average of over 2,120,000 and 912,000 daily transactions on weekdays and weekends, respectively. A schematic view of the relation between the PT system and the compartments of the AFC system in charge of the data collection is depicted in Fig. 5. Smartcard transaction data are recorded through fare gates in train stations and on-vehicle fare

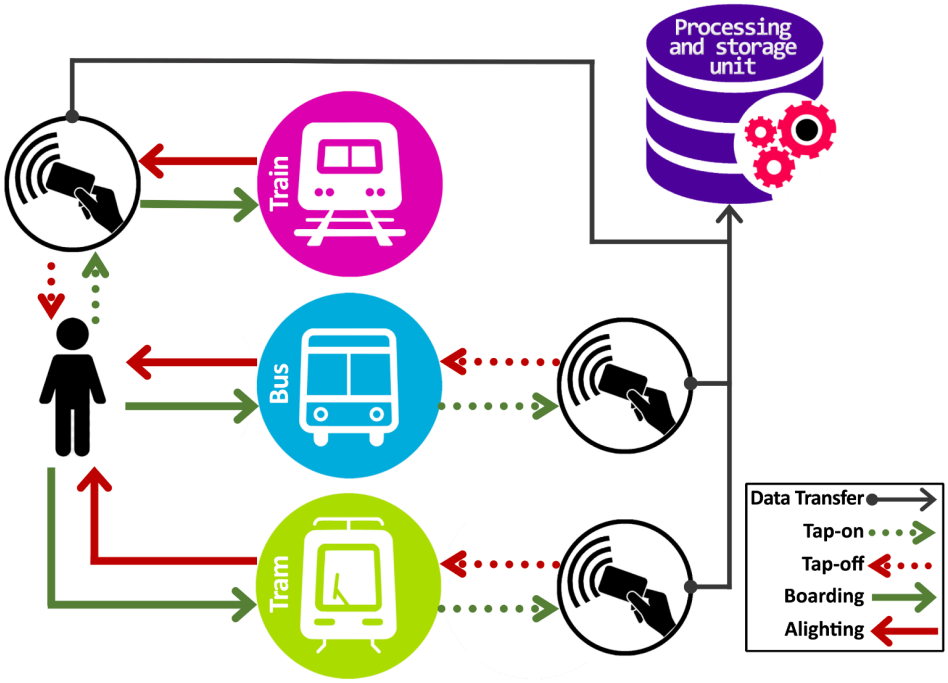


Fig. 5. Schematic view of the smartcard data collection in Melbourne's Public Transportation (PT) system. The graph highlights the difference between the data collection in train mode versus the on-road (bus and tram) PT modes.

collection devices in trams and buses. Each smartcard transaction record consists of several attributes among which we use those listed below:

- Hashed smartcard identifier (card ID): An integer corresponding to a unique physical smartcard being used by a passenger; the physical card serial numbers are anonymized.
- PT mode: A one-digit integer which determines whether the transaction is made on a bus, tram, or at a train station.
- Vehicle identifier (vehicle ID): An integer value corresponding to a unique bus or tram vehicle. This attribute is empty in transaction records associated with the train mode.
- Stop identifier (stop ID): An integer corresponding to a unique stop or station, which is functional for at least one mode in the PT system. Each stop ID corresponds to a unique physical coordinate.
- Timestamp: Date and time of the transaction with the temporal resolution of one second. Transaction time domain is the daily functional period of the system, starting at 4 am on each day and finishing at 4 am on the next day.
- Transaction type: Whether the transaction is a tap-on (scan-on) or tap-off (scan-off).

Stops IDs in transactions recorded for bus and tram modes are inferred from Automatic Vehicle Location (AVL) devices installed on bus and tram vehicles. Transaction type is determined and recorded by the AFC devices, which sometimes requires the aid of the transaction history of smartcards. For a new transaction made by a particular smartcard, if the previous record is a tap-on associated with the same vehicle (in case of bus and tram modes) or same mode (for train mode), then simply a tap-off transaction will be recorded. If the previous transaction is a tap-on on a different mode or vehicle (for bus and tram modes), meaning that there is no tap-off transaction recorded for the previous trip leg, first, a tap-off transaction is recorded (called forced tap-off in the system) before the actual transaction.

During data cleaning process, invalid records with missing values of necessary attributes or attributes with invalid values were removed from the dataset. A transaction record may have attribute values which are within the valid ranges, but along with other transactions, imply unrealistic (e.g., too fast) vehicle/passenger movements between locations; we assume that vehicles/passengers cannot move on the great circle distance between two locations with a speed of higher than 60 km/h. Accordingly, we detect and remove the transactions which imply unrealistic passenger movements in the train network. For the tram and bus modes, we only remove such records if they cannot be rectified. To rectify an unrealistic movement, assuming that the transaction timestamps are error-free, we find the inconsistent records in a set of related transactions. Here, a set of related transactions is either a collection of records from a particular card ID, or records from different cards made on a particular vehicle. We identify inconsistent records in each set of related transactions, and adjust the stop ID (and timestamp) attribute of an inconsistent transaction to the next or previous stop visited by the associated PT vehicle, if it resolves the unrealistic movement. Otherwise, the inconsistent record is removed from the data. This has been successfully applied to rectify a few thousand inconsistent transactions on each day; approximately 0.1% of the

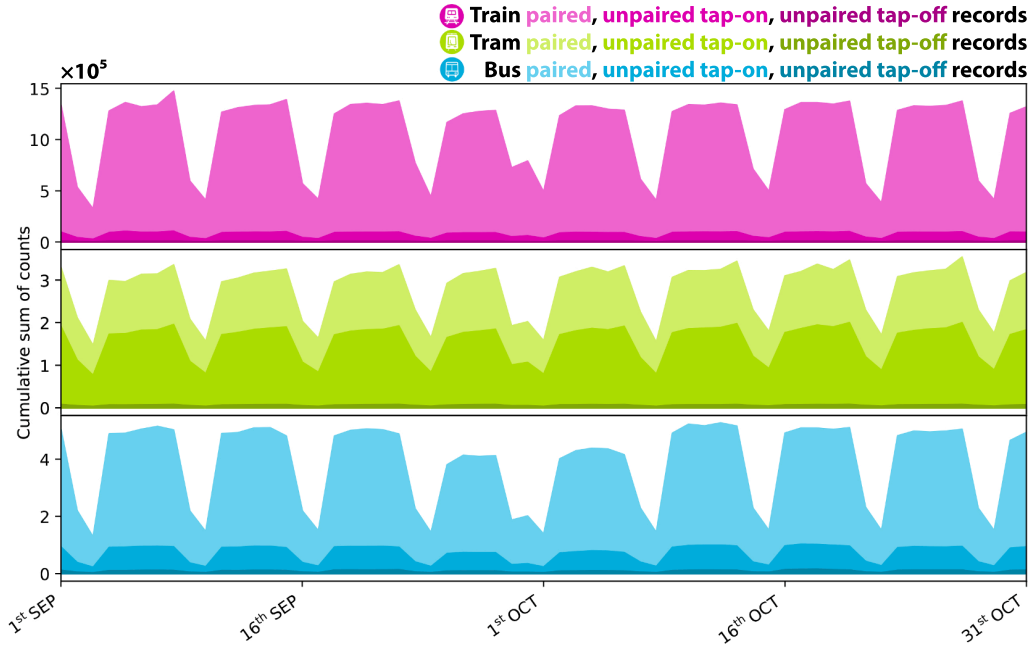


Fig. 6. Paired and unpaired transactions in the cleaned data. The daily counts of paired and unpaired transactions in each mode are shown over the course of the available data. An unpaired tap-on (tap-off) transaction corresponds to a trip with no tap-off (tap-on) information. Note that the unpaired tap-offs are visible only as a line at the bottom of each subplot due to their insignificant count.

daily transactions.

Fig. 6 shows the daily number of transactions in the cleaned data, divided into paired, unpaired tap-on, and unpaired tap-off records in different PT modes. In all PT modes, only for an insignificant number of trips the tap-on record is missing, as demonstrated by the hardly visible portion of transactions colored as unpaired tap-offs in Fig. 6; this is especially the case for the train mode. For the train mode, the number of unpaired tap-on transactions is also insignificant, mainly due to the use of ticket barriers at train station fare-gates and the more frequent ticket control patrols in this mode (Delbosc and Currie, 2016). Trip data in bus and tram modes suffer from a large number of missing tap-off transactions. Especially, most tap-on transactions recorded in tram mode are missing their tap-off pair. Overall, Melbourne's PT smartcard data contains an average of about 1,021,000 tap-on records per day while the number of tap-off records is around 761,000, which indicates the existence of a significant number of missing tap-off transactions.

As train transactions are collected at stations' fare-gates (see Fig. 5), understanding the route choice and tracking the passengers in train mode become very different problems compared to doing the same in bus and tram modes (Kusakabe et al., 2010; Min et al., 2016; Sun et al., 2012). So, considering the scope of this study and the insignificant number of unpaired transactions in train mode, we limit our attention to estimation of missing alighting records in bus and tram modes, where we also take advantage of the information in train transactions.

4.2. *Reconstructing vehicles trajectories*

The last component of our data preparation stage (see Fig. 1) is building the trajectory of every single bus and tram vehicle in the network using all (paired and unpaired) records in the cleaned data. To do so, first, the smartcard transactions made on each vehicle is separated and then, each bus or tram vehicle is tracked using the temporally sorted sequence of stop-timestamp pairs extracted from its associated records. A vehicle trajectory is determined as a sequence of visited stops each with arrival, departure, and dwell time. For a vehicle's single visit to a particular stop, arrival, departure, and dwell time are respectively the timestamp of the first transaction, timestamp of the last transaction, and the time span between them, derived from the uninterrupted sequence of transactions made at that particular stop on the vehicle. If there is only one transaction recorded for a vehicle's stop-visit, the dwell time is considered to be 5 s with the timestamp of that single transaction being the midpoint; this is consistent with most reports in the literature, e.g., (Meng and Qu, 2013).

Transactions can be recorded while the vehicle is moving between two stops as passengers may tap-on after the vehicle departs from a stop, or tap-off before the vehicle arrives at the next stop. These transactions will in effect have a wrong timestamp or may become associated with the wrong stop. By following the sequence of transactions performed by each passenger and on each vehicle, we detect such transactions according to a constraint that restrains a vehicle to travel between two locations with any speed over 60 km/h (maximum speed limit for a collector road). The timestamp of invalid transactions is then adjusted to the midpoint of the dwell interval of its corresponding vehicle if i) the previous (next) valid transaction on the vehicle corresponds to the same stop as the invalid transaction, and ii) the timestamp for the invalid transaction is closer than 15 s to the dwell interval of the vehicle at the previous (next) stop. If the timestamp of an invalid transaction could not be adjusted with the above procedure, it will be eliminated from the data.

5. Results

5.1. *Estimating missing smartcard tap-off records*

The first major task to be addressed in the proposed framework (see Fig. 1) is estimation of missing alighting information in on-road PT modes (bus and tram). As explained in the *Methodology* (*section 3.1*), from trips with both boarding and alighting transactions we

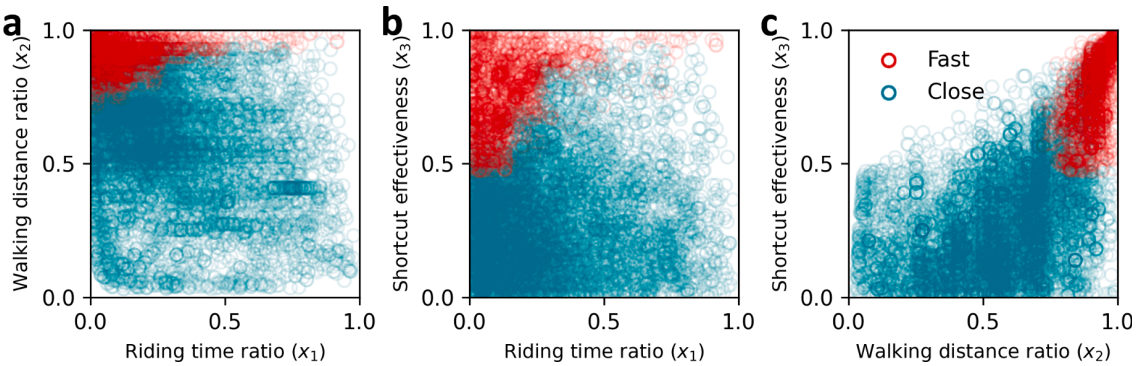


Fig. 7. Classifying passengers' choice between fast versus close alighting options. Each data point in the figure represents a choice between a pair of plausible alighting stops (derived from a regular weekday in the available smartcard data). The three attributes, namely, walking distance ratio, riding time ratio, and shortcut effectiveness describe the choice between the pair of alighting options and the prediction of the trained random forest model is indicated by the color of data points.

extract choices between pairs of fast-close alighting points, i.e., $(s_f, t_f) \leftrightarrow (s_c, t_c)$. Then, passengers' alighting behavior is characterized based on three attributes (Eq. (5)) explaining each binary choice between two plausible alighting points for the passenger's trip. The attributes are calculated according to two given consecutive boarding transactions, namely, the passenger's boarding transaction used to embark the vehicle and the passenger's first boarding transaction after alighting that vehicle. To characterize passengers' alighting behavior, samples of known alighting choices are used to train a random forest model. The random forest model used here is an ensemble of 9 decision trees with maximum depth of 3 corresponding to the number of independent data-point attributes describing passengers' choice between a pair of plausible alighting stops. The model splits the 3-dimensional feature space into two regions one associated with the fast stop being chosen over the close stop and the other one vice versa.

Fig. 7a-c shows the data points from the first day of the smartcard data, each corresponding to a binary choice between a pair of fast-close plausible alighting options. The position of data points in the 3-dimensional feature space is visualized through three 2-dimensional plots, each associated with two out of the three attributes. The color of data points indicates the prediction of the random forest classifier for the binary alighting choice made by passengers. (Regions of different colors indicate the split of the attribute space, done by the random forest classifier.) Testing the model on paired smartcard transactions using 5-fold cross validation, showed that the trained random forest model has a 74.0% accuracy in predicting a passenger's choice between a pair of plausible alighting stops.

In *section 3.1.3*, we discussed the implications of smartcard data only including passengers' tap-on transactions. *Appendix A* presents an unsupervised learning approach based on clustering to train a similar random forest model, when only tap-on transactions are available. This alternative approach leads to a model with 69.0% accuracy in predicting passengers' choices between pairs of fast-close plausible alighting stops.

When riding time to the fast alighting option is much lower than that of the close alighting option, or walking distance to the next boarding point from the fast option is similar to that of the close option, it is more likely that passengers choose the fast option over the close one. Also, when extra walking from the fast option results in a significant time saving compared to alighting at the close option, i.e., when the shortcut effectiveness attribute approaches unity, the passengers tend to take the fast alighting option. The statistical

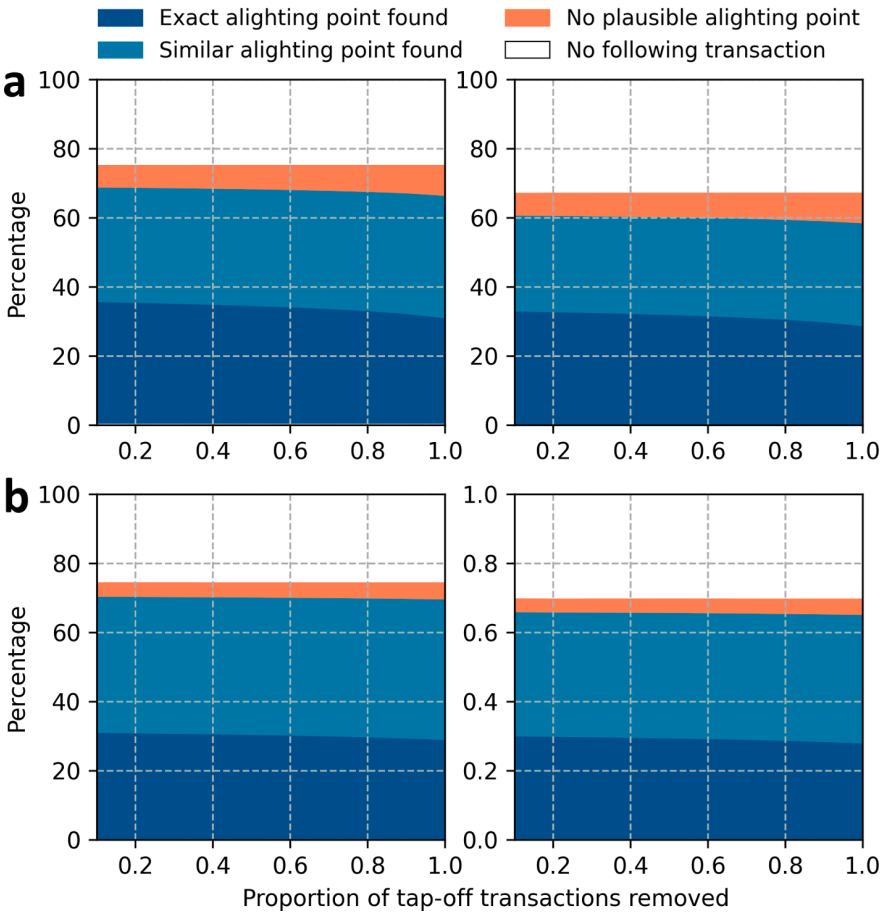


Fig. 8. Evaluating the inference of missing alighting transactions. The results describe the outcome of the proposed method to estimate missing tap-off transactions for (a) bus and (b) tram modes. We start by masking 10% of the tap-offs and then at each step increase this gradually up to 100% (horizontal axis). The results are averaged over 10 realizations of each step (i.e., for a particular proportion of the transactions masked).

pattern in passengers' alighting behavior is therefore well-characterized by the three attributes that we calculate (Eq. (5)) for the binary alighting choices, and the random forest model is able to capture this data pattern in the attribute space. In Fig. 7, it is seen that depending on the value of the three attributes, the random forest model predicts a passenger to take the fast stop over the close one or the opposite, in agreement with the above intuitive explanation about passengers' behavior according to each attribute.

As *section 3.1.3* describes in detail, for a passenger trip missing tap-off and an available consecutive tap-on, first, different plausible alighting points are found by Eq. (4). Then, we start from the first plausible alighting point and pair it with each of those visited later in time by the vehicle; in each pair the point visited earlier is the fast option. If the random forest classifier predicts that the passenger prefers the close options in one of the pairs, the fast alighting point will be ruled out and we repeat this for the rest of the set of plausible alighting points, otherwise the fast option will be picked as the chosen alighting points.

To evaluate the performance of the proposed method in estimating missing transactions, we first exclude all transactions with a missing pair from the data. Then, from the dataset containing both tap-on and tap-off records for each trip, a proportion of the trips is selected at random for which we assume the tap-off transactions are unknown, or in other words we mask those tap-off transactions. For trips with masked tap-off, if we cannot find another tap-on recorded for the same passenger later in time, there is no way of estimating the masked tap-off and these instances will be counted as 'no following transaction.' If a following boarding transaction is available, then we can find the set of plausible alighting stops and use the trained choice-behavior model to estimate the missing tap-off between the two available boarding transactions. The estimated record is then compared with the actual record that was masked, and the result is used to evaluate the performance of the proposed estimation method.

We perform the evaluation process for different numbers of masked transactions, i.e., 10%, 20%,…,100% of all tap-off transactions at each step. Each step is then repeated in 10 independent realizations (random selection of transactions to mask, estimating them, and validating the estimations) and the average results are reported. Note that PT vehicles' trajectories used in the estimation process are derived from all available transactions and we do not use the masked transactions to reconstruct the trajectories. As the number of masked tap-off transactions is increased there would be less information on vehicle movements, which negatively affects the accuracy of reconstructed PT vehicles' trajectories. So, increasing the number of masked tap-off transactions over different steps of the evaluation process, is expected to lower the estimation accuracy.

Fig. 8 evaluates the output of the proposed procedure for inferring missing tap-off transactions, separately for bus (Fig. 8a) and tram (Fig. 8b) modes during weekdays (left panel) and weekends (right panel). Among trips with missing tap-offs, approximately 25% (33%) of those using the bus mode and 26% (30%) of those using trams during weekdays (weekends) are not followed by any other trip recorded from the same passenger, thus, our algorithm has no way of estimating the missing record (white area in graphs of Fig. 8).

When 10% of the tap-off records is masked, the estimation method is able to exactly regenerate 36% (33%) of bus and 31% (30%) of tram records from weekdays (weekends); the blue area in Fig. 8. Note that this means when multiple trips are recorded for a passenger, in approximately 49% (43%) of the cases for the bus (tram) mode the procedure successfully generates the exact missing record. Removing more transactions (moving along the horizontal axes in Fig. 8) only slightly affects this result. When all alighting transactions are masked, the procedure correctly returns 31% (29%) of the bus and 29% (28%) of the tram records during weekdays (weekends). For only 6–8% (3–4%) of the cases in bus (tram) mode (orange bands in Fig. 8), our algorithm is unable to retrieve any estimation of the missing record from other records of the same passenger. This may be a result of data error or can be the case, for example, when a passenger takes a taxi to connect two PT trips and as a result no plausible alighting point is found that allows the passenger to walk to the next boarding point in time.

Depending on the number of masked tap-off transactions, for between 33 and 35% of them in bus mode and 39–41% of them in tram mode during weekdays, the alighting stop-timestamp estimation is not exactly equal to the target (dark-cyan areas in Fig. 8). Figure 9 reports the error for this portion of estimated records, in terms of walking distance (Fig. 9a) and time-gap (Fig. 9b) between

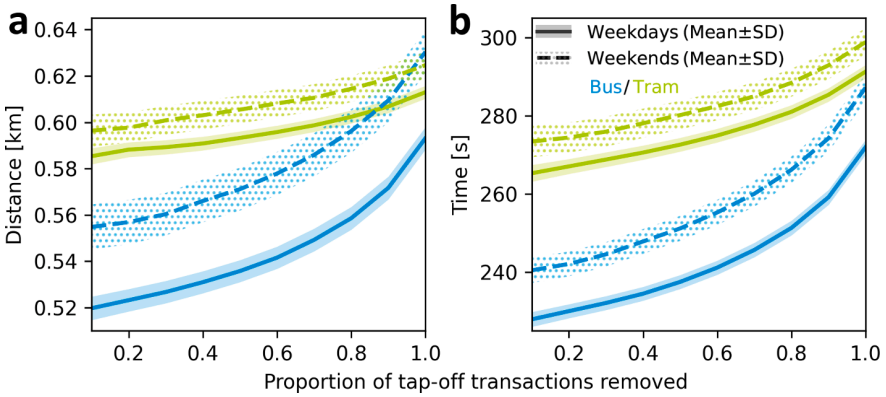


Fig. 9. Performance of the proposed estimation method. For transactions estimated with an incorrect timestamp, stop, or both, (a) distance and (b) time error is calculated over different numbers of masked transactions. For a certain proportion of the tap-off transactions masked, the mean and the standard deviation of estimation error is calculated over 10 independent realizations. Solid lines (dashed lines) indicate mean and shaded areas (dotted areas) indicate standard deviation of the error for weekdays (weekends).

the actual and estimated alighting events. Average estimation error and its standard deviation is smaller for missing tap-off transactions on weekdays compared to weekends. Incorrectly estimated missing transactions for bus (tram) mode are on average around 0.54 km (0.57 km) and 237 s (250 s) far from the target, when 10% of the alighting transactions are masked. As the proportion of the masked tap-off records is increased (moving along the x-axis in Fig. 9), both time and distance error of estimation increase (faster for the bus mode). Yet, even when all the alighting transactions are masked, the proposed method is able to exactly regenerate approximately 32% of them. Furthermore, for over 36% of the masked tap-off transactions (when all tap-off records are masked), the proposed method returns an estimation with an average distance (time) error of approximately 0.61 km (281 s).

The proposed method is also compared with common approaches in state-of-the-art frameworks, where either the closest or fastest alighting point is assumed to always be the passengers' choice. All methods return estimations for the same number of missing transactions. However, the proposed method increases the number of exact estimations by 6.7% (and 9.1%) compared to when the fastest (and closest) stops are deemed as default passenger choices. For Melbourne's smartcard data this improvement means that the number of exactly reconstructed transactions by our proposed method on a regular weekday is between 16,000–21,800 more than that of the conventional approaches.

The error of approximate estimations for the three methods are compared in terms of distance and time in Fig. 10. The estimation errors are reported as a function of the proportion of randomly masked tap-off records in the data. The solid lines indicate the mean and shaded areas indicate the standard deviation of the error. It is seen that taking the closest alighting point as the default choice, results in the worst estimation accuracy among the three approaches. The proposed method shows superior performance to the other methods, and decreases the distance error by approximately 51% (49%), and reduces the time error by approximately 25% (21%) compared to the approach choosing the closest stop (fastest stop) to estimate the missing alighting transactions.

In summary, the results of applying the proposed method show that when a missing passenger tap-off is followed by at least one tap-on from the same card, in approximately 47% of the cases the method is able to find the exact missing alighting point (Fig. 8). For approximately another 46% of the cases, the method estimates the transaction, which on average is only a couple of PT stops away from the target (Fig. 9). Given the large metropolitan area of Melbourne and especially the low density of the network in the majority of the regions covered by its services, retrieving missing transactions with such an average error is an immense enhancement of the passenger trip data. There are records from more than 490,000 (230,000) bus and tram trips per weekday (weekend) collected in Melbourne during September and October 2017, out of which the alighting information is missing in 240,000 (115,000) of the cases. Applying the proposed procedure leads to successful estimation of close to 170,000 (80,000) missing transactions, during weekends (weekdays), which is a significant enhancement to the passengers' travel data.

5.2. *Linking trips at transfers*

To identify transfers linking the passenger trips, an allowable transfer time needs to be determined. We solve this as a binary classification problem (see the *section 3.2*), where the time span between two trips (ITT) performed by a single passenger is used as input to the classifier and the classifier labels the interchange event as either a transfer or an activity. Here, we only deal with consecutive trips recorded for the same passenger where alighting time and location of each trip in the sequence allow the passenger to reach the next boarding point in time.

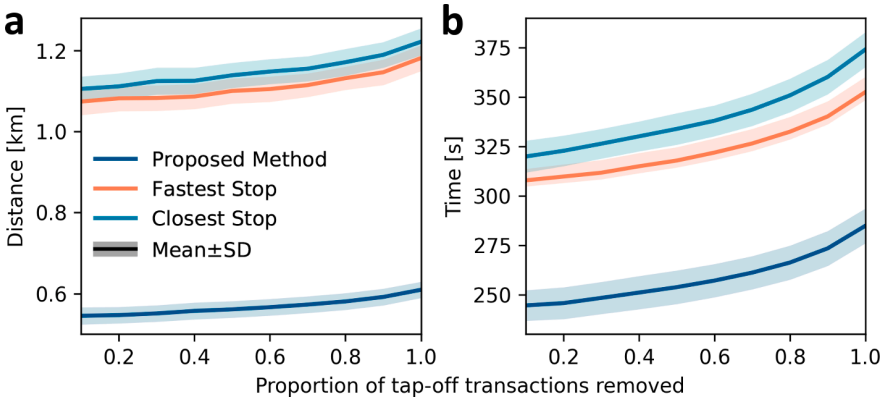


Fig. 10. Comparison between the proposed and the conventional approaches. The (a) distance and (b) time error between the estimated and the actual transactions when the exact transaction could not be regenerated by the estimation methods. The reported results compare the performance of the proposed method (blue) to the conventional approaches in the literature, where it is assumed that passengers always choose the closest (dark-cyan) or the fastest (orange) alighting points with respect to their next boarding point. Solid lines indicate the mean and shaded areas indicate the standard deviation of error, calculated over 10 realizations for each particular proportion of the transactions masked. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

First, using the proposed Alg. 1, we detect a set of return trips performed using any of the three modes in Melbourne's PT network, separately for weekdays and weekends. Let us denote the IIT between consecutive trips with θ. Algorithm 1 returns the histograms of transfers and activities over the variable θ. The histogram of activities and transfers provide a very good estimation for the distribution of the duration θ of transfers ($P(X_t = \theta)$) and activities ($P(X_a = \theta)$) in Melbourne's PT network; see Fig. 11a for weekdays and Fig. 11c for weekends. Histogram of the transfer durations demonstrates a very large number of transfers associated with small θ, while the number of observed transfers decreases drastically as θ increases. The activity duration histogram for weekdays depicts three local maxima, approximately at 1.5, 7, and 8.5 h; see Fig. 11a. Evidently, the first peak is associated with common daily non-occupational or short activities. The other two peaks in the distribution, roughly at 7 (i.e., $\theta = 7$) and 8.5 h long, are in surprising agreement with regular school day and daily working hours, which are 6.5 h (School Policy and Advisory Guide Hours, 2019) and 7.6 h (Fair work ombudsman Employees, 2019), respectively; consider the access/egress time to/from PT services.

The optimal decision threshold is found (see *section 3.2.2* for details) for a classifier with maximal informedness, to identify the transfers and activities in the travel data. Based on the CDF of the transfer durations and CCDF of the activity durations, the optimal ITT threshold for weekdays (weekends) is found to be $\theta_T^{opt} = 47$ min ($\theta_T^{opt} = 52$ min); see Fig. 11b and Fig. 11d associated with weekdays and weekends respectively. The search for the optimal threshold in the threshold domain is performed with 1 min resolution. In order to derive each passenger's journeys (OD trips), first the trip legs should be sorted chronologically. Then, each pair of consecutive trips are assigned to a single journey if the ITT between alighting from the first trip to boarding for the second trip does not exceed the determined time threshold θ_T^{opt}. Each journey is the result of connecting those single trip-legs which are assigned to it via the above rule. Each resulting journey can be described with two location-timestamps pairs associated with the boarding for the first trip-leg (origin) and alighting from the last trip-leg (destination) in the identified chain of linked trips.

We detect an average of approximately 116,400 (48,600) transfers on weekdays (weekends) in Melbourne's PT trips during September and October 2017. Figure 12 shows the difference between the distribution of travel times for single-leg trips (blue) and that of the journeys (red). The distributions have similar shape (approximately log-normal). But aggregating passenger trips at transfers to build journeys, increases the mean travel time (from 28.0 min for single-leg trips to 33.2 min for journeys), hence, the distribution of journey travel times becomes less right-skewed compared to that of the single-leg trips.

5.3. Generating the origin–destination (OD) travel demand

The reconstructed passenger journeys are described by their exact origin/destination locations in the network, which has a downside when this information is used for generating the OD travel demand. The issue is that considering each unique stop as a row and column of the OD matrix leads to an excessively large and sparse matrix which is not an efficient way of storing the information. Also, it does not provide the best representation of the travelling flows across a large area, for further analyses. In particular, in PT networks there may be multiple stops close to each other at different sides of a road serving different route directions, at hubs where different routes meet, or at intersections usually serving different routes. In these cases, a passenger trip or PT service between the

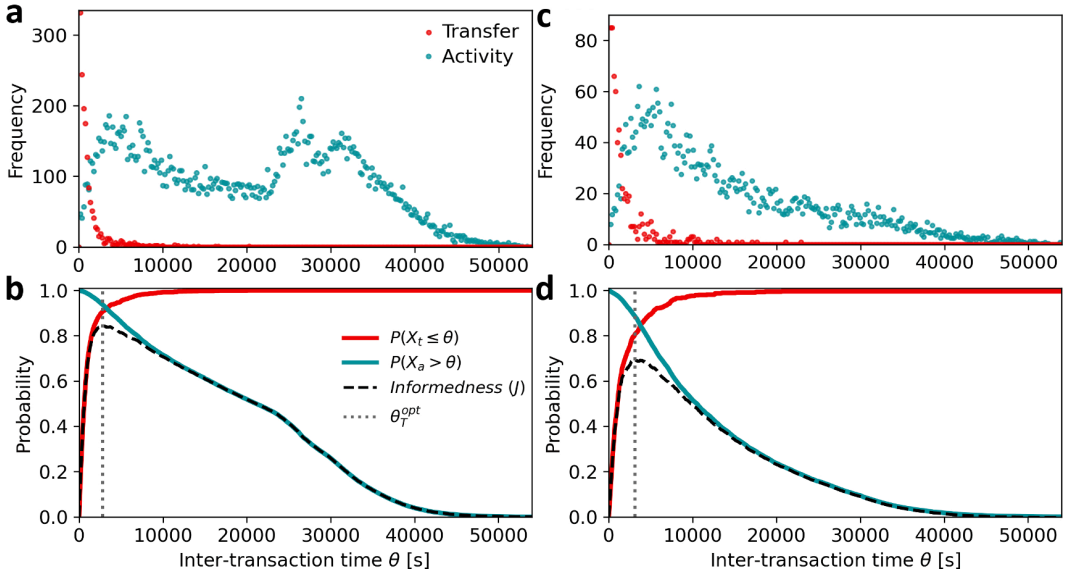


Fig. 11. Classifying the inter-transaction (interchange) events. Distribution of activities and transfers for the inter-transaction time θ, during (a) working days and (c) days off. Using the estimated distributions of transfers and activities for θ, the optimal decision threshold θ_T^{opt} (dotted gray line) for classifying events into transfers and activities is found for (b) weekdays and (d) weekends.

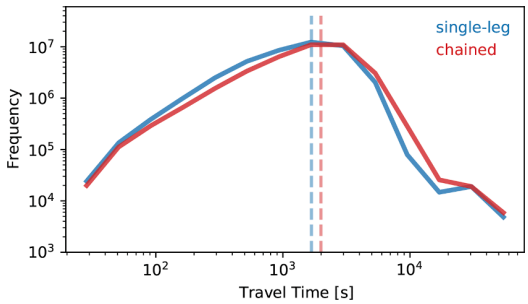


Fig. 12. Trip duration distribution. Histograms show the distribution (solid lines) and mean (dashed lines) of travel time for single-leg trips (blue) and OD trips or journeys (red) during the month of October 2017. Consecutive passenger trip-legs are chained at identified transfers to make journeys. As both distributions have long tails, where both have very small frequencies, the figure is presented in log–log scale to clearly show the discrepancy between the distributions both for low and high travel times. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

closely situated stops is very unlikely to happen, as passengers can conveniently walk between such stops. Therefore, we perform spatial clustering of the stops in the PT network and merge the closely located stops into fewer stop clusters. Stop clusters are then mapped to nodes in the OD demand network of the PT system. See *Appendix B* for details on our suggested method of spatial clustering to be performed on the set of stops in a PT system prior to generating its OD demand matrix.

The OD demand matrix for the PT network can be generated for a desired time period, say, a 3-hour period or a day, using the OD trips starting and ending within the selected time period. Alternatively, for a particular time t, the OD matrix can be generated using the ongoing OD trips, i.e., trips that started before t but ended after that time; this is the approach used here. The OD demand matrix $F_t = [f_{ij}]_{n\times n}$ at a particular time t is an $n \times n$ matrix, where n is the number of origin or destination points in the network. Here, we take the stop clusters found by Alg. B.1 (see *Appendix B*) as origin/destination points. Entry (i, j) of the matrix F (i.e., f_{ij}) shows the number of 'journeys' started at origin i and ended at destination j. The calculated total demand (number of instantaneous journeys) from the smartcard data on the PT network at different times of the day is shown in Fig. 13, separately for weekdays and weekends. Here, the total demand at any time t is calculated as the sum of all entries of matrix F_t, i.e., $\mathbf{1}_n^{\mathrm{T}} F_t \mathbf{1}_n$, where $\mathbf{1}_n$ is a column matrix of all n entries equal to unity.

We finish this section by presenting two visualizations of the OD travel demand matrices generated from Melbourne's smartcard data using the proposed framework in this study. Figure 14 shows all nodes in the PT demand network of Melbourne during weekdays (Fig. 14a) and weekends (Fig. 14b), where the size of each node shows its daily average in-flux (blue) or out-flux (red). A relatively low number of destination points (mostly train stations) have a significantly larger size than the rest, especially over weekdays. Note that the total in-flux is equal to the total out-flux, so larger size of the hotspot destinations compared to hotspot origins suggests that passenger flows are relatively more convergent to the hotspot destination nodes than being divergent from the origin hotspots (compare left and right panels of Fig. 14a).

The OD demand network of Melbourne's multi-modal PT system is illustrated at different times during a regular weekday in Fig. 15. For better clarity, we removed the links with lower than 5 passenger journeys from the network. The approximate location of the central business district (CBD) in Melbourne is easily noticed, as the CBD area happens to be one end of a majority of links in the demand network which has a star-like structure. In order to aid understanding the direction of passenger flows, network links are color-coded based on how they change passengers' distance from the CBD. It is interesting to see the transition in network's total demand and flow direction over the day by looking at different snapshots of the network from 8:00 AM (morning peak) to 5:00 PM (evening

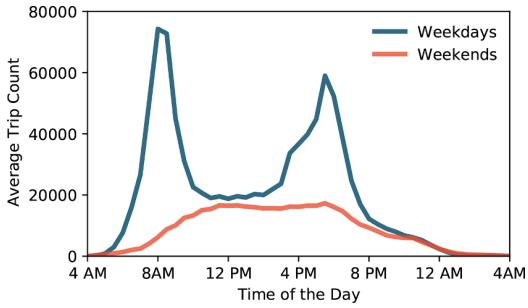


Fig. 13. Temporal network volume. The curves show the total number of passenger journeys happening on the network at each time of the day. The curves depicted here are the result of averaging over all days during the course of the available data.

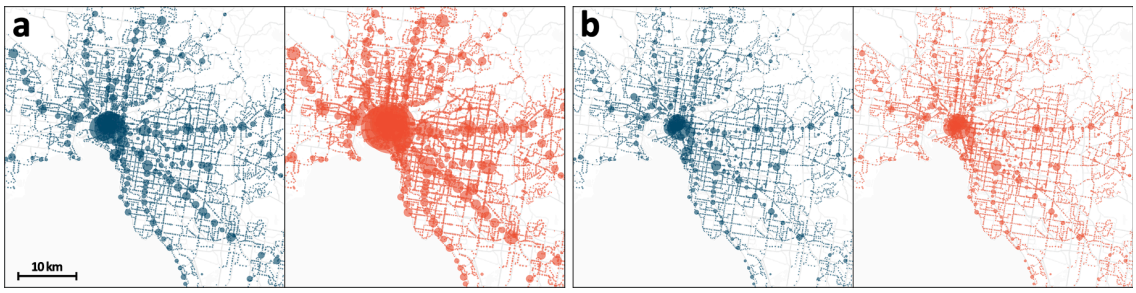


Fig. 14. Spatial distribution of origins and destinations in Melbourne's public transportation (PT) network. Origin (left panel) and destination (right panel) points in Melbourne's PT demand network during (a) weekdays and (b) weekends. The size of the scattered points indicates the daily average number of outgoing (incoming) passenger journeys from (to) network nodes as origins (destinations).

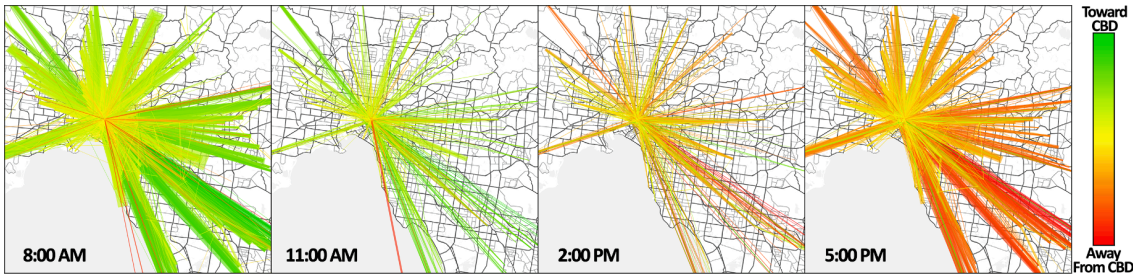


Fig. 15. Temporality of Melbourne's PT Origin-Destination demand network. The Origin-Destination demand network of Melbourne's PT system at different times in a regular weekday. The direction of passenger flow is depicted by link colors, where green indicates moving toward the central business district and red demonstrates the opposite. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

peak). As one may expect, the majority of passengers is headed toward CBD at morning peak, while the direction is gradually changed and becomes almost in the opposite direction at evening peak.

6. Conclusion

In this study, the major problems in the standard process of OD demand estimation from PT smartcard data were tackled: 'inferring missing alighting records' and 'identifying transfers and activities.' The existing methods are often based on a number of predetermined assumptions and parameters, which reduce the accuracy of the outcomes. Even if these parameters are carefully tailored for a particular PT system, they can easily fail to perform well for others. Our approach eliminates the need for devising a system of assumptions and parameters using the expert knowledge or additional data sources. The proposed method here, makes use of statistical pattern recognition tools to extract knowledge from the available data. Derived data patterns are then used to build optimal classifiers, which can predict the parameters required to tackle the main challenges in smartcard data processing. In particular, our approach is flexible to model passengers' choice of alighting point and transfer/activity duration from any smartcard data. Hence, the proposed framework can be applied to various smartcard data settings.

We demonstrated the performance of the proposed procedure by applying it to the large-scale PT smartcard data collected in Melbourne mainly to i) enhance the data by inferring missing tap-off records and ii) identify linked passenger trips via transfers, in order to generate the smartcard-based OD demand matrix of the PT network. Elaborate evaluations showed that substituting conventional assumptions with predictive models effectively improves the performance in estimating missing alighting transactions. We also estimated the distribution of activity durations between PT rides, which apart from capturing short activities was surprisingly consistent with the daily school and business hours. The temporal OD matrix of the network as the final product of the proposed procedure, reflected the expected demand profile of the PT system. Visualization of the generated Melbourne's PT demand network (from the OD matrix), demonstrated the temporal evolution of the demand volume and its directionality. Overall, the results suggest that the framework is effective in processing and enhancing PT smartcard data to extract passenger journeys in PT networks.

The proposed framework is designed to deal with challenges of extracting smartcard-based OD demand for PT networks. In future works and applications, potential use of additional information sources to build on top of the methods presented here, can be investigated. Alternative data sources can used to rectify issues arising from the limited quality of smartcard data and shortcomings of

specific PT systems in recording passenger data. As examples of system-specific defects in the data, Melbourne has a small tram zone where it is not mandatory for passengers to perform transactions, or some PT systems function with both smartcard card and paper tickets, and these result in a portion of trips to be missing from the AFC records. A widely-used solution for such issues is using travel data available from sources such as household travel survey and Automated Passenger Counts (APC) to scale the whole OD demand matrix or the number of trips between some OD node pairs (Kumar et al., 2018). Smartcard-based OD demand can be also corrected using APC and fare evasion report data (Munizaga et al., 2020). Finally, passenger behavior modeling presented in this study can be enhanced by the aid of alternative transportation data sources. For example, characterizing passengers' alighting behavior may be improved with demographic data input, or estimation of transfer/activity durations may benefit from the use of survey data.

In addition to the features used in behavior modeling here, e.g., the shortcut effectiveness and the inter-transaction time θ, new features can be sought for better characterizing the passengers' behavior and enhancing the predictive models used in this study. Future research can also be conducted on augmenting the proposed framework here, with probabilistic models built from historical individual (or collective) smartcard usage.

Appendix

Appendix A. Unsupervised learning to capture alighting behavior

Here, we tackle the problem of capturing the alighting behavior assuming that there are no tap-off transactions available. Consider a passenger's choice between a pair of plausible alighting time-stop points with respect to the next boarding stop, where one alighting point leads to earlier arrival at the next boarding point (fast alighting point) and the other has less walking distance to the next boarding stop (close alighting point). In *section 3.1.1* we explained how to find the set of all plausible alighting points based on two consecutive boarding transactions from a single passenger. In *section 3.1.2* it is shown that the choice between each pair of these plausible alighting points can be viewed as a 'binary choice' between a fast versus a close stop with respect to the passenger's next boarding. There, we map these binary choices to data points with three attributes (see Eq. (5)). These attributes are i) riding time ratio (x_1), ii) walking distance ratio (x_2), iii) and shortcut effectiveness (x_3), calculated to characterize the difference between two alighting options. (Note that here we assume that the outcome of the choices is unknown and thus the data points have no label.)

In Fig. A.1, three graphs each representing two of these attributes, show all binary alighting choices extracted from one day of smartcard data. When the value of one of the above attributes approaches the domain extremes (0 or 1), the fast or the close alighting option becomes significantly more attractive than the other one (see *section 3.1.2*). So, if two binary choices are mapped to similar attribute vectors (the distance between their corresponding data points is small) it is likely that they have the same outcome, e.g., in both choices the fast stop is picked by the passenger. Thus, in effect, the distance between data points in the attribute space indicates whether or not the same option is picked in corresponding choices.

Clustering can be used to find structure and pattern in unlabeled data by categorizing data points into classes, with high similarity among data points in each class and high dissimilarity between data points from different classes. Here in particular, clustering is used to categorize data points into two classes, aiming to separate choices that are likely to have different outcomes. We applied the well-performing BIRCH (balanced iterative reducing and clustering using hierarchies) algorithm (Zhang et al., 1996) to the alighting choices in Melbourne's PT network. BIRCH method allows us to predetermine the number of clusters (i.e., 2 clusters are desired here).

Fig. A.1 shows the result of applying BIRCH algorithm to our data with the goal of clustering similar (dissimilar) data points into same (different) clusters. By comparing the class centroids, we can easily associate each class with one of the possible outcomes for alighting choices; one cluster has lower x_1 and higher x_2 and x_3 which should be associated with choices where the fast stop is significantly more attractive than the close option (and vice versa).

Several approaches can be taken to predict the passengers' choice. BIRCH is an incremental clustering method, meaning that it processes all data points to find the best clusters, but with new incoming data point it does not need to reprocess the existing data and can update the clusters. Using BIRCH to study PT smartcard data allows for implementing procedures that can take the stream of input data as it is recorded and process it online. Here, we perform clustering on one day of the available data and then label the data based on the clustering results and finally, train a random forest model using the labeled data. This model predicts the passengers' choice between fast-close pairs of alighting options with 69% accuracy (random forest trained on data including tap-off transactions has 74% accuracy).

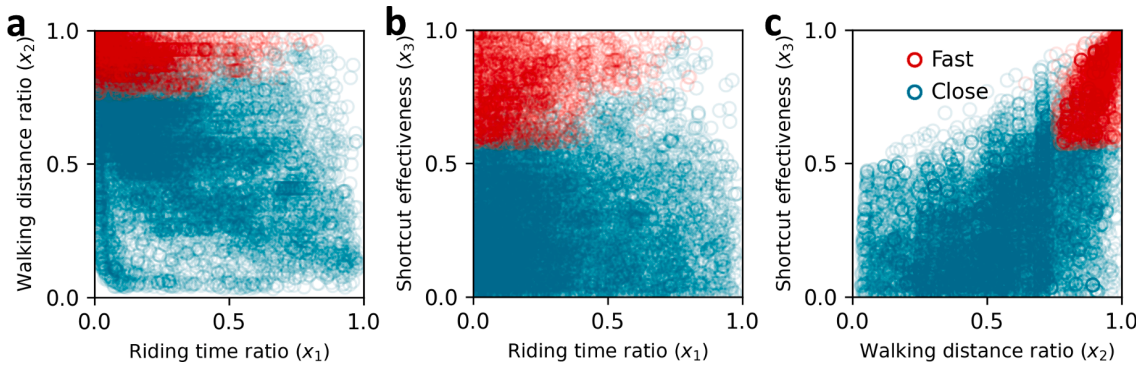


Fig. A.1. Clustering passengers' binary alighting choices between fast-close options. Each data point represents a passenger choice between a pair of plausible alighting stops (derived from Melbourne's smartcard data on a regular weekday). The three attributes, namely, walking distance ratio, riding time ratio, and shortcut effectiveness describe the 'choice' by comparing the pair of alighting options. Clusters of different choice-outcomes found by BIRCH algorithm are color-coded.

Appendix B. Spatial clustering of public transportation stops

The spatial clustering algorithm presented here (Alg. B.1), has a simple implementation and is generally applicable to any set of spatial data points. It uses a key parameter which is easily adjustable for different contexts. The goal here, is to merge closely located public transportation (PT) stops on different sides of the road or in PT stop hubs, into a stop-cluster. Stop clusters mapped to network nodes provide a better representation of the PT system's structure.

Pseudo-code for spatial clustering of stops
Input: $\{s_1, s_2, \cdots, s_m\}$; a set of stop identifiers associated with specific locations.
Output: $\{c_1, c_2, \cdots, c_m\}$; a set of labels where each c_i assigns the stop s_i to a cluster.

(1) Initialize the label variable $l \leftarrow 1$
(2) $S \leftarrow$ Initialize the set of unvisited stops with all the input points $\{s_1, s_2, \cdots, s_m\}$
(3) **While** the set of unvisited stops S is nonempty:
(4) $\quad s_i \leftarrow$ select an unvisited stop from S
(5) $\quad C \leftarrow Form\_Initial\_Cluster(\{s_i\}, S)$
(6) $\quad$ **While** $|\{s_i \in C | d(s_i, centroid(C)) > r_{max}\}| > 0$:
(7) $\quad\quad C \leftarrow C - \underset{s_j \in C}{\operatorname{argmax}}\ d(s_i, centroid(C))$
(8) $\quad$ **End While**
(9) $\quad S \leftarrow S - C$
(10) $\quad$ **For** each s_i in C:
(11) $\quad\quad c_i \leftarrow l$
(12) $\quad$ **End For**
(13) $\quad l \leftarrow l + 1$
(14) **End While**
(15) **Return** $\{c_1, c_2, \cdots, c_m\}$
(16) **Function** $Form\_Initial\_Cluster(C', S')$:
(17) $\quad C'' \leftarrow \{s_i \in S' | \exists s_j \in C' : d_E(s_i, s_j) \leq r_{max}\}$
(18) $\quad$ **Return** $C' \cup Form\_Initial\_Cluster(C'', S' - C'')$
(19) **End Function**

Algorithm B.1. ***Spatial clustering for stops.*** *Given a set of stops with their corresponding location, the algorithm labels the stops with a cluster number. The aim is to cluster the stops in a hub, or different sides of the road together while limiting the radius of the cluster to grow more than r_{max}. Here, $d_E(.,.)$ returns the Euclidean distance between the location of its two inputs.*

The algorithm starts with recursively merging the stops located closer than a predefined distance r_{max}, into a cluster. However, there may be a chain of pairwise closely located stops (especially in high-density areas) forming a very large stop-cluster. To handle this issue, the algorithm limits the radius of each cluster below r_{max}; cluster radius is the maximum distance between the cluster centroid and a member of the cluster, where the cluster centroid is the center of mass in the cluster assuming that the members have equal masses. After each initial cluster is formed with the recursive procedure, the cluster radius is checked and if it is larger than r_{max}, the stop with the highest distance from the centroid is excluded from the cluster. A cycle of i) checking the radius, ii) finding the furthest stop from the centroid, and iii) exclusion of the found marginal stop, is repeated until the stop radius meets the requirement.

References

Adler, T., Ben-Akiva, M., 1979. A theoretical and empirical model of trip chaining behavior. Transportation Research Part B: Methodological 13 (3), 243–257.

Alsger, A., Assemi, B., Mesbah, M., Ferreira, L., 2016. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. Transportation Research Part C: Emerging Technologies 68, 490–506.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., Hickman, M., 2018. Public transport trip purpose inference using smart card fare data. Transportation Research Part C: Emerging Technologies 87, 123–137.

Alsger, A.A., Mesbah, M., Ferreira, L., Safi, H., 2015. Use of smart card fare data to estimate public transport origin–destination matrix. *Transportation Research Record: Journal of the*. Transportation Research Board(2535), 88–96.

Barry, J.J., Freimer, R., Slavin, H., 2009. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transportation Research Record 2112 (1), 53–61.

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., Portugali, Y., 2012. Smart cities of the future. The European Physical Journal Special Topics 214 (1), 481–518.

Delbosc, A., Currie, G., 2016. Four types of fare evasion: A qualitative study from Melbourne, Australia. Transportation Research Part F: Traffic Psychology and Behaviour 43, 254–264.

Estgfaeller, N., Currie, G., De Gruyter, C., 2017. When less is more: Exploring trade-offs in transit route concentration, *Transportation Research Board (USA) Annual Meeting 2017*. Transportation Research Board 1–11.

Ferraro, R., Aktihanoglu, M., 2011. Location-aware applications. Manning Publications Co.

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H., Attanucci, J.P., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transportation research record 2343 (1), 17–24.

Guo, Z., 2009. Does the pedestrian environment affect the utility of walking? A case of path choice in downtown Boston. Transportation Research Part D: Transport and Environment 14 (5), 343–352.

Hamedmoghadam, H., Jalili, M., Vu, H.L., Stone, L., 2021. Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks. Nature Communications 12 (1), 1–10.

Hamedmoghadam, H., Ramezani, M., Saberi, M., 2019. Revealing latent characteristics of mobility networks with coarse-graining. Scientific Reports 9 (1), 1–10.

Hazas, M., Scott, J., Krumm, J., 2004. Location-aware computing comes of age. Computer 37 (2), 95–97.

He, L., Trépanier, M., 2015. Estimating the destination of unlinked trips in transit smart card fare data. Transportation Research Record 2535 (1), 97–104.

School Policy and Advisory Guide: School Hours, 2019. https://www.education.vic.gov.au/school/principals/spag/management/Pages/hours.aspx (Accessed 05/03/2019).

Fair work ombudsman: Full-time Employees, 2019. https://www.fairwork.gov.au/employee-entitlements/types-of-employees/casual-part-time-and-full-time/full-time-employees (Accessed 05/03/2019).

Ho, T.K., 1995. Random decision forests, Proceedings of 3rd international conference on document analysis and recognition. IEEE, pp. 278-282.

Hof, A., Elzinga, H., Grimmius, W., Halbertsma, J., 2002. Speed dependence of averaged EMG profiles in walking. Gait & Posture 16 (1), 78–86.

Hussain, Etikaf, Bhaskar, Ashish, Chung, Edward, 2021. Transportation Research Part C: Emerging Technologies 125, 103044.

Ickowicz, A., Sparks, R., 2015. Estimation of an origin/destination matrix: application to a ferry transport data. Public Transport 7 (2), 235–258.

Kumar, P., Khani, A., He, Q., 2018. A robust method for estimating transit passenger trajectories using automated data. Transportation Research Part C: Emerging Technologies 95, 731–747.

Kurauchi, F., Schmöcker, J.-D., 2017. Public transport planning with smart card data. CRC Press.

Kusakabe, T., Iryo, T., Asakura, Y., 2010. Estimation method for railway passengers' train choice behavior with smart card transaction data. Transportation 37 (5), 731–749.

Li, T., Sun, D., Jing, P., Yang, K., 2018. Smart card data mining of public transport destination: A literature review. Information 9 (1), 18.

Ma, X.-L., Wang, Y.-H., Chen, F., Liu, J.-F., 2012. Transit smart card data mining for passenger origin information extraction. Journal of Zhejiang University Science C 13 (10), 750–760.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies 36, 1–12.

Meng, Q., Qu, X., 2013. Bus dwell time estimation at bus bays: A probabilistic approach. Transportation Research Part C: Emerging Technologies 36, 61–71.

Min, Y.-H., Ko, S.-J., Kim, K.M., Hong, S.-P., 2016. Mining missing train logs from Smart Card data. Transportation Research Part C: Emerging Technologies 63, 170–181.

Mitchell, T.M., 1997. Machine Learning. McGraw Hill.

Mohamed, K., Côme, E., Oukhellou, L., Verleysen, M., 2016. Clustering smart card data for urban mobility analysis. IEEE Transactions on Intelligent Transportation Systems 18 (3), 712–728.

Montufar, J., Arango, J., Porter, M., Nakagawa, S., 2007. Pedestrians' normal walking speed and speed when crossing a street. Transportation Research Record 2002 (1), 90–97.

Munizaga, M., Devillaine, F., Navarrete, C., Silva, D., 2014. Validating travel behavior estimated from smartcard data. Transportation Research Part C: Emerging Technologies 44, 70–79.

Munizaga, M.A., Gschwender, A., Gallegos, N., 2020. Fare evasion correction for smartcard-based origin-destination matrices. Transportation Research Part A: Policy and Practice 141, 307–322.

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies 24, 9–18.

Nassir, N., Hickman, M., Ma, Z.-L., 2015. Activity detection and transfer identification for public transit fare card data. Transportation 42 (4), 683–705.

Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M., 2011. Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. Transportation Research Record 2263 (1), 140–150.

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19 (4), 557–568.

OpenStreetMap Copyright and License, 2020. https://www.openstreetmap.org/copyright (Accessed 18/4/2020).

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Primerano, F., Taylor, M.A., Pitaksringkarn, L., Tisato, P., 2008. Defining and understanding trip chaining behaviour. Transportation 35 (1), 55–72.

Robinson, S., Narayanan, B., Toh, N., Pereira, F., 2014. Methods for pre-processing smartcard data to improve data quality. Transportation Research Part C: Emerging Technologies 49, 43–58.

Shafiei, S., Saberi, M., Vu, H.L., 2020. Integration of Departure Time Choice Modeling and Dynamic Origin-Destination Demand Estimation in a Large-Scale Network. Transportation Research Record 0361198120933267.

Sun, L., Lee, D.-H., Erath, A., Huang, X., 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system, Proceedings of the ACM SIGKDD international workshop on urban computing, pp. 142-148.

Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems 11 (1), 1–14.

Utsunomiya, M., Attanucci, J., Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning. Transportation Research Record 1971 (1), 118–126.

Wang, W., Attanucci, J., Wilson, N., 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems. Journal of Public Transportation 14 (4), 131–150.

Wong, K.-I., Wong, S.C., Tong, C., Lam, W., Lo, H.K., Yang, H., Lo, H., 2005. Estimation of origin-destination matrices for a multimodal public transit network. Journal of advanced transportation 39 (2), 139–168.
Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3 (1), 32–35.
Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Record 25 (2), 103–114.
Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T., 2018. Big data analytics in intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems 20 (1), 383–398.

Chapter 4

Percolation-based Reliability Analysis

Effective approaches towards resolving or mitigating the congestion problem in transportation networks can save an immense amount of time and resources while contributing to the comfort of countless users. In this chapter, a percolation-based analysis is proposed to unpack the organization of congestion at different levels with respect to the heterogeneous passenger flow demand in transportation networks. In the previous chapter, a procedure was elaborated to enhance the smartcard data and extract the travel demand from it. The results are used in this chapter in percolation-based analysis of real-world public transportation networks with the aim of effective measurement of the impact of congestion on traveling flows and identifying bottlenecks to decisively reduce that negative impact. The chapter is presented by a published journal paper [49] which embodies a main article and a supplementary information document, both presented respectively in the following.

4.1 Publication

The following of this chapter includes an article with its accompanying supplementary, associated with the citation information below:

> H. Hamedmoghadam, M. Jalili, H. L. Vu, and L. Stone, "Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks," *Nature Communications*, vol. 12, no. 1, pp. 1–10, 2021.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21483-y OPEN

Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks

Homayoun Hamedmoghadam [1✉], Mahdi Jalili [2], Hai L. Vu [1] & Lewi Stone [3✉]

Whether it be the passengers' mobility demand in transportation systems, or the consumers' energy demand in power grids, the primary purpose of many infrastructure networks is to best serve this flow demand. In reality, the volume of flow demand fluctuates unevenly across complex networks while simultaneously being hindered by some form of congestion or overload. Nevertheless, there is little known about how the heterogeneity of flow demand influences the network flow dynamics under congestion. To explore this, we introduce a percolation-based network analysis framework underpinned by flow heterogeneity. Thereby, we theoretically identify bottleneck links with guaranteed decisive impact on how flows are passed through the network. The effectiveness of the framework is demonstrated on large-scale real transportation networks, where mitigating the congestion on a small fraction of the links identified as bottlenecks results in a significant network improvement.

[1] Institute of Transport Studies, Faculty of Engineering, Monash University, Melbourne, VIC, Australia. [2] Electrical and Biomedical Engineering, School of Engineering, RMIT University, Melbourne, VIC, Australia. [3] Mathematical Sciences, School of Science, RMIT University, Melbourne, VIC, Australia. ✉email: homayoun.hamed@monash.edu; lewistone100@gmail.com

Recent theoretical advances in network science have considerably contributed to our understanding of complex systems, cutting across many disciplines from the social and technological sciences to the fields of ecology and biology[1–12]. In many modern studies, percolation theory[13] has been frequently employed to characterize the structure, functionality, and resilience of network systems. In this approach, link failure is simulated by a percolation model which progressively removes links from the network[14,15]. The impact is usually measured via a reduction in the size of the network's largest connected component, or giant component (GC), as links are gradually removed[16–18]. Different strategies for simulating link failures, e.g., random (error) or targeted (attack)[19], make it possible to study a range of different topological characteristics.

In real infrastructure networks, however, pervasive phenomena such as various forms of congestion (e.g., traffic jams in transportation or packet congestion in communication networks) reduce the quality of flow movement on links in a continuous manner rather than necessarily causing a complete failure. To consider this, link-level flow dynamics on a network G can be modeled by associating each link e_{ij} (connecting node i to node j) with its own "quality" attribute $q_{ij} \in (0, 1]$, which at any time indicates the link performance relative to an observed or predetermined maximum level of performance[20,21]. For example, in a road traffic network with the speed on each road changing temporally, link quality q_{ij} can be defined as the ratio of instantaneous traffic speed to the speed limit of the link e_{ij}[22,23], or in a communication network, quality can be defined as the instantaneous delivery rate of packets flowing along a link[24].

Percolation models have been used to study the organization of link-qualities in networks[23,25,26]. The basic concept requires examining a single network G which may change in time, but at each particular time, the structure and link qualities represent the system's state. The percolation process on G may be seen as a function of a threshold ρ where $0 \le \rho \le 1$. For any specific threshold ρ, the idea is to delete any link in G with quality q_{ij} for which $q_{ij} \le \rho$, leaving the subnetwork G_ρ; see the process on a small network in Fig. 1. We can then gain insights into the network G's properties by monitoring the geometrical phase transitions in G_ρ as ρ varies from $\rho = 0$ to $\rho = 1$. (Note that the whole percolation process is performed on one network snapshot, thus the quality of links representing the state at that snapshot remain fixed during the process.)

Of special interest is the critical percolation threshold $\rho = \rho_c$ at which the GC suddenly fragments into components of smaller size. The percolation threshold ρ_c is an informative measure of the global quality of network structure, indicating that the network fails to provide global connectivity only with paths of links having quality above ρ_c[24,27,28]. While this generic critical phenomenon is of vital importance for characterizing networks, we will show that limiting attention exclusively to the GC and its sudden fragmentation reveals only a part of the full picture when studying real-world problems.

The primary goal in many critical infrastructure networks such as communication, power distribution, and water supply systems is to serve the demand for a certain amount of flow between each pair of nodes; we refer to such systems as "demand-serving networks." In reality, the flow demand is often distributed heterogeneously over the origin–destination (O–D) node pairs in the network. For example, in transportation networks, the passenger travel demand is much larger between O–D points when one or both of them are hotspot locations[29]. The larger the flow demand between two nodes, the more crucial is their connecting paths[30]. When studying percolation in demand-serving networks, although the global connectivity is lost at percolation criticality, yet a substantial proportion of the network's flow demand might be between O–D node pairs that remain connected in the subcritical phase. For example, if the bulk of the flow demand is contained within isolated small and medium-sized clusters (resulting from the GC fragmentation), the network can remain highly functional even after the GC collapse (see the example in Fig. 1b). In other words, the global dynamics in demand-serving networks is not only controlled by the structure and organization of link qualities, but also by the distribution of the flow demand.

The goal of the present paper is to add further realism to percolation-based network analysis by the inclusion of heterogeneous flow demand. We restrict our attention, first to real transportation networks as exemplary instances of demand-serving networks, but then demonstrate the generality of our proposed analysis. We introduce a theoretical framework to quantify the impact of each link's quality (congestion) on flow movements through the network and use it to identify the network bottlenecks. We show that the percolation analysis suggested here can lead to different conclusions compared to those obtained solely from studying structural critical phenomena.

Results

The case of real infrastructure networks. We demonstrate the application of the proposed framework, on the bus and tram

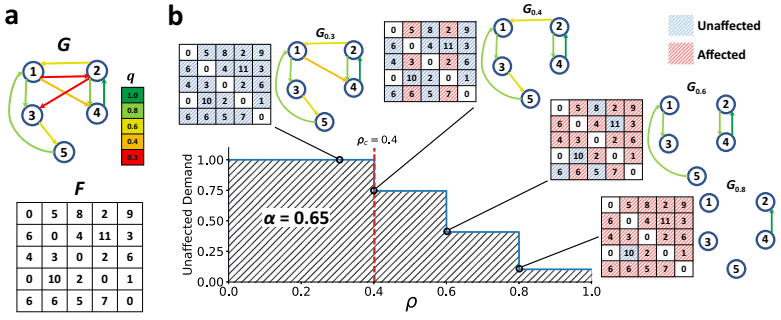


Fig. 1 Percolation on an example demand-serving network. a Network G with size $n = 5$, where quality q_{ij} of each link e_{ij} is color-coded (according to the color-bar). Matrix F quantifies the flow demand between all pairs of nodes which sums up to 100 units in total. **b** The percolation process is simulated by increasing a threshold ρ while removing links e_{ij} with $q_{ij} \le \rho$. Subnetwork G_ρ is visualized at different ρ's with its corresponding affected (red) and unaffected (blue) flow demand color-coded in matrix F. In this example, by definition, the system collapses at $\rho_c = 0.4$, when the 5-nodes strongly connected GC disintegrates into two strongly connected components of sizes 2 and 3, while unaffected demand (UD) is still at 75%. The reliability of the network G is $\alpha = 0.65$, found by calculating the area under the curve of UD versus ρ.

(on-road) public transportation (PT) systems in two major Australian cities, Melbourne and Brisbane, modeled using smart-card transaction data collected during September and October 2017 in Melbourne and over March 2013 in Brisbane. On-road PT systems are in constant conflict with road conditions, such as crowds, traffic, and signals, all negatively affecting the traveling flows by decelerating the PT vehicles. Separation of high demand O–D points by local pockets of congestion is an issue of considerable concern in transportation systems. The concept of travel demand distribution is fundamental to transportation theory[31], but to date, has not been considered in percolation-based analysis of dynamical transportation networks.

We are first interested in the network representation of the transportation system (PT services with disregard to the passenger activity). In this respect, network $G(V,E,t)$ at different times t of each particular day, was generated using the data time-stamped within the 2-h window centered at t (see "Methods" and Supplementary Note 1). Each node $i \in V$ corresponds to a cluster of closely situated bus and tram stops. A directed link $e_{ij} \in E$ connects its source node i to its target node j, if there is at least one PT service visiting node i and then j without any intermediate stops. A directed path from node o to node d is a sequence of links (all in the same direction) joining a sequence of distinct nodes, where the first node is o and the last node is d. In the second step, for each network G, the flow demand matrix $F = [f_{od}]$ was generated with f_{od} counting the number of passengers traveling from node o to node d, respectively, as the origin and destination points. Melbourne's on-road PT network was comprised of approximately an average of 5500 (2800) nodes, 10,500 (4500) links, and a flow demand derived from a part of 470,000 (210,000) trips performed during a normal weekday (weekend day). Brisbane has a relatively smaller network with approximately 1400 nodes and 3400 links on average over a regular weekday.

In order to quantify the link-level road conditions, we assign a quality attribute to each link e_{ij}, calculated as

$$q_{ij}(t) = \frac{\min_{t'}(\tau_{ij}(t'))}{\tau_{ij}(t)}, \quad (1)$$

where $\tau_{ij}(t)$ is the travel time on the link e_{ij} at time t of the day. The quality attribute $q_{ij}(t)$ indicates the effect of temporal link-level congestion on flows passing through e_{ij}. At any point in time, a high-quality link has relatively low travel time (or equivalently high velocity) compared to the rest of that day. In the following, for simplicity, we refer to the network and its attributes without the time parameter t. Figure 2a, b shows the spatial distribution of q_{ij} on the snapshot of the on-road PT network of Melbourne and Brisbane at 8:00 A.M. on a typical weekday. Note that the flow-demand matrix is determined from the passengers' activity data, while the network G and its link qualities are determined from PT vehicles' activity data.

The percolation process on a snapshot of Melbourne's PT network is illustrated in Fig. 2c, indicating a percolation threshold of $\rho_c = 0.39$ when global connectivity is lost. However, as our analysis shows, over 80% of trips are between O–D node pairs that still remain connected even though ρ has reached the percolation threshold (when only the links with quality $q > \rho_c$ are present). This highlights a problem with interpreting ρ_c as a reliability index (as per refs. [26,27,32]) if the main interest is on heterogeneous passenger flow demand. This motivated us to develop a new approach to capture the reliability of heterogeneous demand-serving networks.

Unaffected demand and network reliability. In this study, link removal in the percolation process should be viewed as a hypothetical procedure that unpacks the organization of congestion within a snapshot of the network in time. As explained before, the procedure is built upon constructing the subnetwork G_ρ which inherits all the links from the original network G except the most congested (lowest quality) links with qualities $q \le \rho$. By gradually increasing ρ, and at each step removing the shell of most congested links, the procedure extracts a series of subnetworks G_ρ, each providing a different level of flow movement on the actual network. The impact of different levels of congestion on flows can then be examined by studying the properties of subnetworks G_ρ, $\rho \in (0,1]$.

Our approach is based on monitoring what we refer to as unaffected demand (UD), and requires keeping track of the flow-demand between all O–D node pairs during the percolation process. The network's flow-demand is represented by the matrix $F = [f_{od}]$ of order n equal to the network size, where entry f_{od} is the amount of passenger-flow from origin node o to destination node d (see Fig. 1a). The matrix is normalized by dividing by the total demand $\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n$, to give $F/(\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n)$. (Here, $\mathbf{1}_n$ is a column vector of all n elements equal to one).

Using F that gives the flow-demand between any O–D pair, we can then calculate the UD as the percolation procedure proceeds and as low-quality links are removed. At any threshold ρ, the flow demand between an O–D pair is said to remain "unaffected" by link removals if there is at least one directed path from o to d remaining on G_ρ. To assist in interpreting this, consider a link that is part of a path that begins from origin node o and reaches destination node d. When the link is removed (because it has fallen below threshold in quality), then the fraction of the demand $f_{od}/(\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n)$ remains unaffected by the link removal if and only if there is still at least one other directed path from o to d. We thus define UD_ρ as the fraction of the total flow between all the O–D pairs that remain unaffected at threshold ρ of the percolation process. In other words, UD_ρ is equal to the fraction of the demand on G that can travel between their O–D nodes without having to traverse any link with quality below the threshold ρ. See "Methods" for the formulation of UD_ρ.

It is instructive to examine how UD_ρ varies with increasing ρ on the example network shown in Fig. 1a, where the total volume of flow demand is 100 by some arbitrary unit of measurement and $\mathrm{UD}_0 = 100/100$ initially. When $\rho = 0.3$ (Fig. 1b), two links of the lowest quality (colored red) are removed, but this does not affect the flow between any pair of nodes, and thus $\mathrm{UD}_{0.3} = 1$. When $\rho = 0.4$, however, removal of the link $1 \to 4$ prevents flows from reaching nodes 2 or 4 from either node 1, 3, or 5, by any path on $G_{0.4}$. The proportions of affected flows sum up to 25/100, thus the UD drops to $\mathrm{UD}_{0.4} = 0.75$.

We now present our key index for assessing the reliability of demand-serving networks. We define the demand-serving reliability α, as the area under the curve of UD_ρ over the domain of ρ (hatched area under the curve in Fig. 1b). In compact form, this can be formulated as

$$\alpha = \int_0^1 \mathrm{UD}_\rho d\rho = \int_0^1 \frac{\mathrm{tr}\,(R_\rho F^{\mathrm{T}})}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n} d\rho, \quad (2)$$

where tr(.) is the trace of the $n \times n$ square matrix. As seen in Eq. (2), it is also possible to formulate UD_ρ, and as a result α, in simple mathematical terms making use of the network's so-called reachability matrix R and the flow demand matrix F (see "Methods").

The meaning of UD_ρ and α, becomes clearer from viewing plots as in Fig. 2d. In such plots, if UD_ρ rapidly drops at relatively low ρ values, then most of the flow demand is constrained to traverse low-quality (congested) links. This in turn lowers the area under the curve of UD_ρ, and the reliability α will

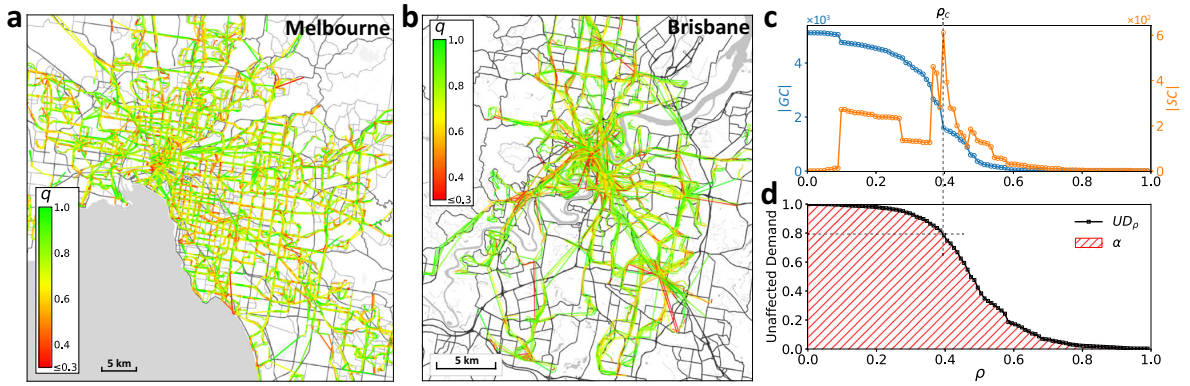


Fig. 2 Real on-road public transportation (PT) networks. a, **b** The network representation of the PT system with color-coded link qualities q at $t = 8$ (8:00 A.M.), for Melbourne on 1 September 2017 (**a**), and for Brisbane on 1 March 2013 (**b**). **c** Percolation process on Melbourne's network shown in (**a**). The size of the giant component |GC| and the size of the second-largest component |SC| are plotted as functions of the threshold ρ. The critical threshold $\rho = \rho_c$ is determined as the point of maximal |SC| (vertical dashed gray line). **d** Percolation process on Melbourne's network shown in (**a**). Unaffected demand is plotted as a function of ρ (UD_ρ) which at percolation critical threshold shows the value of $UD_{\rho_c} \approx 0.8$ (marked by dashed gray lines). The area hatched in red corresponds to the reliability α of the network in (**a**). Streetmap layers in **a** and **b** ©OpenStreetMap contributors[44].

consequently be low. If UD_ρ does not drop rapidly until much larger ρ values, then most of the demand is between node pairs that are connected via paths of high-quality links, and the reliability α will be high. Hence, reliability α gives an indication of how well the flows pass between their O–D points given the organization of congestion on the network. (See Supplementary Note 2 on the relevance of the links' flow-capacity to our reliability analysis.)

Let $|GC_\rho|$ be the size (number of nodes) of the GC in G_ρ. In the "Methods", we show that when flow demand distribution is homogeneous (i.e., the passenger flow f_{od} is the same between all reachable pairs of nodes o and d), then on any large-enough undirected network, we have $|GC_\rho| \approx n.\sqrt{UD_\rho}$ at any threshold ρ during the percolation. Thus, only by assuming a uniform flow demand over the network, UD is able to replicate the percolation analysis based on monitoring the GC; this is also confirmed numerically later in the paper. Second, with heterogeneous flow demand, the above relation no longer holds, and the fall-off of UD as a function of ρ provides its unique description of the system dynamics. By aggregating UD's description of the system, α provides a simple and useful indication of network reliability.

Bottleneck identification. Improving the infrastructure networks via protection or enhancement of a minimal set of links is currently receiving intense research interest[20,33,34]. Our framework suggests a new approach for identifying network bottlenecks. Here, inspired by the work on the maximum capacity paths problem[35], we introduce the link criticality score s_{ij}, which quantifies the overall role of each link e_{ij} in impeding the network flows.

Suppose there is a set of different directed paths Ψ_{od} that connect node o to node d (see Fig. 3a). On each path $\psi \in \Psi_{od}$, we search for the link with the minimum quality (Fig. 3b). Among those particular links, we choose the link with the maximum quality (Fig. 3c), denote it by e^*_{od}, and refer to it as the "limiting link" associated with the O–D node pair (o,d). For simplicity, let us assume that each link quality value on the network is unique. Then, there will be only a single limiting link between any reachable pair of nodes. For a link e_{ij}, if it is never found to be the limiting link between a node pair, it will have a criticality score of zero. If $e_{ij} = e^*_{od}$ for only a single pair (o,d), then the link criticality score s_{ij} will be the fraction of the total demand that flows from o to d, i.e.,

$$s_{ij} = \frac{f_{od}}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n}. \tag{3}$$

The index relies on the feature that, for a given O–D pair, during the hypothetical percolation process, as soon as the threshold ρ reaches the quality of the associated limiting link, removal of the latter causes complete rupture of all paths between the O-D pair on G_ρ. This means the limiting link has the lowest quality, that flows are constrained to traverse in order to travel between their origin and destination nodes on the actual network G. If the link e_{ij} is the limiting link between several node pairs (see Supplementary Fig. 3A), Eq. (3) extends to

$$s_{ij} = \sum_{o,d \in V, e^*_{od} = e_{ij}} \frac{f_{od}}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n}. \tag{4}$$

We have identified an important relationship that connects the link quality (q_{ij}), the link criticality score (s_{ij}), and the network reliability (α), namely

$$\sum_{e_{ij} \in E} s_{ij}.q_{ij} = \alpha, \tag{5}$$

as proven in "Methods" (and illustrated in Supplementary Fig. 3B). It can be rigorously shown that for any link e_{ij}, increasing q_{ij} within a non-empty range will increase the network reliability α, with the magnitude of increase being proportional to s_{ij} (see Supplementary Note 3). (This is a nontrivial problem since alteration of the quality of any link in the network can change the criticality score of multiple links.) Therefore, after ranking the links according to their criticality scores, a desired number of the top-ranked links can be identified as network bottlenecks.

Numerical simulations were used to test how accurately the ranking of links based on link criticality scores (CS ranking) can identify network bottlenecks. To this end, first, a simple intuitive method was used to find the true bottleneck links, i.e., the ground truth. The method requires perturbing the quality q_{ij} of individual links by a small positive amount ε (we chose this to be $\varepsilon = 0.01$), one by one, and then ranking the links according to their ability to perturb the reliability score α. The link whose perturbation increases the reliability α the most is deemed to be the most

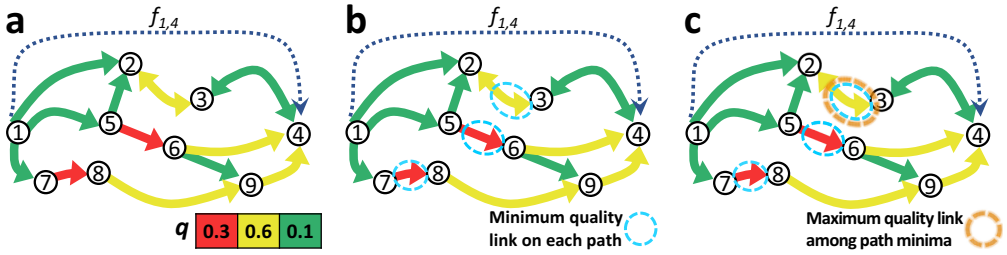


Fig. 3 Finding the limiting link between an origin-destination node pair. a A small network with color-coded link qualities q, where as an example, we demonstrate the process to identify the limiting link between the O–D node pair (1,4) having a directed flow demand of $f_{1,4}$. **b** The available paths from node 1 to node 4 (and path's minimum-quality link) are $1 \to 2 \to 3 \to 4$ ($e_{2,3}$), $1 \to 5 \to 2 \to 3 \to 4$ ($e_{2,3}$), $1 \to 5 \to 6 \to 4$ ($e_{5,6}$), $1 \to 5 \to 6 \to 9 \to 4$ ($e_{5,6}$), $1 \to 7 \to 8 \to 9 \to 4$ ($e_{7,8}$). **c** Among the minimum-quality links on these paths, $e_{2,3}$ has the maximum quality. Just below the threshold $\rho = 0.6$, still, two paths connect node 1 to node 4, but then with $e_{2,3}$ removed, node 4 becomes unreachable from node 1 on $G_{0.6}$. The limiting link associated with node pair (1,4) is $e_{2,3}$, thus, an increase in $q_{2,3}$ will increase the lowest quality that the flow from node 1 to node 4 is constrained to interfere with. The ratio of $f_{1,4}$ to the total demand, is added to criticality score $s_{2,3}$ of the link $e_{2,3}$ to reflect the importance of its quality $q_{2,3}$ for flow movement over the network.

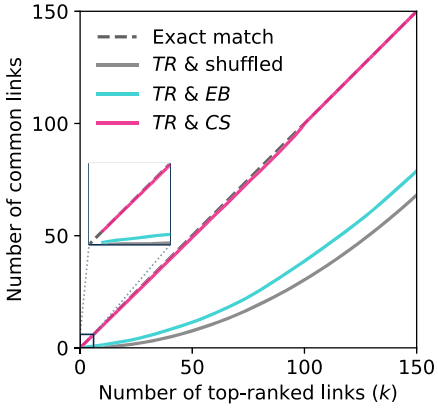


Fig. 4 Assessing the accuracy of link criticality score index in identifying the true bottlenecks. The true ranking of links (TR) in terms of their improvement effect on network reliability α, is compared to rankings based on link criticality score (CS), edge betweenness centrality (EB), and randomly shuffled rankings of links. Each curve shows the number of common links between the set of top-k true bottlenecks and top-k bottlenecks of another ranking scheme. A ranking equal to TR leads to a line lying on the diagonal dashed line.

critical link etc. Through this brute-force procedure, the true ranking (TR) of the criticality of all links are obtainable.

We applied the ranking schemes on random geometric graphs (RGGs) with $n = 100$ nodes spread over the space $[0,10]^2$ uniformly at random, and links connecting any pair of nodes with distance less than $r_0 = 1.5$ (which ensures connectivity and having over 300 links[36]).

To compare CS and TR rankings, we took the set of k top-ranked links in each ranking and counted the number of common links between them. Figure 4 shows the number of common links between the CS and true top-bottlenecks of the network for $k = 1, 2, \ldots, 150$, averaged over 500 realizations. We also compared against the ranking obtained by the conventional index edge betweenness centrality[37] (EB), and a randomly shuffled ranking. The set of CS bottlenecks was found to be almost exactly the same as the set of true bottlenecks (TR) with (on average) 98–100% of their elements matching for different k values. The EB and the shuffled rankings were by far inferior to the CS scheme as Fig. 4 shows, although as might be expected, the EB ranking had a higher accuracy compared to the shuffled ranking. Note that unlike the brute-force approach used to find TR, the criticality score s of all network links can be calculated via scalable algorithms, e.g., our suggested modified Dijkstra's algorithm (see Supplementary Note 3).

Application to public transportation networks. We return now to using the above tools to study the PT networks of Melbourne and Brisbane. Figure 2c illustrates the percolation process on Melbourne's bus and tram (on-road) PT network (at 8:00 A.M. on 1 September 2017) through |GC| and the size of the second-largest component (|SC|) as functions of ρ. In practice, the percolation threshold is determined as the threshold $\rho = \rho_c$ at which |SC| is maximal[38]. In Fig. 2c, the point of maximal |SC| captures the GC collapse, however, this was not always the case at other times and dates. The GC fragmentation during the percolation process was often blurred out rather than demonstrating a drastic change in |GC|, or in other cases, appeared as multiple peaks in |SC| which makes it difficult (if not impossible) to identify the critical threshold (Supplementary Fig. 4); ref. [39] reports similar observations in the road network of multiple cities. The index α evaluates the network according to the whole percolation process and does not depend on the existence of a clear phase transition, making the above issue irrelevant.

Figure 2d demonstrates the percolation process shown in Fig. 2c, but this time with UD as a function of ρ. As pointed out before, at the critical percolation threshold $\rho_c = 0.39$ where the global connectivity on G_ρ breaks down, we see that $\mathrm{UD}_{0.39} = 0.8$. Thus, 80% of all the trips on the network G are between O–D node pairs that remain connected after the breakdown of the GC, and only via paths of links with $q > 0.39$. This empirically demonstrates how characterizing a network based on ρ_c alone can be misleading when flow demand distribution is heterogeneous. In effect, during the percolation process, UD does not necessarily decline with the same rate as pairwise connectivity (see Supplementary Note 4 and Supplementary Fig. 5). For Melbourne's PT network, the number of connected node pairs on G_ρ decreases faster than UD_ρ, meaning that demand is higher within clusters of high-quality links in the network.

We also examined both reliability α and ρ_c on Melbourne's (Brisbane's) PT network over the main functioning hours of the system during September and October 2017 (March 2013), separately for weekdays and weekends. Temporally, ρ_c had relatively large fluctuations over the day, and there appeared to be no repeating pattern on a day to day comparison (see Fig. 5a, c for Melbourne and Brisbane networks, respectively). In contrast, the proposed reliability measure α followed a clear daily pattern (see Fig. 5b for Melbourne and Fig. 5d for Brisbane's PT network)

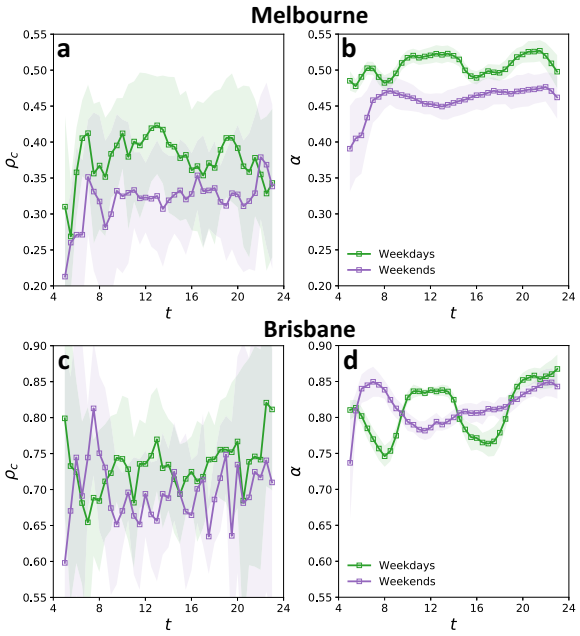


Fig. 5 Reliability of the on-road public transportation (PT) networks. **a**, **b** Temporal evolution of ρ_c (**a**) and α (**b**) for Melbourne's PT network, during weekdays (green) and weekends (purple). At each time t, curves show the mean, and shaded areas indicate the standard deviation of values around the mean, over September and October 2017. **c**, **d** Temporal evolution of ρ_c (**c**) and α (**d**) for Brisbane's network averaged over the days in March 2013, separately for weekdays and weekends.

with variations that have a relatively small standard deviation. (Supplementary Note 5 and Supplementary Fig. 6A, C provide more details concerning ρ_c and α and their comparison.) The approximately 10% drops in α at 8:00 and between 16:00 and 18:00 are associated with weekdays' morning and evening peak commuting periods when high rates of congestion and large numbers of commuters predictably increase the conflict between PT system and road conditions. Consistency of the daily evolution of α (for both Melbourne and Brisbane networks) with the circadian rhythm of urban human mobility and its low variability over different days indicate its success in unraveling the repeating daily pattern in complex interactions between major constituents of the system, namely, supply network structure, link-level congestion, and passenger flow demand (see Supplementary Note 5 for more detail). The results also suggest that Melbourne's PT network is relatively stable over a day, despite multiple periods of intense traffic, which is partially due to more available PT services during the rush hours which increase the number of links and thus network density (see Supplementary Fig. 7).

Despite the larger flow demand and more extensive congestions during weekdays, α was larger for weekdays compared to weekends in Melbourne (Fig. 5b). This is because Melbourne's PT network is fine-tuned for weekday demand, operating with a higher number of services during weekdays as compared to weekends. The larger number of PT services not only resulted in a larger number of network links but also led to a significantly higher link density during weekdays when compared to weekends (see Supplementary Fig. 7B). Higher link density of the network on weekdays means the availability of more paths between nodes and that if a path between two nodes includes congested links, it is generally more likely that an alternative less congested path exists. We also observed that in Melbourne's PT network during weekends a significantly larger proportion of the trips are to/from the central business district (CBD) area, where the links are often subject to a higher level of congestion than elsewhere in the network. Lower link density of the network together with the large proportion of the passengers traveling to/from CBD on weekends, results in more conflict between flows and congestion (that is what α measures) which is reflected with the lower network reliability α during weekends. (From UD's perspective, a larger proportion of the network demand has to pass through lower-quality links during weekends compared to weekdays.) In Brisbane, however, although the network has more links during weekdays, links (PT services) are supplying the transportation between a larger number of nodes, which keeps the link density of the network approximately the same between weekdays and weekends. As a result, unlike Melbourne, α fluctuated within approximately the same range during both weekdays and weekends for Brisbane's PT network (Fig. 5d). Yet, similar to the case of Melbourne's PT network, the daily evolution of Brisbane's PT network reliability α on weekdays had distinct patterns from that of weekends.

Bottlenecks of real transportation networks. Link criticality scores vary over time in temporal on-road PT networks. Therefore, we calculated the mean criticality score of each link over the course of the available data, and identified the network bottleneck links as those with the largest mean criticality scores, separately for weekdays and weekends. The identified bottlenecks were found to be robust, appearing with high criticality scores on most days (Supplementary Fig. 8).

The spatial distribution of link criticality scores over Melbourne's weekday PT network is portrayed in Fig. 6a (see also Supplementary Fig. 9A for Melbourne's weekends and Supplementary Fig. 10A for Brisbane). Pockets of traffic congestions and crowds, which decrease the quality of PT network links, are usually formed around the high-demand urban hotspots. As a result, links with large criticality scores were found to be situated in urban hotspots and the areas surrounding them, making the spatial distribution of link criticality scores in surprising alignment with the urban morphology. Specifically, Melbourne's biggest urban shopping center was surrounded by links with high criticality scores, and the top bottlenecks were mostly distributed around the single most significant hotspot of Melbourne which is the CBD. Furthermore, universities are good examples of urban hotspots that are only fully active on weekdays. Among the top bottlenecks of Melbourne's network, we observed links to and from major universities (Fig. 6b) emerging only on weekdays (see Supplementary Fig. 9B). Given that the proposed method does not incorporate any geospatial information from the network, the surprising alignment between the locations pinned by identified bottlenecks and the urban hotspots, suggests that the method is capturing the actuality.

We also observed that four out of the top ten pain points on Melbourne's road network reported in the media[40] are overlapping with or in very close proximity to our identified top bottlenecks at morning rush hour. Since almost half of the reported ten points do not have bus or tram services in conflict with the road conditions, the results suggest that our methodology does indeed work well.

Bottleneck amelioration. It is interesting to compare the effectiveness of our proposed CS-based bottleneck identification scheme, to other well-established bottleneck identification schemes. In particular, we compare against the bottlenecks

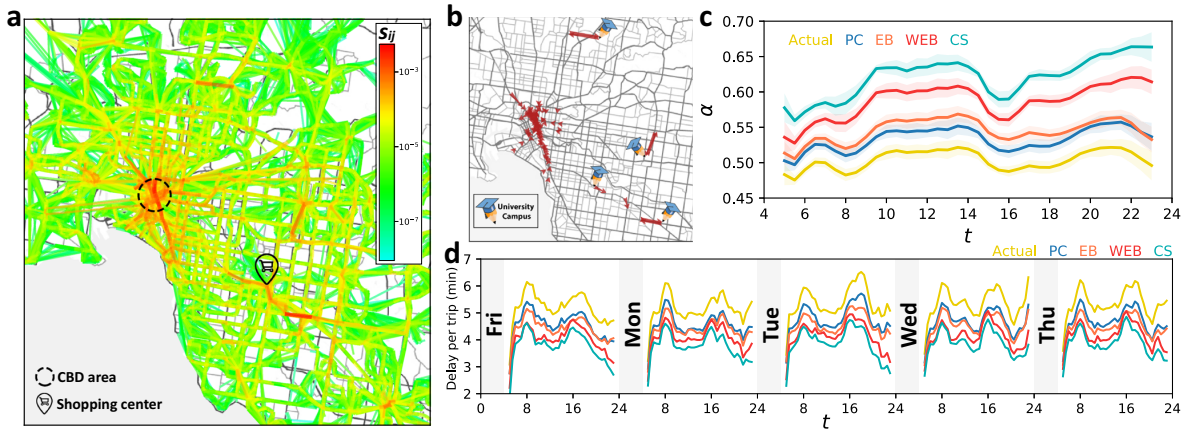


Fig. 6 Bottleneck identification and amelioration on a real-world network. a Spatial distribution of the link criticality scores s_{ij} over Melbourne's public transportation (PT) network during weekdays. The central business district (CBD) and the biggest shopping center in Melbourne are pinned on the map. **b** Top 100 weekday bottlenecks of Melbourne's PT network, identified based on link criticality scores. Major university campuses outside Melbourne's CBD area are pinned on the map. **c**, **d** The impact of ameliorating perturbations on bottlenecks identified by different approaches, i.e., criticality score (CS), edge betweenness centrality (EB), demand-weighted edge betweenness centrality (WEB), and percolation criticality (PC). The number of identified bottlenecks by each approach is equal to 2% of the average number of links that appear on the network. **c** Daily evolution of α calculated for the actual (yellow) and ameliorated networks associated with different bottleneck identification approaches. Results show the average (solid line) and standard deviation (shaded area) over the weekdays of September and October 2017. **d** Delay per trip (in minutes) caused by road congestions, on the actual and improved networks. Streetmap layers in **a** and **b** ©OpenStreetMap contributors[44].

identified based on the widely used edge betweenness (EB) centrality measure, here referred to as EB bottlenecks. We also use an extended version of the EB scheme, which incorporates the demand distribution by weighting the O–D node pairs when calculating the EB centrality of links, here referred to as Weighted EB or simply WEB. Alternatively, bottlenecks can be identified among the links removed at percolation criticality as used in ref. [26], which we refer to as PC bottlenecks. These bottlenecks termed "red bonds" in percolation theory[41], glue the GC together by connecting the communities of higher-quality links. (For a more detailed description of the above approaches, see Supplementary Note 6.)

To compare these approaches, we separately ameliorated the bottlenecks of each type and monitored the response of the network in terms of changes to the demand-serving reliability α. In practice, the most obvious proposal for enhancing the reliability of an on-road PT network is to reduce the conflict of PT vehicles with road conditions at network bottlenecks, which can be achieved, for example, by giving signal priority to PT vehicles or allocating segregated (exclusive) PT lanes. Here, the bottlenecks are taken to be the top 2% most critical links in the network over time, according to each approach. Let B denote the set of bottlenecks identified by one of the schemes. We ameliorated the bottlenecks by synthetically increasing the qualities of bottleneck links $e_{ij} \in B$, to unity ($q_{ij} = 1$). Figure 6c (Supplementary Fig. 9A) compares the impact of ameliorating the bottlenecks identified by the four different approaches, as functions of time during weekdays (weekends) in Melbourne; see Supplementary Fig. 10B for Brisbane's PT network. Amelioration of the CS bottlenecks resulted in more than 23% (26%) improvement in reliability α of Melbourne's PT network, on average during weekdays (weekends). However, on average over both weekdays and weekends, amelioration of PC, EB, and WEB bottlenecks, only increased α by approximately 16%, 8%, and 6%, respectively. See Supplementary Fig. 10B, C for comparison between the effectiveness of different types of bottlenecks for Brisbane's PT network.

The investigation was extended by verifying the impact of bottleneck amelioration on reducing the delay in passenger travel times. In order to calculate the delay caused by congestion, we first generated a congestion-free copy of the network at each time of a day by synthetically changing the actual travel time on each link to the minimum travel time observed on that link during the day. We assumed that each trip took place on the directed path with the minimum sum of the link travel times, between its origin and destination nodes. Then, for any particular network, the total delay was calculated as the absolute difference between the total travel time on the actual and the congestion-free copy of the network. Delay indicates the extent of the impeding effect of link congestions on passenger trips.

Separately for weekdays and weekends, we simulated the amelioration of the top CS, EB, WEB, and PC bottlenecks (the top 2% most critical links based on each scheme) of Melbourne's PT network. The delay per passenger trip of 5.3 min (5.7 min) decreased to 3.8 min (4.2 min) by ameliorating the CS bottlenecks of weekdays (weekends). Figure 6d shows the delay per passenger trip on the actual and ameliorated networks at different times during the first five weekdays of September 2017; Supplementary Fig. 11B extends the results to two months of data. The time saved by amelioration of CS bottlenecks was 25% more than that of WEB bottlenecks while it was twofold compared to those of EB and PC bottlenecks. Ameliorating the top CS bottlenecks saved close to 2,000 hours of passenger travel time during a single morning peak period (7:00–9:00 A.M.), and approximately 11,000 hours of passenger travel time over a normal weekday.

The generality of the proposed framework. In order to emphasize the generality of the proposed framework, we used undirected RGGs as a generic proxy of spatial networks and showed that the framework is able to reflect the true global flow-properties of the network. Here, RGG structures were generated by first distributing $n = 2500$ nodes uniformly at random on the plane $[0, \sqrt{n}]^2$, and then connecting any pair of nodes with

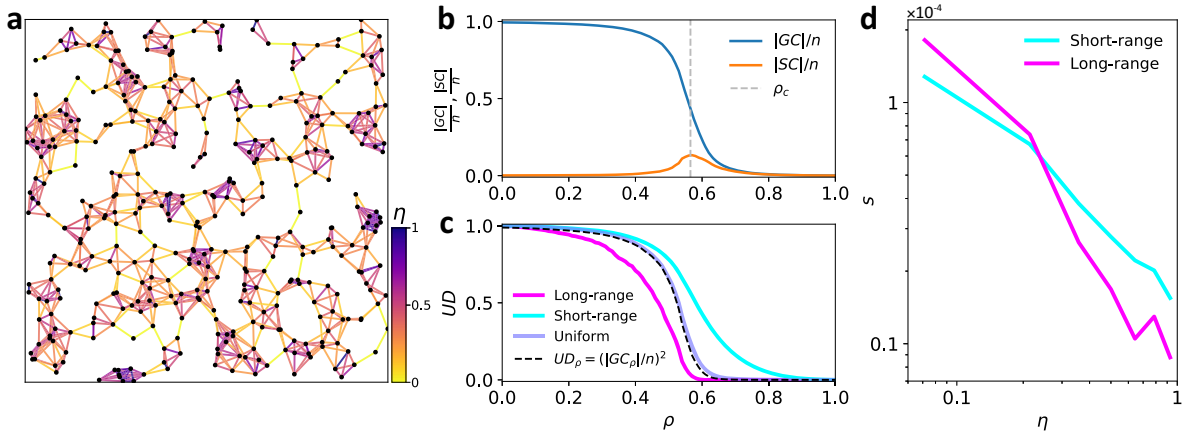


Fig. 7 Capturing true properties of the demand-serving networks. a A sample RGG of size $n = 400$ with color-coded link overlap index η. **b** Normalized |GC| and |SC| during the percolation process averaged over 100 realizations of RGG structure with $n = 2500$ nodes and random link qualities. **c** Unaffected demand (UD) versus ρ for different flow demand scenarios on the RGG structure, averaged over 100 realizations. As predicted, the evolution of the GC size during the percolation process is approximately equal to $n.\sqrt{\mathrm{UD}_\rho}$ (see Methods) when the flow demand is uniform, which is the reason for the similarity between the blue and the dashed black curves. **d** Link criticality score s versus link overlap η, compared for short-range and long-range flow demand scenarios. See Supplementary Note 7 for extension of this analysis to the square grid and random graph structures.

Euclidean distance below $r_0 = 1.6$. We chose r_0 to be greater than the threshold $r_0^c \approx \sqrt{\ln(n)/\pi} \approx 1.58$ for which it is known[42] that the network will be a.a.s. connected. The quality of each link was drawn uniformly at random from (0,1], making percolation a random link removal process depending only on the network topology. RGGs are built of clusters with high intra-connectivity, glued together by bridging links (Fig. 7a). This structure demonstrates a clear phase transition during the percolation process, as removal of a sufficient number of intercluster links causes an abrupt fragmentation of the GC (Fig. 7b).

Over each RGG instance, we distributed a fixed volume of flow demand, according to three different scenarios, namely, uniform, short-range, and long-range. In the uniform demand scenario, the total flow demand volume was divided equally among all reachable (o,d) node pairs; i.e., all entries of F, which correspond to a reachable node pair, are equal to a constant. Let D_{od} be the Euclidean distance between nodes o and d, and $D_{\max}$ the distance between the most distant node pair in the network. Then, to generate the short-range (long-range) flow demand scenarios, we picked a node pair (o,d) uniformly at random and then with probability $0.2e^{-0.2D_{od}}$ ($0.2e^{-0.2(D_{\max}-D_{od})}$) added one unit to the volume of flow demand between that O–D pair f_{od}, and repeated this until the fixed total flow volume was completely allocated to the node pairs over the network.

We simulated the percolation on 100 realizations of RGG structure for each one of the above flow demand distributions. During the percolation, we monitored the GC and SC, which are independent of the demand distribution, and also monitored the UD for different demand distribution scenarios (Fig. 7b, c). Remarkably, in Fig. 7c for the case of uniform flow, the percolation diagram as a function of ρ is the same for UD as it is for the square of |GC| (normalized by the network size). Thus, simulation results confirm the previously discussed theoretical relationship $\mathrm{UD}_\rho \approx (|\mathrm{GC}_\rho|/n)^2$ between evolution of the GC and UD when demand is uniformly distributed over the network. This shows that by assuming a uniform flow demand over the network, our method can provide an analogous analysis to that of monitoring the GC. Furthermore, UD shows logical sensitivity to the nonuniformity of flow demand distributions over the network. Long-range flows are more likely to get caught up in lower-quality links because each time they have to pass between clusters their choices become limited to a few bridging links. This resulted in lower reliability ($\alpha = 0.43$) compared to when the flow-demand is uniformly distributed (α=0.50). In contrast, short-range flows are more likely to stay within the well-connected clusters of RGG, where there are more alternative paths available to bypass low-quality links. Hence, the network is more reliable for a short-range flow demand, which was fairly characterized by a higher $\alpha(= 0.58)$.

Here, we use RGG networks with different flow demand scenarios to verify the success of link criticality score in identifying network bottleneck links. We use the link overlap $\eta \in [0,1]$ to determine whether a link belongs to a community (high overlap) or acts as an intercommunity bridge (low overlap); overlap of a link e_{ij} is defined as $\eta_{ij} = \frac{|\Gamma(i)\cap\Gamma(j)|}{|\Gamma(i)\cup\Gamma(j)|-2}$ where $\Gamma(i)$ is the neighborhood set of node i. In Fig. 7a, links are color-coded according to their overlap index. The criticality score of intra-community (high overlap) links was found to be higher for the short-range flow demand scenario compared to the long-range scenario (Fig. 7d). This is consistent with the fact that short-range flows are more likely to have their origin and destination within a community, which makes the flow-carrying role of intra-community links more critical. Inter-community (low overlap) links have a stronger role in bridging between the remote points of the network, thus, the larger the proportion of the demand flowing between the distant nodes, the more critical these links become for the network. As expected, the criticality score of inter-community links was higher in the long-range flow scenario compared to the short-range flow scenario.

Discussion

Percolation analysis is a powerful tool for understanding the global flow properties of networks. However, most conventional percolation-based analyses become less effective in the presence of a heterogeneous flow demand between different node pairs over the network. We have developed a method that makes use of a newly introduced percolation-driven property, namely, UD, in order to quantify network reliability. Based on the concept of UD, we presented a bottleneck identification scheme, that proved more effective than other state-of-the-art methods reported in the

literature, in terms of both improving the reliability and reducing the delay imposed on flows by congested links. Note that the direct effect of congestion organization on travel time delay cannot be studied using the existing percolation models, because the removal of a congested link simply cannot help quantify the effect of its congestion on flow travel times. But it is an intriguing problem that suggests an important direction for future research.

Our proposed ideas are generally applicable to demand-serving networks including most physical infrastructures where there is an inherent demand for movement of an uneven amount of flow between different pairs of nodes in the network. With the ever-increasing availability of detailed data from real-world critical infrastructure networks, this study can be a helpful starting point for new research avenues and the development of more sophisticated theoretical tools to analyze flow demand, in order to achieve a more profound understanding of these complex systems.

Methods

Smart-card data. The data used in the real-world case study, are the smart-card transaction records, collected by the automated fare collection system for PT in Melbourne and Brisbane, Australia. Passengers are supposed to perform a scan-on transaction at the start and a scan-off at the end of their trip. Every smart-card transaction record contains multiple attributes, namely, anonymized card identifier, PT mode (bus, tram, or train), vehicle identifier (a unique number for each bus or tram vehicle), stop identifier, time-stamp, and transaction type (scan-on/off). For Melbourne's network, we used an average of over 2,120,000 and 912,000 daily transactions associated with all PT modes on weekdays and weekends, respectively, collected during 61 days of September and October 2017. Brisbane data was collected during March 2013. After applying a cleaning process, we used the data to generate the temporal network of on-road PT supply and its corresponding passenger travel flow demand (see Supplementary Note 1 for details).

Network and demand matrix construction. To generate the network representation of the on-road PT system on a particular day at time t, the structure and link attributes were estimated from the smart-card transactions time-stamped within the window $[t-\delta/2, t+\delta/2]$. The time window length δ, was set to 2 hours for experiments presented in the main article. First, we clustered the closely located PT stops and mapped each cluster to a node. Using information of smart-card transactions we derived the trajectory of every vehicle on the network, and if there was at least one vehicle traveling from one of the stops associated with node i to a stop associated with node j without stopping, we added a direct link e_{ij} starting at node i and pointing at node j. For each link e_{ij} the average travel time τ_{ij} over the time window was also calculated based on the information from the tracked vehicles. For a network of time t, demand matrix F measures the flow demand volumes by the number of O–D trips between nodes, within the time window used for the construction of the network. An O–D trip is a chain of one or more trip legs with transfers (but no activities) in between them. See Supplementary Note 1 on how single trip legs are chained to obtain O–D trips.

Unaffected demand. To formulate the UD calculation, we use the so-called reachability matrix $R=[r_{od}]$ (the transitive closure of the network adjacency matrix) which is a square matrix of order n. Each entry r_{od} is equal to 1 if there is at least one directed path from node o to node d on the network, and $r_{od}=0$ otherwise. Let R_ρ be the reachability matrix of network G_ρ. At any threshold ρ, the amount of flow from o to d (f_{od}) is deemed to be "unaffected" by link qualities q below the threshold ρ ($q \le \rho$) if there is at least one directed path from o to d remaining on G_ρ, i.e., $r^\rho_{od}=1$. So, UD_ρ (defined as the unaffected proportion of the demand at threshold ρ) will be the sum of $r^\rho_{od}.f_{od}$ for all (o,d) pairs of nodes, normalized by the total flow demand

$$\mathrm{UD}_\rho = \frac{\mathbf{1}_n^{\mathrm{T}}(R_\rho \circ F)\mathbf{1}_n}{\mathbf{1}_n^{\mathrm{T}}F\mathbf{1}_n} = \frac{\mathrm{tr}(R_\rho F^{\mathrm{T}})}{\mathbf{1}_n^{\mathrm{T}}F\mathbf{1}_n}, \tag{6}$$

where $\circ$ is the entry-wise product of matrices, tr(.) is the trace of the $n\times n$ square matrix, and $\mathbf{1}_n$ is a column vector of all n elements equal to one.

The relation between the evolution of UD and GC during the percolation. Let $|\mathrm{GC}_\rho|$ be the size of the GC as a function of ρ, then $|\mathrm{GC}_\rho|/n$ is called the incipient order parameter which is sometimes used to describe the connectivity of a fragmented network. If we assume a uniform flow demand distribution then on any undirected network, UD_ρ equals the proportion of connected node pairs in G_ρ, which approaches $(|\mathrm{GC}_\rho|/n)^2$ as $n\to\infty$[43]. So, for large enough networks, monitoring the GC during the percolation is a special case of monitoring UD when flow demand is uniform. Therefore, we can accurately predict the evolution of |GC| during the percolation by assuming a uniform flow demand over the network and using $|\mathrm{GC}_\rho| \approx n.\sqrt{\mathrm{UD}_\rho}$. This is confirmed numerically in Fig. 7 and Supplementary Fig. 12.

Considering the above relation, when the demand is homogeneous (or unknown but assumed to be homogeneous), instead of the definition in Eq. (2) one may choose to use the area under the curve of $\mathrm{UD}_\rho^{1/2}$ as a reliability indicator that reflects the rate at which size of the connected components decline over the percolation process. However, our original definition in Eq. (2) has a simpler interpretation and it is mathematically tractable, allowing for theoretical analysis of network links in the simplest possible way.

Link criticality score and its relation to network reliability. Suppose there exists a non-empty set of different directed paths Ψ_{od} that route between an origin node o and a reachable destination node d. During the percolation process on the network (whereby ρ is increased from zero to unity), each pathway $\psi\in\Psi_{od}$ breaks up when the threshold ρ reaches to the minimum link-quality on that path. The "limiting link" associated with the flow from o to d (e^*_{od}), when removed during the percolation process at $\rho=q^*_{od}$, breaks the last path(s) connecting o to d and affects the flow between them (f_{od}). Using the definition of link criticality score in Eq. (4), we can expand the left-hand-side of Eq. (5) as

$$\sum_{e_{ij}\in E} s_{ij}.q_{ij} = \sum_{e_{ij}\in E}\sum_{\substack{o,d\in V,\\ e^*_{od}=e_{ij}}} \frac{f_{od}}{\mathbf{1}_n^{\mathrm{T}}F\mathbf{1}_n}.q_{ij}, \tag{7}$$

and for any pair $o,d\in V$ with non-zero f_{od} there exist a single limiting link $e^*_{od}\in E$ with quality q^*_{od}, so

$$= \frac{1}{\mathbf{1}_n^{\mathrm{T}}F\mathbf{1}_n}\sum_{o,d\in V} f_{od}.q^*_{od}. \tag{8}$$

During the percolation process, each entry in the reachability matrix R_ρ switches from 1 to 0 as soon as the last path(s) between its corresponding O-D nodes break. So, we can write

$$r^\rho_{od} = \begin{cases} 1, & \rho < q^*_{od} \\ 0, & \rho \ge q^*_{od} \end{cases}, \tag{9}$$

where r^ρ_{od} is the (o,d) entry of the reachability matrix R_ρ associated with the network G_ρ. Note that the integral of r^ρ_{od} with respect to ρ between the limits $\rho=0$ and $\rho=1$ is equal to q^*_{od}. So, from Eqs. (8) and (9) we can write

$$\sum_{e_{ij}\in E} s_{ij}.q_{ij} = \frac{1}{\mathbf{1}_n^{\mathrm{T}}F\mathbf{1}_n}\sum_{o,d\in V} f_{od}.\int_0^1 r^\rho_{od}d\rho, \tag{10}$$

where the right-hand-side can be simplified with matrix operations to obtain Eq. (2) which is the definition of the reliability index α, so we can conclude that Eq. (5) holds. In Supplementary Note 3, the definition of the criticality score and the proof of Eq. (5) are generalized further, requiring no assumption on the link quality values.

Data availability

Two weeks of Melbourne's public transportation network data used in this study, are available at https://gitlab.com/homayoun/demand-serving-networks. Raw passenger smart-card data from Melbourne's public transportation network were made available for research purposes by the associated transportation authority, which retains ownership over the data.

Code availability

Source codes for the algorithms proposed in this study are available at https://gitlab.com/homayoun/demand-serving-networks. Specific codes that produce the results presented in this paper are available upon request.

Received: 25 July 2020; Accepted: 13 January 2021;
Published online: 23 February 2021

References

1. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025 (2010).
2. Jalili, M. & Perc, M. Information cascades in complex networks. *J. Complex Netw.* **5**, 665 (2017).
3. Duan, D. et al. Universal behavior of cascading failures in interdependent networks. *Proc. Natl Acad. Sci. USA* **116**, 22452 (2019).
4. De Domenico, M. & Baronchelli, A. The fragility of decentralised trustless socio-technical systems. *EPJ Data Sci.* **8**, 2 (2019).
5. Askarisichani, O. et al. Structural balance emerges and explains performance in risky decision-making. *Nat. Commun.* **10**, 1 (2019).
6. Barja, A. et al. Assessing the risk of default propagation in interconnected sectoral financial networks. *EPJ Data Sci.* **8**, 32 (2019).

7. Akbarzadeh, M. & Estrada, E. Communicability geometry captures traffic flows in cities. *Nat. Hum. Behav.* **2**, 645 (2018).
8. Stone, L. The google matrix controls the stability of structured ecological and biological networks. *Nat. Commun.* **7**, 12857 (2016).
9. Hill, S. M. et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **13**, 310 (2016).
10. Stone, L. The feasibility and stability of large complex biological networks: a random matrix approach. *Sci. Rep.* **8**, 1 (2018).
11. Santolini, M. & Barabási, A.-L. Predicting perturbation patterns from the topology of biological networks. *Proc. Natl Acad. Sci. USA* **115**, E6375 (2018).
12. Stone, L., Simberloff, D. & Artzy-Randrup, Y. Network motifs and their origins. *PLoS Comput. Biol.* **15**, e1006749 (2019).
13. Stauffer, D. & Aharony, A. *Introduction to Percolation Theory* (Taylor & Francis, 2018).
14. Ganin, A. A. et al. Resilience and efficiency in transportation networks. *Sci. Adv.* **3**, e1701079 (2017).
15. Latora, V. & Marchiori, M. Vulnerability and protection of infrastructure networks. *Phys. Rev. E* **71**, 015103 (2005).
16. De Domenico, M., Solé-Ribalta, A., Gómez, S. & Arenas, A. Navigability of interconnected networks under random failures. *Proc. Natl Acad. Sci. USA* **111**, 8351 (2014).
17. Mirzasoleiman, B., Babaei, M., Jalili, M. & Safari, M. Cascaded failures in weighted networks. *Phys. Rev. E* **84**, 046114 (2011).
18. Jalili, M. Error and attack tolerance of small-worldness in complex networks. *J. Informetr.* **5**, 422 (2011).
19. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378 (2000).
20. Halu, A., Scala, A., Khiyami, A. & González, M. C. Data-driven modeling of solar-powered urban microgrids. *Sci. Adv.* **2**, e1500700 (2016).
21. Wang, F., Li, D., Xu, X., Wu, R. & Havlin, S. Percolation properties in a traffic model. *Europhys. Lett.* **112**, 38001 (2015).
22. Saberi, M. et al. A simple contagion process describes spreading of traffic jams in urban networks. *Nat. Commun.* **11**, 1 (2020).
23. Zeng, G. et al. Switch between critical percolation modes in city traffic dynamics. *Proc. Natl Acad. Sci. USA* **116**, 23 (2019).
24. Echenique, P., Gómez-Gardenes, J. & Moreno, Y. Dynamics of jamming transitions in complex networks. *Europhys. Lett.* **71**, 325 (2005).
25. Gallos, L. K., Makse, H. A. & Sigman, M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *Proc. Natl Acad. Sci. USA* **109**, 2825 (2012).
26. Li, D. et al. Percolation transition in dynamical traffic network with evolving critical bottlenecks. *Proc. Natl Acad. Sci. USA* **112**, 669 (2015).
27. Li, D., Zhang, Q., Zio, E., Havlin, S. & Kang, R. Network reliability analysis based on percolation theory, reliability engineering. *Syst. Saf.* **142**, 556 (2015).
28. Cohen, R. & Havlin, S. *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, 2010).
29. Hamedmoghadam, H., Ramezani, M. & Saberi, M. Revealing latent characteristics of mobility networks with coarse-graining. *Sci. Rep.* **9**, 1 (2019).
30. Smith, A. M. et al. Competitive percolation strategies for network recovery. *Sci. Rep.* **9**, 1 (2019).
31. Oppenheim, N. *Urban Travel Demand Modeling: From Individual Choices to General Equilibrium* (John Wiley and Sons, 1995).
32. Zhang, L., Zeng, G., Guo, S., Li, D. & Gao, Z. Comparison of traffic reliability index with real traffic data. *EPJ Data Sci.* **6**, 19 (2017).
33. Yang, Y., Nishikawa, T. & Motter, A. E. Small vulnerable sets determine large network cascades in power grids. *Science* **358**, eaan3184 (2017).
34. Schneider, C. M., Moreira, A. A., Andrade, J. S., Havlin, S. & Herrmann, H. J. Mitigation of malicious attacks on networks. *Proc. Natl Acad. Sci. USA* **108**, 3838 (2011).
35. Pollack, M. Letter to the editor—the maximum capacity through a network. *Oper. Res.* **8**, 733 (1960).
36. Estrada, E. & Sheerin, M. Random rectangular graphs. *Phys. Rev. E* **91**, 042805 (2015).
37. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821 (2002).
38. Callaway, D. S., Newman, M. E., Strogatz, S. H. & Watts, D. J. Network robustness and fragility: percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468 (2000).
39. Olmos, L. E., Çolak, S., Shafiei, S., Saberi, M. & González, M. C. Macroscopic dynamics and the collapse of urban traffic. *Proc. Natl Acad. Sci. USA* **115**, 12654 (2018).
40. 2018 Redspot Survey. https://www.redspotsurvey.com.au (2018).
41. Ben-Avraham, D. & Havlin, S. *Diffusion and Reactions in Fractals and Disordered Systems* (Cambridge University Press, 2000).
42. Penrose, M. *Random Geometric Graphs*, Vol. 5 (Oxford University Press, 2003).
43. Chen, Y. et al. Percolation theory applied to measures of fragmentation in social networks. *Phys. Rev. E* **75**, 046107 (2007).
44. Openstreetmap Copyright and License. https://www.openstreetmap.org/copyright (2021).

Acknowledgements

We thank Prof. Shlomo Havlin for his extremely helpful discussions of the ideas presented in this paper and for his suggestions. H.H. thanks Dr. Meead Saberi for the fruitful discussions, Dr. Nora Estgfaeller for providing valuable feedback throughout this study, and Public Transport Victoria for their support and providing the access to Melbourne's Public Transportation (PT) smart-card data. We are especially grateful to Prof. Mark Hickman and Dr. Zhenliang Ma for their help and running our codes on the Brisbane's PT smart-card data. M.J. and L.S. are supported by the Australian Research Council (ARC) through project No. DP170102303. M.J. is also supported by the ARC through project No. DP200101119. H.V. is supported by the ARC through project Nos. DP180102551 and DP190102134. L.S. is supported by the ARC through project No. DP150102472.

Author contributions

H.H. conceived the original idea and conducted the experiments and analyses. H.H., M.J., and H.V. designed the study. H.H. and L.S. developed the methodology and wrote the paper. All authors interpreted the results, reviewed the paper, and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21483-y.

Correspondence and requests for materials should be addressed to H.H. or L.S.

Peer review information *Nature Communications* thanks Maksim Kitsak, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

4.2 Supplementary information for the publication

Supplementary Information for

Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks

Homayoun Hamedmoghadam*, Mahdi Jalili, Hai L. Vu and Lewi Stone*

Homayoun Hamedmoghadam. E-mail: homayoun.hamed@monash.edu
Lewi Stone. E-mail:lewistone100@gmail.com

The SI includes:

Supplementary Notes 1 to 7
Supplementary Figures 1 to 12
Supplementary References

Supplementary Note 1: Smart-Card Data Processing

Tracking vehicles. We tracked bus and tram vehicles based on temporally sorted sequence of stop ID-timestamp pairs associated with smart-card transactions made on each vehicle. Each vehicle trajectory corresponds to a sequence of visited stops with arrival, departure, and dwelling time at each stop, which are respectively the timestamp for first transaction, last transaction, and the time span between the first and last transactions within the uninterrupted sequence of transactions made on the vehicle at a particular stop. If there was only one transaction recorded for a vehicle's visit to a stop, the dwelling time was considered to be from 10 seconds before to 10 seconds after the timestamp of that transaction.

Spatial clustering of stops. To construct the network representation of the Public Transportation (PT) system, we first performed a simple spatial clustering on the stops, and then mapped each stop cluster to a node on the network. This allowed us to deal with the existence of closely located stops with no direct transit service connecting them. This can be the case at intersections, PT hubs where different routes meet, and different sides of a street that often serve the same route but in different directions. The goal is to make stop clusters consisting of the stops so close to each other that no route would be designed to serve more than one of them and no passenger would use any transportation mode other than walking, to move from one to the other. To do so, we considered a convenient walking distance threshold of 200 m and in a recursive process we merged close stops together so to include every pair of stops with a distance of 200 m or less into the same cluster. As there is a chance that a chain of spatially close stops leads to a large cluster, we limited the clusters to have a maximum of 100 m radius; i.e. the maximal distance of the cluster members from the cluster centroid. Then, for clusters of radii larger than 100 m, we repeatedly removed the point with the largest distance from the centroid until the radius satisfies the condition. Then, we applied the clustering process to the points removed from the clusters in the previous step, and repeated the whole procedure until all points are assigned to clusters.

Network construction. To generate the network representation of the PT system, we first mapped each stop cluster to a node. Then, if there was at least one vehicle traveling from node i to node j, we added a directed link e_{ij} from node i and pointing to node j. Each link represents a service in the PT system between two immediate nodes. For any particular time t during the day, the average travel time attribute τ_{ij} of the link e_{ij} can be calculated based on the data from the vehicles travelling on the link during the interval $[t - \delta/2, t + \delta/2]$ where δ is length of the time window, discussed further in the followings. Travel time attribute of links was derived from vehicle trajectories as the average elapsed time for vehicles since departure from the source node until arrival time at the target node.

Trip chaining. An Origin-Destination (O-D) trip is a journey between two anchor points consisting of a single trip leg, or multiple trip legs entailing transfers in-between. The process of connecting trip legs at transfer points to estimate O-D trips (also called *journeys*) is called *trip chaining*. The O-D flow demand in transportation networks, often called O-D travel demand or O-D passenger flow demand, is explained by the number of O-D trips between each O-D pair of nodes in the network over a certain time window. Trip chaining for PT passenger trips is usually performed based on some assumptions about the behavior of PT users. The only assumption we make here is that PT users do not use any mode of transportation other than walking between two consecutive trip legs, i.e. at the interchange point. Thus, transfers should include nothing more than a walk from the alighting stop to the next boarding stop and waiting for the connection. The problem is to find an optimal *allowable transfer time* to determine whether a transfer or an activity is carried out by the passenger between consecutive trip legs (from alighting time to the next boarding time).

We developed an unsupervised learning approach to classify the times spent by passengers at interchange points (inter-transaction times) into transfers and activities. Let us define a return trip as a sequence of trip legs which starts with boarding at a reference point and ends with alighting at a close proximity of the reference point, while the maximum inter-transaction time corresponds to the interchange point with the largest geodesic distance from the reference point. The idea is that each return trip includes undertaking an activity at that particular target point. We first detected the return trips separately on the first five weekdays and the first five weekend days in September 2017. Then, we built the histogram of transfer and activity duration from the inter-transaction times extracted from the return trips, separately for weekdays (Supplementary Figure 1*A*) and weekends (Supplementary Figure 1*C*). The histogram shows a large number of transfers associated with small inter-transaction times, while the number of observed transfers decreases drastically with increasing inter-transaction

duration. The activity duration histogram for weekdays depicts three local maxima, approximately at 1.5, 7, and 8.5 hours; see Supplementary Figure 1*A*. The first peak is associated with common daily non-occupational activities. The two other peaks roughly at 7 and 8.5 hours long, are in surprising agreement with regular school day and daily work duration that are 6.5 hours (1) and 7.6 hours (2), respectively. The peak associated with school disappears in weekend inter-transaction time distribution while a small local maximum remains at the duration associated with work.

Next, we find the allowable transfer time θ, to label inter-transaction times equal or less than the threshold as transfers, and the rest as activities. Labeling the inter-transaction durations can be viewed as a dichotomous (binary) classification problem where the positive samples, i.e. samples that should have been labeled as positive, are the transfers and activities are the negative samples. In order to choose the optimum threshold θ to perform the classification, we used the Youden's J statistic (also called *Informedness*) (3), which measures the accuracy of a dichotomous (binary) classifier and is defined as below:

$$Informedness = Recall + Inverse\ Recall - 1, \tag{1}$$

where $Informedness \in [0, 1]$ is the probability of an informed decision as opposed to a random guess taking into account all predictions made by the classifier. Assuming that the actual labels for transfers are positive and activities are negative, *Recall* and *Inverse Recall* can be calculated as:

$$Recall = P(W \leq \theta), \tag{2}$$

$$Inverse\ Recall = P(V > \theta), \tag{3}$$

where W and V are the random variable corresponding to derived transfer and activity durations, respectively.

The optimal θ maximizing the *Informedness* was found to be 43 minutes for weekdays (Supplementary Figure 1*B*) and surprisingly the exact same threshold was found for weekends (Supplementary Figure 1*D*). Therefore, to build O-D trips for each passenger, first the trip legs associated with a particular card ID were sorted chronologically. Then, each pair of consecutive trips belong to a single O-D trip if the inter-transaction time between the subsequent legs did not exceed the determined inter-transaction time threshold $\theta = 43$ min, implying a transfer at the interchange point. There was an average of more than 33,000 transfers per working day and about 10,000 daily transfers during weekends. Finally, an O-D trip was described with two location-timestamp pairs associated with the boarding of the first leg (origin) and alighting of the last leg (destination) for a chain of trip legs.

Estimating missing alighting transactions. The information of a passenger trip is complete only when there is full information from a boarding scan-on record and an alighting scan-off record. Unpaired transactions are a common problem for AFC data, due to the nature of AFC systems as they depend on human actions which involve errors, and also their various components, e.g. reading and recording, which can malfunction. However, missing transactions can be estimated with high accuracy using the information of immediate previous and following transactions (4, 5).

We deployed a procedure to estimate the missing alighting transactions, where for a particular card ID, there is a scan-on transaction on a bus or tram without a valid paired scan-off later on the same vehicle (missing alighting), but the next transaction recorded for that card is a scan-on (second boarding) made that day or the next day on any PT mode (see Supplementary Figure 1*E*). By following the boarded vehicle's trajectory starting from the first boarding timestamp until the second boarding timestamp, we generated a set of candidate alighting points by filtering the stops visited by that vehicle. Candidate alighting stops are within 2 km radius of the second boarding point, and it is possible to walk from them to the second boarding point with the speed of 4.5 km/h within the inter-transaction duration. An allowable transfer time is already calculated, which classifies the purpose of the alighting-then-boarding events into transfers and activities. If alighting at the candidate stop with the shortest Euclidean distance to the second boarding stop, allows undertaking an activity, it was identified as the missing alighting point. Otherwise, the candidate stop which allows the earliest arrival of the passenger to the second boarding stop was identified as the missing alighting point.

Choice of time window length and O-D flow demand generation. Here, we discuss the effect of time-window length on the generated network and the O-D travel flow demand matrix F. For each time window within the day, the travel time of a link was calculated by aggregating the travel time information of multiple vehicles traversing the link. The aggregation time window should be long enough, so the link qualities fairly represent the impact of phenomena, such as signals, which might affect each vehicle differently. Furthermore, if the window is too small, the network links associated with low frequency services will be intermittent on the temporal network structure. However, during a long interval, conditions such as a transient traffic congestion might change and the link qualities will not reflect the temporary conditions accurately if the time window is too wide. To count the number of trips between O-D pair of nodes on the network, the time window should be large enough to encompass the trips on the network.

A 2-hour time window was found to be large enough to observe at least one vehicle on the links with low service frequency, while encompassing almost all O-D trips on the network (see Supplementary Figure 1*F*). Additionally, it is not too large to conceal or smooth out the transient road conditions in resulting link qualities. As such, we chose the aggregation time window length of 2 hours for the experiments presented in *Main Text* and also here, unless otherwise stated. It is worth mentioning that a 2-hour time window as maximum duration of most trips, is also recognized by Melbourne's PT authority (Public Transport Victoria), and passengers do not have to pay additional fares for 2 hours after each payment made on a scan-on transaction.

After generating the network representation of the on-road PT system at time t on a particular day based on PT services running within the time interval $[t - \delta/2, t + \delta/2]$, we followed the next steps, namely, determining the maximum allowable transfer time, estimating the missing alighting information, and applying trip chaining process, to obtain O-D trips during the same time interval. The $n \times n$ matrix F for network of size n, was generated, where each entry (o, d) of the matrix, denoted as f_{od}, is the number of O-D trips (chained trips or passenger journeys) form node o to node d within the target time window.

Supplementary Note 2: Involving flow-capacity of links in network reliability analysis

We proposed a new approach to monitor the percolation process, which provides a quantitative insight on network reliability, in terms of the ability to provide paths of high-quality links for its flows with respect to the demand distribution. Our network analysis in the *Main Text* is not concerned with capacity of the network. The reason is that our analysis pinpoints the links with problematic congestion levels, and any capacity-related problem can be seen independent of that and may be solved completely in parallel. To make this clear with an example related to PT networks, let us imagine that the capacity of some PT vehicles is increased. This increases the maximum flow-capacity of the network, yet it does not result in any change to the demand and link qualities over the network, thus, the result of our reliability analysis will remain unchanged. So, one can use our analysis to study the network congestion in relation to demand distribution and pinpoint the problems with the network, but then the any problem related to capacity of links can be attended completely in parallel or independently. Nevertheless, here we demonstrate that the proposed framework can be extended to involve the flow-capacity of network links in reliability analysis. In particular, definition of the Unaffected Demand (UD) can be extended to study the ability of networks to provide high-quality paths and "accommodate" the demand on such paths.

The amount of demand on the network under percolation. Recall that at any threshold ρ during the percolation process, our proposed $UD(\rho)$ is the proportion of the network's flow-demand between the Origin-Destination (O-D) node pairs that remain connected only by links of quality above the threshold ρ ($q_{ij} > \rho$). At each threshold ρ during the percolation process, subnetwork G_ρ is generated by inheriting all the links with quality q above the threshold ($q > \rho$) from the network G. We defined $R_\rho = [r^\rho_{od}]$ as the reachability matrix of subnetwork G_ρ (see *Methods* in the *Main Text*). Using the reachability matrix, for the network G with total demand of $\sum_{o,d\in E} f_{od}$ the amount of remaining demand on its subnetwork G_ρ is $\sum_{o,d\in E} r^\rho_{od}.f_{od}$, where f_{od} is the volume of demand from node o to node d; note that $UD(\rho) = (\sum_{o,d\in E} r^\rho_{od}.f_{od})/\sum_{o,d\in E} f_{od}$.

Calculating the flow-capacity of the network under percolation. From the data, we derived the number of buses/trams running on each link at each snapshot of the network in time. We assumed the capacity of each on-road PT vehicle to be 50 passengers which is a conservative choice (a normal sized bus can practically take 70-80 passengers). Then, at each threshold ρ in the percolation process we approximated the capacity of the subnetwork G_ρ as explained in the following. Take a very small λ (close to zero) so that the network G_ρ has the capacity for concurrent movement of $\lambda.r^\rho_{od}.f_{od}$ passengers between all (o,d) node-pairs (r^ρ_{od} is zero if o and d are disconnected). The variable λ can be increased until it reaches a maximum before it becomes impossible for network links to match the amount of flows. We denote the maximum possible value of λ for the network G_ρ by $\lambda_{max}(\rho)$. The problem of finding λ_{max} is known as "maximum concurrent multicommodity flow" problem, which is strongly NP-complete but can be approximated in polynomial time by a number of algorithms. Here, we used a modified version of Fleischer's algorithm (6) which is fast and accurate enough for our purpose, and provides a lower approximation of the maximum flow-capacity of the network. For any network with given link capacities and O-D flow demand, Fleischer's algorithm chooses a priority path between each O-D pair, assigns a small proportion of the demand between the O-D pair to all the links on that path, and updates the capacity of those links. The algorithm iteratively augments flows to the network links (which corresponds to increasing λ), and terminates when augmentation becomes impossible and returns the final λ as λ_{max}. At any threshold ρ if the algorithm returns, say, $\lambda_{max}(\rho) = 0.1$, it means that G_ρ can accommodate the concurrent flow of 10% of the demand between each O-D pair, but if $\lambda_{max}(\rho) \geq 1$ then G_ρ is capable of accommodating the whole demand between its connected O-D pairs. We define $C(\rho) = \lambda_{max}(\rho).UD(\rho)$ to simply compare the capacity $C(\rho)$ and the demand $UD(\rho)$ on G_ρ, both normalized by the total amount of demand on network G_0.

We checked the change in capacity of the real on-road PT networks during the percolation process and observed that as congested links are progressively being removed, the capacity of the network never falls below the amount of demand corresponding to UD. This is not surprising, as although some routes can become very crowded at peak hours, on average the utilization of on-road PT vehicles are generally low even in large cities (vehicles are not operating close to their capacity). As an example, in Supplementary Figure 2 the black curve shows the evolution of $UD(\rho)$ during the percolation process on a snapshot of the Melbourne's network at rush-hour (the same curve as in Fig. 2d of the *Main Text*), and the capacity $C(\rho)$ depicted via the red curve demonstrating that the network can handle 1.5 to 3 times of the $UD(\rho)$ at any point in the percolation process. The network during non-rush hours has even a higher capacity relative to its demand, as the drop in the number of PT services is less than the decline of total passenger flow-demand from rush to non-rush hours.

Involving the capacity into percolation analysis. If a demand-serving network functions close to its flow-capacity and one wants to study the problems with capacity of the network in addition to the conflict between flows and congestion, then, our definition of UD can be extended to $UD_c(\rho) = \min\{UD(\rho), C(\rho)\}$. The new capacity-aware unaffected demand (UD_c), monitors the proportion of the total demand that can be "accommodated" between O-D pairs only on links with quality above the threshold ρ; during the percolation always $UD_c(\rho) \leq UD(\rho)$, and $UD_c(\rho) < UD(\rho)$ if O-D paths on G_ρ cannot match the remaining demand over the subnetwork. Accordingly, the reliability measure α can be extended to capacity-aware reliability α_c defined as the area under the curve of $UD_c(\rho)$ over $\rho \in [0,1]$, i.e. $\alpha_c = \int_0^1 UD_c(\rho)\, d\rho$.

Supplementary Note 3: Link Quality, Link Criticality Score, and Network Reliability

In this note we first provide the general definition of the link criticality score, given that it is possible for links to have equal quality attributes. Then, we prove that the relationship between link quality q, link criticality score s, and the proposed percolation-based demand-serving reliability of networks α, given by Eq. 5 in the *Main Text*, still holds. We end this note by giving analytical proof to the effectiveness of using link criticality score for the problem of network bottleneck identification. In the following we try to provide sufficient examples to assure the comprehensibility and reproducibility of the proposed method.

Link criticality score. The definition of link criticality score in Eq. 4 of the *Main Text* assumes unique link quality values over the network. To be generally applicable though, it needs to be modified to take into account the possibilities that i) for a single O-D pair there are multiple connecting pathways each include a different link with quality equal to q^*_{od}, and ii) there are multiple minimum-quality links on a single connecting pathway.

Let ψ be a sequence of links corresponding to a directed path on the network. For any path ψ, we define the *path quality* (q_ψ) as the minimum quality among all links on that path; i.e. $q_\psi = \min_{e_{ij}\in\psi} q_{ij}$. We deem a network as reliable, where despite presence of local (link-level) perturbations, reflected as lowered link qualities, the network provides alternative paths with high-quality links for the flows between O-D nodes. Therefore, for an origin node o and a reachable destination node d on the network ($o, d \in V$), from a non-empty set of all directed paths connecting them, denoted by Ψ_{od}, we take the maximum quality path(s) as the primary (optimal) path(s) for flows between (o, d) pair. The optimal path maximizes the minimum link-quality on the pathways from o to d. Let us define the set of all optimal paths between a pair of O-D nodes as:

$$\sigma_{od} = \underset{\psi\in\Psi_{od}}{\operatorname{argmax}} \; q_\psi. \tag{4}$$

The minimum link-quality on all paths (path quality) in σ_{od} are equal to q^*_{od}. So, paths in σ_{od} break at threshold $\rho = q^*_{od}$ and make node d unreachable from node o; i.e. the flow volume f_{od} becomes affected. As an example, in Fig. 3a-c of the *Main Text*, $\sigma_{1,4}$ includes $1 \to 2 \to 3 \to 4$ and $1 \to 5 \to 2 \to 3 \to 4$ with $q^*_{1,4} = q_{2,3} = 0.6$.

Considering the quality of links as their weights, finding the set $\sigma_{od} \subseteq \Psi_{od}$ on the weighted network is known as the maximum capacity paths problem (7). The optimal path(s) between each node pair can be found via a modified version of a conventional shortest path algorithm; e.g., Dijkstra's algorithm (8) should be modified to compare paths based on their minimum link-quality and then prefer the path with the largest value of the minimum link-weight. With optimal paths between all O-D pairs of nodes found, the generalized definition of criticality score s_{ij} for the link e_{ij}, free from any assumption, can be formulated as:

$$s_{ij} = \sum_{o,d\in V} \sum_{\psi\in\sigma_{od}} \frac{f_{od}.\lambda(e_{ij},\psi)}{(\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n).|\sigma_{od}|.|\epsilon_\psi|}, \tag{5}$$

where $\lambda(e_{ij}, \psi)$ is equal to 1 if e_{ij} is (one of) the minimum-quality link(s) on the path ψ, and 0 otherwise. And $\epsilon_\psi = \{e_{kl} \in \psi | q_{kl} = q_\psi\}$ is the set of all links e_{kl} on the path ψ, that have equally the minimum link-quality on the path ψ. In practice, all (o, d) pair of nodes can be visited to first find q^*_{od} according to all paths in Ψ_{od}, and then the set of optimal directed paths $\sigma_{od} \subseteq \Psi_{od}$ connecting each pair, and finally the set of minimum-quality link(s) ϵ_ψ on each path $\psi \in \sigma_{od}$, to allow calculation of all criticality scores s_{ij} using Supplementary Equation 5. If a link e_{ij} is not found as the minimum-quality link on the optimal path between any (o, d) pair with $f_{od} > 0$, or in other words if it never found as the limiting link associated with a (o, d) pair with $f_{od} > 0$, then s_{ij} will be 0.

The proportion of the total demand flowing from node o to node d is $f_{od}/(\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n)$, which indicates the importance of the connectivity between them. If there is a single optimal path connecting o to d with a single minimum-quality link on the path, the proportion of the total flow demand between the pair (o, d) is fully added to the criticality score of that link. For the example of Fig. 3 in the *Main Text*, $s_{2,3}$ is the proportion of the total demand that is between six node pairs, as shown in Supplementary Figure 3*B*. In case there are multiple optimal paths from node o to d, i.e. $|\sigma_{od}| > 1$, the flow demand proportion $f_{od}/(\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n)$ is divided equally between those paths and share of each path is added to the criticality score of the limiting link on that path. For the network seen in Fig. 3 (*Main Text*), $|\sigma_{1,4}| = 2$ but the single limiting link $e_{2,3}$ for the node pair $(o, d) = (1, 4)$ is on both optimal paths, thus ultimately $f_{od}/(\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n)$ is fully contributed to $s_{2,3}$. If there are multiple minimum-quality links on an optimal path $\psi \in \sigma_{od}$, i.e. $|\epsilon_\psi| > 1$, the share of each optimal path, i.e. $f_{od}/((\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n).|\sigma_{od}|)$, is divided equally between those equally-minimum-quality links.

Deriving the key identity for α. Here, we show that the relationship between link qualities, link criticality scores, and the demand-serving reliability of a network, expressed by Eq. 5 of the *Main Text*, also holds for the general definition of the criticality score in Supplementary Equation 5. The significance of this identity is that it allows us to derive the impact of increasing the quality of links (improving links) on network reliability. Using Supplementary Equation 5 we can write:

$$\sum_{e_{ij}\in E} s_{ij}.q_{ij} = \frac{1}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n} \sum_{o,d\in V} \frac{f_{od}}{|\sigma_{od}|} \sum_{\psi\in\sigma_{od}} \sum_{e_{ij}\in E} \frac{q_{ij}.\lambda(e_{ij},\psi)}{|\epsilon_\psi|} \tag{6}$$

and as for each path ψ, when iterating over all links in the network, $\lambda(e_{ij}, \psi)$ becomes 1 for only links from the set of minimum-quality links on ψ,

$$= \frac{1}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n} \sum_{o,d\in V} \frac{f_{od}}{|\sigma_{od}|} \sum_{\psi\in\sigma_{od}} \sum_{e_{ij}\in\epsilon_\psi} \frac{q_{ij}}{|\epsilon_\psi|} \tag{7}$$

$$= \frac{1}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n} \sum_{o,d\in V} \frac{f_{od}}{|\sigma_{od}|} \sum_{\psi\in\sigma_{od}} \sum_{e_{ij}\in\epsilon_\psi} \frac{\min\limits_{e_{kl}\in\psi} q_{kl}}{|\epsilon_\psi|} \tag{8}$$

and as $|\epsilon_\psi|$ is the number of links which are equally the minimum-quality links on the path ψ,

$$= \frac{1}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n} \sum_{o,d\in V} \frac{f_{od}}{|\sigma_{od}|} \sum_{\psi\in\sigma_{od}} \min_{e_{kl}\in\psi} q_{kl} \tag{9}$$

and for any optimal directed path ψ connecting o to d ($\psi \in \sigma_{od}$), the minimum link-quality is equal to q^*_{od}, so

$$= \frac{1}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n} \sum_{o,d \in V} f_{od}.q^*_{od} \tag{10}$$

Supplementary Equation 10 is identical to Eq. 8 of the *Main Text* which is then manipulated in a couple of steps to conclude (see *Methods* section of the *Main Text*):

$$\sum_{e_{ij} \in E} s_{ij}.q_{ij} = \alpha \tag{11}$$

Let us validate the above equation in an example. For the toy network of Fig. 1 (*Main Text*), paths from node 1 to node 2 are $1 \to 2$ and $1 \to 4 \to 2$. Links with the minimum quality on these paths are $e_{1,2}$ and $e_{1,4}$, among which $e_{1,4}$ has the maximum quality; thus $1 \to 4 \to 2$ is the optimal path. Having $f_{1,2} = 5$ and the total flow demand of 100, contributes 0.05 to $s_{1,4}$. The two paths $2 \to 3 \to 5$ and $2 \to 1 \to 3 \to 5$ connect nodes 2 and 5, and their minimum link-qualities are respectively $q_{2,1} = q_{3,5} = 0.6$ and $q_{2,3} = 0.3$. Therefore, the former is the optimal path, but as it has two equally minimum-quality links, the flow demand proportion $f_{2,5}/100 = 0.03$ is divided by two and then 0.015 is added to both $s_{2,1}$ and $s_{3,5}$. Link $e_{1,2}$ is not the minimum-quality link of any optimal path thus $s_{1,2} = 0$. Link $e_{1,3}$ is the minimum-quality link on the optimal path connecting node 1 to node 3, and also together with $e_{5,1}$, they are the minimum-quality links on the optimal path connecting node 5 to 3, thus $s_{1,3} = (f_{1,3} + f_{5,3}/2)/100 = 0.105$. Other non-zero link criticality scores on the network are $s_{1,4} = 0.25$, $s_{2,1} = 0.14$, $s_{2,4} = 0.11$, $s_{3,5} = 0.21$, $s_{4,2} = 0.1$, and $s_{5,1} = 0.085$. Therefore, the sum of $s_{ij}.q_{ij}$ over all links (left-hand side of Supplementary Equation 11) equals 0.65, which is equal to the area under curve of UD_ρ which equals α (right-hand side of Supplementary Equation 11) as expressed by Eq. 2 of the *Main Text*; see also Fig. 1b of the *Main Text* for the area under curve of UD_ρ for this example network. Supplementary Figure 3*B* visualizes the relationship between link qualities q, link criticality scores s, and the demand-serving reliability α using this example network.

Implication of link criticality score in reliability of the network. Consider increasing the quality q_{ij} of a chosen link e_{ij} to q'_{ij} ($q'_{ij} > q_{ij}$). Here, we give the full details to prove that there exists a non-empty range of values for q'_{ij}, for which criticality score of the whole network will remain unchanged. Thus, according to Supplementary Equation 11, increasing the quality of the link e_{ij} from q_{ij} to q'_{ij} ($\Delta q_{ij} = q'_{ij} - q_{ij}$) within the abovementioned range, increases the reliability of the network, directly proportional to the criticality score of the link s_{ij}, by exactly $s_{ij}.(q'_{ij} - q_{ij}) = s.\Delta q_{ij}$.

Let us assume again that for two different links $e_{ij} \neq e_{kl}$ on the network, the probability of having equal qualities is zero, i.e. $P(q_{ij} = q_{kl}) = 0$. This in theory is correct assuming that link qualities come from a continuous distribution, and is approximately correct in practice when link qualities are calculated as decimals with high precision. Based on this assumption, link criticality scores can be calculated as Eq. 4 (in the *Main Text*), which simply expresses that s_{ij} is the proportion of total flow demand between the O-D pairs for which e_{ij} is the limiting link. Let us denote the set of all ordered pairs of network nodes (o, d), where $o, d \in V$ and $o \neq d$, with non-zero flow demand ($f_{od} > 0$), as V_p. Our approach is to investigate the impact of increasing the quality q_{ij} of link e_{ij}, on three mutually disjoint subsets of V_p, denoted as V^p_c, $c = 1, 2, 3$, with the property:

$$V^p = \bigcup_{c=1,2,3} V^p_c(e_{ij}), \quad \forall e_{ij} \in E. \tag{12}$$

From all (o, d) pairs in V^p:

- $V^p_1(e_{ij})$ includes those for which e_{ij} is the minimum-quality link on the optimal path(s) $\psi \in \sigma_{od}$ connecting o to d ($q_{ij} = q_\psi = q^*_{od}$),
- $V^p_2(e_{ij})$ include those for which e_{ij} does not appear on optimal paths, but it is part of at least one connecting path within which q_{ij} is the minimum link-quality,
- $V^p_3(e_{ij})$ includes those for which e_{ij} does not appear on any path connecting o to d, or if it does, q_{ij} is larger than the path quality (minimum link-quality on the path).

For any link e_{kl} same as or different from e_{ij}, criticality score s_{kl} can be calculated as:

$$s_{kl} = \sum_{c=1,2,3} \sum_{\{(o,d) \in V^p_c(e_{ij}) | e^*_{od} = e_{kl}\}} \frac{f_{od}}{\mathbf{1}_n^{\mathrm{T}} F \mathbf{1}_n}. \tag{13}$$

Note that criticality score of any link can be written as the summation of three components associated with $c = 1, 2, 3$. For a network with initial demand-serving reliability of α, after increasing the quality of e_{ij} from q_{ij} to q'_{ij}, the new reliability will be α' ($\alpha' \geq \alpha$). Next, we show that under certain conditions for q'_{ij} all link criticality scores will remain unchanged, which allows the exact calculation of α'.

Increasing q_{ij} does not change the limiting link of any (o, d) pairs in $V^p_3(e_{ij})$, hence for all links the component associated with $c = 3$ in Supplementary Equation 13 remains unchanged. Let $\hat{V}^p_2(e_{ij})$ be a subset of (o, d) pairs in $V^p_2(e_{ij})$, for which there is at least one path connecting o to d including e_{ij} where quality of all other links is above q^*_{od}, formally defined as:

$$\hat{V}^p_2(e_{ij}) = \{(o, d) \in V^p_2(e_{ij}) | \exists \psi \in \Psi_{od} : (e_{ij} \in \psi) \wedge (q_{kl} > q^*_{od})\}. \tag{14}$$

Then, as long as $\forall(o,d) \in \hat{V}_2^p : q'_{ij} < q^*_{od}$, e_{ij} (with its new quality q'_{ij}) cannot take over the role of limiting link between any O-D pair and the component associated with $c = 2$ of Supplementary Equation 13 for criticality score of all links remains unchanged. Any link e_{ij} with $s_{ij} = 0$, is not the limiting link between any node pair with non-zero flow demand, i.e. $V_1^p(e_{ij}) = \emptyset$, thus we can conclude that:

$$s_{ij} = 0 \wedge q'_{ij} < \min(\{q^*_{od} | (o,d) \in \hat{V}_2^p(e_{ij})\} \cup \{1\}) \Rightarrow \alpha' = \alpha. \tag{15}$$

Let q^{**}_{od} be the second lowest link-quality on all optimal paths connecting o to d. For a link e_{ij} with $s_{ij} > 0$, we have $V_1^p(e_{ij}) \neq \emptyset$, and the component associated with $c = 1$ in Supplementary Equation 13 will remain fixed as long as e_{ij} maintains its role as the limiting links between all pairs in $V_1^p(e_{ij})$, i.e. if $\forall(o,d) \in V_1^p(e_{ij}) : q'_{ij} < q^{**}_{od}$. Therefore, with the help of proved identity in Supplementary Equation 11 we can write:

$$s_{ij} > 0 \wedge q'_{ij} < \min(\{q^*_{od} | (o,d) \in \hat{V}_2^p(e_{ij})\} \cup \{q^{**}_{od} | (o,d) \in V_3^p(e_{ij})\}) \Rightarrow \alpha' - \alpha = s_{ij}.(q'_{ij} - q_{ij}). \tag{16}$$

To summarize all the above, for any link with criticality score of zero there exists a non-empty range of values for increased link criticality score q'_{ij} that network reliability will certainly remain unchanged $\alpha' = \alpha$. However, if the link criticality score is larger than zero, there exists a non-empty range of q'_{ij} values, for which no link criticality score changes in the network. Thus, according to the key relationship identified between q, s, α (Supplementary Equation 11 here or Eq. 5 in the *Main Text*), increasing q_{ij} to q'_{ij} within a certain non-empty range, changes the reliability of the network to $\alpha' - \alpha = s_{ij}.(q'_{ij} - q_{ij})$, that is directly proportional to the criticality score of the ameliorated link. Thus, one can generally expect further improvement on the network reliability from amelioration of the links with higher criticality scores. Accordingly, network bottlenecks can be identified as links with the highest criticality scores. For the network in Fig. 1 of the *Main Text*, the link with the highest criticality score is $e_{1,4}$ which if completely ameliorated ($q_{1,4} = 0.4$ and $q'_{1,4} = 1$), will improve the reliability of the network from $\alpha = 0.65$ to $\alpha' = 0.768$.

Supplementary Note 4: Implication of Heterogeneous O-D Flow Demand

A drawback of the percolation threshold ρ_c as a reliability index can be noticed through consideration of heterogeneous distribution of flow demand, where the volume of in-demand flow is not equal between all pairs of nodes. In presence of a heterogeneous flow demand, the pathways connecting O-D pairs with high demand, are of more importance than those with low or no flow demand. After percolation criticality (subcritical phase), when the Giant Component (GC) is fragmented into small and medium-sized clusters, heterogeneous flow demand might allow a significant portion of flows to still be preserved within isolated clusters. Or on the contrary, even before criticality (supercritical phase) a significant portion of the in-demand flows might be unable to reach their destination due to, for example, insignificant fragmentations of the GC leading to isolation of O-D nodes with very large demand volume.

Here, we investigate the implication of the heterogeneous demand, by demonstrating that during the percolation process on the Melbourne's on-road PT network, actual in-demand O-D trips do not break with the same rate as O-D pairs of nodes become disconnected. In the proposed framework, Unaffected Demand (UD) is defined as the proportion of total flow demand that can reach the destination node from the origin at any point during the percolation process; see *Main Text* for details. On any network, if the O-D flow demand is uniformly distributed over all reachable O-D pairs of nodes, then at any threshold ρ, UD will be equal to the proportion of reachable node pairs on the network. Having a uniform demand means that the total flow demand is divided up equally between all reachable O-D pairs of nodes on the network. Let us denote the UD of the special case of having a uniform flow demand, as UD', which as a function of ρ can be calculated as below:

$$UD'_\rho = \frac{\mathbf{1}_n^{\mathrm{T}} R_\rho \mathbf{1}_n}{\mathbf{1}_n^{\mathrm{T}} R_0 \mathbf{1}_n} \tag{17}$$

Supplementary Equation 17 can be derived from Eq. 6 of the *Main Text*, when the flow demand matrix F has a constant value in all its entries associated with reachable O-D nodes on the network.

Generally, during the percolation, UD' decreases as O-D pairs of nodes, separated by lower quality links, become progressively disconnected. Given that , if $\gamma < 1$, then UD decreases faster than UD' during the percolation, meaning that the flow demand is relatively more between O-D node pairs separated by lower quality links. And if $\gamma > 1$, then there is relatively more demand to flow between O-D pairs connected by paths of higher quality links. Supplementary Figure 5 compares the evolution of UD_ρ (the proportion of not-yet-broken O-D trips) with UD'_ρ (the proportion of not-yet-separated node pairs) during the percolation on Melbourne's PT network, separately for weekdays (Supplementary Figure 5*A*) and weekends (Supplementary Figure 5*B*). The figure demonstrates that $\gamma > 1$ during both weekdays and weekends, meaning that there was a relatively high passenger travel flow demand between the O-D pairs connected through pathways made up of higher quality links. Furthermore, the observed phenomenon was magnified during weekdays ($\gamma = 4.64$) compared to weekends ($\gamma = 3.66$). This finding is consistent with higher demand-serving reliability α of weekdays compared to weekends, as it shows that the loss of connectivity during the percolation affected the demand with a slower rate in networks of weekdays compared to those of weekends.

Supplementary Note 5: Percolation on Melbourne's PT Network

We calculated the critical threshold ρ_c and demand-serving reliability α of the on-road PT network of Melbourne, during the daily active period of the system, i.e. 4:00-24:00, in steps of 30 min length. In Supplementary Figure 6, ρ_c (Supplementary Figure 6*A*&*B*) and α (Supplementary Figure 6*C*&*D*) are depicted as a function of the time t of the day. Supplementary Figure 6*A*&*C* show the evolution of the indices over each single weekday (curve color shows the date) during the two months of September and Octorber 2017. Supplementary Figure 6*B*&*D* plot ρ_c and α versus time t, averaged at each time-point over 43 (18) working days (days off) during the course of the available data. Also, the envelope of a single standard deviation of ρ_c

and α values (calculated at each time of the day over two months) is shown by the shaded areas, below and above the mean. The standard deviation of α values is particularly small, making it possible to easily discern a general trend in the average α, as it changes over the day (the signal to noise ratio is high). Because of the small standard deviation, the same trend will observable in the evolution of α over any single day. This is corroborated by the plots of the raw timeseries over each single day in Supplementary Figure 6*C*. In contrast, in Supplementary Figure 6*B*, we see that the standard deviation of ρ_c is larger than fluctuations in the average trend for ρ_c (averaged at each time t over the two months), plotted over a day. Thus, the trends in the average signal of ρ_c, are dominated by the fluctuations of the noise (the signal to noise ratio is very low). The temporal evolution of ρ_c over a single day will not resemble the hourly averaged data seen in Supplementary Figure 6*B*. This is corroborated by the plots of the raw timeseries during each single day in Supplementary Figure 6*A*, where the effect is accentuated even more strongly. Thus, ρ_c exhibits large fluctuations in comparison to the smooth temporal evolution of α. In the real world under normal conditions, it is unlikely that the global dynamics of the transportation network would repeatedly undergo drastic changes between closely-taken snapshots from hour to hour. Therefore, we conclude that α provides a better and more informative picture of the network's evolution over time.

Also, note the approximately 10% drops in α at 8:00 and between 16:00-18:00 on weekdays (Supplementary Figure 6*D*) is consistent with circadian rhythm of urban human mobility, as these times are associated with morning and evening peak commuting periods, when high rates of congestion and large numbers of commuters predictably increase the conflict between PT system and road conditions.

We also tested the sensitivity of the two reliability indices (ρ_c and α), to the choice of parameter δ (time window length). The proposed α shows the same temporal trend in reliability of the network with different choices of δ for constructing the network structure and its corresponding flow demand matrix F (Supplementary Figure 6*C*&*D*). Overall, the results of our simulations suggest that the choices of δ from 1 to 3 hours do not have a significant impact on the temporal evolution of α and its consistency in a day-to-day comparison, or in other words, the daily trend is robust against alteration of δ within a wide range of reasonable values. Also, the relationship between α in weekday and weekend mode is robust against different choices of δ. Note that ρ_c is noticeably more sensitive to δ, and both its temporal trends during a day and the relation between its values in weekday and weekend modes change with different choices of δ.

There are also other facts about the actual Melbourne's on-road PT network which validate the results achieved by α and suggest that it works well. The relatively low reliability of the network during early morning and late evening is mainly due to fewer PT services running on the network which decreases the number of links and weakens the connectivity (Supplementary Figure 7). The reliability α is relatively stable over a day, despite multiple periods of intense traffic. This is partially explained by the fact that the average degree $\langle k \rangle$ is higher during the rush hours (Supplementary Figure 7*B*) indicating more available services or higher frequency of the existing services during those hours. Generally, larger number of links implies better connectivity which generally should increase the reliability α. If there were no additional PT services during rush hours (manifested by peaks in $\langle k \rangle$), we would see a larger drop in reliability α during rush hours. Thus, planning of the system contributes to stability of the network reliability α during the day by increasing the PT services around rush hours.

Supplementary Note 6: Evaluating the Bottlenecks of Melbourne's PT Network

Our proposed bottleneck identification approach leads to bottleneck links that are consistent with the formation of the hotspots in Melbourne urban area. Generally, there is a high passenger travel demand to and from important activity centers in urban areas, while due to traffic congestion and crowding, link-level quality of service can decrease to a great degree in the surroundings of such areas. Accounting for passenger flow demand, link dynamics, and structure of the network, the calculated link criticality scores were observed to be higher in proximity of known activity centers in Melbourne which is in agreement with the above facts (see Supplementary Figure 9*A*). Network bottlenecks were concentrated mainly around the Central Business District (CBD) and other commercial hubs in suburban areas. Bottlenecks associated with major university campuses outside CBD were among the top bottlenecks only on weekdays, which is interesting because universities are obvious hotspot points while they are only active during weekdays; compare left and right panels of Supplementary Figure 9*B*.

In order to assess the effectiveness of our identified bottlenecks in improving the network's reliability, we applied three well-established bottleneck identification methods on Melbourne's on-road PT network. We then compared the response of the network reliability to improving each of the four different types of bottlenecks. The bottlenecks identified based on our proposed approach are referred to as Criticality Score-based (CS) bottlenecks. The other three bottleneck identification approaches used in comparisons are explained below:

- Bottleneck identification based on Edge Betweenness centrality (EB bottlenecks): Edge betweenness centrality (9) of a link is defined as the number of shortest paths between all pairs of nodes in a network that traverse the given link. Betweenness centrality indicates the influence of the link on flow circulation when the optimal path for flow between a pair of nodes is assumed to be the shortest path on the network. Here, we used the normalized edge betweenness centrality calculated based hop count between the node pairs. The centrality of a link according to this measure, is the number of shortest paths between different O-D node pairs that the link is a part of; centralities are normalized by the total number of O-D pairs on the network. Separately for weekdays and weekends, links with the largest mean edge betweenness centrality scores (averaged over the two months of September and October 2017) were identified as EB bottlenecks of the network.

- Bottleneck identification based on Weighted Edge Betweenness centrality (WEB bottlenecks): Here, we also extend the standard edge betweenness centrality measure, to incorporate the flow-demand over the network when choosing the bottlenecks. In order to do so, the importance of the shortest path between each O-D node pair is weighted by the volume of the flow demand between that pair. The resulting scores of the links are then normalized by the total flow-demand

over the network. The centrality of a link according to this measure is proportional to the volume of flow (i.e., number of passenger trips in our case study of a PT network) passing through the link as a part of the shortest path between their associated O-D pair of nodes. Thus, the links on the shortest path connecting a pair of nodes with higher flow demand volume become relatively more central. Separately for weekdays and weekends, links with the largest mean weighted edge betweenness centrality scores (averaged over the two months of September and October 2017) were identified as WEB bottlenecks of the network.

- Bottleneck identification based on Percolation Criticality (PC bottlenecks): Classical percolation-based reliability analysis views the links bridging between the clusters of higher quality links, as network bottlenecks (10, 11). In this approach, the key to bottleneck identification is in the study of the network at percolation criticality $\rho = \rho_c$, when the GC fragments into smaller sized components formed of links with quality higher than ρ_c. From the set of links removed at criticality (all of which have a quality equal to ρ_c), only a subset is actually responsible for the fragmentation of the GC (11). Let us refer to the set of all links removed at ρ_c as the candidate bottleneck set. The actual bottleneck links can be identified through an exhaustive search among all possible combinations of different number of bottlenecks in the candidate set, to discover the minimal subset that actually glue the GC together. However, this brute-force approach is impractical for a large candidate set. For Melbourne's on-road PT network, the size of the candidate set was over 100 during most of the times on weekdays. In order to identify the real bottlenecks, we counted the occurrences of each link in all candidate sets (associated with networks of different times) and the most frequently observed links were identified as the network bottlenecks.

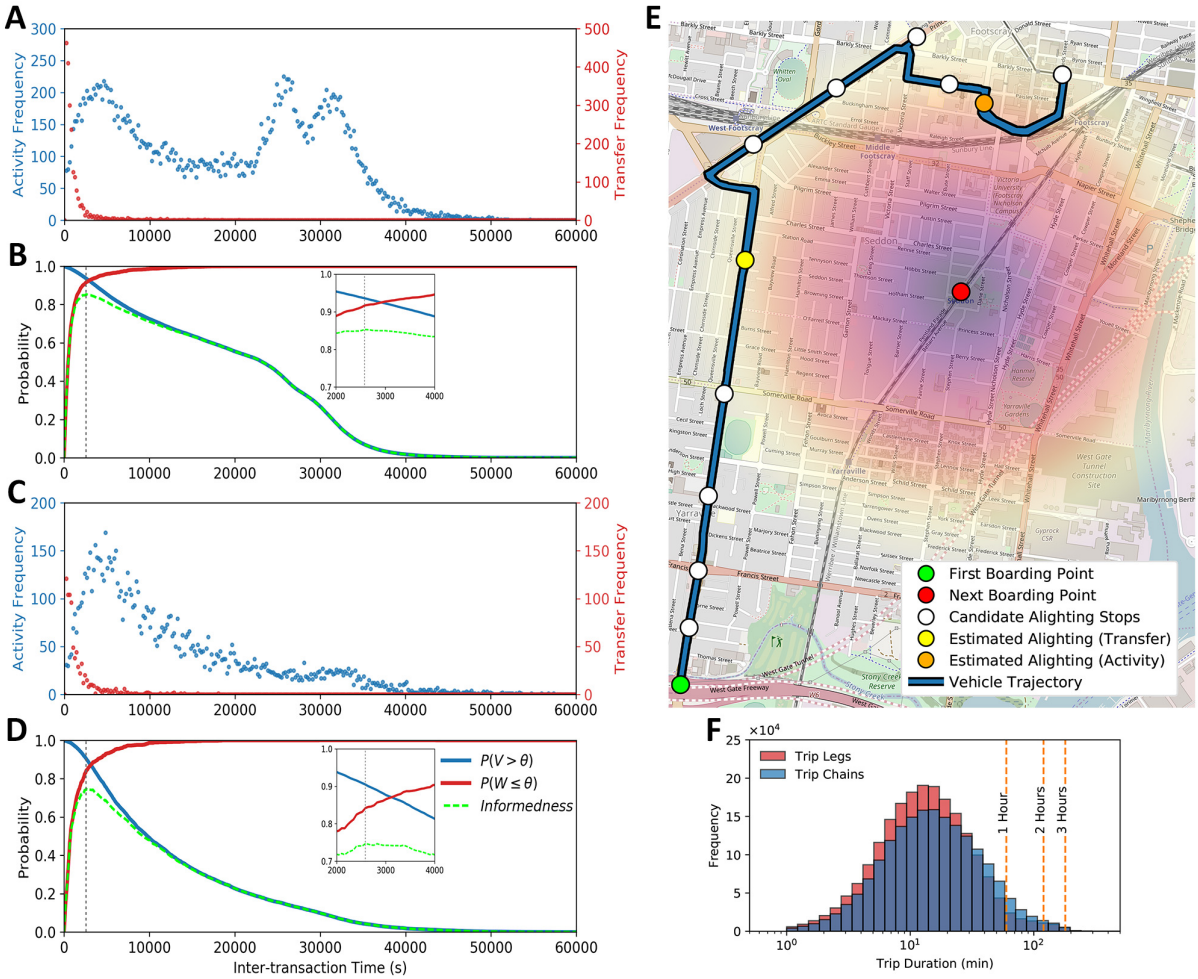
Supplementary Note 7: Appendix to Generality of the Proposed Framework

Random Geometric Graph (RGG) structure (investigated in the *Main Text*) was the perfect choice for testing the generality of our framework as it shares common properties with many spatial infrastructure networks. Links on these networks either belong to relatively small well-connected local communities or they bridge between these communities. This allowed us to characterize our proposed bottlenecks which can shift toward either of the two above roles, depending on the flow demand distribution over the network. Here, we extend the analysis of different flow-demand scenarios to square grid and random network (ER) structures (Supplementary Figure 12). We analyzed the square grid and ER structures with three different flow-demand scenarios, namely, uniform, long-range, and short-range flow-demand. Unaffected Demand (UD) showed logical sensitivity to non-uniformity of flow-demand distributions on network.

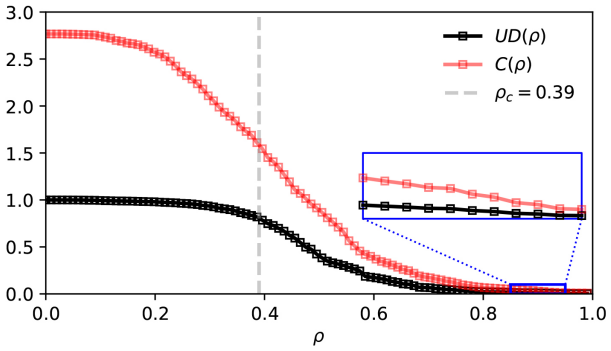
Grid structure showed similar results to that of RGG when demand distribution changes between long-range, short-range, and uniform scenarios. These two structures, are both locally well-connected but do not have long-range links, thus, availability of more alternative paths between closely situated nodes makes them more reliable for serving short-range flows. In a grid network structure, long-range (short-range) flows need to traverse more (less) links to reach destination thus they are more (less) likely to be affected by low-quality of links when link-qualities are distributed randomly over the network. This was reflected by UD with a faster (slower) decrease as a function of increasing ρ when flows tend to be long-range (short-range), resulting in a lower (higher) demand-serving reliability α compared to the uniform flow demand scenario; note the area under the curves marked by up (down) -pointing triangles in Supplementary Figure 12*C*, .

In ER networks, however, connectivity between nodes is independent from their Euclidean distance, thus, connectivity properties are similar between all pairs of nodes on the network. This, expectedly, resulted in an almost identical average UD_ρ over the number of realizations, for all three demand scenarios. Thus, the network is equally reliable when serving any of the three demand distributions.

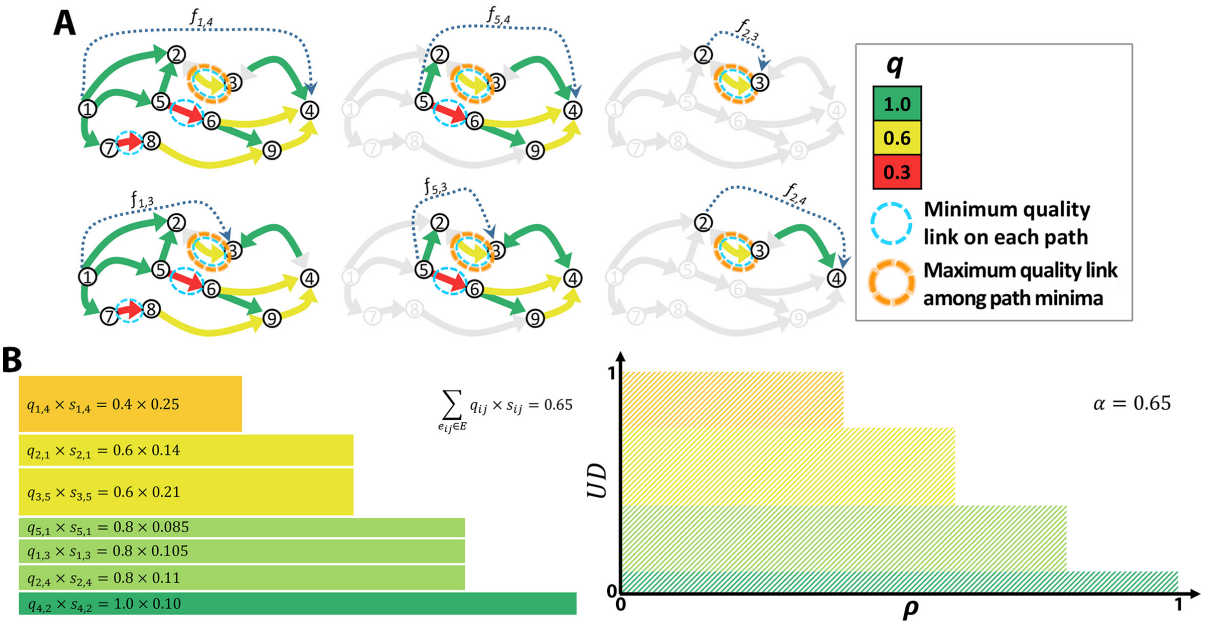
Last but not least, for uniformly distributed flow demand over the network, the theoretical relationship $UD_\rho \approx (|GC_\rho|/n)^2$ (where n is the network size) between evolution of the Giant Component (GC) and UD as functions of the threshold ρ, is confirmed on ER and grid networks in addition to RGG which was studied in the *Main Text*; compare the blue curve and the dashed black curve in Supplementary Figure 12*C*&*F*.



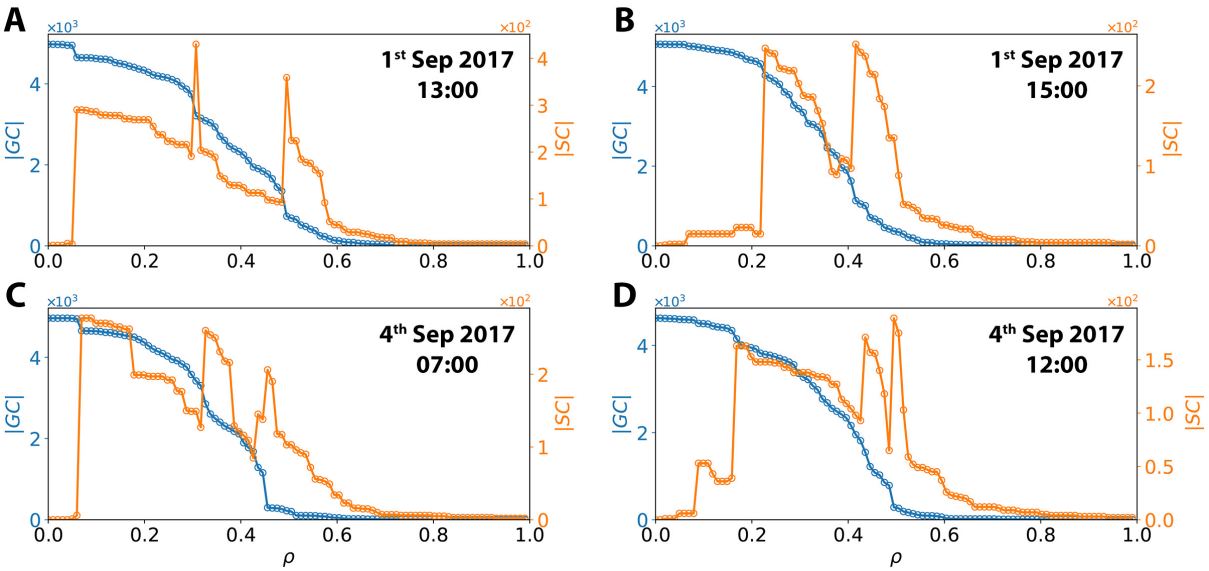
Supplementary Figure 1. Smart-card data processing. (*A*&*C*) Distribution of activity duration (blue) and transfer duration (red) within first five working days (*A*) and first five weekend days (*B*), in September 2017. (*B*&*D*) The cumulative distribution function (CDF) of transfer duration (red curve) and complementary cumulative distribution function (CCDF) of activity duration (blue curve) on workdays (*B*) and weekends (*D*). *Informedness* (green dashed curve) of the CDF and CCDF manifests the accuracy of discerning activities from transfers for different allowable transfer times. (*E*) An example of estimating a missing alighting transaction, where a boarding transaction (green) on a bus misses a valid scan-off pair. However, a scan-on transaction (red) is recorded for the same smart-card identifier later at a train station, with a location and timestamp that allows a non-empty set of plausible alighting stops (white). Among the candidate bus visits, the visited stop with the smallest Euclidean distance to the next boarding point, and the visited stop leading to the earliest arrival of the passenger to the next boarding location, are colored orange and yellow, respectively. Radial color gradient is depicted to aid comparing the distances from the final boarding stop to different candidate stops. Street map layer © OpenStreetMap contributors (12). (*F*) Histogram of single trip-leg (red) and O-D trip (blue) duration during September 2017.



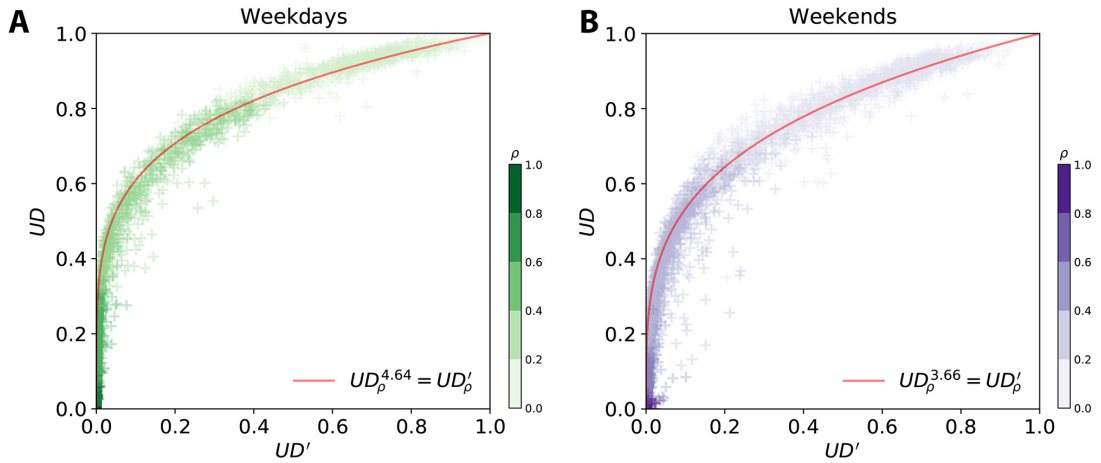
Supplementary Figure 2. Monitoring both the unaffected demand and capacity of the network under percolation. The plot depicts the percolation process on one snapshot (at 8:00 AM on 1 September 2017) of the Melbourne's public transportation (PT) network monitored by $UD(\rho)$ (same as in Fig. 2d of the *Main Text*); the percolation critical point is at $\rho_c = 0.39$. The red curve shows the capacity $C(\rho)$ of the subnetwork G_ρ at different thresholds ρ during the percolation process. For both $UD(\rho)$ and $C(\rho)$, units are relative to the total flow-demand on the network, that is UD at threshold $\rho = 0$ ($UD(0)$) when no link is removed. The inset shows that the capacity remains above the demand at the end of the percolation process.



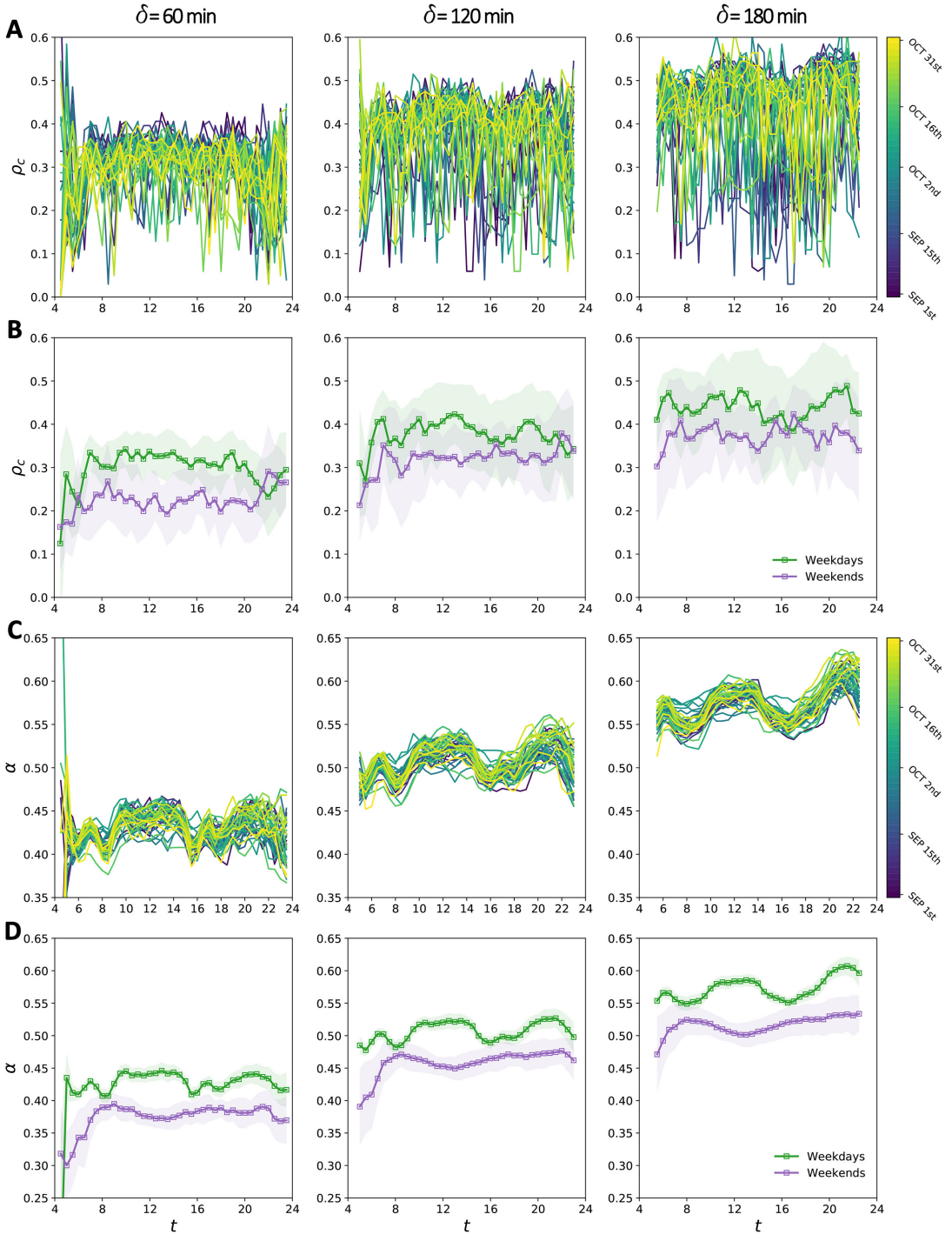
Supplementary Figure 3. Visualization of link criticality score calculation and its relationship with link quality and network reliability. (*A*) In the example provided in Fig. 3 (of the *Main Text*), the "limiting link" associated with the node pair $(1, 4)$ is found to be $e_{2,3}$; i.e. $e^*_{1,4} = e_{2,3}$. However, to calculate the criticality score of a link we need to know all pairs of origin-destination nodes between which the link acts as the limiting link. In the toy network of Fig. 3a (*Main Text*), $e_{2,3}$ is the limiting link for six different (o, d) node pairs, i.e., $e^*_{1,4} = e^*_{5,4} = e^*_{2,3} = e^*_{1,3} = e^*_{5,3} = e^*_{2,4} = e_{2,3}$. In *A*, each of those (o, d) pairs is indicated with a dashed arrow from node o to d on a separate copy of the network, where the links which are not part of the paths connecting o to d are colored light gray. By definition of link criticality score, $s_{2,3}$ is the sum of flow demand between these node pairs, divided by the total flow demand on the network. (*B*) Illustration of relationship between the reliability α, link qualities q_{ij}, and link criticality scores s_{ij} for the network in Fig. 1a of the *Main Text*. On the left, link quality multiplied by link criticality score is visualized for network links with non-zero criticality scores. Links are sorted in ascending order of their quality from top to bottom. On the right, the area under curve of Unaffected Demand (UD) as a function of the threshold ρ (which is seen in Fig. 1b of the *Main Text*) is partitioned into multiple rectangles (each corresponding to a network link) with different colors to show the relationship between the proposed reliability α, link qualities q_{ij}, and link criticality scores s_{ij}.



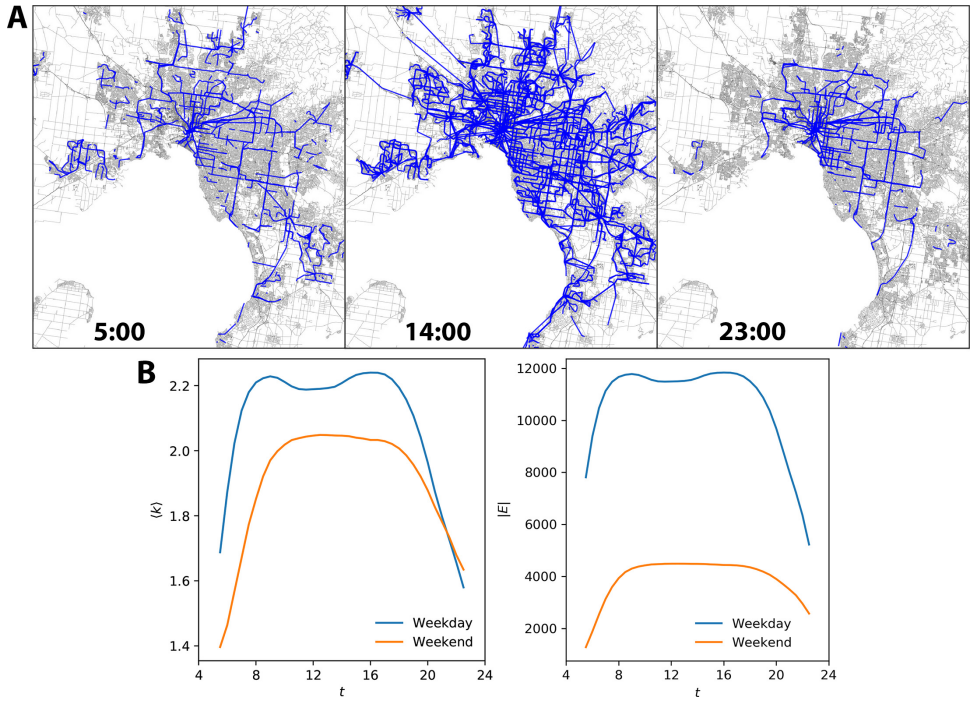
Supplementary Figure 4. Absence of a single abrupt phase transition. Four examples are illustrated from Melbourne's Public Transportation (PT) network from the first two weekdays of September 2017, where it is difficult to find the critical threshold ρ_c. Evolution of $|GC|$ (blue) and $|SC|$ (orange) as functions of ρ are depicted for Melbourne's on-road PT network at (*A*) 13:00 on 1st, (*B*) 15:00 on 1st, (*C*) 7:00 on 4th, and (*D*) 12:00 on 4th of September 2017. In *A* and *C* the three peaks in $|SC|$ demonstrate the detachment of a component of substantial size from the GC at least at three different thresholds. However, none of the three fragmentations is highly distinctive from the rest, and the most severe reduction in $|GC|$ does not happen at the point of maximum $|SC|$. In *B*, the two peaks in $|SC|$ have only 5 nodes difference, yet the percolation criticality is marked by the second peak, which occurs at a threshold ρ approximately 0.2 higher than the first one. During the percolation shown in *D*, $|SC|$ always remains insignificant (< 200) relative to $|GC|$ (> 4500), which suggests that practically the percolation process only gradually erodes the GC and the fragmentation is blurred out over a range of ρ values. These examples demonstrate that at times there is no clear fragmentation of the GC at a single threshold on the Melbourne PT network, due to its finite size and non-random character.



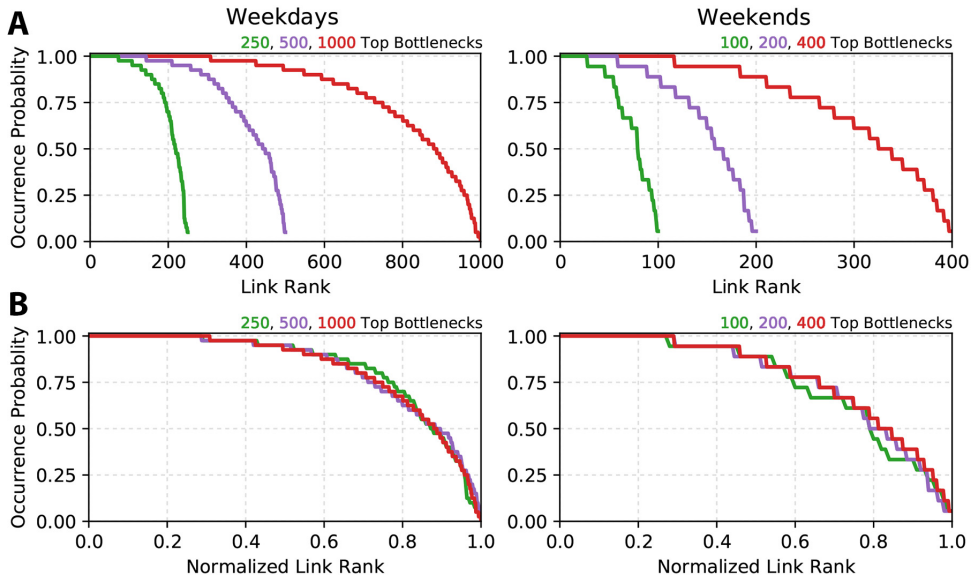
Supplementary Figure 5. Capturing the heterogeneity of flow demand via Unaffected Demand (UD). UD on Melbourne's PT network during the percolation process, in the presence of actual travel flow demand versus a synthetic uniform travel flow demand, during (*A*) weekdays and (*B*) weekends. Data points show the results for networks of the first two weeks in September 2017, i.e. 148 networks for weekends and 370 networks for weekdays. For network of each particular time, 26 data points are scattered for threshold ρ values (depicted by color intensity) between 0 and 1 with steps of 0.04. With increasing ρ, UD' (the proportion of connected node pairs) decreases faster than UD (the proportion of unbroken trips). Furthermore, the decrease in UD' is faster in weekdays compared to that of weekends.



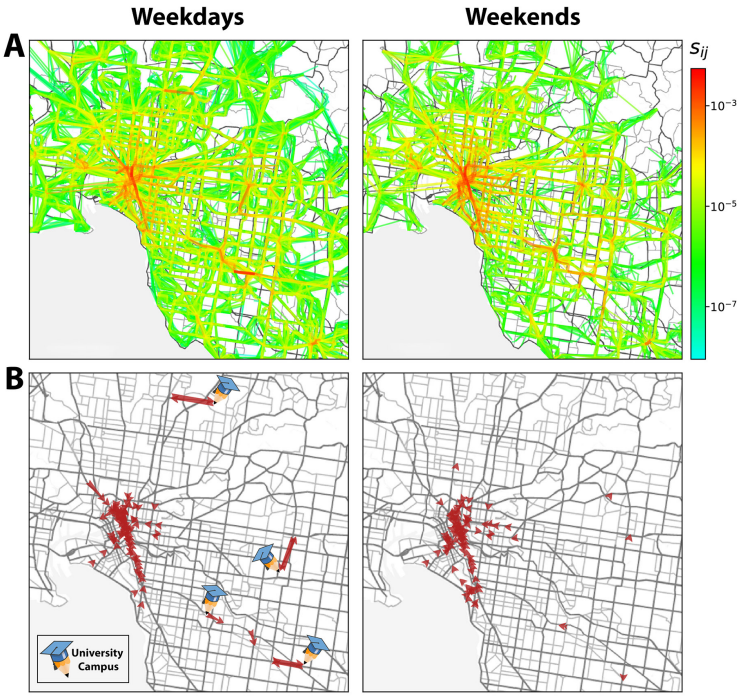
Supplementary Figure 6. Temporal reliability of Melbourne's on-road Public Transportation (PT) network. Temporal evolution of reliability indices ρ_C and α during the day is depicted for Melbourne's PT network with δ set as 60 min (left column), 120 min (middle column), and 180 min (right column) while time window is moved in 30 min steps over the day. Temporal evolution of (*A*) ρ_c and (*C*) α, during the day for all weekdays (each day has a unique color) in September and October 2017. Mean (*B*) ρ_c and (*D*) α, as a function of time of the day, averaged separately over weekdays (green) and weekends (weekends). In *B* and *D*, thickness of the shaded area around the curve is equal to two standard deviations at each particular time of the day.



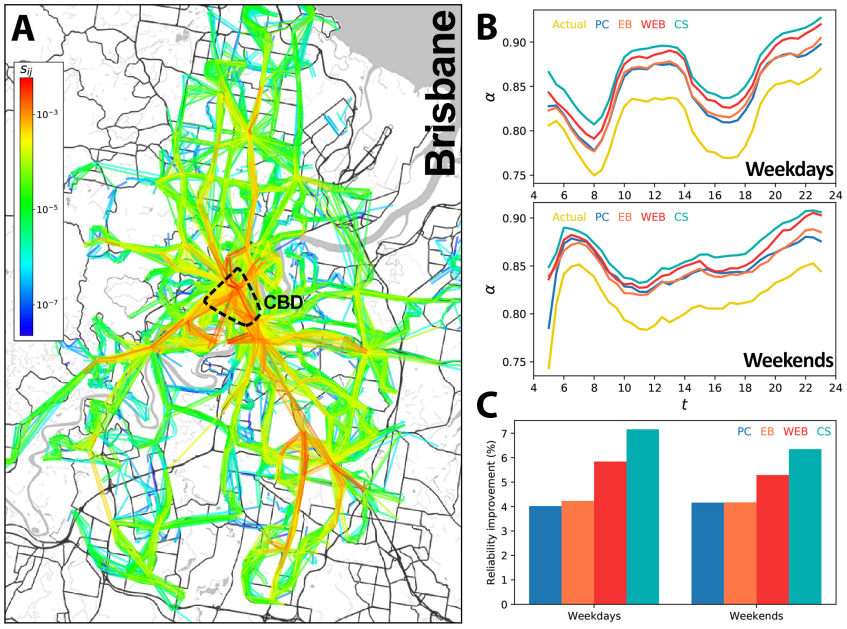
Supplementary Figure 7. Temporal structure of Melbourne's on-road Public Transportation (PT) network. (*A*) The maps show the structure of the network at three points in time over a normal weekday. Each link on the network is depicted with a straight blue line connecting its source and target nodes (corresponding to stops). Street map layers © OpenStreetMap contributors (12). (*B*) Temporal average degree $\langle k \rangle$ of the network and the number of its links $|E|$ versus time t of the day, separately for weekdays (blue) and weekends (orange).



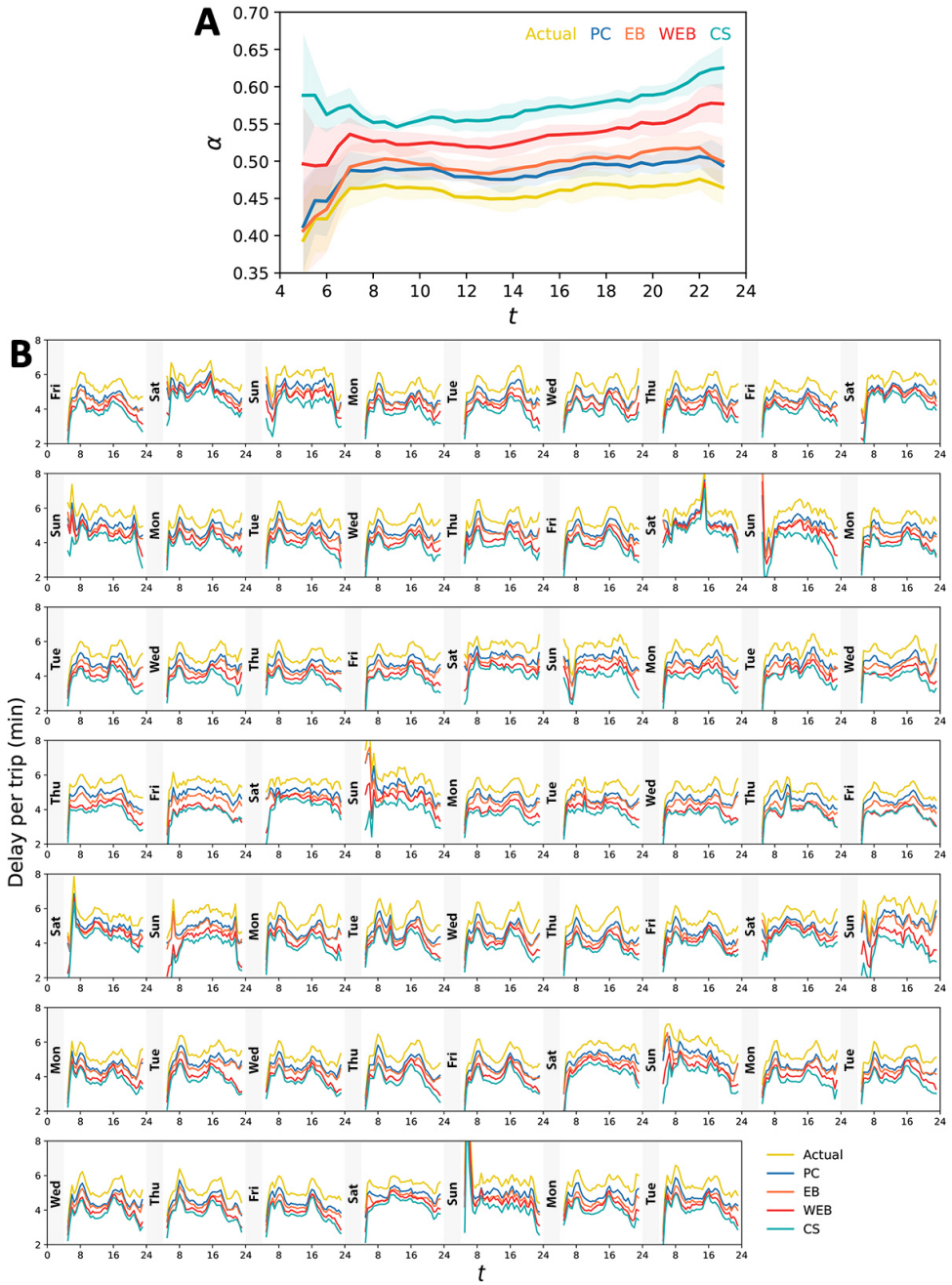
Supplementary Figure 8. Robustness of the identified bottlenecks in Melbourne's PT network. Here, we investigate the extent to which the identified network bottlenecks, persist as the most critical links on the networks of different days. So, we calculated the average criticality score of each link over different times of each particular day to identify the most critical link on the day. Then, we counted the number of the occurrences for each identified bottleneck among the top most critical links of each day. (*A*) Fraction of days on which top bottlenecks appear in the set of most critical daily links versus the link rank while links are sorted in descending order of their robustness over the different. (*B*) Same as *A*, but the link ranks are normalized between 0 and 1 (most and least persisting links are associated with 0 and 1 respectively) to show that the results are not sensitive to the number of selected bottlenecks. In separated experiments, the number of top bottlenecks of the network and most critical links on each day are both limited to 250 (100), 500 (200), and 1,000 (400) for networks of weekdays (weekends). Regardless of the number to which we limit number of to daily critical links and network bottlenecks, almost 80% of the top bottlenecks, appear as the most critical link on approximately 75% of the days, supporting the robustness of the identified network bottlenecks.



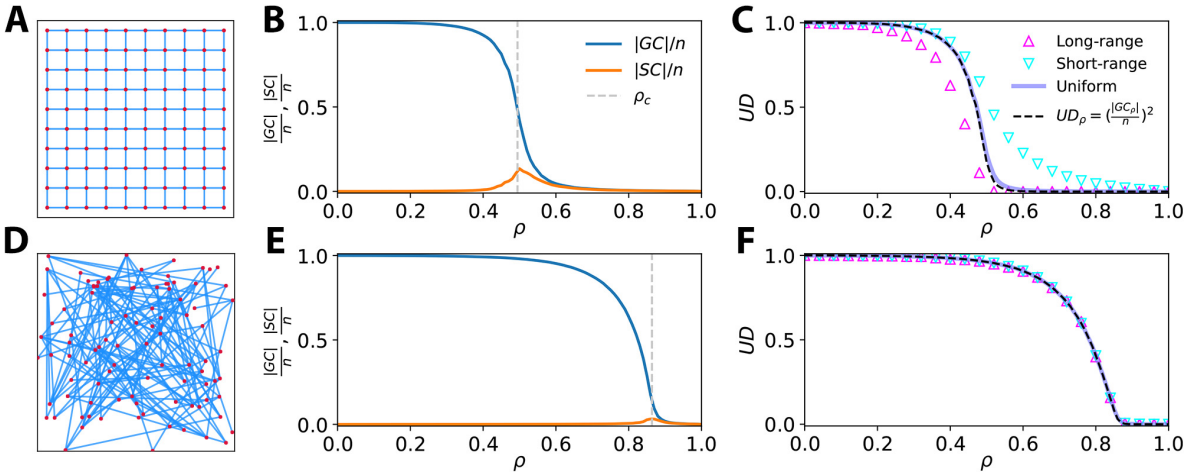
Supplementary Figure 9. Spatial distribution of critical links. (*A*) Spatial distribution of the link criticality scores over the Melbourne's on-road PT network, separately for weekdays (left panel) and weekends (right panel). (*B*) Top 100 bottlenecks identified according to link criticality scores, separately for weekdays (left) and weekends (right). The major visible distinction between the top 100 bottlenecks of weekdays and weekends is emergence of the links associated with major university campuses (pinned on the map) outside Melbourne CBD area on weekdays. This is an evidence on alignment of the identified bottlenecks with urban activity centers, as large university campuses are obvious hotspot points during the weekdays which cease to act as hotspots during the weekends. Street map layers © OpenStreetMap contributors (12).



Supplementary Figure 10. Improving Brisbane's on-road public transportation network. (*A*) Spatial distribution of the link criticality scores over the Brisbane's on-road PT network of weekdays. Street map layer © OpenStreetMap contributors (12). (*B*) Reliability α over the day calculated for the actual (yellow) network, and its synthetically improved versions obtained by ameliorating CS bottlenecks (cyan), EB bottlenecks (orange), WEB bottlenecks (red), and PC bottlenecks (blue); each curve shows the average over the days of March 2013. (*C*) Overall improvement in network reliability α achieved by amelioration of different types of bottlenecks; the results correspond to the average of α over all snapshots of the network, separately, during weekdays and weekends.



Supplementary Figure 11. Improving Melbourne's Public Transportation (PT) network by ameliorating its bottlenecks. (*A*) Daily evolution of α calculated for the actual (yellow) and improved networks obtained by ameliorating CS bottlenecks (cyan), EB bottlenecks (orange), WEB bottlenecks (red), and PC bottlenecks (blue). Results show the average (solid line) and standard deviation (shaded area) over the weekend days during September and October 2017. (*B*) Effect of improving bottlenecks on the delay caused by road conditions. Additional delay per passenger trip imposed by the road conditions on the PT system is shown for the actual network and improved networks. Improved networks are simulated by synthetically increasing the quality of bottlenecks identified using three well-established approaches in addition to our proposed approach.



Supplementary Figure 12. Capturing properties of the demand-serving networks with grid and random graph structure. (*A*) A sample square grid graph. (*B*) Normalized $|GC|$ and $|SC|$ during the percolation process averaged over 100 realizations of random link qualities on a square grid of size 2,500 nodes. (*C*) Unaffected Demand (UD) versus ρ for different flow demand scenarios on the grid structure, averaged over 100 realizations. (*D*) A sample random graph generated using Erdős-Rényi (ER) model (13) where nodes are attributed with a random spatial position. (*E*) Normalized $|GC|$ and $|SC|$ during the percolation process averaged over 100 realizations of random link qualities on ER networks of size 2500 and average degree of $\langle k \rangle \approx 8$. (*F*) UD versus ρ for different flow demand scenarios on ER network structures, each averaged over 100 realizations.

Supplementary References

1. School policy and advisory guide: School hours. (https://www.education.vic.gov.au/school/principals/spag/management/Pages/hours.aspx) (2019) [Accessed 21/07/2020].
2. Fair work ombudsman: Full-time employees. (https://www.fairwork.gov.au/employee-entitlements/types-of-employees/casual-part-time-and-full-time/full-time-employees) (2019) [Accessed 21/07/2020].
3. WJ Youden, Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
4. A Alsger, B Assemi, M Mesbah, L Ferreira, Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transp. Res. Part C: Emerg. Technol.* **68**, 490–506 (2016).
5. M Munizaga, F Devillaine, C Navarrete, D Silva, Validating travel behavior estimated from smartcard data. *Transp. Res. Part C: Emerg. Technol.* **44**, 70–79 (2014).
6. LK Fleischer, Approximating fractional multicommodity flow independent of the number of commodities. *SIAM J. on Discret. Math.* **13**, 505–520 (2000).
7. M Pollack, Letter to the editor—the maximum capacity through a network. *Oper. Res.* **8**, 733–736 (1960).
8. EW Dijkstra, A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
9. M Girvan, ME Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).
10. YN Kenett, et al., Flexibility of thought in high creative individuals represented by percolation analysis. *Proc. Natl. Acad. Sci.* **115**, 867–872 (2018).
11. D Li, et al., Percolation transition in dynamical traffic network with evolving critical bottlenecks. *Proc. Natl. Acad. Sci.* **112**, 669–672 (2015).
12. Openstreetmap copyright and license. (https://www.openstreetmap.org/copyright) (2019) [Accessed 21/07/2020].
13. B Bollobás, B Béla, *Random graphs*. (Cambridge university press) No. 73, (2001).

Chapter 5

Percolation-based Traffic Signal Control

Perimeter control coordinates the vehicular movement in road networks, by timing signals on the boundaries separating regions with distinctive traffic dynamics. The method is proved to be effective in optimizing urban road traffic and is a crucial component in modern intelligent transportation systems. A potential drawback of controlling the incoming traffic of a network region is the susceptibility to formation of congested queues outside the perimeter. Here, we propose a multi-perimeter control scheme that alleviates the propagation of congestion using a time-varying perimeter underpinned by percolation analysis.

In the proposed scheme, a classic fixed perimeter control is implemented to optimize traffic flows within a target region of particular importance, i.e., a hotspot region of the network. Queue development at the boundary of the fixed perimeter can lead to a cascade propagation of congestion, from congested links approaching the perimeter from outside the region to their upstream links [152]. The propagation dynamics depend on several factors such as travel demand, road network structure, and physical properties of road segments (e.g. number of lanes) [153, 154]. The complexity and stochasticity of the factors involved make it difficult to predict the shape of congested component(s) emerging due to control at the fixed perimeter [4, 154]. Therefore, we aim at identifying the smallest region that contains the pockets of congestion formed outside the first perimeter, that are likely to merge as a result of congestion propagation. A second perimeter will be implemented at the boundary of this identified region, thus, the region will act as a buffer space between the two perimeters, i.e., where congestion propagation is treated.

Well-established requirements for a suitable buffer space, are low variance in the density of its links and its compactness [155]. The former guarantees a well-defined MFD explaining the flow dynamics in the component, and the latter allows for effective traffic manage-

ment via signals on the boundary of the component. We study the percolation of congestion outside the fixed perimeter to identify a compact buffer space (made up of only congested links) that encompasses the merging high-level congestion pockets. As the organization of congestion outside the protected region evolves, whether due to demand temporality or by control at the perimeters, we update the buffer space using the proposed percolation-based procedure, to keep the control at the second perimeter focused on mitigating the current situation. The application of percolation theory allows for controlling the traffic to/from the percolating cluster of links with higher congestion levels to prevent the imminent integration of small congestion pockets into a congested cluster of substantial size.

As queue spillback occurs in a (spatio-temporally) heterogeneous manner, the control scheme coordinates the traffic signals to balance the queues based on the organization of congestion unpacked by percolation analysis. The dynamic percolating cluster is identified at different points in time and the signal control setting will be adjusted and executed accordingly. We demonstrate the performance of the proposed controller in a typical grid network. Our simulation results show that i) the evolution of the congestion can be well-characterized through percolation approaches, and ii) the percolation-based dynamical perimeter control significantly improves the network performance, compared to the classic fixed perimeter control.

5.1 Methodology

We consider the road network system of a generic mono-centric city where a primary hotspot region generates and attracts a substantial portion of the traffic. The boundary of this hotspot region can be determined so that the flow dynamics within the region follows a well-defined Macroscopic Fundamental Diagram (MFD) [155]. This basically means that the region's flow movement (e.g., in veh.km/h) can be explained as function of the traffic density (e.g., in veh/km), with a critical density separating free-flow and congested regimes where the function has respectively positive and negative derivatives with respect to density. (Figure 2.2 depicts a well-defined MFD.) A perimeter signal control is implemented at the intersections on the determined boundary of this region, protecting the entire region from congestion by regulating the flow entry and exit according to the region's MFD. Let us refer to this as the first or fixed perimeter as this perimeter remains spatially fixed over time, although the timing of its signals is altered in real time depending on the region's density.

The queues generated by this perimeter may propagate toward the upstream of the protected region. This propagation of congestion can lead to heterogeneous spillback over the periphery space of the region and hinder the desired operation of the fixed perimeter control, reducing the global traffic efficiency in the network. To this end, we propose an

extended control scheme where a boundary-adaptive second perimeter circumscribes the traffic propagated from the fixed perimeter to balance the queues and smooth the traffic flows approaching the hotspot region of the network. The second perimeter is implemented at the boundary of a buffer space determined by percolation analysis of congestion outside the fixed perimeter. This strategy differs from those reacting to the queues formed at gated links by adjusting the gating at the first perimeter [156, 157, 158], and it is more proactive in the sense that it attempts at preventing the formation of those queues via the second perimeter.

The flowchart in Fig. 5.1 illustrates the proposed control framework. In the remaining of this section, we first explain the proposed percolation approach to characterize the evolving congestion around a target region (Section 5.1.1). Next in Section 5.1.2, the proposed multi-perimeter control method is formulated in detail, and main components of the control strategy are explained in detail.

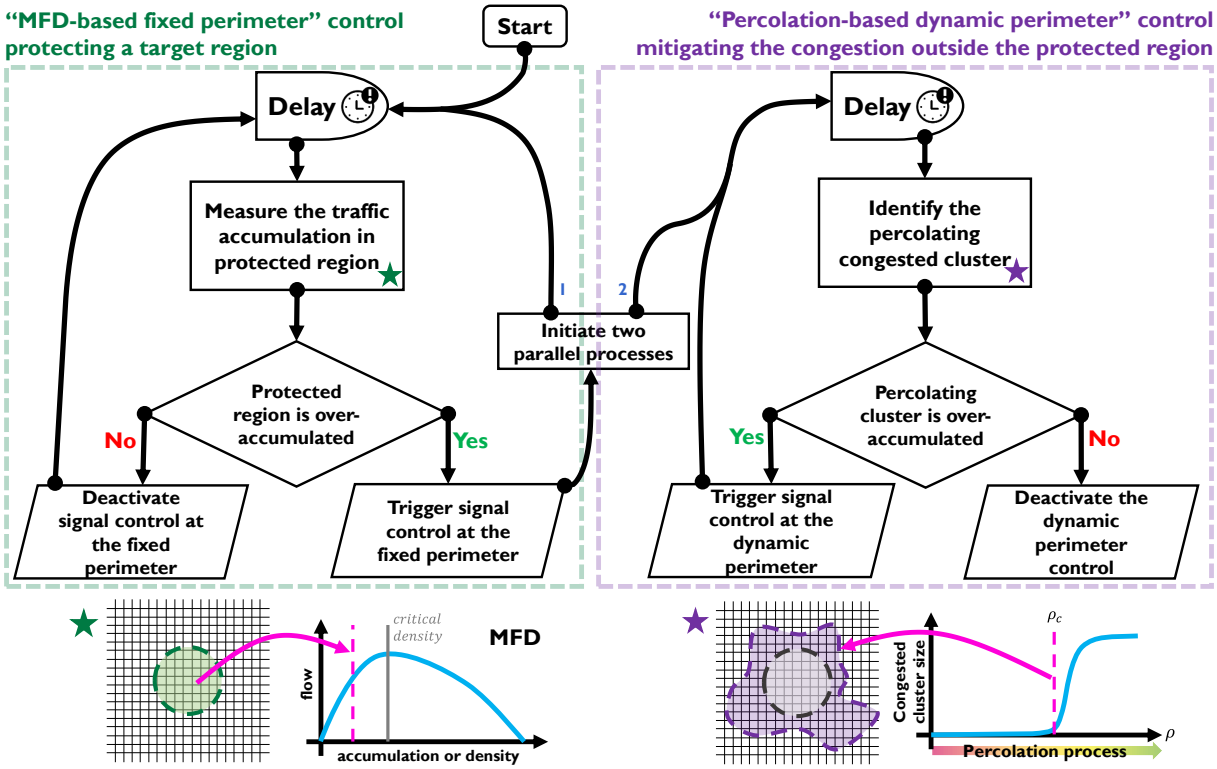


Figure 5.1: Control scheme flowchart. The flowchart illustrates the procedure used by the proposed multi-perimeter control. Two major modules in this procedure are highlighted, one controlling the fixed perimeter (dashed green line) and the other related to the dynamic second perimeter (dashed purple line). At the bottom of the figure it is illustrated that i) the region encompassed by the fixed perimeter is monitored to determine the traffic state according to the region's MFD (green star), and ii) percolation analysis of the congestion outside the fixed perimeter leads to identification of the boundary of the second (dynamic) perimeter (purple star).

5.1.1 Congestion propagation analysis

Let us denote the graph representation of the road network by G, where nodes represent intersections and links represent road segments connecting intersections. The level of congestion on a link i at time t can be indicated by its relative density $k_i(t) \in [0, 1]$, calculated as the instantaneous density (number of vehicles per one kilometer lane) divided by the jam density of the link i. Let us consider a single snapshot of the actual network in time $G(t)$ with a particular distribution of congestion. A simple percolation process can be simulated using a threshold ρ and starting from an empty network G_ρ at $\rho = 1$, by gradually decreasing the threshold ρ and simultaneously adding any link i (from the original network G) with a relative density above the threshold $k_i \geq \rho$ to the network G_ρ. At the beginning of the process (ρ close to unity), G_ρ is comprised of only highly congested links, and as the threshold decreases, less congested links are progressively added to the network. The evolution of the network G_ρ resembles the propagation of congestion from highly congested links to their neighborhood. The process terminates when $G_\rho = G$ at $\rho = 0$.

We are interested in finding a compact buffer space that encompasses local congested communities around the fixed perimeter. In case the congestion around the first perimeter is concentrated in separate highly congested communities, our desired buffer space should include the links which are transitioning into the congested state, as growing congested communities are merging. Such a buffer space does not include links with low congestion-level and thus the variance of its links' densities is not expected to be high. To identify this desired buffer space, we use the concept of percolation criticality which is the point of phase transition from isolated small congested clusters to a connected congested cluster of substantial size in the aforementioned percolation process.

Monitoring the size (number of nodes) of the connected components during the percolation process reveals important properties of the network. Monitoring the size of the largest connected component, also called Giant Component (GC), or the second-largest connected component (SC) is of special importance when studying the network properties from its percolation behavior [159, 160]. Let us denote the size of the GC and SC of a network, with $|GC|$ and $|SC|$, respectively. The percolation criticality can be identified as the point where a GC of significant size emerges. In practice, a convention is to determine the percolation criticality as the point in the percolation process where $|SC|$ is maximal, marked by the percolation threshold ρ_c [43, 159]. Percolation criticality separates the phase where no significant GC exists on the network (subcritical phase) and the phase where small pockets of congestion are merged and form a GC of significant size (supercritical phase). In this sense, percolation threshold ρ_c can be formally defined as:

$$\rho_c = \underset{\rho \in [0,1]}{\operatorname{argmax}} |SC_\rho|, \tag{5.1}$$

where $|SC_\rho|$ denotes the size of the second-largest connected component of G_ρ (network under percolation at threshold ρ).

To illustrate the procedure of finding the percolation-based buffer space, it is helpful to imagine the inverse percolation process, i.e., starting from the complete network, links with relative densities falling below the threshold are removed as the threshold ρ is increased from 0 to 1. The inverse percolation process is illustrated in an example network in Fig. 5.2a-e. At first, the GC of the network contains all network links, as in a functional road network there is at least one path connecting any pair of intersections (see Fig. 5.2a). During the inverse percolation process, first, low-density links are removed from the network thus, eventually, the shrinking GC is only made of medium- and high-density links. Just below the percolation threshold $\rho \to \rho_c^-$ all remaining links in the network (under percolation) have densities above the threshold, i.e. for any remaining link i on the network $k_i \geq \rho_c$ (Fig. 5.2c). Then, at percolation threshold link(s) with the least density $k_i = \rho_c$ are removed causing the GC to suddenly collapse into small- and medium-sized connected components comprised of congested links (Fig. 5.2d). The GC of the network just before the phase transition (area shaded in light-blue in Fig. 5.2c) contains the pockets of congestion produced by the control at the first perimeter; this component is often called 'percolating cluster.' We take the percolating cluster as our candidate buffer space. The buffer space B is a subnetwork of G ($B \subseteq G$), which can be formally defined in our terms as:

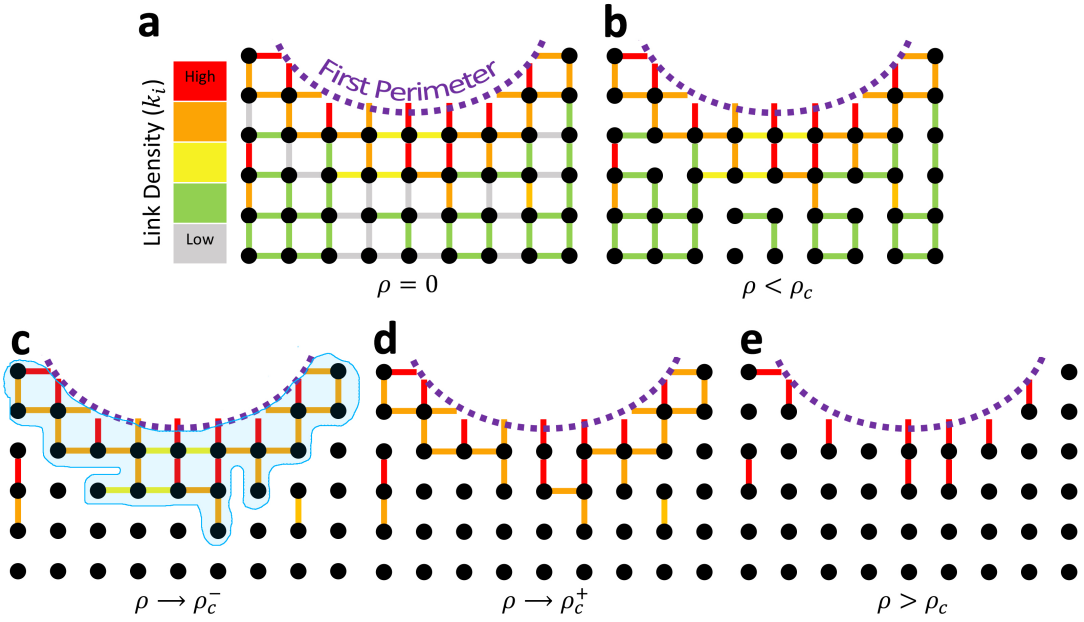


Figure 5.2: Schematic illustration of an inverse process of congestion percolation. The dashed curve (purple) marks the fixed controlled perimeter around a hotspot region. Low to high link-densities are color-coded in the toy road network. The light-blue shaded area in **c** highlights the identified buffer space according to our percolation-based approach, i.e., the GC at percolation criticality.

$$B = \lim_{\rho \to \rho_c^-} |GC_\rho|, \tag{5.2}$$

where ρ_c can be calculated as in Eq. (5.1) and GC_ρ denotes the GC of G_ρ (network at threshold ρ).

At any point in time if the congestion is not propagated over the network, by definition, the percolation criticality occurs at one end of the percolation process; this is the same as if all the network links are in the same state. However, if a percolating congested cluster exists, we trigger the signal control at the intersections marked by the boundary of the identified buffer space. The goal of the signal control at this perimeter is to mitigate the traffic congestion inside the buffer space and also optimize the flow exchange between the buffer space and the hotspot region. As congestion evolves (whether it dissipates or propagates) we identify the percolating congested cluster and update the buffer space at different points in time. If the overall density in the new buffer space was over a predefined threshold, signal control at the perimeter of the buffer will be triggered to reduce the in-flow. Otherwise, control at the second perimeter will be disabled until the next update.

5.1.2 Multi-perimeter traffic signal control

The proposed multi-perimeter control algorithm essentially determines the activation and the degree of flow control at the perimeters upstream to a protected hotspot region. Fundamentally, the control at the first perimeter is performed according to the accumulation of vehicles inside the protected region and also the region's traffic dynamics described by its MFD. Control strategy at the second (dynamic) perimeter is governed by the dynamic growth of queues in every direction from the protected region. The flowchart in Fig. 5.1 illustrates an overview of the algorithmic procedure and different components of the proposed control scheme.

The first component of our proposed control scheme is an MFD-based fixed perimeter control at the boundary of a *hotspot* region of the network which generates and attracts substantial traffic. This single-region (sometimes referred to as a single-reservoir) perimeter control can effectively protect and improve the performance in a region of the network. The effectiveness of this control scheme is analytically proved [161] and tested under simulations [162]. As the first step to designing an MFD-based perimeter control for a part of the network, one can derive the MFD of a region by monitoring the traffic and measuring i) the accumulation (i.e., the number of vehicles) or density (i.e., number of vehicles per unit length of a lane) and also ii) the corresponding overall movement of vehicles or trip completion rate (the rate at which trips arrive or exit).

We find the MFD of the hotspot region, describing the trip completion rate $C(t)$ as a function of accumulation $N(t)$, i.e., $C(t) = F(N(t))$, at any time t. A useful and informative

MFD can be extracted from data collected by loop detectors installed in a small proportion of the streets [137, 163], nevertheless, this study is based on known road conditions (vehicular density/flow) in all network links. At any time t by measuring the out-flow of the hotspot region at the perimeter intersections, $F(N(t))$ can be divided into outgoing $n_o(t).F(N(t))$ and circulating $n_i(t).F(N(t))$ flows, with $n_i(t) = 1 - n_o(t)$. We perform this measurement by aggregating the data from the past 1-minute period and update $n_o(t)$ at every step of the simulation; more traffic simulation settings are described in Section 5.2. Similarly, the arriving flow to the hotspot region via all links connected to the perimeter intersections from outside the region is measured and denoted by $I(t)$.

The fixed-perimeter component of our proposed control strategy (see the green module in Fig. 5.1) is concerned with keeping the accumulation in the hotspot region of the network close to the critical accumulation, where the traffic movement is at its maximum, but restraining the accumulation from exceeding N_c. The critical accumulation is the point marking the transition between free-flow and congested regimes in the MFD of the region (see Fig. 2.2 in Chapter 2) and can be derived as:

$$N_c = \underset{0<N<N_{jam}}{\operatorname{argmax}} \; F(N). \tag{5.3}$$

The objective of the controller at the boundary of the hotspot region is to eliminate the difference between the steady-state accumulation and the accumulation maximizing the flow (i.e., N_c), which requires satisfying the following:

$$\lim_{t\to\infty} \varepsilon(t) = \lim_{t\to\infty} N(t) - N_c = 0. \tag{5.4}$$

The control problem is subject to the system state dynamics in terms of the change in vehicle accumulation per unit time, which can be formulated as:

$$\dot{N}(t) = \beta(t) + q(t) - F(N(t)), \tag{5.5}$$

where at any time t, $q(t)$ denotes the trip generation rate (the number of trips initiated per unit time) inside the region and $\beta(t)$ is the entry flow rate allowed by the perimeter control to the region at time t.

The system dynamics are regulated by the following control law to obtain the steady-state error of zero:

$$\dot{\varepsilon}(t) + k\epsilon(t) = 0, \tag{5.6}$$

where k is the control gain parameter [164]. Substituting Eqs. (5.4) and (5.5) into Eq. (5.6),

we obtain the optimal controlled entry flow $\beta(t)$:

$$\beta(t) = k(N_c - N(t)) + F(N(t)) - q(t). \tag{5.7}$$

The control gain parameter k can be calibrated through trial-and-error process [161], however, we simply use $k = 1$, as our focus in this work is more on assessing the percolation-based analysis of congestion. The critical accumulation is often adjusted to a smaller value than the critical accumulation N_c according to the MFD, so that the controller functions in a proactive manner. In our simulations, we trigger the control at 80% of the critical accumulation, which means substituting N_c for $0.9.N_c$ in Eq. (5.7).

The trip generation rate $q(t)$ does not vary drastically within a small time step in any urban network. So, we initially estimate $q(0)$ to calculate $\beta(0)$, and with each update of the accumulation $N(t+\Delta t)$ the trip generation of the last step can be estimated by rearranging Eq. (5.5) as below, to calculate β for the current step:

$$\hat{q}(t) = \frac{N(t+\Delta t) - N(t)}{\Delta t} + F(N(t)) - \beta(t). \tag{5.8}$$

With this estimation, the simple controller defined here does not require information such as the network's demand and only requires monitoring the network links to derive $N(t)$, $n_o(t)$, and $I(t)$, which is feasible using the existing technologies such as loop detectors and traffic cameras [165].

We implement a four phase traffic signal at each road intersection with a fixed cycle length of S, where flow from each approach is put into a single phase avoiding all conflicts. At every perimeter intersection let g_0 and $1-g_0$ be the proportion of the signal cycle already allocated to all phases associated with the flow incoming to and outgoing from the hotspot region, respectively. To optimize the traffic within the hotspot region, we should find the optimal g and $1-g$ to adjust the proportion of signal cycles serving the in- and out-flow of the region. Assuming that the flows enter and exit the region homogeneously via all perimeter intersections, on average the entry flow allowed to the region by the controller should be $\beta(t) = g.I(t)/g_0$, but determining g affects the out-flow of the region that is $(1-g).n_o.F(N(t))/(1-g_0)$. To find the new optimal timing of the signals, we rewrite Eq. (5.7) as:

$$\frac{g}{g_0}.I(t) - \frac{1-g}{1-g_0}n_o.F(N(t)) = k(N_c - N(t)) + n_i.F(N(t)) - q(t), \tag{5.9}$$

which can be rearranged as follows, determining the green time proportion which should be given to the in-flow of the region:

$$g = \frac{k(N_c - N(t)) + n_i.F(N(t)) + n_o.F(N(t))/(1-g_0) - q(t)}{I(t)/g_0 + n_o.F(N(t)/(1-g_0)}. \tag{5.10}$$

At each perimeter intersection, we divide the time $g.S$ equally between all approaches feeding the region and divide $(1-g).S$ equally between the rest of the approaches. A predefined small positive value can be assigned to g in case Eq. (5.10) resulted in a negative value for the allowed in-flow green time.

As it is seen in Fig. 5.1, when the fixed perimeter around the hotspot region is active, the control strategy checks for a percolating congested cluster formed outside the fixed perimeter. As soon as a percolating congested cluster is identified (as explained in detail in Section 5.1.1), a second perimeter is triggered at the boundary of that cluster which will be updated over time as shown in Fig. 5.1. The aim of the dynamic perimeter is to limit the in-flow of the buffer space (i.e., the identified percolating cluster) and balance the length of queues formed in different directions away from the fixed perimeter around the hotspot. So, for each intersection i on the second perimeter, we reduce the green time of the approaches serving the in-flow by a factor of $1/(d_M(i)+1)$, where $d_M(i)$ is the Manhattan distance between the intersection i and the closest intersection on the fixed perimeter.

Updating the time at the approaches of concern in the intersections lying on the dynamic perimeter, simply deters drivers from choosing the paths that lead to increasing the length of longer congested queues and encourages drivers to change their route to their destination to paths that end up at the boundary of shorter queues. Note that as soon as one of the perimeters is deactivated or the second perimeter is updated, intersections previously located on the perimeter will be set back to their default signal timing. Also, since reduced green times are always calculated as a proportion of the default green time, no approach is given an all-red signal during a full cycle.

5.2 Results

5.2.1 Simulation settings

To test the proposed control scheme, we use a realistic agent-based model to perform micro-simulations in a 20×20 square grid network where each pair of neighboring nodes are 300 m apart. The network is generated by connecting each pair of neighbor nodes via two oppositely directed links at first, and then randomly removing 15% of the links[4]. The configuration of signals at all intersections (nodes) follow a classic four-phase cycle, with pre-timed signal settings. In the simulation, traffic movement with respect to signals is idealized in the sense that no time is lost when switching between red and green signals at different approaches of an intersection.

[4]We ensure that the final network remains strongly connected, i.e., there is at least one path in each direction between any pair of nodes on the network.

For each link, the number of lanes is picked uniformly at random from the set $\{1, 2, 3\}$, and the free-flow speed is selected randomly between 40 and 60 km/h with probabilities of 0.8 and 0.2, respectively. Traffic on a single road lane follows a triangular speed-density relation shown in Fig. 5.3a. As it is seen in Fig. 5.3a, flow dynamics for the two different types of lanes corresponding to different free-flow speeds $u_f = 40$ or 60 km/h, are respectively characterized by different jam densities $k_{jam} = 150$ or 170 veh/km, and different maximum flows $q_{max} = 1200$ or 1800 veh/h. In real road networks, speed limit of roads can be correlated with their number of lanes, e.g. a 3-lane road with low speed limit is rare. For our purposes, however, this is not a concern and we are introducing complexity to the network structure to assure that the proposed analysis does not depend on the structural properties of the network.

Trips are generated with the demand profile shown in Fig. 5.3b, which mimics a typical peak period in real urban areas. The trip generation rate increases rapidly until it reaches a peak at about half an hour into the simulation, and then decreases with a slower rate

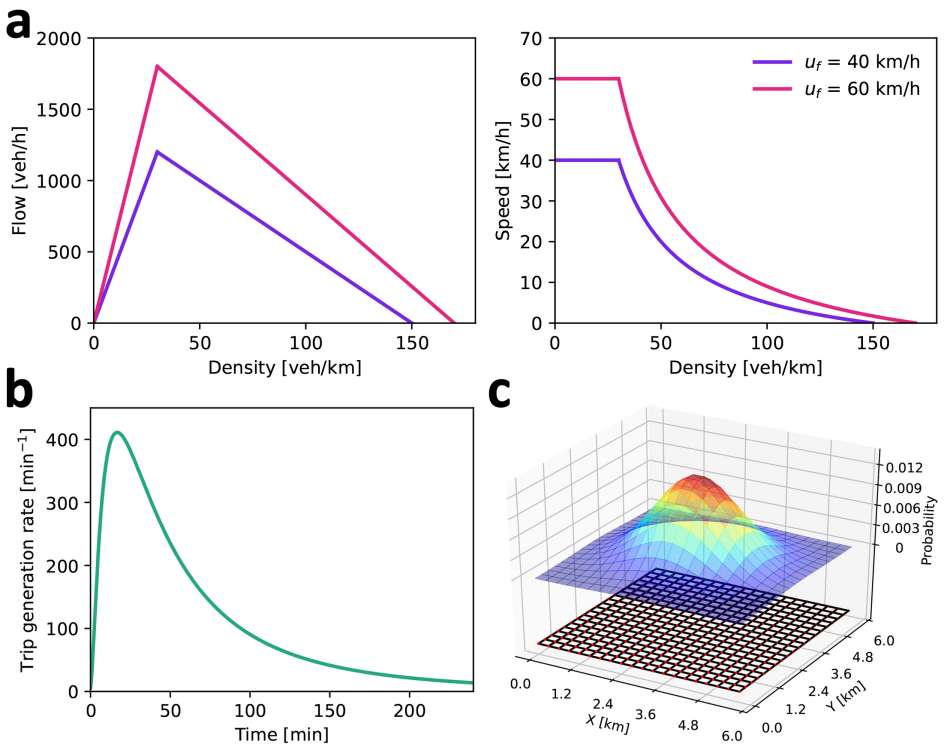


Figure 5.3: Basic settings of the road traffic network micro-simulations. **a** The simple triangular flow-density (left) and its consequent speed-density (right) relation for the two different types of links in the network, i.e., $u_f = 60$ km/h and 80 km/h free-flow speed. **b** Profile of the trip generation rate used in the simulations over a 4-hour period. **c** Probability distribution function (the Gaussian surface) used for selecting the destination of each trip in the grid network. The probability distribution function used to select trip origins only has a larger standard deviation. In both distributions, the probability is larger than zero for the points situated furthest away from the grid's center.

until the end of the simulation. This demand profile is used in all traffic simulations in this chapter, generating a sum of $30,000$ trips during the 4-hour period. Spatial distribution of the trip origin and destination is dictated by two Gaussian distributions[5]. We consider the central region of the network to be the network's hotspot, thus, generating and attracting more trips than the fringes of the network. In Fig. 5.3c a symbolic grid road network is shown, and the surface hovering over it, depicts the probability of each node being selected as the destination of each trip; origins are picked from a similar distribution only with larger variance. We assume that 80% of all drivers are informed about the link travel times and make their choices en route, i.e., at each intersection they choose the next link of their pathway with the objective of minimizing their travel time to the destination. The rest of the drivers (i.e., 20% of all vehicles), move along the shortest path between their origin and destination on the network.

Figure 5.4a shows the MFD of the network based on the flow movement and accumulation recorded during the loading phase, averaged over 10 runs of the traffic simulation. We monitored the flow dynamics inside regions with different radii at the center of the network, to find a region that accommodates a substantial traffic volume while the congestion remains fairly homogeneous over its links. Simulations showed that the region within the 0.75 km radius of the grid's geographical center has a well-defined MFD (see Fig. 5.4b for MFD of this region). This region is therefore identified as the 'hotspot' region of the network to be protected via signal control. The boundary of this region is thus fixed as the first (MFD-based) perimeter to be controlled (associated with the green module in Fig. 5.1). The fixed perimeter consists of 12 intersections along the boundary of the central region. As marked by the dashed line in Fig. 5.4b, the critical accumulation for the central region is $N_c \approx 1200$ vehicles, which divided by the total road lane length inside the hotspot region gives the critical density of approximately 29.5 veh/km.

5.2.2 Percolation of congestion from the perimeter

To improve the traffic flow inside the hotspot region of the network, we implemented an MFD-based fixed perimeter control at its boundary which functions based on Eqs. (5.3-5.10) in Section 5.1.2. The accumulation of vehicles inside the hotspot region is counted every $S = 1$ min (length of a signal cycle) and the control mechanism is triggered if the density of the region exceeds 80% of the critical accumulation of $N_c = 1200$ veh (equal to the critical density of 29.5 veh/km). The control mechanism reduces the in-flow of the region by

[5]In particular, selection of an origin-destination pair for each trip is performed according to $\|p_o - p_0\|_2 \sim \mathcal{N}(0, 5)$ and $\|p_d - p_0\|_2 \sim \mathcal{N}(0, 2.5)$ distributions, where p_o, p_d, and p_0 are positions of the origin, destination and grid's center and $\|.\|_2$ returns the Euclidean norm of the input vector. $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with average μ and variance σ^2, here in kilometers.

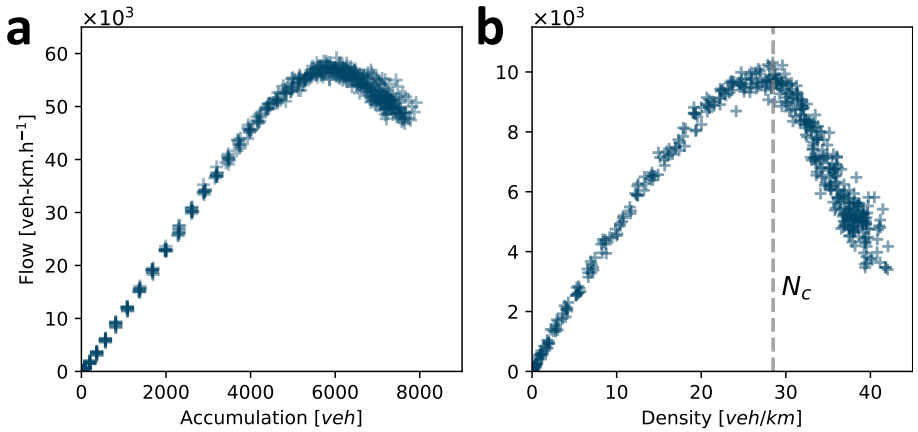


Figure 5.4: Traffic dynamics in the simulated network. The MFDs describing the traffic flow dynamics of the whole network (in **a**) and within the hotspot region of the network (in **b**). The 'hotspot region' consists of network links falling within the 0.75 km radius from the grid's geographical center.

decreasing the green time given to the incoming flows at all perimeter intersections, which in effect also increases the green time serving the region's outgoing flows and thus increases the region's out-flow.

While the in-flow of the protected region is being reduced by the fixed perimeter, queues build up in the upstream of the perimeter. As pockets of congestion grow and shape small congested clusters, a portion of traffic flows will successfully avoid the clusters by seeking the optimum routes to their destination. But even this contributes to the density of links between pockets of congestion and boosts the congestion propagation. One can expect that in the presence of enough traffic directed to the protected region of the network, links bridging between the small clusters of highly congested links become more accumulated. These congested clusters formed outside the fixed perimeter will eventually merge and form a congested component of significant size around the hotspot region. As soon as a connected congested component forms outside the hotspot region, the cost of detour to avoid the congestion increases and leads to even faster growth of queues. Using percolation analysis of the link-level congestion we detect the forming congested cluster and deploy the dynamic (or second) perimeter control at the boundary of this cluster to mitigate its propagation and improve the network's overall performance.

Figure 5.5 illustrates the location of the fixed perimeter control at the boundary of the hotspot region (the dashed purple circle in Fig. 5.5a-c) and the built-up queues in the upstream of the perimeter. The figure also depicts how the percolation process leads to identification of the congested component around the fixed perimeter. Starting from low-density links and gradually removing the shell of least congested links on the network (Fig. 5.5a), the process reaches the critical percolation threshold (Fig. 5.5b) where removing the next link causes the GC comprised of congested links to fragment into smaller pieces (Fig. 5.5c).

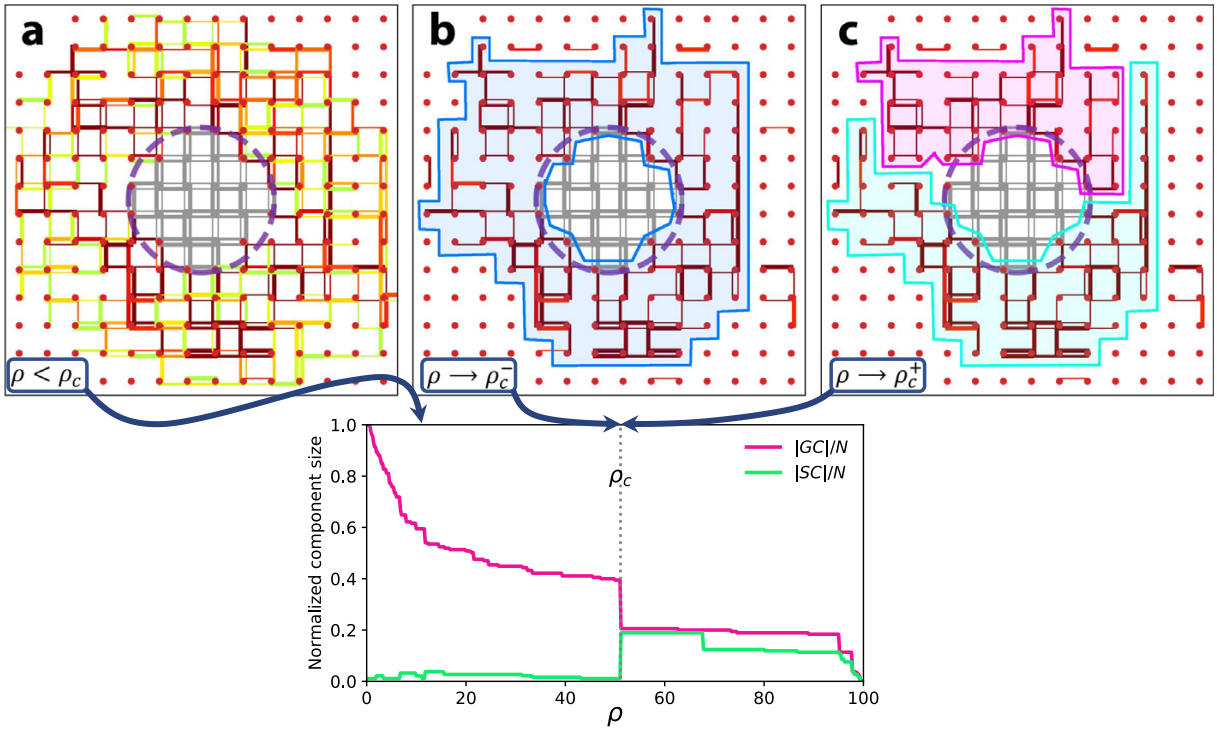


Figure 5.5: Identifying the congested cluster outside the hotspot region of the network. Here, the inverse percolation process on one snapshot of the network is illustrated at three different thresholds (ρ), namely, at a low threshold below the criticality (**a**), approaching the criticality from below (**b**), and approaching the criticality from above (**c**). Link colors indicate their relative densities, with green, yellow, and red being indicative of low, medium, and high traffic congestion. **d** The same percolation process as in (**a**-**c**) is demonstrated via the normalized size of the GC and the SC as functions of the threshold ρ.

As explained in Section 5.1.1, percolation threshold ρ_c is characterized by an abrupt drop in $|GC|$ and a maximal $|SC|$ during the percolation process. The percolation process shown at three thresholds in Fig. 5.5a-c is fully demonstrated in Fig. 5.5d via $|GC|$ and $|SC|$ of the network under percolation as functions of the threshold ρ (the critical threshold ρ_c is marked with a dashed grey line). The connected component at threshold ρ_c (i.e., the percolating cluster) contains the propagated congestion outside the fixed perimeter.

As time passes, the propagated congestion around the fixed perimeter evolves as a result of either the ongoing traffic or the signal control at the perimeters. So, over time, we identify the instantaneous percolating cluster of congested links, update the boundaries of the second perimeter, and control its signals to mitigate the congestion propagation outside the protected hotspot region. Figure 5.6 shows the network at three different points in time during a simulation. In the figure, vehicular density of links is color-coded (green to red representing free-flow to congested states), and the boundaries of both fixed and dynamic perimeter control are marked. The performance of our proposed control scheme is investigated in the next section.

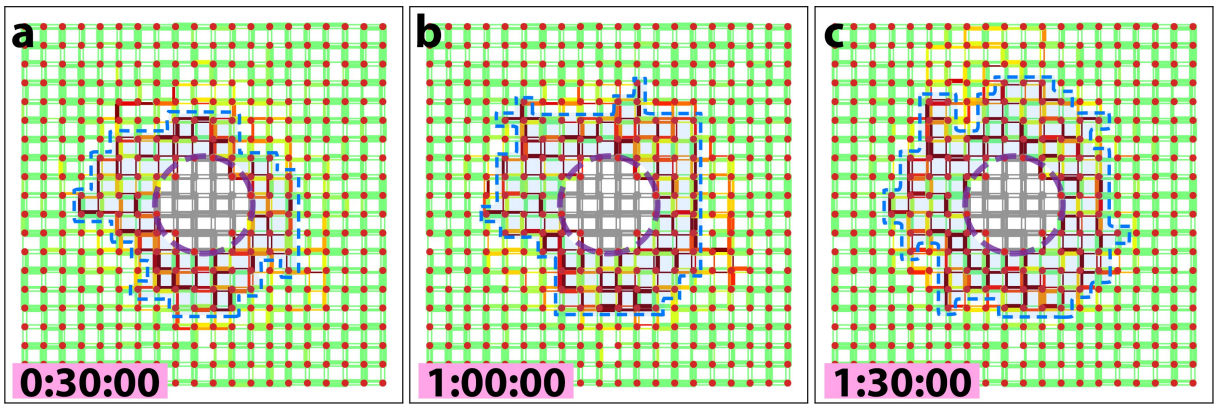


Figure 5.6: Evolution of the dynamic perimeter. Simulation environment is shown at three different snapshots: 30 min (**a**), 60 min (**b**), and 90 min (**c**) into the simulation. Link colors indicate their relative densities, with green, yellow and red being indicative of low, medium, and high relative vehicular densities. The dashed purple circle shows the first fixed perimeter and the dashed blue line shows the time-varying second perimeter identified using the proposed percolation-based methodology, at different points in time during the simulation. Links in the central region are colored gray as identification of the buffer space is not concerned with their state.

5.2.3 Control performance evaluation

We simulate a base scenario, with all the signalized intersections working according to their default setting, i.e., the green time allocated to each approach is proportional to the number of its lanes. The simulation results are recorded, first with no control strategy, then with the fixed perimeter control implemented around the hotspot region, and finally, with a time-varying percolation-based perimeter implemented (i.e., the dynamic multi-perimeter control) on top of the fixed perimeter. We refer to these control strategies as *pre-timed control*, *fixed perimeter*, and *dynamic perimeter*, respectively. For each control scheme, the simulation is repeated 10 times and the average outcome of those independent realizations is presented in the following. Averaging over multiple realizations diminishes the potential effects of randomness in picking origin-destination pairs from the predetermined distributions (see Section 5.2.1).

Overall performance

The two control approaches and the base pre-timed control scenario (i.e., no-control strategy) are compared in Fig. 5.7. Figure 5.7a depicts the network accumulation over time for three control strategies, and Fig. 5.7b presents the MFD of the whole network for the same simulation scenario under each of the three different control strategies. The fixed and dynamic perimeter controls are triggered at approximately 18 and 26 minutes (on average) into the simulations, respectively.

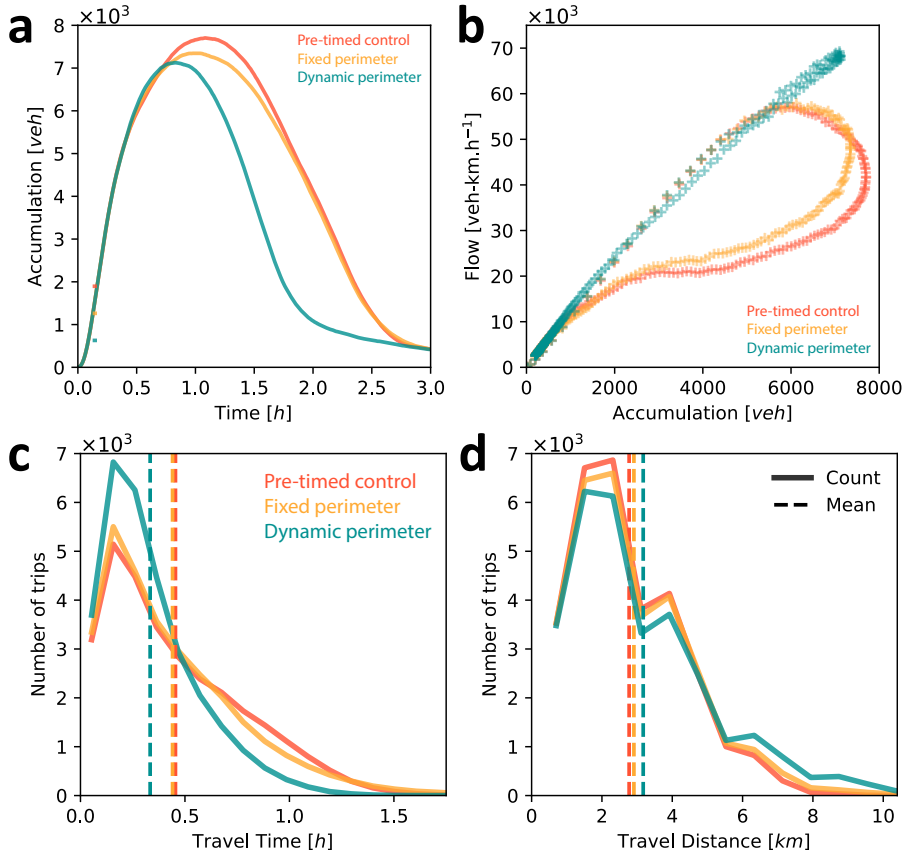


Figure 5.7: Comparison between different traffic control strategies. a The number of vehicles in the network (network accumulation) over time. **b** Flow versus accumulation in the network at different points in time during the simulation. Each data point corresponds to the sum of all vehicle movements and average number of vehicles obtained over one minute. **c**,**d** Histograms of travel time (**c**) and travel distance (**d**) for all network trips (solid lines). Mean travel time and travel distance of trips is indicated via dashed lines. In (**a**-**d**) the data associated with each control strategy is derived by averaging over 10 independent realizations.

Both perimeter control schemes improve the service capacity of the network. This is seen in Fig. 5.7a with the yellow (fixed perimeter control) and green (dynamic perimeter control) curves varying below the red (pre-timed control) curve. The network accumulation increases rapidly at the beginning of the simulation (during the network's loading phase) in all control scenarios. Compared to the pre-timed control strategy, the fixed perimeter control slightly increases the overall flow of the network during the loading phase, and reduces the hysteresis in the MFD (compare red and yellow data points in Fig. 5.7b). This allows the network to transition earlier from loading to unloading phase. Comparing red and yellow data in Fig. 5.7a demonstrates this, as the peak network accumulation occurs earlier in time when the fixed perimeter control strategy is implemented.

Congestion is significantly mitigated under the proposed dynamic perimeter control and the result achieved by the dynamic perimeter strategy is far superior to the results of the

fixed perimeter. The efficiency of the dynamic perimeter control compared to the other two cases is seen in Fig. 5.7a from the earlier onset of the unloading phase (peak accumulation occurs earlier in time) and the significantly higher rate of trip arrival, leading to a significantly lower accumulation in the network for most of the times. When the dynamic perimeter control is in place, network flow barely drops during the loading phase, and the network unloads with no delay in flow recovery (see the green data points in Fig. 5.7b).

The dynamic perimeter control strategy redirects vehicles from longer to shorter queues formed toward the upstream of the first perimeter. This forces the congestion to spread more evenly around the network's main basin, i.e., the pre-determined hotspot of the network. The result is a significant reduction in the average travel time of all trips at the cost of an increase in travel distance of some trips. This is seen in the distribution of travel time (Fig. 5.7c) and distance (Fig. 5.7d) for trips associated with the dynamic perimeter scenario compared to the fixed perimeter and pre-timed control scenarios. The dynamic perimeter control strategy increases the average travel distance of trips by 14.4% but reduces the average travel time over all trips by 26.6%, compared to the case when no control strategy is in place (pre-timed signaling at intersections).

Intra- and inter-region traffic performance and dynamics

Here, we investigate the dynamics of the traffic flow exchange between the hotspot region (center) and the rest of the network (outer). To do so, we compare the performance of different control strategies in terms of their effect on traffic flows with different travel directions. With respect to the division of the network into the center and outer subregions, trips are divided into four travel direction categories, including two intra-region categories, i.e. center to center and outer to outer, and two inter-region categories, i.e., center to outer and outer to center. Figure 5.8 illustrates the trip completion rate of trips in these four directions (over time) in the same simulated traffic scenario but under different control strategies.

When intersection signals are functioning according to their pre-timed default setting, i.e. no control scheme is in place, trips with both origin and destination outside the protected region are completed at a higher rate, compared to when one of the control strategies is implemented (Fig. 5.8b or 5.8c). This is especially the case during the first hour of the simulation. However, with pre-timed signals the hotspot region becomes overly accumulated early into the simulation, causing the completion rate of trips toward the center to drop quickly.

The fixed perimeter control keeps the density of the central region below the critical point by reducing the outer-to-center flows and increasing the center-to-outer flows, initially. This regulates center-to-center and outer-to-center traveling flows, preventing the sudden drop in their completion rate (when no control is in place), which can be seen by comparing Fig.

5.8b to Fig. 5.8c. As a result, the overall performance of fixed perimeter control strategy is much better than the pre-timed control in terms of trip completion, especially up until the peak demand is over. These results are consistent with those reported in the literature [147].

The control at the fixed perimeter is determined independent from the queue lengths behind the gating and traditionally relies on the traffic dynamics inside the perimeter. Since drivers consistently try to minimize their individual travel time to destination (in a selfish manner), depending on the location of their destination with respect to different congested queues, the may end up adding to the length of a longer queue. Thus, it is not expected from the fixed perimeter control to balance the length of queues formed outside the perimeter. This is one of the main reasons that adding the dynamic perimeter to the control scheme results in the substantial improvement to the traffic performance.

Implementation of the dynamic perimeter in addition to the fixed perimeter allows for achieving a better performance by increasing the trip completion rate of the flows originated outside the hotspot region. The increase in outer-to-outer and outer-to-center flows triggers the build-up of pockets of congestion outside the protected region, which eventually hinders the traveling flows circulating the outer region and those moving toward the hotspot region. The percolation-based dynamic perimeter holds up a portion of these flows further away from the protected region and creates a buffer space containing the congestion queues formed toward the upstream of the hotspot region. Simultaneously, the dynamic perimeter increases the outflow of the buffer space allowing the inside congestion to dissipate faster. By protecting this buffer space during the peak demand period, the dynamic perimeter is able to prevent the huge drop in the trip completion rate of trips originated in the outer re-

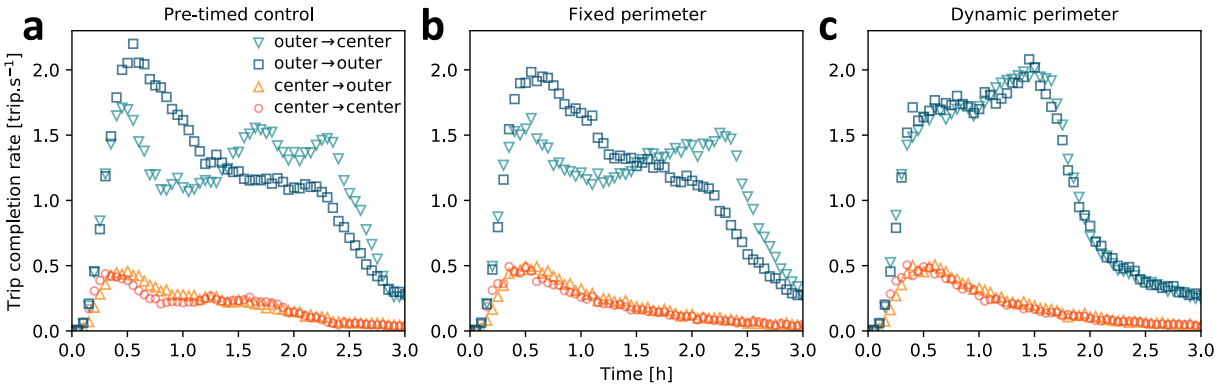


Figure 5.8: Comparison between trip completion rates of traffic flows in different directions. Here, the trip completion rate over time is shown separately for 4 different classes of traffic flows, with respect to the location of their origin and destination being within or outside the hotspot region. **a-c** The results show the trip completion rate of separated flow classes over time when the network is using pre-timed control (**a**), fixed perimeter (**b**), and dynamic perimeter (**c**).

gion (both outer-to-outer and outer-to-center flows). After the peak, the dynamic perimeter shrinks, and eventually it is deactivated following the dissipation of the congestion outside the protected region, which in effect allows the outer-to-outer and outer-to-center trips to be completed at higher rates. The overall performance of the proposed dynamic perimeter control is substantially higher than the fixed perimeter (protecting the hotspot region) alone, as it is demonstrated by Fig. 5.8c when compared to Fig. 5.8a,b.

Sensitivity analysis

The important manually set parameter in our proposed approach is the *update interval* for the time-varying perimeter control, i.e., the time interval-length between two consecutive updates of the dynamic perimeter involving identification of the percolating congested cluster and adjustment of the signals at its boundary. The sensitivity of the traffic control performance to this parameter is tested and results are displayed in Fig. 5.9. Here, the experiments are conducted using different interval-lengths to update the dynamic perimeter, from every 0.5 to every 32 minutes (frequencies of 2 to ~ 0.03 per minute). The presented results are averaged over 10 simulation runs for each particular value of update interval. For better comparison, the performance of a fixed-perimeter-only control is depicted as a baseline (dashed black line) in each graph of Fig. 5.9.

Figure 5.9a shows the cumulative trip completion over time and Fig. 5.9b shows the network accumulation over time, associated with different update intervals used by the dynamic perimeter. The time is limited between 45 to 120 min into the simulations for better visibility of the variation in the dynamic perimeter's performance when using different update frequencies. It is seen that regardless of the choice of update interval (at least within a wide range of values), an addition of the proposed time-varying perimeter improves the traffic flow in the network compared to when only a single fixed perimeter is used to protect the hotspot region (compare solid lines to the dashed line in Fig. 5.9). Also, it is seen that a good choice of update interval can emphasize the improving effect of the dynamic perimeter control. Especially, the choice of update interval is shown to have a substantial impact on how early the network is able to make the transition to the unloading phase and how quickly the network is able to unload the traffic. This is seen clearly in Fig. 5.9b with peaks occurring at different times for each curve and different slopes for the curves after they peak.

Figure 5.9c shows the sensitivity of the proposed dynamic perimeter control to the choice of update interval, in terms of average trip arrival (completion) rate during the first two hours of the simulation[6]. A first observation is that a very short update interval of 0.5 min

[6]As shown in Fig. 5.3b, later in the simulation the demand becomes very low and the network becomes

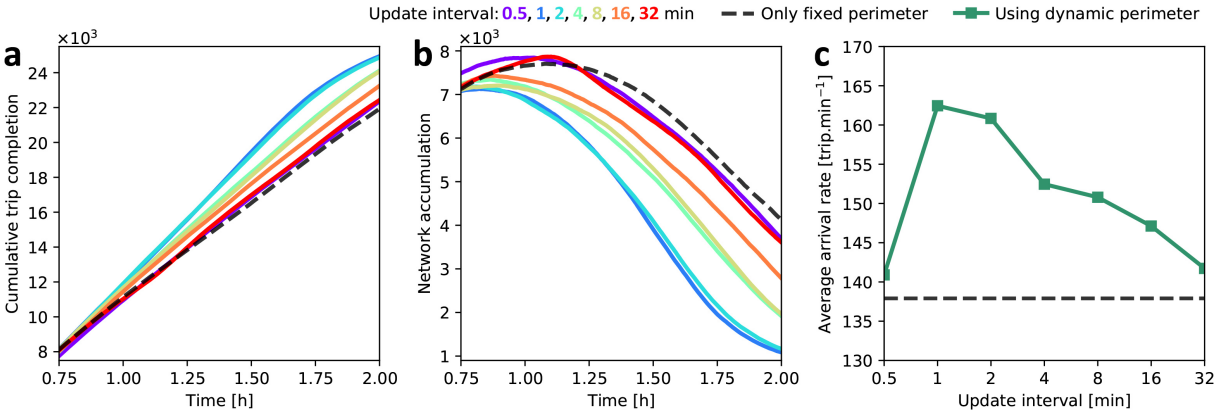


Figure 5.9: Sensitivity of the proposed percolation-based dynamic perimeter control to the choice of update interval parameter. a-c Graphs show the result of updating the dynamic perimeter with different interval-lengths between 0.5-32 minutes, in terms of cumulative trip completion over time (**a**), network accumulation over time (**b**), and average trip completion rate during two hours of the simulation (**c**). The dashed black line marks the performance of the fixed perimeter control as a baseline for comparisons.

negatively affects the performance of the dynamic perimeter control. However, increasing this slightly to 1-3 minutes leads to a substantial increase in average trip completion rate achieved by the proposed control strategy. Lowering the update frequency from this point on, results in a decreasing trip completion rate, albeit the trip completion rate will be maintained at a good level compared to using the fixed perimeter control alone. With longer choices of update interval, the controller does not respond quick enough to the change in network congestion, which adversely affects the control performance. It is seen in Fig. 5.9c that the traffic performance gradually declines with increasing the update interval above 4 min.

To apply the proposed dynamic perimeter control in real networks, one may determine an optimal or near optimal update interval using historical traffic data and by considering factors such as the network structure and the level of congestion at different times of the day. This is due to the fact that the level of congestion directly affects the rate at which the congestion propagates and different network topologies have different percolation properties [44, 49, 60, 105], which may result in the dynamic perimeter control responding differently to the choice of update interval. Nevertheless, the general pattern is expected to be the same, meaning that, it would be beneficial to update the percolation-based dynamic perimeter relatively frequent, but choosing an extremely short update interval may be avoided as it does not allow for the effect of the control at the perimeter to properly take place between the updates.

almost congestion free.

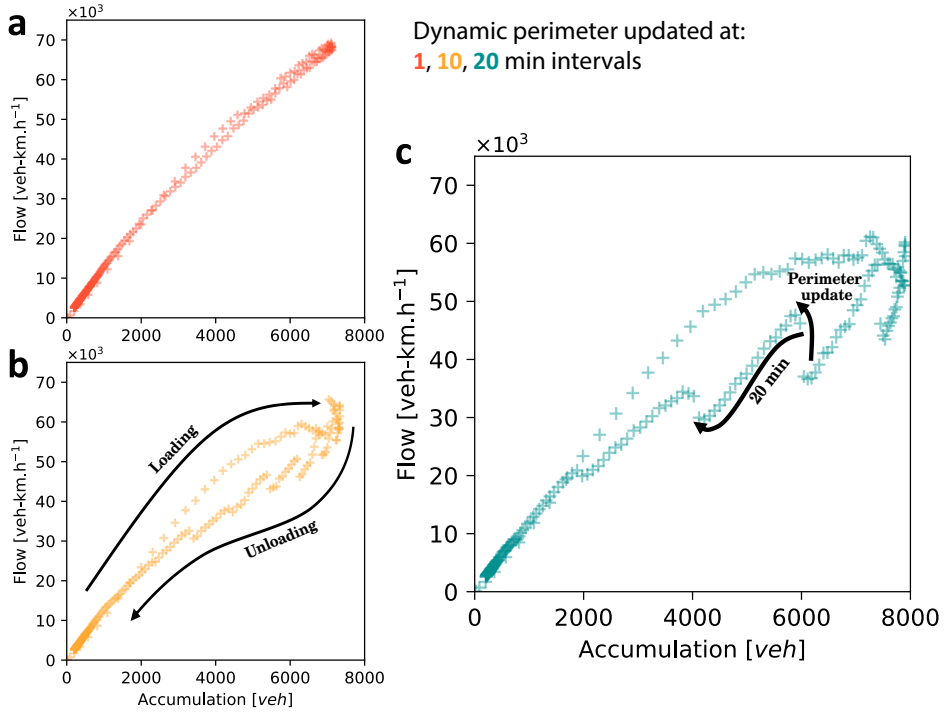


Figure 5.10: Network traffic dynamics for different perimeter update intervals. **a-c** Data points indicate the macroscopic traffic dynamics during simulations using the proposed dynamic perimeter control with the update-interval parameter set to 1 min (**a**), 10 min (**b**), and 20 min (**c**). The macroscopic traffic dynamics are demonstrated via total traffic flow movement as a function of the accumulation in the network.

By further investigations into the traffic dynamics for different update intervals, we can gain more insights regarding the effect of this parameter on the performance of the dynamic perimeter. In Fig. 5.10a-c the MFD of the network during simulations is depicted for three different choices of the update interval. Lower update frequencies for the time-varying perimeter, better reveals how the proposed control strategy enhances the network traffic performance. When the perimeter is updated by finding the percolating congested cluster, control at the new perimeter redirects traffic flows from the longer queues. Rerouting the traffic to shorter queues increases the traffic movement toward a state where congestion is more balanced in the network (i.e., queues formed at the upstream of the hotspot region have similar lengths). This is manifested in the overall flow movement bouncing back to higher values when the time-varying perimeter is updated (see Fig. 5.10c).

Comparing the MFDs shown in Fig. 5.10 (especially the unloading phase in Fig. 5.10a,b), reveals that between two consecutive updates of the dynamic perimeter, network traffic becomes more heterogeneous (this is seen from the excessive drop in flow relative to the reduced accumulation). Thus, with longer update intervals there is more time for the state of the network traffic to deteriorate, e.g., queues propagate or dissipate disproportionately, and it will be more difficult for the network to bounce back. A short update interval keeps

the evolution of queues in check and prevents increase in congestion heterogeneity as a result of the dissipation of congestion pockets. For a sufficiently short update interval, the proposed control strategy makes it possible for the network to load and unload a substantial amount of traffic with an almost negligible hysteresis in the MFD, as it is seen in Fig. 5.10a.

5.3 Concluding notes

In this chapter, we applied percolation theory to study the propagation of congestion in urban road networks. We used a percolation analysis to identify a critical congested cluster for the network's traffic and utilized the analysis in a traffic signal control strategy. The control strategy proposed here, uses a classic MFD-based fixed perimeter to protect a hotspot region in the network and the strategy is also equipped with a percolation-based perimeter which evolves over time to mitigate the congestion propagating as a result of control at the fixed perimeter. Using simulations we demonstrated that the dynamic perimeter manages to reroute traffic in a manner that boosts the traffic capacity of the network during the unloading phase. The approach proposed in this chapter analyzes a rarely studied aspect of the congestion, that is the organization of congestion at different levels over the network structures. By studying the organization of congestion we were able to characterize how it is propagating from heavily congested pockets into a congested cluster of substantial size. Although congestion dynamics at the link and network level are well exploited in signal control area, there is still a great opportunity for improving the existing or developing new control strategies by further studying different aspects of congestion in urban road networks.

Chapter 6

Conclusions

The work presented in this dissertation was an effort toward developing methodologies to study the less understood aspects of congestion in transportation networks and devising mitigating solutions to increase the efficiency of these systems. Here, the main technical chapters (Chapters 3,4, and 5) of this manuscript are summarized and implications of their results are further discussed. We also communicate the important limitations of the conducted studies and explore the potential for future works based on these studies. The chapter is divided into three sections, each corresponding to one of the technical chapters.

We applied percolation approaches to tackle two main problems: i) the conflict between congestion and passenger flows and ii) propagation of congestion in transportation networks. Although processing and mining raw transportation data is not the focus of this project, we processed large-scale detailed passenger smartcard data and used them to study the first problem (mentioned above) in real-world Public Transportation networks. To retrieve the desired information for our analysis, we developed a procedure that automates the extraction of passenger travel demand from raw PT smartcard data (see Chapter 3). The results were then used to address the first problem in real on-road PT networks (see Chapter 4). The two chapters that cover our investigations on reliability of PT networks are presented by publications, and the publications already include a discussion of limitations and future works. Nevertheless, a brief summary along with concluding remarks for each one of these chapters is provided here. The second problem addressed in Chapter 5 is summarized and concluded in the last section of the present chapter.

6.1 Smartcard data processing

To analyze the reliability of on-road PT networks against congestion on road, we used available passenger smartcard transaction data, and processed them to extract the passenger

travel demand and digitally reconstruct the enriched network representation of the real-world PT systems. In Chapter 3, a procedure was proposed to address two common challenges in travel demand extraction from smartcard data, namely, 'inferring missing alighting records' and 'identifying transfers and activities.' The existing methods often make a number of assumptions and require manual tuning of multiple parameters. This in effect can reduce the accuracy of the extracted demand. Thus, the main contribution of our study is that it eliminates the dependency of the smartcard data processing procedure to the expert knowledge of the transportation system under study or additional data sources and to develop an automated procedure with more flexibility.

To estimate the missing alighting transactions, the common approaches use rules governing the passengers' choice of the stop/location to disembark the PT vehicle. Similarly, to identify whether two consecutive passenger trips are linked by a transfer or separated by an activity, the common approaches require manual parameter setting. We mapped these tasks to classification problems. Classifiers were then trained using the available data to model passengers' behavior in the system. Using simple classifiers we modeled passengers' alighting choice and transfer/activity behavior. The first model was used to predict the passenger's alighting time-location when the transaction was missing, and the second model was used to classify the interchange events between consecutive trips into transfer and activity classes. The proposed procedure is flexible and can be applied to various smartcard data settings used in different cities.

The proposed procedure was applied to PT smartcard data from Melbourne, Australia. First, data were enhanced by inferring a substantial number of missing records. Then, we identified transfers linking consecutive trips of every single passenger. Chained single-leg trips were aggregated to generate the smartcard-based O-D demand matrix of the PT network, representing the actual passenger demand for movement between places. Elaborate evaluations regarding the estimation of missing alighting transactions estimation results showed that our approach has significantly improved the estimation performance, by substituting conventional assumptions and manual parameter setting with predictive models. The temporal OD matrix of the network as the final product of the proposed procedure reflected the expected demand profile of the PT system in terms of temporal evolution of volume and geographical direction of the demand. Overall, the results suggested that the framework is effective in enhancing PT smartcard data and is able to extract a reliable passenger O-D demand matrix for PT networks.

Future works can investigate the applications of additional data sources in improvement and building upon the methods introduced in our work. As the focus of this dissertation is not processing transportation data, we restricted our attention to one source, i.e., raw smartcard records. Yet, alternative data sources can be used to improve the quality of the smartcard data and also to rectify issues with data caused by the methods adopted in the

PT systems to record passenger data. For example, in Melbourne, it is not mandatory for passengers to perform transactions on trams in a particular region of the city (called the free tram zone), or in some PT systems passengers use both paper tickets and smartcard. These examples both show how some features, specific to a PT fare collection system, can result in a portion of transactions to be missing from the records.

A widely-used solution for such issues is using travel data sources such as household travel surveys and Automated Passenger Counts (APC) to correct the whole O-D demand matrix or the number of trips between some O-D node pairs extracted from the smartcard [166]. Apart from APC data, fare evasion reports are shown to be useful in improving the accuracy of smartcard-based O-D demand [167]. Our proposed modeling of passenger behaviors can be enhanced by the aid of alternative transportation data sources. For example, passengers' alighting behavior can be modeled more accurately by taking advantage of demographic data, or the distributions of transfer and activity duration estimated in our study can be adjusted using relevant information in travel survey data. More advanced machine learning algorithms may be able to better characterize the passengers' behavior, and thus, enhance the predictions and estimations done by our proposed procedure. Future research can also be conducted on augmenting the proposed framework in this manuscript, with probabilistic models built from historical individual (or collective) smartcard usage.

6.2 Percolation-based reliability analysis

In Chapter 4, we developed a framework to study the reliability of demand-serving networks under congestion. The focus of the study was measuring and improving the reliability of on-road PT networks, in terms of their ability to carry passenger flows via non-congested pathways between demanded O-D points. The major contribution of this study is that it theoretically extends percolation-based approaches and makes it possible to examine the organization of congestion while accounting for the passengers' heterogeneous demand for mobility. We proposed a reliability measure, α, which quantifies the ability of demand-serving networks to provide pathways of high-quality (low congestion level) links for the movement of passengers between places. The measure was applied to temporal on-road PT networks of two cities (Melbourne and Brisbane, Australia), where link qualities indicate the relative velocity of transportation which constantly varies due to actual adversarial road conditions, e.g., traffic congestion, signals, and pedestrian crowds. The measured temporal demand-serving reliability α of the network exhibited a strong daily periodicity with two distinctive patterns during weekdays and weekends.

Similar to most PT systems worldwide, on-road PT network structures of Melbourne and Brisbane are less dense during the weekends, as there are fewer PT services available compared to weekdays. However, the percolation process reveals that there is relatively more

demand to travel between places separated by congested links during weekends. Thus, network reliability α was found to be lower on weekends compared to the weekdays. In contrast, without accounting for travel demand, previous percolation-based analyses found the road networks to be more reliable during the weekends compared to weekdays [62, 111]. The contradictory conclusions arise from two major differences, one between road networks and on-road PT networks and the other between our method and the conventional analyses. Firstly, road networks have different structural characteristics compared to PT networks (e.g., bus-tram networks). Road networks represent a physical infrastructure of time-invariant nature. For PT networks, on the other hand, the structure represents the available PT services at the time and this is usually time-varying. Secondly, the proposed reliability measure α, accounts for the volumes of passenger flows from node to node on the dynamical network with actual link-level congestion. Hence, the analysis can lead to a different conclusion as the level of congestion on each link is not the only concern, and the volume of passenger flows affected by congestion on each link becomes the focus of the analysis.

Furthermore, during early morning and late evening, road networks usually show higher reliability than other non-rush hours. Considering Melbourne's daily traffic patterns, the low PT network reliability α during early mornings and late evenings arises because of the high demand in these periods for long-range movements although the availability of PT services is at its minimum. As a result, there are fewer alternative routes to choose from in these periods. Moreover, during the early morning hours in Melbourne, PT services occasionally face adversarial road conditions (early morning traffic of the commuting workforce).

We developed a theoretical framework to relate the link-level congestion to the network-level reliability, and in effect, measure the effect of congestion on each link on flow-carrying ability of the network. Thereby, we were able to analytically identify the most critical links (or bottlenecks), where the adversarial effect of congestion on the network's flow circulation is maximal. A number of interesting features were found among Melbourne's and Brisbane's most critical bottlenecks which tended to be located around urban hotspots where the large demand volume for movement is unable to cope with the impeding road conditions. Among the top overall critical bottlenecks, we found a number of links to and from major universities in Melbourne which carry an extremely large amount of passenger flows on buses and trams over weekdays. These bottlenecks were close to the central railway stations (Melbourne Central and Flinders St.) and were also in the vicinity of major hospitals. Most of these bottlenecks were not found on weekends which is consistent with the association of bottlenecks with urban hotspots and the fact that universities act as hotspots only during weekdays.

A significant number of top bottlenecks at each time during the day were associated with the CBD area where large volumes of passenger flows start or end, while the presence of

pockets of congestion is common. We also observed that four out of the top ten 'pain points' on Melbourne's road network [168] reported in the media, are in very close proximity to links among our identified top bottlenecks at morning rush hour. Since almost half of these ten pain points do not have bus or tram services in conflict with the road conditions, the results showed that our methodology does indeed work well.

Through numerical simulations, we demonstrated that amelioration of a relatively small number of identified bottlenecks can significantly improve the network, both in terms of the demand-serving reliability and the total delay imposed on passenger travel times by disruptions. Simulating the separation of PT vehicles and eliminating their conflict with road conditions on 2% of the top bottleneck links, saved close to 2,000 hours of passenger travel time during a single morning peak period (7:00–9:00 A.M.), and approximately 11,000 hours of passenger travel time over a normal weekday.

When studying networks with link-level dynamics (such as congestion) described as *link qualities*, the state-of-the-art percolation-based measures [60, 62, 88, 108] will remain the lead for understanding the topological properties of the network and even the organization of congestion in relation to the structure (i.e., network's *global quality*). However, when dealing with demand-serving networks, the proposed reliability measure α effectively accounts for the heterogeneity of node-to-node flow demand, and thus it unveils the previously obscure *global flow-quality* provided by the network against congestion. Similarly, the introduced criticality score represents a more comprehensive approach to identify the bottlenecks of such networks and proves more effective than other schemes reported in the literature.

It is worth mentioning that the direct effect of congestion organization on travel time delay cannot be studied using the existing percolation models, because the removal of a congested link simply cannot help quantify the effect of its congestion on flow travel times. But it is an intriguing problem that suggests an important direction for future research. Our proposed ideas are generally applicable to demand-serving networks including most physical infrastructures where there is an inherent demand for movement of an uneven amount of flow between different pairs of nodes in the network. With ever-increasing availability of detailed data from real-world critical infrastructure networks, our work can be a good starting point for new research avenues and the development of more sophisticated theoretical tools to analyze flow demand, which we hope it leads to achieving a more profound understanding of these complex systems.

6.3 Percolation-based traffic signal control

Chapter 5 covers our study on the second problem highlighted in this dissertation, namely, characterizing congestion propagation in transportation networks. There, we proposed the application of percolation theory in analyzing road congestion propagation and used it to develop a new traffic signal control strategy for urban road networks. Traffic signal control[7] is an essential part of urban intelligent transportation systems. The topic is well-studied in transportation engineering and the existing methods in the literature are proved effective in mitigating the congestion and improving the travel experience in urban road networks. Development and application of new tools to analyze different aspect of congestion propagation dynamics toward achieving a better understanding of congestion phenomena can still be constructive for designing new traffic control strategies. In Chapter 5, we applied percolation theory to investigate new aspects of congestion propagation, beyond the existing literature on traffic signal control. In particular, the main contribution of the presented work is applying percolation theory to characterize the spatial propagation of congested queues in road networks and designing a traffic signal control scheme that leverages this analysis to effectively prevent the propagation of congestion.

Our proposed traffic signal control strategy was developed to improve the traffic dynamics in a network with respect to a region of particular importance. Such a hotspot region can be easily recognized in real city networks by its relatively high level of trip attraction or generation. A fixed perimeter was implemented at the boundary of the hotspot region of our simulated network. Intersections at this fixed perimeter were controlled by monitoring the traffic and according to the Macroscopic Fundamental Diagram (MFD) describing the traffic dynamics within the hotspot region. A characteristic drawback of protecting a hotspot region with perimeter control is the possibility that the control causes congestion by hindering the flows at perimeter intersections.

The second component of our proposed control scheme was designed to resolve the above issue. As soon as the fixed perimeter was activated to protect the hotspot region, we monitored the traffic dynamics outside the perimeter. A percolation analysis was applied to each snapshot of the network in time to unpack the organization of different levels of congestion around the boundary of the fixed perimeter. Thereby, based on the concept of percolation criticality, a component (i.e., percolating congested cluster) containing the small pockets of congestion merging to form a large congested cluster can be identified. We controlled the traffic at the boundary of this percolating congested cluster with the aim of balancing the queues formed at the upstream of the first perimeter and mitigating the congestion within

[7] A strategy to modify the timing of traffic signals at intersections with the aim of controlling the traffic dynamics in the network.

the identified cluster. This second perimeter creates a buffer space around the hotspot region and resists the propagation of congestion as a result of control at the first perimeter. The result of our numerical simulations demonstrated the effectiveness of the proposed approach in boosting the capacity of the network by mitigating the propagation of congestion around a hotspot region of interest in the network. The outcome suggests that the proposed control can be used in city road operation and planning, especially to locally treat and improve the traffic of a hotspot zone, such as a city center or a shopping center, independent from the rest of the network.

In our simulations we introduced heterogeneity to the link dynamics and introduced complexity to the traffic dynamics through simple approaches such as considering drivers with different routing behaviors[8]. Yet, some aspects of the experiments can be delved into more, to investigate the effect of higher levels of complexity on traffic congestion dynamics and the effectiveness of percolation approaches in characterizing them. Most obvious steps are experimenting with more complicated demand scenarios and with networks having more complex topologies. Ultimately, with increasing availability of detailed empirical urban traffic data [169], the proposed dynamic perimeter control approach may be tested on road network structures extracted from real-world data to verify its strengths and identify its downsides to further extend and improve this control strategy.

As mentioned above, complex and uncertain demand is an important feature of real transportation networks. In real-world scenarios the complex demand and its variations throughout the day can easily affect the signal control performance [170]. Thus, to establish the practical value of the proposed control strategy, its robustness should be tested under more realistic simulations. The second element of more elaborate experiments is the topological complexity of the network. Percolation theory has gained its popularity in network science due to its ability to accurately characterize networks with various topological properties, thus, it is generally expected that our proposed idea works well in realistic network. However, the percolation-based signal timing scheme as proposed here, can be tested in more irregular networks to establish the method's flexibility or seek possible improvements. Generally, our efforts in this work were focused more on development of analysis tools and their integration into traffic control schemes, thus improving the signal timing methods to achieve optimality in network traffic can be a direction for future works.

[8]We divided drivers into 80% of informed and 20% of uninformed agents. Informed agents were assumed to have access to real-time condition of roads over the network and taking the path with shortest travel time between their origin-destination points, while uninformed agents choose the path that minimizes their travel distance.

6.4 Final remarks

In this dissertation, percolation approaches were applied to different problems related to congestion in transportation networks. Due to rapid urbanization, congestion has become a growing issue in urban transportation systems around the world. Although the underlying dynamics of congestion in road networks has been heavily studied, there is still room to study other aspects of congestion in transportation systems from a complex network perspective. Here, we attempted at demonstrating the ability of percolation analyses to unpack the organization of different levels of congestion on network structures and used it for different purposes aiming at improving transportation networks. We hope the work presented in this manuscript opens new avenues for future research on transportation networks and congestion phenomena using new concepts and analysis tools.

References

[1] L. M. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G. B. West, "Growth, innovation, scaling, and the pace of life in cities," *Proceedings of the National Academy of Sciences*, vol. 104, no. 17, pp. 7301–7306, 2007.

[2] M. Batty, "The size, scale, and shape of cities," *Science*, vol. 319, no. 5864, pp. 769–771, 2008.

[3] V. Snieška and I. Zykiene, "The role of infrastructure in the future city: Theoretical perspective," *Procedia-Social and Behavioral Sciences*, vol. 156, pp. 247–251, 2014.

[4] S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas," *Nature Communications*, vol. 7, no. 1, pp. 1–8, 2016.

[5] A. Cartwright, "Better growth, better cities: Rethinking and redirecting urbanisation in africa," *The New Climate Economy*, 2015.

[6] M. Treiber, A. Kesting, and C. Thiemann, "How much does traffic congestion increase fuel consumption and emissions? applying a fuel consumption model to the ngsim trajectory data," in *87th Annual Meeting of the Transportation Research Board, Washington, DC*, vol. 71, pp. 1–18, 2008.

[7] J. Currie and R. Walker, "Traffic congestion and infant health: Evidence from e-zpass," *American Economic Journal: Applied Economics*, vol. 3, no. 1, pp. 65–90, 2011.

[8] "Tomtom traffic index report." https://corporate.tomtom.com/node/26026/pdf, 2018. [Accessed 21/04/2021].

[9] "Tomtom traffic report for Melbourne, Australia." https://www.tomtom.com/en_gb/traffic-index/melbourne-traffic/, 2018. [Accessed 20/09/2020].

[10] P. Cramton, R. R. Geddes, and A. Ockenfels, "Set road charges in real time to ease traffic," 2018.

[11] E. Ferguson, *Travel demand management and public policy*. Routledge, 2018.

[12] R. Lindsney and E. Verhoef, *Traffic congestion and congestion pricing*. Emerald Group Publishing Limited, 2001.

[13] T. Gärling and S. Fujii, "Travel behavior modification: Theories, methods, and programs," *The expanding sphere of travel behaviour research*, pp. 97–128, 2009.

[14] M. Ben-Akiva, M. Cyna, and A. De Palma, "Dynamic model of peak period congestion," *Transportation Research Part B: Methodological*, vol. 18, no. 4-5, pp. 339–355, 1984.

[15] M. L. Anderson, "Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion," *American Economic Review*, vol. 104, no. 9, pp. 2763–96, 2014.

[16] J. R. Kuzmyak, "Land use and traffic congestion," tech. rep., Arizona Department of Transportation Research Center, 2012.

[17] T. Moore and P. Thorsnes, "The transportation/land use connection," tech. rep., American Planning Association, 1994.

[18] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama, "Dynamical model of traffic congestion and numerical simulation," *Physical Review E*, vol. 51, no. 2, p. 1035, 1995.

[19] T. Tsekeris and N. Geroliminis, "City size, network structure and traffic congestion," *Journal of Urban Economics*, vol. 76, pp. 1–14, 2013.

[20] A.-L. Barabási and M. Pósfai, *Network science*. Cambridge University Press, 2016.

[21] I. Scholtes, "Understanding complex systems: When big data meets network science," *it-Information Technology*, vol. 57, no. 4, pp. 252–256, 2015.

[22] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[23] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[24] L. Stone and A. Roberts, "Competitive exclusion, or species aggregation?," *Oecologia*, vol. 91, no. 3, pp. 419–424, 1992.

[25] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda, "Error and attack tolerance of complex networks," *Physica A: Statistical mechanics and its applications*, vol. 340, no. 1-3, pp. 388–394, 2004.

[26] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, no. 7291, pp. 1025–1028, 2010.

[27] L. Stone, "The google matrix controls the stability of structured ecological and biological networks," *Nature Communications*, vol. 7, p. 12857, 2016.

[28] M. Jalili and M. Perc, "Information cascades in complex networks," *Journal of Complex Networks*, vol. 5, no. 5, pp. 665–693, 2017.

[29] M. Akbarzadeh and E. Estrada, "Communicability geometry captures traffic flows in cities," *Nature Human Behaviour*, vol. 2, no. 9, pp. 645–652, 2018.

[30] L. Stone, "The feasibility and stability of large complex biological networks: a random matrix approach," *Scientific Reports*, vol. 8, no. 1, pp. 1–12, 2018.

[31] M. De Domenico and A. Baronchelli, "The fragility of decentralised trustless sociotechnical systems," *EPJ Data Science*, vol. 8, no. 1, p. 2, 2019.

[32] A. Barja, A. Martínez, A. Arenas, P. Fleurquin, J. Nin, J. J. Ramasco, and E. Tomás, "Assessing the risk of default propagation in interconnected sectoral financial networks," *EPJ Data Science*, vol. 8, no. 1, p. 32, 2019.

[33] R. Lambiotte, M. Rosvall, and I. Scholtes, "From networks to optimal higher-order models of complex systems," *Nature physics*, vol. 15, no. 4, pp. 313–320, 2019.

[34] L. Stone, D. Simberloff, and Y. Artzy-Randrup, "Network motifs and their origins," *PLoS Computational Biology*, vol. 15, no. 4, 2019.

[35] U. Alvarez-Rodriguez, F. Battiston, G. F. de Arruda, Y. Moreno, M. Perc, and V. Latora, "Evolutionary dynamics of higher-order interactions," *arXiv preprint arXiv:2001.10313*, 2020.

[36] M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, no. 1-3, pp. 1–101, 2011.

[37] R. Ding, N. Ujang, H. B. Hamid, M. S. Abd Manan, R. Li, S. S. M. Albadareen, A. Nochian, and J. Wu, "Application of complex networks theory in urban traffic network researches," *Networks and Spatial Economics*, vol. 19, no. 4, pp. 1281–1317, 2019.

[38] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[39] H. Badia, J. Argote-Cabanero, and C. F. Daganzo, "How network structure can boost and shape the demand for bus transit," *Transportation Research Part A: Policy and Practice*, vol. 103, pp. 83–94, 2017.

[40] D. Stauffer and A. Aharony, *Introduction to percolation theory*. CRC press, 2018.

[41] R. Cohen and S. Havlin, *Complex networks: structure, robustness and function*. Cambridge university press, 2010.

[42] S. Havlin and D. Ben-Avraham, "Diffusion in disordered media," *Advances in Physics*, vol. 36, no. 6, pp. 695–798, 1987.

[43] D. S. Callaway, M. E. Newman, S. H. Strogatz, and D. J. Watts, "Network robustness and fragility: Percolation on random graphs," *Physical review letters*, vol. 85, no. 25, p. 5468, 2000.

[44] M. Li, R.-R. Liu, L. Lü, M.-B. Hu, S. Xu, and Y.-C. Zhang, "Percolation on complex networks: Theory and application," *Physics Reports*, 2021.

[45] C. F. Daganzo, "Remarks on traffic flow modeling and its applications," in *Traffic and Mobility*, pp. 105–115, Springer, 1999.

[46] W. Jifeng, L. Huapu, and P. Hu, "System dynamics model of urban transportation system and its application," *Journal of Transportation Systems engineering and information technology*, vol. 8, no. 3, pp. 83–89, 2008.

[47] H. Wei, G. Zheng, V. Gayah, and Z. Li, "A survey on traffic signal control methods," *arXiv preprint arXiv:1904.08117*, 2019.

[48] H. Hamedmoghadam, H. L. Vu, M. Jalili, M. Saberi, L. Stone, and S. Hoogendoorn, "Automated extraction of origin-destination demand for public transportation from smartcard data with pattern recognition," *Transportation Research Part C: Emerging Technologies*, vol. 129, p. 103210, 2021.

[49] H. Hamedmoghadam, M. Jalili, H. L. Vu, and L. Stone, "Percolation of heterogeneous flows uncovers the bottlenecks of infrastructure networks," *Nature Communications*, vol. 12, no. 1, pp. 1–10, 2021.

[50] B. Greenshields, J. Bibbins, W. Channing, and H. Miller, "A study of traffic capacity," in *Highway research board proceedings*, vol. 1935, National Research Council (USA), Highway Research Board, 1935.

[51] M. J. Lighthill and G. B. Whitham, "On kinematic waves ii. a theory of traffic flow on long crowded roads," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 229, no. 1178, pp. 317–345, 1955.

[52] H. Greenberg, "An analysis of traffic flow," *Operations research*, vol. 7, no. 1, pp. 79–85, 1959.

[53] M. Treiber and A. Kesting, "Traffic flow dynamics," *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg*, 2013.

[54] L. Elefteriadou *et al.*, *An introduction to traffic flow theory*, vol. 84. Springer, 2014.

[55] D. Helbing, "Traffic and related self-driven many-particle systems," *Reviews of modern physics*, vol. 73, no. 4, p. 1067, 2001.

[56] C. F. Daganzo, "A behavioral theory of multi-lane traffic flow. part i: Long homogeneous freeway sections," *Transportation Research Part B: Methodological*, vol. 36, no. 2, pp. 131–158, 2002.

[57] A. Kirkley, H. Barbosa, M. Barthelemy, and G. Ghoshal, "From the betweenness centrality in street networks to structural invariants in random planar graphs," *Nature Communications*, vol. 9, no. 1, pp. 1–12, 2018.

[58] A. Lampo, J. Borge-Holthoefer, S. Gómez, and A. Solé-Ribalta, "Multiple abrupt phase transitions in urban transport congestion," *Physical Review Research*, vol. 3, no. 1, p. 013267, 2021.

[59] A. Lampo, J. Borge-Holthoefer, S. Gómez, and A. Solé-Ribalta, "Emergence of spatial transitions in urban congestion dynamics," *arXiv preprint arXiv:2103.04833*, 2021.

[60] G. Zeng, D. Li, S. Guo, L. Gao, Z. Gao, H. E. Stanley, and S. Havlin, "Switch between critical percolation modes in city traffic dynamics," *Proceedings of the National Academy of Sciences*, vol. 116, no. 1, pp. 23–28, 2019.

[61] G. Zeng, J. Gao, L. Shekhtman, S. Guo, W. Lv, J. Wu, H. Liu, O. Levy, D. Li, Z. Gao, *et al.*, "Multiple metastable network states in urban traffic," *Proceedings of the National Academy of Sciences*, vol. 117, no. 30, pp. 17528–17534, 2020.

[62] D. Li, B. Fu, Y. Wang, G. Lu, Y. Berezin, H. E. Stanley, and S. Havlin, "Percolation transition in dynamical traffic network with evolving critical bottlenecks," *Proceedings of the National Academy of Sciences*, vol. 112, no. 3, pp. 669–672, 2015.

[63] R. Arnott and A. Yan, "The two-mode problem: Second-best pricing and capacity," *Review of urban & regional development studies*, vol. 12, no. 3, pp. 170–199, 2000.

[64] Y. Xu and M. C. González, "Collective benefits in traffic during mega events via the use of information technologies," *Journal of The Royal Society Interface*, vol. 14, no. 129, p. 20161041, 2017.

[65] P. Goodwin, "The economic costs of road traffic congestion," tech. rep., UCL (University College London), The Rail Freight Group, 2004.

[66] E. J. Gonzales and C. F. Daganzo, "Morning commute with competing modes and distributed demand: user equilibrium, system optimum, and pricing," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1519–1534, 2012.

[67] C. Scheepers, G. Wendel-Vos, J. Den Broeder, E. Van Kempen, P. Van Wesemael, and A. Schuit, "Shifting from car to active transport: a systematic review of the effectiveness of interventions," *Transportation research part A: policy and practice*, vol. 70, pp. 264–280, 2014.

[68] P. J. Flory, "Molecular size distribution in three dimensional polymers. i. gelation1," *Journal of the American Chemical Society*, vol. 63, no. 11, pp. 3083–3090, 1941.

[69] W. H. Stockmayer, "Theory of molecular size distribution and gel formation in branched-chain polymers," *The Journal of chemical physics*, vol. 11, no. 2, pp. 45–55, 1943.

[70] A. Coniglio, H. E. Stanley, and W. Klein, "Site-bond correlated-percolation problem: a statistical mechanical model of polymer gelation," *Physical Review Letters*, vol. 42, no. 8, p. 518, 1979.

[71] S. R. Broadbent and J. M. Hammersley, "Percolation processes: I. crystals and mazes," in *Mathematical proceedings of the Cambridge philosophical society*, vol. 53, pp. 629–641, Cambridge University Press, 1957.

[72] M. Newman, *Networks*. Oxford university press, 2018.

[73] M. E. Newman and R. M. Ziff, "Fast monte carlo algorithm for site or bond percolation," *Physical Review E*, vol. 64, no. 1, p. 016706, 2001.

[74] A. A. Saberi, "Recent advances in percolation theory and its applications," *Physics Reports*, vol. 578, pp. 1–32, 2015.

[75] A. Allard, B. M. Althouse, S. V. Scarpino, and L. Hébert-Dufresne, "Asymmetric percolation drives a double transition in sexual contact networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 34, pp. 8969–8973, 2017.

[76] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.

[77] Z. Wang, A. Szolnoki, and M. Perc, "Percolation threshold determines the optimal population density for public cooperation," *Physical Review E*, vol. 85, no. 3, p. 037101, 2012.

[78] Z. Wang, A. Szolnoki, and M. Perc, "If players are sparse social dilemmas are too: Importance of percolation for evolution of cooperation," *Scientific reports*, vol. 2, no. 1, pp. 1–6, 2012.

[79] L. K. Gallos, H. A. Makse, and M. Sigman, "A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks," *Proceedings of the National Academy of Sciences*, vol. 109, no. 8, pp. 2825–2830, 2012.

[80] A. Goodarzinick, M. D. Niry, A. Valizadeh, and M. Perc, "Robustness of functional networks at criticality against structural defects," *Physical Review E*, vol. 98, no. 2, p. 022312, 2018.

[81] S. Porta, P. Crucitti, and V. Latora, "The network analysis of urban streets: a primal approach," *Environment and Planning B: planning and design*, vol. 33, no. 5, pp. 705–725, 2006.

[82] B. Maitra, P. Sikdar, and S. Dhingra, "Modeling congestion on urban roads and assessing level of service," *Journal of Transportation Engineering*, vol. 125, no. 6, pp. 508–514, 1999.

[83] T. J. Lomax, *Quantifying congestion*, vol. 398. Transportation Research Board, 1997.

[84] J. A. Lindley, "A methodology for quantifying urban freeway congestion," *Transportation Research Records*, vol. 1132, pp. 1–7, 1987.

[85] J. Gao, B. Barzel, and A.-L. Barabási, "Universal resilience patterns in complex networks," *Nature*, vol. 530, no. 7590, pp. 307–312, 2016.

[86] R. Cohen, K. Erez, S. Havlinl, M. Newman, A.-L. Barabási, D. J. Watts, *et al.*, "Resilience of the internet to random breakdowns," in *The Structure and Dynamics of Networks*, pp. 507–509, Princeton University Press, 2011.

[87] Y. Chen, G. Paul, R. Cohen, S. Havlin, S. P. Borgatti, F. Liljeros, and H. E. Stanley, "Percolation theory and fragmentation measures in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 378, no. 1, pp. 11–19, 2007.

[88] D. Li, Q. Zhang, E. Zio, S. Havlin, and R. Kang, "Network reliability analysis based on percolation theory," *Reliability Engineering & System Safety*, vol. 142, pp. 556–562, 2015.

[89] M. Jalili, "Error and attack tolerance of small-worldness in complex networks," *Journal of Informetrics*, vol. 5, no. 3, pp. 422–430, 2011.

[90] O. Artime and M. De Domenico, "Percolation on feature-enriched interconnected systems," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021.

[91] B. Mirzasoleiman, M. Babaei, M. Jalili, and M. Safari, "Cascaded failures in weighted networks," *Physical Review E*, vol. 84, no. 4, p. 046114, 2011.

[92] M. De Domenico, A. Solé-Ribalta, S. Gómez, and A. Arenas, "Navigability of interconnected networks under random failures," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8351–8356, 2014.

[93] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[94] S. N. Dorogovtsev, A. V. Goltsev, and J. F. Mendes, "Critical phenomena in complex networks," *Reviews of Modern Physics*, vol. 80, no. 4, p. 1275, 2008.

[95] S. Shao, X. Huang, H. E. Stanley, and S. Havlin, "Percolation of localized attack on complex networks," *New Journal of Physics*, vol. 17, no. 2, p. 023049, 2015.

[96] C. P. Stark, "An invasion percolation model of drainage network evolution," *Nature*, vol. 352, no. 6334, pp. 423–425, 1991.

[97] X. Huang, I. Vodenska, S. Havlin, and H. E. Stanley, "Cascading failures in bi-partite graphs: model for systemic risk propagation," *Scientific Reports*, vol. 3, no. 1, pp. 1–9, 2013.

[98] S. M. Rinaldi, J. P. Peerenboom, and T. K. Kelly, "Identifying, understanding, and analyzing critical infrastructure interdependencies," *IEEE Control Systems Magazine*, vol. 21, no. 6, pp. 11–25, 2001.

[99] A. Bashan, Y. Berezin, S. V. Buldyrev, and S. Havlin, "The extreme vulnerability of interdependent spatially embedded networks," *Nature Physics*, vol. 9, no. 10, pp. 667–672, 2013.

[100] D. Y. Kenett, M. Perc, and S. Boccaletti, "Networks of networks–an introduction," *Chaos, Solitons & Fractals*, vol. 80, pp. 1–6, 2015.

[101] A. A. Ganin, M. Kitsak, D. Marchese, J. M. Keisler, T. Seager, and I. Linkov, "Resilience and efficiency in transportation networks," *Science Advances*, vol. 3, no. 12, p. e1701079, 2017.

[102] V. Latora and M. Marchiori, "Vulnerability and protection of infrastructure networks," *Physical Review E*, vol. 71, no. 1, p. 015103, 2005.

[103] S. Dong, A. Mostafizi, H. Wang, J. Gao, and X. Li, "Measuring the topological robustness of transportation networks to disaster-induced failures: A percolation approach," *Journal of Infrastructure Systems*, vol. 26, no. 2, p. 04020009, 2020.

[104] A. Halu, A. Scala, A. Khiyami, and M. C. González, "Data-driven modeling of solar-powered urban microgrids," *Science Advances*, vol. 2, no. 1, p. e1500700, 2016.

[105] M. Saberi, H. Hamedmoghadam, M. Ashfaq, S. A. Hosseini, Z. Gu, S. Shafiei, D. J. Nair, V. Dixit, L. Gardner, S. T. Waller, and M. C. González, "A simple contagion process describes spreading of traffic jams in urban networks," *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020.

[106] F. Wang, D. Li, X. Xu, R. Wu, and S. Havlin, "Percolation properties in a traffic model," *EPL (Europhysics Letters)*, vol. 112, no. 3, p. 38001, 2015.

[107] P. Echenique, J. Gómez-Gardenes, and Y. Moreno, "Dynamics of jamming transitions in complex networks," *EPL (Europhysics Letters)*, vol. 71, no. 2, p. 325, 2005.

[108] Y. N. Kenett, O. Levy, D. Y. Kenett, H. E. Stanley, M. Faust, and S. Havlin, "Flexibility of thought in high creative individuals represented by percolation analysis," *Proceedings of the National Academy of Sciences*, vol. 115, no. 5, pp. 867–872, 2018.

[109] J. Borge-Holthoefer, Y. Moreno, and A. Arenas, "Topological versus dynamical robustness in a lexical network," *International Journal of Bifurcation and Chaos*, vol. 22, no. 07, p. 1250157, 2012.

[110] H. Hamedmoghadam, M. Ramezani, and M. Saberi, "Revealing latent characteristics of mobility networks with coarse-graining," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[111] L. Zhang, G. Zeng, S. Guo, D. Li, and Z. Gao, "Comparison of traffic reliability index with real traffic data," *EPJ Data Science*, vol. 6, pp. 1–15, 2017.

[112] N. Oppenheim *et al.*, *Urban travel demand modeling: from individual choices to general equilibrium.* John Wiley and Sons, 1995.

[113] L. He and M. Trépanier, "Estimating the destination of unlinked trips in transit smart card fare data," *Transportation Research Record*, vol. 2535, no. 1, pp. 97–104, 2015.

[114] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007.

[115] X.-l. Ma, Y.-h. Wang, F. Chen, and J.-f. Liu, "Transit smart card data mining for passenger origin information extraction," *Journal of Zhejiang University Science C*, vol. 13, no. 10, pp. 750–760, 2012.

[116] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014.

[117] A. Alsger, B. Assemi, M. Mesbah, and L. Ferreira, "Validating and improving public transport origin–destination estimation algorithm using smart card fare data," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 490–506, 2016.

[118] W. Wang, J. P. Attanucci, and N. H. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, pp. 131—-150, 2011.

[119] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.

[120] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[121] F. Kurauchi and J.-D. Schmöcker, *Public transport planning with smart card data*. CRC Press, 2017.

[122] E. Hussain, A. Bhaskar, and E. Chung, "Transit od matrix estimation using smartcard data: Recent developments and future research challenges," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103044, 2021.

[123] Z. Guo, "Does the pedestrian environment affect the utility of walking? a case of path choice in downtown boston," *Transportation Research Part D: Transport and Environment*, vol. 14, no. 5, pp. 343–352, 2009.

[124] N. Estgfaeller, G. Currie, and C. De Gruyter, "When less is more: Exploring tradeoffs in transit route concentration," in *Transportation Research Board (USA) Annual Meeting 2017*, pp. 1–11, Transportation Research Board, 2017.

[125] J. B. Gordon, H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in london using fare-transaction and vehicle location data," *Transportation Research Record*, vol. 2343, no. 1, pp. 17–24, 2013.

[126] N. Nassir, M. Hickman, and Z.-L. Ma, "Activity detection and transfer identification for public transit fare card data," *Transportation*, vol. 42, no. 4, pp. 683–705, 2015.

[127] F. Primerano, M. A. Taylor, L. Pitaksringkarn, and P. Tisato, "Defining and understanding trip chaining behaviour," *Transportation*, vol. 35, no. 1, pp. 55–72, 2008.

[128] T. Adler and M. Ben-Akiva, "A theoretical and empirical model of trip chaining behavior," *Transportation Research Part B: Methodological*, vol. 13, no. 3, pp. 243–257, 1979.

[129] S. Robinson, B. Narayanan, N. Toh, and F. Pereira, "Methods for pre-processing smartcard data to improve data quality," *Transportation Research Part C: Emerging Technologies*, vol. 49, pp. 43–58, 2014.

[130] T. Li, D. Sun, P. Jing, and K. Yang, "Smart card data mining of public transport destination: A literature review," *Information*, vol. 9, no. 1, p. 18, 2018.

[131] A. A. Alsger, M. Mesbah, L. Ferreira, and H. Safi, "Use of smart card fare data to estimate public transport origin–destination matrix," *Transportation Research Record*, vol. 2535, no. 1, pp. 88–96, 2015.

[132] N. Nassir, A. Khani, S. G. Lee, H. Noh, and M. Hickman, "Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system," *Transportation Research Record*, vol. 2263, no. 1, pp. 140–150, 2011.

[133] A. Alsger, A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman, "Public transport trip purpose inference using smart card fare data," *Transportation Research Part C: Emerging Technologies*, vol. 87, pp. 123–137, 2018.

[134] H. K. Lo, "A novel traffic signal control formulation," *Transportation Research Part A: Policy and Practice*, vol. 33, no. 6, pp. 433–448, 1999.

[135] P. Mirchandani and L. Head, "A real-time traffic signal control system: architecture, algorithms, and analysis," *Transportation Research Part C: Emerging Technologies*, vol. 9, no. 6, pp. 415–432, 2001.

[136] C. F. Daganzo and N. Geroliminis, "An analytical approximation for the macroscopic fundamental diagram of urban traffic," *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 771–781, 2008.

[137] T. Tsubota, A. Bhaskar, and E. Chung, "Macroscopic fundamental diagram for brisbane, australia: empirical findings on network partitioning and incident detection," *Transportation Research Record*, vol. 2421, no. 1, pp. 12–21, 2014.

[138] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 759–770, 2008.

[139] M. Keyvan-Ekbatani, A. Kouvelas, I. Papamichail, and M. Papageorgiou, "Exploiting the fundamental diagram of urban networks for feedback-based gating," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1393–1403, 2012.

[140] N. Geroliminis, J. Haddad, and M. Ramezani, "Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 348–359, 2012.

[141] J. Haddad, "Robust constrained control of uncertain macroscopic fundamental diagram networks," *Transportation Research Procedia*, vol. 7, pp. 669–688, 2015.

[142] J. Haddad, M. Ramezani, and N. Geroliminis, "Cooperative traffic control of a mixed network with two urban regions and a freeway," *Transportation Research Part B: Methodological*, vol. 54, pp. 17–36, 2013.

[143] K. Aboudolas and N. Geroliminis, "Perimeter and boundary flow control in multi-reservoir heterogeneous networks," *Transportation Research Part B: Methodological*, vol. 55, pp. 265–281, 2013.

[144] M. Ramezani, J. Haddad, and N. Geroliminis, "Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control," *Transportation Research Part B: Methodological*, vol. 74, pp. 1–19, 2015.

[145] A. Kouvelas, M. Saeedmanesh, and N. Geroliminis, "Enhancing feedback perimeter controllers for urban networks by use of online learning and data-driven adaptive optimization," in *95th Annual Meeting of the Transportation Research Board (TRB 2016)*, Transportation Research Board (TRB), 2016.

[146] N. Chiabaut, "Evaluation of a multimodal urban arterial: The passenger macroscopic fundamental diagram," *Transportation Research Part B: Methodological*, vol. 81, pp. 410–420, 2015.

[147] K. Ampountolas, N. Zheng, and N. Geroliminis, "Macroscopic modelling and robust control of bi-modal multi-region urban road networks," *Transportation Research Part B: Methodological*, vol. 104, pp. 616–637, 2017.

[148] J. Haddad, "Optimal perimeter control synthesis for two urban regions with aggregate boundary queue dynamics," *Transportation Research Part B: Methodological*, vol. 96, pp. 1–25, 2017.

[149] K. Yang, N. Zheng, and M. Menendez, "Multi-scale perimeter control approach in a connected-vehicle environment," *Transportation Research Procedia*, vol. 23, pp. 101–120, 2017.

[150] W. Ni and M. Cassidy, "City-wide traffic control: modeling impacts of cordon queues," *Transportation research part C: emerging technologies*, vol. 113, pp. 164–175, 2020.

[151] R. Mohajerpoor, M. Saberi, and M. Ramezani, "Analytical derivation of the optimal traffic signal timing: Minimizing delay variability and spillback probability for undersaturated intersections," *Transportation research part B: methodological*, vol. 119, pp. 45–68, in press.

[152] M. Keyvan-Ekbatani, X. Gao, V. V. Gayah, and V. L. Knoop, "Traffic-responsive signals combined with perimeter control: investigating the benefits," *Transportmetrica B: Transport Dynamics*, vol. 7, no. 1, pp. 1402–1425, 2019.

[153] A. Mazloumian, N. Geroliminis, and D. Helbing, "The spatial variability of vehicle densities as determinant of urban network capacity," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1928, pp. 4627–4647, 2010.

[154] Y. Ji, J. Luo, and N. Geroliminis, "Empirical observations of congestion propagation and dynamic partitioning with probe data for large-scale systems," *Transportation Research Record*, vol. 2422, no. 1, pp. 1–11, 2014.

[155] Y. Ji and N. Geroliminis, "On the spatial partitioning of urban transportation networks," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1639–1656, 2012.

[156] M. Keyvan-Ekbatani, R. C. Carlson, V. L. Knoop, and M. Papageorgiou, "Optimizing distribution of metered traffic flow in perimeter control: Queue and delay balancing approaches," *Control Engineering Practice*, vol. 110, p. 104762, 2021.

[157] M. Keyvan-Ekbatani, R. C. Carlson, V. L. Knoop, S. P. Hoogendoorn, and M. Papageorgiou, "Queuing under perimeter control: Analysis and control strategy," in *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1502–1507, IEEE, 2016.

[158] Y. Li, J. Xu, and L. Shen, "A perimeter control strategy for oversaturated network preventing queue spillback," *Procedia-Social and Behavioral Sciences*, vol. 43, pp. 418–427, 2012.

[159] L. Hébert-Dufresne and A. Allard, "Smeared phase transitions in percolation on real complex networks," *Physical Review Research*, vol. 1, no. 1, p. 013009, 2019.

[160] P. Zhang, "Spectral estimation of the percolation transition in clustered networks," *Physical Review E*, vol. 96, no. 4, p. 042303, 2017.

[161] C. F. Daganzo, "Urban gridlock: Macroscopic modeling and mitigation approaches," *Transportation Research Part B: Methodological*, vol. 41, no. 1, pp. 49–62, 2007.

[162] N. Geroliminis and C. F. Daganzo, "Macroscopic modeling of traffic in cities," in *86th Annual Meeting of the Transportation Research Board, Washington, DC*, 2007.

[163] L. Ambühl and M. Menendez, "Data fusion algorithm for macroscopic fundamental diagram estimation," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 184–197, 2016.

[164] P. Kachroo and K. Özbay, *Feedback ramp metering in intelligent transportation systems*. Springer Science & Business Media, 2003.

[165] A. Nantes, D. Ngoduy, A. Bhaskar, M. Miska, and E. Chung, "Real-time traffic state estimation in urban corridors from heterogeneous data," *Transportation Research Part C: Emerging Technologies*, vol. 66, pp. 99–118, 2016.

[166] P. Kumar, A. Khani, and Q. He, "A robust method for estimating transit passenger trajectories using automated data," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 731–747, 2018.

[167] M. A. Munizaga, A. Gschwender, and N. Gallegos, "Fare evasion correction for smartcard-based origin-destination matrices," *Transportation Research Part A: Policy and Practice*, vol. 141, pp. 307–322, 2020.

[168] "Redspot survey." https://www.redspotsurvey.com.au, 2018. [Accessed 21/07/2020].

[169] S. Respati, A. Bhaskar, and E. Chung, "Traffic data characterisation: Review and challenges," *Transportation Research Procedia*, vol. 34, pp. 131–138, 2018.

[170] C. Shirke, N. Sabar, E. Chung, and A. Bhaskar, "Metaheuristic approach for designing robust traffic signal timings to effectively serve varying traffic demand," *Journal of Intelligent Transportation Systems*, pp. 1–17, 2021.