FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Adding intelligence to a smartphone application prototype for exchanging public transport information among travelers

Luís Carlos Moreira Dias

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Teresa Galvão

Second Supervisor: António Nunes

January 20, 2014

Adding intelligence to a smartphone application prototype for exchanging public transport information among travelers

Luís Carlos Moreira Dias

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Doctor Rui Camacho
External Examiner: Penousal Machado
Supervisor: Doctor Teresa Galvão
January 20, 2014

Abstract

The increasing number of private owned cars in urban areas is contributing to the increase of urban mobility problems. Thus, the need for expanding of the public transport utilization is imminent, and one of the main decisions to be made centres in improving the traveller's experience.

At the same time, the massification of the social networks' use and the increase on the number of smartphones available is changing the way we share information and everyday the mobility of the user is less of a problem.

The increase of users in public transports connected to social networks in their journey allows them to share in real time information regarding their travel. This information can be useful to other users, giving relevant knowledge that can influence behaviour, and to the transport company managers, providing reliable data to be used on decision making.

With a better understanding of the travellers' travel patterns and the similarities with other travellers' routines, the relevance and efficiency of the information shared in a public transport can be largely improved. At the same time, the behaviour of the user can be understood with this information, providing the identification of an unique user profile.

In the context of this problem, this dissertation aims to propose the implementation of algorithms that identifies travel patterns of public transport users, which allows the sharing of information with improved efficiency and relevance. These goals are achieved by inferring travellers' destinations and predicting future travels, modelling travel patterns and creating networks of users who share similar travel routines.

The results of this work will improve an existing prototype of a *smartphone* application for the exchange of public transport related information between travellers that was developed in the context of the same project. It will allow users to automatically connect with travellers with similar routines to get and share relevant information to their current or predicted future travel.

Resumo

O crescente número de carros privados em áreas urbanas está a contribuir para o o crescimento de problemas de mobilidade urbana. Assim, a necessidade de expandir a utilização de transportes públicos é iminente, e uma das principais decisões a ser feita centra-se em melhorar a experiência de viagem do utilizador.

Ao mesmo tempo, a massificação do uso das redes sociais e o crescimento do número de *smartphones* utilizados está a mudar a forma como trocamos informação e a cada dia a mobilidade da comunicação é menos um problema.

O aumentar de utilizadores ligados a redes sociais dentro dos transportes públicos durante a sua viagem permite-lhes partilhar em tempo real informação acerca da sua viagem. Esta informação pode ser útil para outros utilizadores, dando conhecimento relevante que pode influenciar o seu comportamento, e para os administradores de empresas de transportes, providenciando informação fiável que pode ser útil para tomadas de decisão.

Percebendo melhor os padrões de viagem dos viajantes e as semelhanças com as rotinas de outros, a relevância e eficiência da informação partilhada em transportes públicos pode ser altamente melhorada. Ao mesmo tempo, o comportamento do utilizador pode ser percebido com esta informação, permitindo a identificação de um perfil de viagem.

No contexto deste problema, esta dissertação tem como alvo propor a implementação de algoritmos que identifiquem padrões de viagem em utilizadores de transportes públicos, permitindo assim a partilha de informação com melhor de eficiência e relevância. Estes objectivos são atingidos inferindo destinos de viagens dos utilizadores e prevendo futuras viagens com base nestas inferências, utilizando os seus dados de viagem, modelando assim padrões de viagem e criando redes de utilizadores cujas rotinas são semelhantes.

Estes resultados levarão ao melhoramento de um protótipo existente duma aplicação para *smartphones* existente para troca de informação relacionada com transportes publicos entre utilizadores, que foi desenvolvida previamente no contexto do mesmo projecto. Irá permitir aos utilizadores conectar-se automaticamente com outros viajantes com rotinas semelhantes, obtendo e partilhando informação relevante à sua rota actual ou prevista futura viagem.

Acknowledgements

This work wouldn't be possible without help of a lot of people I am grateful.

First of all, I would like to thank my family. Without their comprehension, support and motivation the result surely wouldn't be the same. They were who took all my frustrations in home, and I'm sure it wasn't easy to deal with at some hard times. For that, I am immensely grateful. To my dog also, for never obeying me.

For my friends, for always being there, even when I wasn't. They were the ones who helped me having strength to carry on on my course, and without them I probably wouldn't be writing this now.

To my supervisor, Prof. Teresa Galvão, for all the support, to get time on her schedule to help me in my difficult times, giving me valuable suggestions and understanding some difficulties in the management of my time due to my part-time job.

To Eng. António Nunes, my co-supervisor, for always being there to help me, giving great advices and great ideas for the development of this work. His project idea has great potential and I hope this work can help him achieving good results.

To the rest of the people in CAS for taking me and making me comfortable, giving me good moments and providing a great place to work.

To Strongstep, my contractors at the time of this work, that always were available for me and had an incredible understanding of my need of time to finish this work. Their comprehension had a great impact on the success of the final results, and I am grateful.

And to STCP and OPT, for gently providing the data used for this research and making it possible.

Luís Dias

2

"A ship in port is safe...
but that is not what ships are made for."

Grace Murray Hopper

Contents

List of Figures

List of Tables

Abbreviations

AFC	Automatic Fare Collection
CO_2	Carbon Dioxide
CSV	Comma Separated File
ETA	Estimated Time of Arrival
GPS	Global Positioning System
IR	Information Retrieval
JVM	Java Virtual Machine
NFC	Near Field Communication
PoI	Points of Interest
PTAL	Public Transport Accessibility Level
REST	Representational State Transfer
STCP	Sociedade Portuguesa de Transportes do Porto
SQL	Structured Query Language
TIP	Transportes Intermodais do Porto
USA	United States of America

Chapter 1

Introduction

1.1 Context

The massification of social networking is undoubtedly one of the main reasons for the actual evolution on communication between people. Just in the USA, 81% of adults (ages 30-49) go on-line, and 47% use social networks [LPSZ10]. At the same time, the evolution of comunication technology and devices has allowed this communication to be made anywhere. 81% of adults (ages 30-49) are wireless Internet users, and 93% of that age class have a cell phone [LPSZ10].

While we communicate easier, the mobility of people in urban areas keeps getting harder due to the number of private owned vehicles in these areas. This along the escalation in CO_2 emissions, that suggests that a 50% reduction is needed by 2050 "to avoid a 2 degrees increase in global temperatures and sea level rise" [GB13], have been one of the main reasons of the growing importance of public transports. This ultimately calls for actions to be made regarding the use of these vehicles and improving the travellers' experience should be one of the main focus to successfully promote the expansion of their use.

Communication through social networks can improve the user experience while travelling on public transports. Major transport companies are already *on-line*, offering information regarding events or vehicle schedules and are present on the main social networks providing a way for their users to exchange information on their travels. However, this information is highly scattered and in big quantities [NGCP11], and consequently finding relevant information for one specific traveller riding on one specific route in real-time is extremely difficult as users need to search for what is relevant to them.

This work appeared in the the context of a current PhD project concerning the co-creation of value in urban public transport between travellers [Nun12], and following an MSc thesis that consisted of the implementation of a prototype of a smartphone application for the exchange of public transport related information between travellers [Gon12], extending it to improve network creation intelligence and exchanged information relevance.

The data used came from Porto's public bus transport network, gently provided by STCP and OPT for investigation purposes, and its attributes quality regarding vehicle, time and location information benefit this thesis final results quality.

1.2 Motivation and Goals

Travellers who use public transports on their lives tend to develop some routines in their daily journeys. Using the potential of easy communication, public transport companies try to help their travellers providing information in social platforms like *Twitter* and *Facebook* that helps them in their travels. This kind of procedure works both ways, since company managers can use data shared by travellers in these social networks to improve their service.

During the travels, a big portion of users are connected to the Internet [Gon12] and use it, among others, to share information regarding their current travel. This information is usually sparse and hard to deliver to friends to whom that information could be valuable if delivered timely and efficiently.

Information capture regarding passenger transit information on an open system has been historically associated to an expensive task [Gor12]. Individual information regarding origin-destination pairing for one travel on a bus route could be gathered through the combination of boarding counts and traveller surveys samples [BAMH85]. However, survey response rates are shown to be declining, thus increasing the cost and losing reliability [Sto08][Sim10] and it has been studied that public transports' card systems can "provide similar OD [Origin-Destination] information at larger scales and at lower cost [PYKL08] [Gor12].

Origin-destination pairing for riders' travels can provide enough data to predict future travels from the passengers given a time interval of the day. Information about regular public transport users can thus, through this information, be clustered though similarity among historical travels or travel paths in a way that allows merging the most similar travels in networks which provide a mean to find passengers with similar travel patterns and to interchange valuable real-time information among them.

This dissertation aims to use the previously dissertation work on a *smartphone* application to share public transport information among travellers [Gon12] and improve it so that the application uses travel data to detect user travel patterns, infer previous destinations and predict future travels. Through this adaptation, it is expected that, in the future, this *smartphone* application will be able to integrate the algorithms developed in this work and be tested in the real world, with real passengers.

The scientific objectives of this work are the creation of a framework that efficiently determines users' travel destinations, allowing inference of users' travel patterns and the identification of users' travel profiles recurring to users' travel data mining, using different techniques and algorithms to infer the best solution to the problem. These techniques, along with the concept of

relevance among user travels, will allow the creation of networks of users with similar and relevant travel routines, providing automatic insertion on these networks and feeding them with more relevant and efficient information.

1.3 Document Structure

Along with this introduction, this dissertation has 4 more chapters and is organized as follows:

In chapter 2, the state of the art for the fields related to this work is defined. Contains a detailed explanation of the previous work on which this thesis is based alongside with literature about destination inference, relevance searching and major clustering algorithms and their attributes.

In chapter 3, the problem of inferring the passengers' destinations is presented, along with the methods proposed to solve it. It contains the implementation and explanation of several main decisions to estimate travel destinations, along with a proposed evaluation method based on the inference of a simulated dataset.

Chapter 4 details the implementation of the second phase of this work, the creation of temporary networks. In this chapter the concept of relevance among travellers and its measurement methods are deeply explained, and from it the creation of temporary networks is detailed and the results analysed, along with the conclusions obtained from them.

Chapter 5 presents the conclusions and the future work to be done on the context of the project [Nun12] this work relates to.

Chapter 2

State of The Art

2.1 Introduction

In the last years, due to the evolution in communication and social interaction, some work related to social networking for travellers has appeared. The application on which this work revolves is one of those works, and will be detailed in the following section 2.2.1.

The amount of Geo-tagged data existent on networks like *Flickr* or *Twitter* [FS10] provides also the possibility of detection of user patterns in multiple areas. Studies to detect patterns on travels, mobility or even points-of-interest visit in some regions have been made, and this work will present its main concepts and the possibilities of improvement to provide a good background.

Public transports that run on open systems, in which users only validate their travel when boarding, exist in many cities on the world and thus transport companies have with them big sets of information regarding those travels. Using this data, studies have been made in order to obtain passenger flows, travel behaviour or peaks of travel in certain areas. Work related to the inference of origin-destination matrices exists, providing ways to obtain destinations from data on which only boarding is described, and is analysed on this section.

This chapter presents the most relevant concepts to this thesis and the related work performed in this area.

2.2 Social networks for travelers

2.2.1 A smartphone application prototype for exchanging public transport information among travelers

To respond to the needs of improvement of user experience in public transports, a platform was created and implemented [Gon12] in the context of a PhD work by António Nunes [Nun12] [NGCP11]. This prototype service "enables collaborative exchanges on information in real-time among public transport travellers and operators" [NGG13].

Being a social network a set of "connections between groups of people" [Mit69], with the massification of Internet it became defined as a platform in which users are allowed to communicate and share information [Dic14]. These networks are, therefore, formed by groups of people who know each other in some way, or that have some kind of personal connection or even mutual interests. Hereupon, to travellers who ride public transports, this kind of connections are rarely found, since in their journeys they share the different vehicles with different people every day. Regarding this foundation, the implemented application uses the concept of temporary network [NGCP11][NGC12], being this network one with temporary connections, in which for each travel a user may have different sets of "friends", i.e, connections. This temporary attribute is defined by the current location of each user, by grouping all users currently in a given route and direction in the same network, dissimilar from users in different routes at the same time. The improvements proposed by this thesis work are directly related with this component, aiming to add intelligence to this process of temporary networks creation.

With this application the traveller can rate aspects of his current journey. After the manual check-in in the vehicle, where the user indicates the route in which the vehicle operates, he can "rate aspects of the current journey, rate recent comments provided by others in the same route, read spatially referenced comments that match his or her travel profile and intentions, check points and available rewards, see which users are in the same temporary network, and plan a journey according to his or her recorded travel profile and intentions" [NGCP11]. This information may be of different types such as seating availability, crowding, progress, punctuality, temperature, noise or driver skills.

Furthermore, there are two kinds of information: (i)structured by type and restricted to the options on the application or (ii) free content input, with the possibility of writing comments, referenced by route and type [NGCP11]. The access to this data is only available to travellers in the same network, providing effective information in a way that only users to whom that information is relevant can see it and rate it as good or bad. The effectiveness of this information is improved by a rating system, which consists in that, for a given information being shared, it must be rated by a small random set of users and validated before being shared with the remaining network elements, improving its reliability.

Travellers are encouraged to share relevant data and rate other travellers' information through a rewards model [NGCP11]. This system would "enhance the game-like nature of the proposed model of social network interaction, making it more appealing as as serious real-life game" [NGCP11]. Users that provide good information and rate others accordingly are rewarded with virtual points, based on parameters defined by transport network managers. Thus, good information and rating earns points, whereas irrelevant input incurs in a penalty. With enough points, transport managers could reward the users that provide best information with discounts or other kinds of gain for the traveller. On the other side, from a business model perspective, this information can help the same managers to improve service quality, possibly increasing revenue. With the gathered data, transport companies can cut off in satisfaction surveys, since with the rewarding system they are already "buying" relevant data from their customers.

2.3 Origin-Destination Inference

Inference represents the process of deriving conclusions based on premises assumed true. In the context of this work, these assumptions allow the inference of origins and/or destinations, when only one or none of them is previously known, for each trip segment of one user's day in the public transport system. Inferring the trip segment origin and destination becomes mandatory to reveal each user's full travel path, allowing future work on the detection of patterns among travels.

Regarding related work, Barry et al. [BNRS02] use New York metro data (MetroCard[1]) with the objective of infer destinations for each travel, since New York's is, as in Porto, an open system, i.e, one system in which the passengers are only required to register their entry on the system through their card, being allowed to navigate freely until the final destination. The main application of this work is "to describe travel patterns for service planning and to create Origin-Destination trip tables" [BNRS02] to determine volume of crowdedness on trains at peak load points, using the MetroCard's gathered data. This dataset has information regarding time and location of the sequence of trips' origins performed by each user in the system, and a set of algorithms is applied to this data to estimate each trip segment destination. As an inference, this work follows two premisses: first, "a high percentage of travellers return to the destination station of their previous trip to begin their next trip" [BNRS02] and second, "a high percentage of travellers end their last trip of the day at the station where they began their first trip of the day" [BNRS02].

Gordon [Gor12] uses London Oyster[2] *farecard* validations and iBus[3] automatic vehicle location (AVL) system to infer origin and interchange locations between trips of various public modes, constructing full origin-interchange-destination matrices. The dataset is composed of spatial coordinates relative to the travellers path. Using this data, this work calculates distances between travellers coordinates and stop locations. Gordon's [Gor12] goal is to estimate passengers' flows in the London's public transport network for each time period (early morning, morning peak, evening peak), and based on the assumption that the sample's time spans for the travels are a proxy for those of the population.

The automatic fare collection (AFC) system of the Chicago Transit Authority[4] is used by Zhao et al. [ZRW07] to develop a method to infer passengers' origin-destination for each trip. Its ticketing system runs, as in Barry et al. [BNRS02], on an open system, which means that the data gathered also provides only the boarding location for each trip. The dataset contains spatial and temporal information for each boarding validation on each card, as the route and sequence number of the boarding stop on that route. To infer the origin-destination matrix, this work follows the premise that "a high percentage of users stay at, or return to, the destination station of their previous trip segment to begin their next trip segment" [ZRW07], taking advantage of each person's consecutive trip segments. For that, three assumptions are made: there is no private transportation

[1] http://web.mta.info/metrocard/
[2] http://www.tfl.gov.uk/tickets/14836.aspx
[3] http://www.tfl.gov.uk/static/corporate/media/newscentre/archive/11573.html
[4] http://www.transitchicago.com/

mode between each consecutive trip, passengers don't walk long distances (assuming the acceptable walking distance as 1320 feet [400 meters]) and passengers end their last trip of the day at the station where they began their first trip of that day. This works tries to obtain results with the objective of demonstrate the potential of this study to replace expensive origin-destination surveys and help improving decision making.

Summarizing, there's some work done on the inference of origin-destination matrices, providing some reliable and useful information to this document's work. However, none of the literature contains zonal validation for the inferred origin and/or destination, and as detailed in chapter 3 this component provides an useful improvement on the inference final quality assurance.

2.4 Knowledge discovery from data

Regarding this work, a big set of validations data was provided by STCP public transport system for research purposes. This consisted in nearly 30 million records for the months of January, April and May of 2010, from which we could extract information regarding the vehicle internal identifier, journey start time, location and time of boarding from the passenger, among other important data.

Hereupon, since we were dealing with such amounts of data, work around discovering the relevant information from data was studied for this thesis, and the next sections present the state of the art regarding these concepts.

2.4.1 Data Pre-Processing

Data analysis is considered to be the base of investigation in many fields of knowledge [FS97]. Through this analysis, researchers can get a better understanding of the problems and improve decision making to solve them. Thus, perfecting data quality is of major importance to its reliability, and in most cases [FS97] imperfections in data are not noticed until the analysis begins.

Considering this problem, the main objectives of data pre-processing are to improve data quality and solve problems such as missing attributes, duplicates, corrupt data or data structure modification. Furthermore, this phase's results are indispensable to the final success of the work, as can be found in some work described below.

In Prasad et al. [PRA10], data pre-processing is shown to be critical to successfully extract useful and reliable information in large volumes of data from web-based organizations. The impact of data pre-processing on the analysis and optimization of web-based educational systems and its learning content is analysed in Sato et al. [SMS+11], comparing the extracted results. Alcalá-Fdez et al. [AFSG+08] presents KEEL, a software tool specialized in Data Mining problems along with the integration of different pre-processing techniques. The importance of this preparation is discussed as one of the main contributions to the algorithms results and it is shown as a complex component of knowledge discovery from data. Gordon [Gor12] works on the the estimation of origin and destination of user travels based on spatial data from the users' cards. This data is referred as propitious to invalid information such as unfinished entries, duplicate transactions and

corrupt travel segments, and requires special handling of data, making data pre-processing one of the major factor of success for the inference.

Summarizing, when facing large datasets the impact of data quality is severe. Thus, pre-processing and elimination/transformation of corrupt, incomplete or unstructured data reveals itself as one of the most important components to ensure good results. Because this work uses large volumes of data from travellers' validations, pre-processing this information is the first step towards the proposed solution.

2.4.2 Data Mining

Data mining is "the entire process of applying computer-based methodologies", through either automatic or manual methods, with the main goal of "knowledge discovery" from "large collections of data" [Kan11]. This knowledge if often hidden in the data, and "to act on that knowledge is becoming increasingly important in today's competitive world" [Kan11], as we live in data-driven times, in which data mining" is a fast growing application area in business" [BM01].

The process of data mining tends to have two main objectives:

- Prediction: used to predict unknown of future values of variables of interest. *Predictive mining* produces a model for the system based on the given data set, and this model can be used to classification, prediction or estimation, among other tasks.

- Description: used to find patterns described in the data, in a way that can be interpreted by humans. *Descriptive mining* produces new, non-trivial information based on the given set. This information uncovers patterns and relationships in the data set, with the goal of deeply understand the system.

The goals mentioned above are achieved using different data mining techniques. These are applied, as showed in [Kan11], [Jia06], [MH09], for the following tasks:

- Classification and Regression: Given a range of predefined classes, discovers a predictive function classifying data into on of the classes (classification) or mapping the data to a real-value prediction variable (regression).

- Clustering: Technique that groups similar objects in the same sub-group. Different objects belong to different sub-groups.

- Summarization: Finds and discriminates a compact description for characteristics or features of subsets of data.

- Dependency Modelling: Models dependencies between features in a data set, predicting values of some attributes based on another.

- Change and Deviation Detection: Discovers items that exhibit unexpected changes and deviations in a data set.

2.4.3 Clustering

Clustering is "the process of grouping a set of physical or abstract objects into classes of similar objects" [Jia06]. This groups of data have the name of cluster, and a single cluster is formed by objects "that are similar to one another" and at the same time "are dissimilar to the objects in other clusters" [Jia06].

Regarding the proposed problem, with the need to find patterns from unlabelled data such as the location of the travellers' journeys, clustering becomes one approach to analyse. Grouping of similar patterns among travellers could answer the need of creating networks of similar travels, provided a data set of travel intentions. Furthermore, the techniques of spatial and temporal clustering, as described ahead in this document, could provide algorithms capable of geo-spatial aggregations.

There is some work that can be analysed regarding the finding of travel patterns using clustering methods. Xiaolei et al.[MWW+13] uses density-based clustering to identify travel patterns, along with partition based clustering to classify travel regularities. Travellers' information is obtained from smart card data collected, which contains spatio-temporal characteristics of each trip chain. Based on these chains, this work applies clustering algorithms along with rough-set theory to cluster and classify travel pattern regularities.

A density-based approach is also used by Zheng et al.[ZLZC11] to discover tourist regions of attraction based on photography location patterns. Photos from Flickr have temporal and geographical meta-data, and so this geo-tagged data is used to find photographers' spatio-temporal paths. Using a dataset with geo-tagged photos, with the aim to find people's travel patterns within a local tour destination, results in a statistical dataset of peoples' trails to identify regions of attraction.

Lee [LCL13] also uses geo-tagged photographs from Flickr and its spatio-temporal characteristics to investigate association between points-of-interest(PoI) through the analysis of peoples' travel paths from one PoI to another, resulting in a study that reveals PoI that are frequently visited along with other PoI. This work is also based on the use of a density based method.

Summarizing, the literature suggests that the main approach to find spatio-temporal pattens is related to partitioning and density based algorithms. Thus, in the following section we will describe the most relevant types of cluster analysis used in the literature regarding travel patterns and travel prediction, along with a detailed examination and description of the most relevant algorithms.

Partitioning Clustering

Partitioning algorithms organize a given data set into a previously chosen number of uniform clusters. The partitioning method creates an initial partitioning, and then iteratively "attempts to improve this partition by moving objects from one group to another" [Jia06]. After the last iteration, each one of these clusters is circular and its data is similar inside the cluster and dissimilar

to another clusters. This similarity is calculated based on a distance function and the cluster is formed to optimize a partition criteria.

The most commonly used and well-known algorithm for this method is k-means. [Jia06]. "The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low" [Jia06]. This method defines a centroid or center of gravity as a mean value of the objects in a cluster, and thus comparing the distance of each point to the centroid, assigns it to the most similar cluster based on that distance. This process iterates until the criterion function converges, and the calculus of the variation inside the cluster to calculate cluster quality in each iteration is defined by a *square-error criterion*, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i}^{k} |p - m_i|^2 \tag{2.1}$$

Above, E is the sum of the square error for the whole data set; p is an object of cluster C_i. m_i is the mean of the cluster, therefore "for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed" [Jia06]. This results in k clusters, each as inner compact and separated from each other as possible.

Density-Based Clustering

Clustering methods based on density are those that search areas of higher density, and define a cluster as being the objects on those areas. To separate clusters, these methods search for regions of low density. Thus, the algorithms that follow these methods can be used to filter noise (outliers), allowing the modelling of clusters of arbitrary shape, being suitable to problems where the limit to circular clusters can be a constraint.

DBSCAN [SjRkG13] is the most well-known and used density-based methods "that grow clusters according to a density-based connectivity analysis" [Jia06]. This method is based on data intensity. It searches for areas of high density, and makes clusters of objects of those areas. This search is done with two parameters, a given area (*Eps*) of a ε-neighborhood and the minimum of points contained in the ε-neighborhood (*MinPts*) [LCL13]. If the ε-neighborhood of an object contains at least *MinPts*, then this object is named core object [Jia06] and it's part of the cluster.

A point p_1 is *directly density-reachable* from a point p_2 if it is within the *Eps* area of the ε-neighborhood and if *Eps* has at least *MinPts*. If there is a sequence p_1, p_2... p_n, where p_1=p and p_n=q, and each p_{i+1} is directly density-reachable from p_i, then p is *density-reachable*. A set of density-reachable object forms a cluster.

If the input *Eps* and *MinPts* are appropriately defined, this algorithm is effective finding clusters of arbitrary-shape. The computational complexity of DBSCAN is $O(n^2)$, where n is the number of database objects. Best results are achieved with the use of a spatial index, where it performs

with *O(n log n)*.

2.5 Relevance in information retrieval

Information retrieval (IR) is an interdisciplinary research field that includes, among others, relevance rankings, search engines and evaluation measures. IR systems are the "predecessors of Web and search engines" [ZN14], designed to retrieve documents and digital collections. In IR, similarity is used to measure semantic and syntactic similarity, comparing meanings or syntax [JRK13], and its ranking algorithms are used to obtain "high-recall documents" [ZN14]. These measures constitute a classical approach [JRK13] to information retrieval, being applied for the course of many years with increase in the spatial information (geographic) domain [SCH08].

Information retrieval is "about computing the degree of relevance between a set of objects and the search parameters" [JRK13], and its major appliance is related to web search engines [ZN14] using user-specified keywords and returning lists of web pages sorted by relevance to the user query [ZN14]. This information can be entered directly, through keywords, or inferred from implicit data and used for the relevance rankings [SCH08].

These concepts can be aligned with this work's problem. The degree of relevance could be measured between travel validations provided from STCP for this research, given that the full travel path could be inferred to be able to match against the ones from the remaining travellers. Through the bus network data collected in the context of this project [Nun12], with data from bus routes and bus stops present on the bus network, and taking advantage of the attributes on the validations data, like boarding location and number of zones allowed to travel with that card, the full journey of the passenger can be estimated and an algorithm to measure the relevance between the travels can thus be created. This way, the concept of relevance could rank each travel path, matching against the others and score through relevance measures, resulting in a list of high-recall passenger travels for each computed travel path, i.e, the "input-query" of the relevance scoring.

The adaptation of the information retrieval relevance ranking concept to our work and the detailing of its main features and attributes are presented further on this document, on chapter 4.

2.6 Summary and Conclusions

Data Pre-Processing is of major importance to the final efficiency of the results obtained, and its methods application needs to be carefully studied according to the work's dataset.

Trip Inference is the subject of inferring, for each trip segment of one traveller, its origin or destination (or both if both needed). This process, as an inference, often relies on premisses assumed true to obtain effective results. This work's data is entry-only, thus to estimate trip destinations, some assumptions had to be made based on the literature. This process will be detailed further down on this document in chapter 3.

The process of clustering, i.e, grouping of similar data within the same cluster, and at the same time dissimilar from data in the other clusters has wide applications, and can be used as a stand-alone data mining tool or to pre-process data for other algorithms.

Clustering methods used to travel patterning found in literature can be classified as partitioning or density-base. Some may belong to more than of of these categories.

Partitioning methods divide a set in k partitions, and iterates to improve partitioning moving objects among groups.

Density-based methods groups data based on density, and creates clusters according to the density of its neighborhood.

The problem of pattern identification and the need to find similarities between the patterns of travellers was studied and considered as one where clustering techniques would be used. To use these techniques, a distance function is used in order to define the similarity between the different objects in the dataset, to result in the final clusters. In particular, to cluster based on geographical distance - through DBSCAN - this distance signified the distance between the location of two points.

Furthermore, using this clustering technique for the two columns "origin location" and "destination location" would answer the question "Which travellers board and arrive in similar locations?", and would need two distance functions, one for each column (origin and destination).

This approach involves two main issues. First, the answer to the previous question does not provide the wanted relation between travellers full paths, but only clusters from each passenger boarding and destination locations. This happens mainly because through clustering the main goal is to group a set of points, whereas in our problem we have a sequence of points that result in a path for each traveller.

Second, each distance function results in the calculation of a dissimilarity matrix, that consists on the set of distances between the observations on the dataset that is used to find the most similar (less distant) objects to create one cluster. The size of these matrices is the number of records being clustered, and in our work we have sets of millions of records, with thousands for each day of the month. Computational and memory wise, computing this matrices would result in huge amounts of memory being used and thus not realistically possible for our dataset.

These two main problems resulted in an approach to the problem that would not be directly related with the clustering algorithms. Along with the similarities found in the concept of relevance ranking on the Web search engines, a new concept was decided as the way to solve our problem: relevance among user travels. This concept is detailed in chapter 4.

Chapter 3

Travel path inference

This chapter presents in detail the first stage of development on this thesis, the inference of travel paths from riders validations in an open system.

The next sections provide deeper analysis on the needed data and its essential attributes, data pre-processing and transformation needed to guarantee data quality and the proposed solution along with its main decisions.

In the last section, the obtained results are analysed and a evaluation solution is proposed, based on a simulated dataset with previously known real destinations and a comparison with the ones obtained through the destination inference procedure provided.

3.1 Introduction

Being an open system, Porto public transport system Andante requires validation only at entrance on the system, meaning that riders board vehicles through the validation of their cards and then leave at their desired destination. Consequently, information regarding passengers travel path is found incomplete, since we don't have the destination, and thus obtaining similar travel patterns becomes a task not doable in these conditions.

Regarding this problem, a solution is proposed on the next sections. Given a day, the data from Andante system for bus riders validations is able to provide all the travels for each traveller, for that day. Knowing this, the goal is to obtain an origin-destination matrix that will provide, for each travel validation found in the training dataset, its estimated destination.

To achieve this, an algorithm is proposed based on [BNRS02], with some innovative steps developed in the context of the project this work relates to [Nun12], and implemented as part of this thesis, such as zonal verification, and other additional work not found in the literature [NDGC14a]. Through the proposed solution the goal of obtaining riders' travel paths is achieved and therefore the possibility of finding similar travel paths, not doable before, becomes a possible task.

The next section provides detailed explanation on the architecture of the proposed solution to the travel path inference.

3.2 Architecture

Three major components are part of the travel path inference, seen in figure 3.1: the datasets with validations and Andante network data, data pre-processing and the origin-destination matrix inference algorithm.

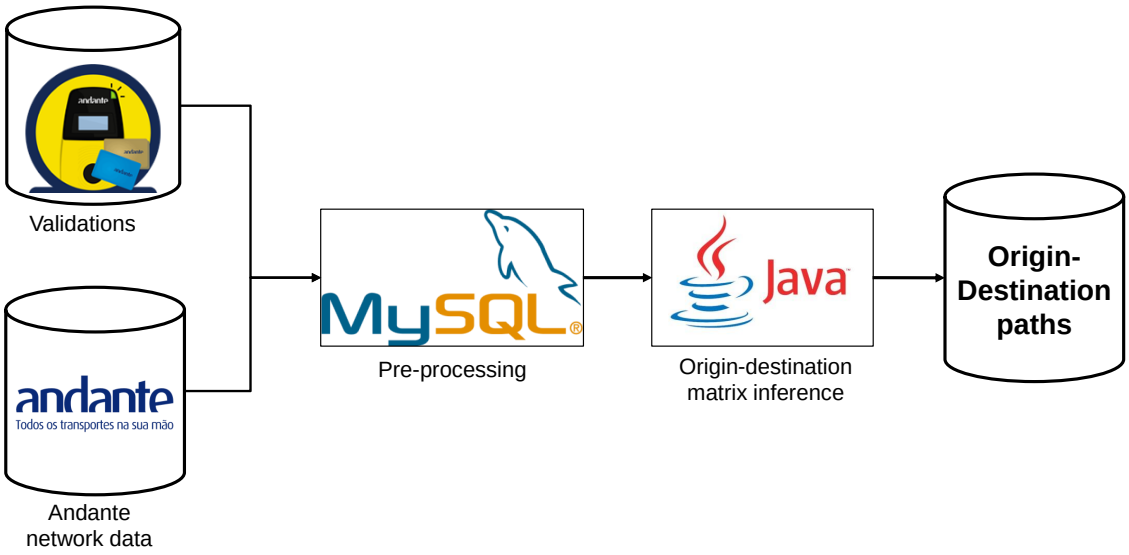

Figure 3.1: Travel inference architecture

The datasets of validations and Andante network are part of a MySQL[1] local server, used in this work to feed the algorithms. The database used was MySQL since, in addition to being an open source technology, has great scaling capabilities and is able to support stored procedures.

To pre-process the data MySQL procedures were created, allowing transformation of the datasets. Using SQL allowed working with the database in a more efficient way in order to improve the following implementation of the origin-destination inference algorithm. This algorithm, developed to infer each travel destination, was implemented using the Java programming language. This choice was mainly influenced by its class-based and object-oriented structure and since it runs in any Java Virtual Machine (JVM), it is cross-platform and it compiles regardless of the computer architecture, making it very flexible. Using Java, the algorithm can use its capabilities to collect and store in memory all the data from the database as objects, speeding the process comparing to the execution time it would take using a procedure language through SQL.

[1] http://www.mysql.com/

The final result is a new dataset with mandatory validation attributes and the inferred destination as a new one. To store the inferred set of travels the same local MySQL database was used.

The next two sections provides detailed description for the datasets used and each of its essential attributes in order to allow the reader to be able to have a better understanding of the proposed methods, along with the necessary data transformation procedures to guarantee incorruptions on the identified attributes.

3.3 Data description

Each time one traveller validates his card, a series of data is registered and recorded in the STCP system's database. This is data containing information regarding a set of components, many of them needed in the context of this work. In order to use this data for research purposes, STCP gently provided a dataset of nearly 30 million ticketing validations, each record obtained in real-life travels with the needed attributes.

Below are listed the main attributes extracted from each of the dataset's records, from vehicle data to card information, along with a succinct explanation of its role in our work:

- Vehicle number: serial number of the vehicle, one of the attributes necessary to identify one full journey along with journey start time;
- Journey start time: starting time of the vehicle's current journey, one of the attributes necessary to identify one full journey along with the vehicle's number;
- Serial Number: card's unique serial number, responsible to identify anonymously each passenger in the system;
- Ticketing type: defines what kind of ticketing this card follows. At the time when these validations were registered in the system, there was more than one ticketing system in service. In this work, only the Andante zonal system was analysed.
- Number of zones: gives how many zones this card's current travel is allowed to run. As described in appendix A, the number of zones can be 2 (Z2), 3 (Z3), until a maximum of 9 (Z9) zones [STC14];
- Route: vehicle's current route, composed by the current line and direction, necessary to identify the current travel;
- Start stop: stop where the traveller validated his card, composing the current travel along with route;
- Validation time: time of validation at the start stop, provides data necessary to infer destination time;

- Zone: zone of the current validation, i.e, zone where the start stop belongs. Used to validate the destination stop zone;

All this data concerns users' validations and is of critical importance for this work. Furthermore, information on the routes and stops of the Andante network has already been collected in the scope of this project [Nun12] and was used to map with the validations' data, namely:

- Lines: Contains the total of 85 lines of the Andante network and its unique identifiers;
- Stops: Set of 2534 stops in the Andante network, including bus, metro and train, with information regarding its name, location, unique code and Andante zone.
- Route stops: Mapping of all the stops with the lines to which they belong, with the sequence of each stop, allowing to obtain the ordered path of bus stops for each route. One route is defined by the line name and its direction, with at most two directions per line. Contains 5664 records, including bus, metro and train.

The above data is mandatory to get results from the implemented algorithms. Furthermore, data with corruption in any of these attributes or irrelevant information like data regarding metro and train systems, provided imperfections on future stages of work. The following section analyses deeper this constraints.

3.4 Data pre-processing

From the nearly 30 million records provided by STCP for this research, from this stage onwards we will only consider data from January, mainly due to the high amount of data.

At the time the data was collected, Andante was not the only mandatory card system at STCP. Because the proposed algorithm points towards the system running nowadays, early data analysis and comparison was mandatory - data from other systems couldn't be verified by the proposed zoning verification step on the inference algorithm, and so imperfect data would lead to imperfect results. This way, a study was made for the validations of January - seen in appendix A - in order to be able to assure that using only the Andante records would still provide enough quality data to feed the algorithms. From this study, the conclusion was that 35% of January's travels belonged to bus riders using the Andante ticketing system and those validations provided good quality data. Thus, validations from other ticketing systems were discarded as they were not relevant to this study.

The Andante sample needed proper examination and pre-processing to find any possible flaws in the validations that could lead to incorrect results and conclusions. One of the factors of greatest concern was related with the volatility of the routes and stops at the STCP system, since routes paths and/or stops from those lines are subject to frequent changes. This impacts our work because the collected data dates from 2010, while the sets of routes and stops were gathered at the time of this work, in 2013, resulting in possible inconsistencies among the data.

Therefore, a pre-processing procedure was implemented, as seen on the diagram on figure 3.2, aiming to efficiently process the data in a way that, in the end, the resulting validations would be the most correct for the algorithm training. The entire process is now described and its result on the validations dataset can be seen on figure 3.3.

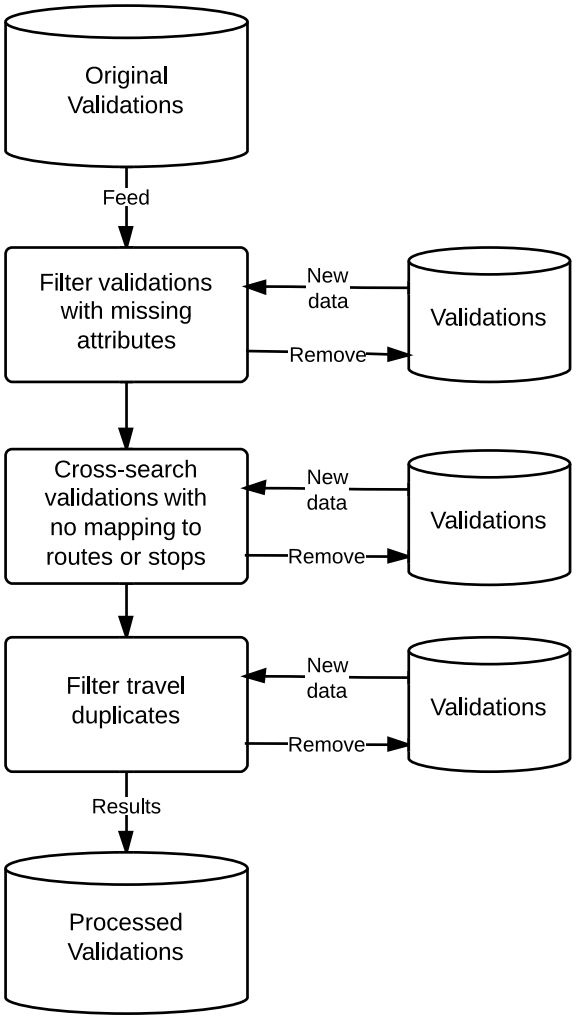

Figure 3.2: Pre-processing procedure flowchart

The first step of the dataset processing passed by the verification of missing attributes. After reviewing the dataset, the key fields were defined - as seen on 3.3 - and so records with any of those attributes missing or corrupted were discarded from further processing. At the end of this step, 3% of January's validations were discarded.

After these verifications the datasets were cleaned of imperfect and missing data. Afterwards, two more situations were identified on the validations set: unmapped bus stops and routes. This could happen because, as explained before, the STCP system is quite volatile, meaning that on a short time interval routes could change its course, bus stops could be added to another route or removed from previous ones or even bus lines or stops set aside from the network. This happens due to network updates or because of temporary occurrences like road construction/maintenance or

other short/medium-time impediments on the roads that happen frequently in a city. This results in outdated datasets from one year to another (or even in shorter spans), and in this work the difference is of major importance - from 2010 to 2013. To achieve this verification, a cross search between the 2010 validations set from January and the 2013 data for bus lines and bus stops was performed. Validations whose stop or line were no longer found in the network were thus discarded. This stage cleaned 18% of incorrect records relatively to the original dataset.

Finally, when all the verifications defined to clean the data were performed, the last data needing filtering was related to duplicate validations. Duplicates are cases in which users validate more than once while in the same vehicle in the same journey. One of these situations is related to the time of travel, since when validating the card the machine presents the time left, and so the passengers sometimes validate more than one time to check on the remaining time. This way, was decided to previously remove that data in the pre-processing stage since these records were providing incorrect results. Although the result was a total of only 0,3% removal on the January set, this records provided unnecessary complications for the origin-destination algorithm.

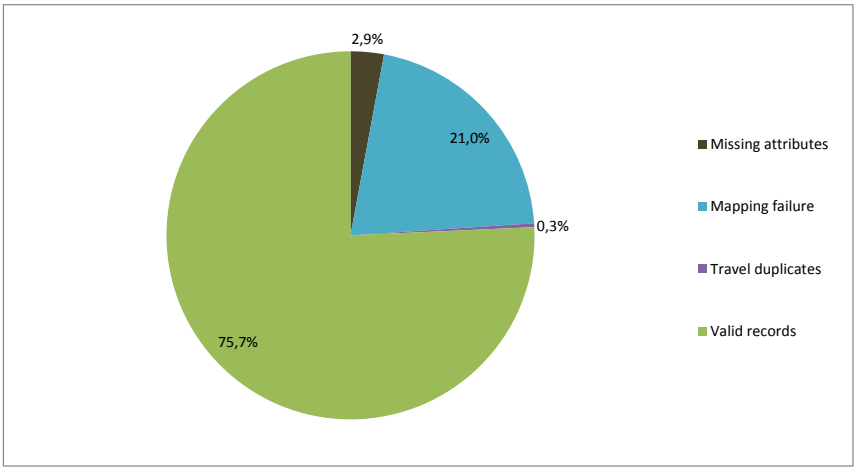

Figure 3.3: Pre-processing results for the validations dataset

Now that the data was pre-processed and cleaned of record imperfections and potential misinterpretations, the final result was a total of around 2.4 million validations on the January month, supported by 74 bus lines and its 2120 bus stops.

3.5 Origin-Destination matrix inference

3.5.1 Class diagram

To implement the algorithm to infer travellers' destinations, a set of necessary classes were implemented in order to follow the object-oriented programming defined as the approach to this problem. Using Java's class-based and object-oriented capabilities, and recurring to the data from the datasets to create the objects, the classes seen on 3.4 formed the main part of the program.

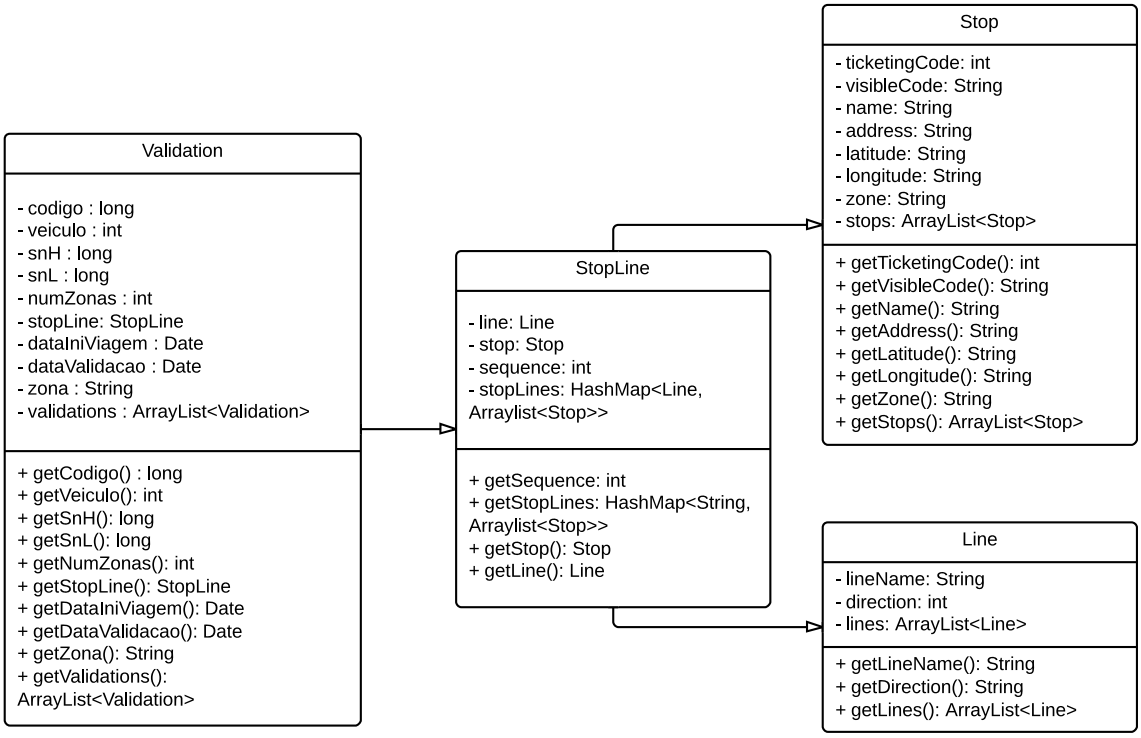

Figure 3.4: Origin-Destination inference class diagram

Four major classes were created: Validation, Line, Stop and StopLine. Follows a brief description for each of them:

- **Validation**: A Validation object was created for each traveller, with its mandatory members and the respective *getters*. An *array* of validations is created to hold all the validations in memory, and through *getStopLine()* we obtain the respective StopLine from the *linhaPublico*, *sentido* and *paragem* fields originally on the validation dataset.

- **StopLine**: Because each line may have multiple stops and each stop is assigned to multiple lines, this class helps to map them to one another and define the sequence of the stop on that route. An HashMap is used to hold all the "StopLines" in memory, mapping for each line its respective array of stops.

- **Stop**: Each stop from the bus network is loaded to memory creating an object with its *ticketingCode*, that maps to the code system present on the Validation, its *visibleCode* which defines the "known" code on the Andante network, the zone in which it is present and its location coordinates.

- **Line**: Each line object contains the line name and direction, being that each line may have at most two directions (some have only one as they are circular).

Furthermore, to help in some methods two other classes were created, *MySQLConnection* and *GenerateCSV*. The first one was responsible for reading data from the database and write it to the

several *Arrays* in memory. Recurring to the MySQL packages from Java, this class executed all the SQL queries and thus was essential to create the all the objects for the validations, lines and stops.

The second one was needed to write the final set of inferred validations to a CSV file to import later to an MySQL table. This decision was made due to the high number of transactions that would be made while computing the algorithm if the travels were immediately inserted in the database as soon as computed, slowing the process. Writing to an CSV in the hard drive beforehand became a more efficient way to handle the high amount of data.

3.5.2 Nested loop structure

To run through all the validations in the dataset, this nested loop structure was adapted from a study by Zhao et al. [ZRW07] in the context of this project, and was used to aid the implementation of the algorithm. For each serial number (SN), for each day, the algorithm should be able to infer all its validations, one at a time. To that purpose, three nested loops were determined: the outer loop cycles through each SN, the middle loop through each day and the inner loop through each validation (flowchart in figure 3.5). In other words, for each SN, for each day, we can infer all its validations, one at a time.

For this principle, the entire validations dataset was sorted with two main keys: using the SN as primary key and the day as secondary key. Furthermore, all the validations were sorted by date of validation, resulting in a dataset of validations in which for each serial number, and for each day, the validations are sorted from the beginning to the end of the day, providing the needed daily travel history. This sorting became mandatory because of the *consecutive trip segments* method used in the algorithm. This method has as base idea that "a high percentage stay at, or return to, the destination station of their previous trip segment to begin their next trip segment" [ZRW07].

Following this, four assumptions were established, adapted from [BNRS02], in order to take into account the specificities of this work:

- There no other transport type (public or private), except bus, between trip segments;
- There is a limited amount of distance that can be made by passengers on foot, i.e, after leaving one bus, the passenger will not walk a too long distance to the next station (trip segment). In this work, it was assumed as 640 meters, or 8 minutes on foot at 4,8 km/h, speed used to calculate the public transport accessibility level (PTAL) [fL10]. This distance is considered as the maximum walking distance for bus stops in the Great London and is used as a reference for this work;
- The best estimate destination for one trip segment is the stop closest to that passenger's next origin (if within walking distance);
- The last trip of the day has as destination the origin of the first trip of the day. In other words, the assumption is that at the end of the passenger's day, his last trip is the return to where his day began.

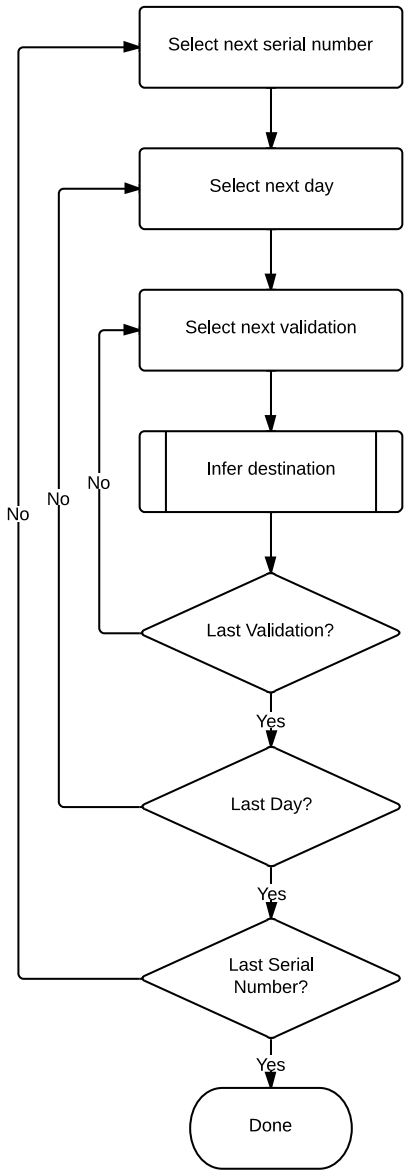

Figure 3.5: Nested loop structure adapted from Zhao et al. [ZRW07]

3.5.3 Distance-to-Stop matrix pre-calculation

The destination inference procedure studied in this work requires knowledge of the closest stop to some identified target location. Each stop is identified mainly by its unique network identifier and from the geo-spatial coordinates (latitude and longitude). Calculating this closest stop for each of the validations can become costly, and if this distance calculation would be computed for each of these validations the computation would not be nearly efficient [Gor12].

Rather than doing this for each validation, distance between stops are calculated only once - from each stop to all stops. If the calculated distance is superior to the maximum on-foot distance (640 meters), it is discarded. This way, previously identified non-selectable stops won't delay the algorithm computation. The valid distances are then sorted in ascending order, optimizing the

```
preComputeDistanceBetweenStops(ArrayList<Stops> stops):
// stops: stops read from the database
for each stop s1 in stops:

for each stop s2 in stops:

if s1 == s2 continue;

distance = GetDistance(s1.latitude, s1.longitude ,s2.latitude, s2.longitude)
// distance gets the geographic distance between the two stops coordinates

if distance < MAX_DISTANCE_ON_FOOT
// MAX_DISTANCE_ON_FOOT: distance walkable on foot = 640

if distanceBetweenStops.contains(s2, (s1, distance) //distanceBetweenStops: matrix
    with all the walkable distances between all the stops
  continue; //do nothing because the map already has the distance between the two
      of them
else
  distanceBetweenStops.put(s1, (s2, distance));

distanceBetweenStops.sort();
//after inserting all the distances, for each one it is sorted in ascending order
    to provide the closer ones on top of the list
```

Listing 3.1: Function that creates the map of distances between stops

future search for the closest stop. Since the distance between A and B is the same between B and A, only one distance between one pair of stops is stored in memory, creating a matrix stored in memory for later reference. The *pseudo-code* can be seen on Listing 3.1.

3.5.4 Destination inference

Once all the validations, stops, lines and zones are loaded to each respective map in memory, the algorithm follows the pattern described before: for each SN and for each day, cycles through all its validations. The flowchart in figure 3.6 below describes the flow of the entire process. The architecture of the algorithm depicted in the flowchart was adapted from Gordon [Gor12], to which some validation steps were designed and added in the context of this project [Nun12][NDGC14a], and implemented as part this thesis. The aim of the aforementioned additional validation steps is to take advantage of the rich set of attributes of the dataset to give greater confidence to the destination inference results.

If one passenger has only one validation in a day, then the algorithm has no reference to the possible location where he may have left the bus. This way, every validation that is the unique one for that person for that day is immediately discarded. Here, it is important to clarify that one day in the bus network circulation is the time interval between 05:30 of the current day and 1:30 of the next day.

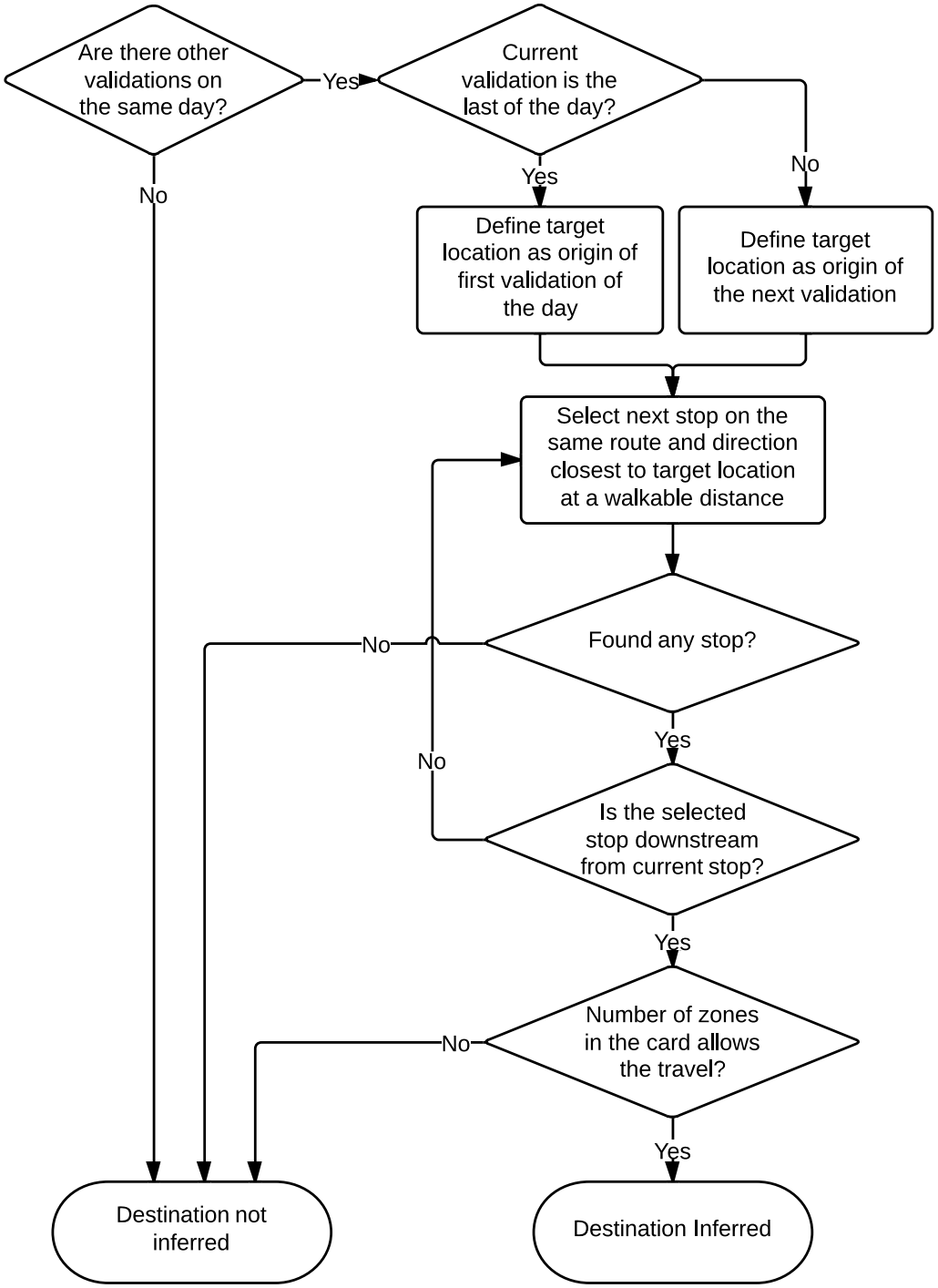

Figure 3.6: Destination Inference algorithm flowchart adapted from Gordon [Gor12]

When this is not the case, the *consecutive trip segments* [ZRW07] method is followed, joined with the the assumption that the last travel of the day has as destination the first origin of the day. Thus, if the current validation is the last of its day, the target location is assumed to be the origin of the first computed validation of that day. Otherwise, there are more trip segments before the last one, and so the target location is assigned as the origin of the next validation.

Following, the validation has a target location assigned, working as a temporary destination. However, when a passenger leaves a bus, one of two scenarios is considered: leaves and in the next validation re-enters in the same route or changes to a different one (figure 3.7). This means that the target location currently assigned to a trip segment is not certain to be the real destination on the current route.

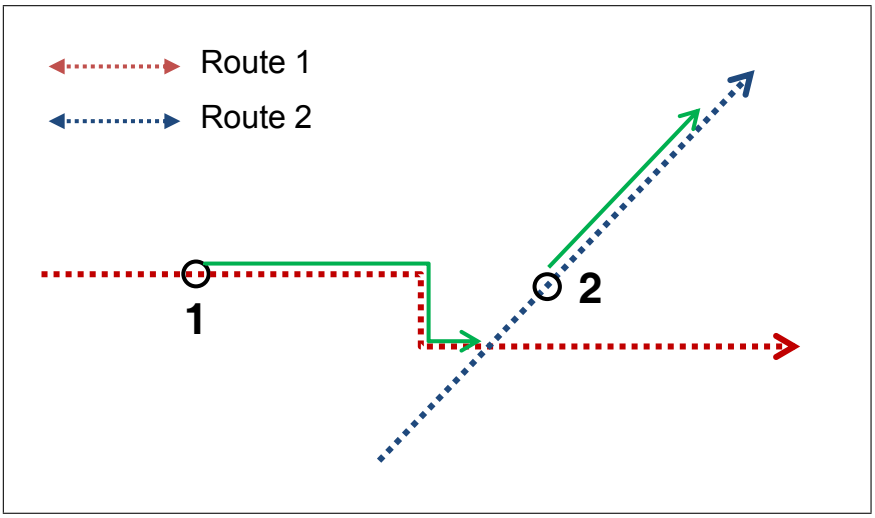

Figure 3.7: Passenger changes route between two (1 and 2) consecutive travels

Therefore, the next step in the algorithm is to find the closest stop to the current target location that is in the same route of the trip segment. The previously computed distance-to-stop matrix is now referenced by the target location, finding the closest stop belonging to the current route. If it is not possible to find the closest stop, meaning that there is not a stop closer to 640 meters, the algorithm cannot progress and so the validation is discarded. This may be the final inferred destination if it was correctly assumed. Following, the next steps validate this still temporary inference to be able to assume it as correctly inferred.

First, it is verified if the inferred destination is downstream from the origin. In other words, checking if in the current route, in the current direction, the destination's sequence is higher than the origin's. This prevents possible errors when finding the closest distance to the target location. If it is not downstream, the algorithm backtracks to the previous step, referencing again the distance-to-stop matrix, finding the next closer stop. If found, verifies if this new stop is downstream from origin. This step repeats until a downstream location is found or until there's no more stops closer than the minimum distance on foot.

Finally, it remains to check if the number of zones in the card allows the current predicted origin-destination segment. Assuming that the passenger only travels to a valid destination, this step assures that the current inference is not outside of the card's range. As seen in the study of the Andante system in the appendix A, for each card there's a maximum number of zones in which the passenger is allowed to travel. These zones, referred in A, have n number of zones between each other, being $n => 2$, $n <= 9$. If the n distance between the origin and destination stops exceeds the distance allowed in the card, the validation is taken as a bad inference and is discarded.

Concluding, after the validation passes through all the algorithm's steps, it is added to the currently inferred set of validations. When all the nested loops are passed over, all the now not necessary maps in memory are cleared and the new map of successfully inferred validations is written to a CSV file for later use.

3.5.5 Time of arrival inference

Along with the inference of destination, the arrival time was also estimated. Here again, the architecture of the algorithm depicted in the flowchart of figure 3.8 was adapted from Gordon [Gor12] in the context of this project, and implemented as part this thesis, and adds to the previous inference steps on the flowchart on 3.6.

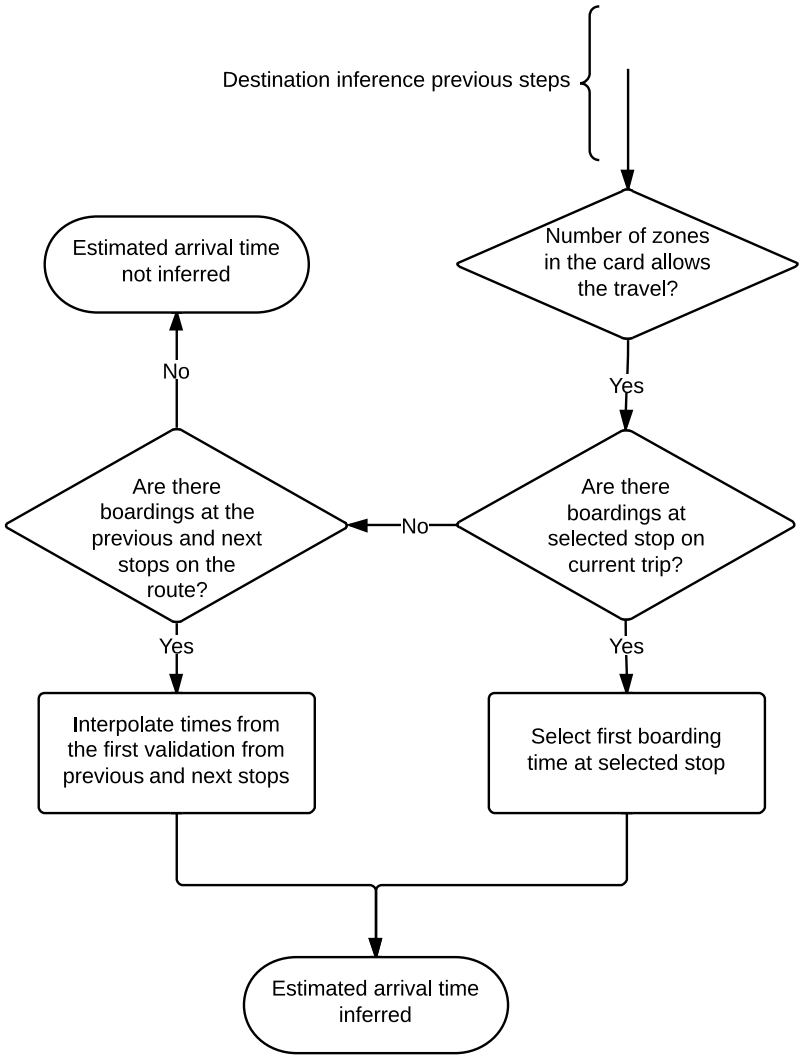

Figure 3.8: Time of arrival inference steps adapted from Gordon [Gor12]

The first step for time of arrival inference is checking if there are any boardings at the currently inferred destination in the current trip. In other words, the algorithm identifies the trip of the validation whose destination is being inferred through its vehicle identification and time of journey

start and then looks for any validation made in that vehicle, in that journey, in that selected stop. In case it finds any, the time of validation of its first boarding is accepted as the estimation of arrival time at the inferred stop, as

$$ETA(Selected) = TimeOfFirstBoarding(Selected) \tag{4.1.5.1}$$

since it's assumed that the passenger left that vehicle at that stop at roughly the same time that the first person entered it.

If it can't find any, the next step is to do the same procedure but to the previous and next stops in that journey's route. This means that, with this information, the algorithm can find times of boarding at these stops in a way that, in the end, it gets first passenger boarding times for them. If it can find those times, then an interpolation is made in order to get the estimated time of arrival (*ETA*). The interpolation procedure was added in the scope of this project since it was not found in the literature.

$$ETA(Selected) = \frac{|ETA(Prev) - ETA(Next)|}{2} \tag{4.1.5.2}$$

The *ETA* for the inferred destination is, thus, roughly calculated as the absolute mean value between the time of arrival of the vehicle at the stops that are positioned before and after in the current route, given by:

$$ETA(Prev) = TimeOfFirstBoarding(Prev) \tag{4.1.5.3}$$

$$ETA(Next) = TimeOfFirstBoarding(Next) \tag{4.1.5.4}$$

If any of the previous methods is able to identify the time of arrival, the result is a complete inference with location and time. Otherwise, although the time is discarded, the destination is still inferred and returned as successfully inferred.

3.6 Results

Using the dataset prepared for this case study, the previous algorithm was computed and its results analysed. First, it is important to refer the weight of the validations computational-wise. In total, after all the pre-processing, there were around 2,4 million validations to compute in January. Knowing this, and because the set of validations contained too much information to run in memory, it was divided in two sets: the first contains the first two weeks (days 1-15) and the second contains the last ones (days 16-31).

Table 3.1 shows the inference rates for the main stages of verification, along with the total time spent computing them for the destination inference experience both for the first two weeks and the last ones.

Step	Valid	Total	Percentage (%)	Total time (s)
One-validation days	1853452	2400682	77%	0,463 s
Selecting closest stop	1130620	2400682	47% (61% from previous)	893,2 s
Zoning verification	1123197	2400682	46,7% (99,3% from previous)	0,391 s
Time inference	302019	2400682	12,5% (26,8% from previous)	22,3 s

Table 3.1: Results for running the inference algorithm on the case study data for January

Analysing the results, it is possible to verify that the first step has serious consequences on the inference, since 23% of the passengers only validate once in one day . The next stage relatively to destination inference is getting the closest stop to the one the passenger has boarded to find the destination of the previous trip segment. This step had the most critical impact, resulting in 47% of the total dataset with 61% "survival" from the previous one-day-validations stage. This happens because for that particular bus stop there were not stops in a walking distance(640 meters) that belong to the same route of the boarding stop and are downstream from it.

Although this may occur because of the passengers' behaviour, one of the other possible reasons why this happens is related to the difference between the dates of the validations and the date when network data was gathered. Table 3.2 demonstrates that, of the total of non-inferred validations, 11% of them occurred directly because nowadays the connection between the validation boarding stop and bus route does not exist (and thus also not present in the current network data). The full indirect impact of these problem can only be seen through a future different study with validations and network data with the same timespan, since failures in one validation may impact the inference of the following ones of the same traveller.

Total non-inferences	Due to change in network (%)
1277485	138172 (11%)

Table 3.2: Non-inferences due to network changes

Further analysing the results on 3.1, the zonal verification proposed by this work results in a drop of only 0,3% relatively to the complete set of validations and a 99,3% "survival" from the previous step of closest stop inference. This result, representing a small change in the outcome, adds the verification that the estimated destination is at a zone reachable from the passenger's Andante card, and thus confirms the good results on the previous stage of destination inference.

Finally, the time of arrival computation results were not promising, with a total of 12,5% of stops with estimated arrival time. Giving the low results on the time of arrival estimation, this estimate was not included in the subsequent work of creating the networks of passengers. Future

work should be carried out in order to provide a better algorithm to estimate time of arrival at the inferred destination.

Figure 3.9 shows the path inferred from all the validations from one of the passengers on the 6th of January, with the inferred data shown on table 3.3. The different colors define the different paths taken on the course of the day, with two significant time intervals: morning, most possibly being the journey for work (or other) and evening with the return to the first boarding location. The number of zones travelled is legal according to the zone system, and thus is assumed correct for the zonal verification. The times are relative to the boarding validation time as this only provides a destination estimate.

Figure 3.9: Inferred travels for one passenger on the course of the day

Route	Origin	Destination	Time of Validation	# Zones authorized	# Zones travelled
704	GIT2	ARS5	09:29:49	3	2
701	ARS3	PLIM2	09:43:21	3	2
701	PLIM1	ARS2	17:48:57	3	2
706	ARSM1	TNG1	18:10:43	3	2

Table 3.3: Journeys made by one random passenger on January 6^{th}

3.7 Evaluation

In order to evaluate the proposed algorithm and its results, it became mandatory to be able to create our own travels with already foreseen destinations, since the main goal is to infer the correct one to each travel. To achieve this we propose evaluation through the use of one simulated sample that contains, for one unique user, a group of a maximum of 10 trip segments from the beginning of the day until the end of it. The number 10 was chosen only as a base value to ease calculations.

As seen in figure 3.10, the proposed sample is built assuming a controlled random behaviour of one passenger. This means that, for that hypothetical person, a maximum of 10 validations were generated based on the following assumptions:

1. A random stop is selected to be the passenger's first boarding location. This follows the premise that we don't know each passenger's behaviour in a given day. The selected stop must have at least one route for which it isn't the last stop to guarantee that the traveller has any vehicle to enter;

2. In that random stop, the passenger enters one vehicle in one of the random routes that still has stops ahead;

3. In that route, the passenger leaves the bus at a random stop provided that it is downstream from the origin bus stop, meaning that the virtual traveller is controlled to only leave at a stop in which the vehicle hasn't passed before the passenger has boarded. This stop is kept as the foreseen destination for this trip segment, if there is one, otherwise is the last of the day;

4. After leaving the bus, the traveller can board a new vehicle in any of the surrounding bus stops, provided that it isn't farther than 640 meters away. This distance is based on the assumption that the maximum distance on foot that one person travels between two stops is, at most, 640 meters (or 8 minutes). If there isn't any selectable stop for this passenger, the algorithm cannot generate any more destinations and stops generating journeys;

5. In that selected new boarding stop, the passenger enters a new vehicle following one of the routes of which that stop is part of. The selected route is also randomly chosen, following the same premise that the traveller's behaviour is not known;

6. The process repeats from 2 for each of the remaining trip segments, until a maximum of 10.

Following this method, a dataset was generated to test the efficiency of the destination inference algorithm proposed. Each one of the travellers would contain a max of 10 travels, with the possibility of being less in cases where, when following our assumptions, the algorithm couldn't generate any more validations for that user. A total of 250000 travellers were generated, resulting in around 2 million validations created. It is important to note that these validations are completely simulated based on the methods aforementioned, and thus created on a controlled environment,

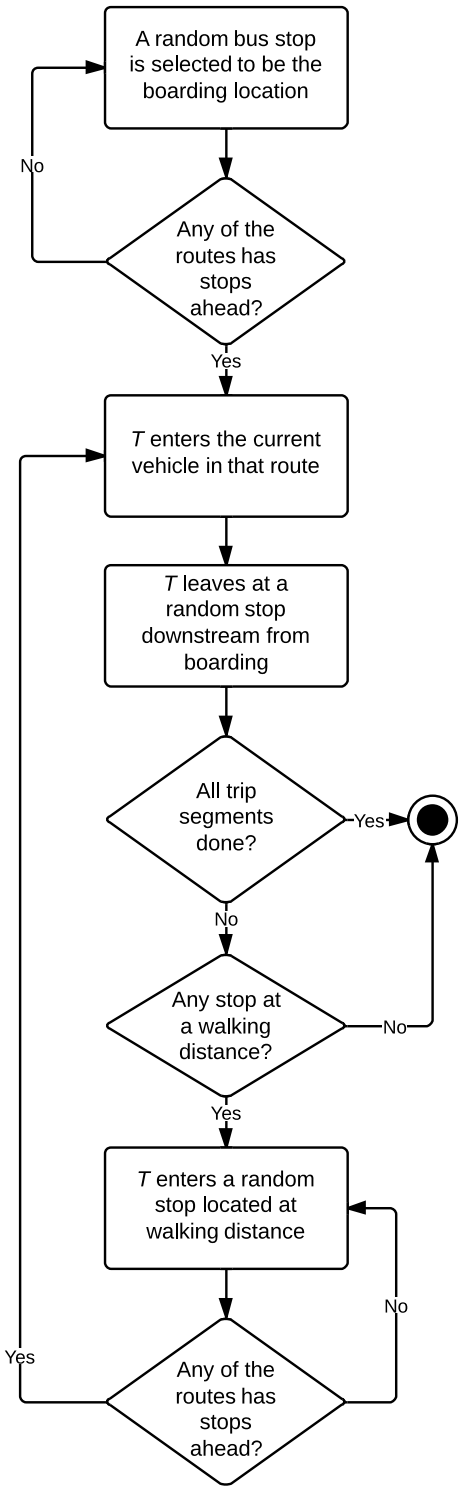

Figure 3.10: Simulation of trip segments for one the traveller T during one day

with the needed randomness in the travellers since we were simulating the worst case scenario on which we have no idea of the next journey for each traveller.

Figure 3.11 below provides the average final results after applying the destination inference algorithm to the simulated dataset. The criteria for evaluation was how distant was the predicted

destination stop compared to the foreseen one, where the distance is measured as the number of stops between them on that route. "Correct" means it was the right prediction and "> 3" stops refers to cases where the prediction exceeded the distance of 3 stops.

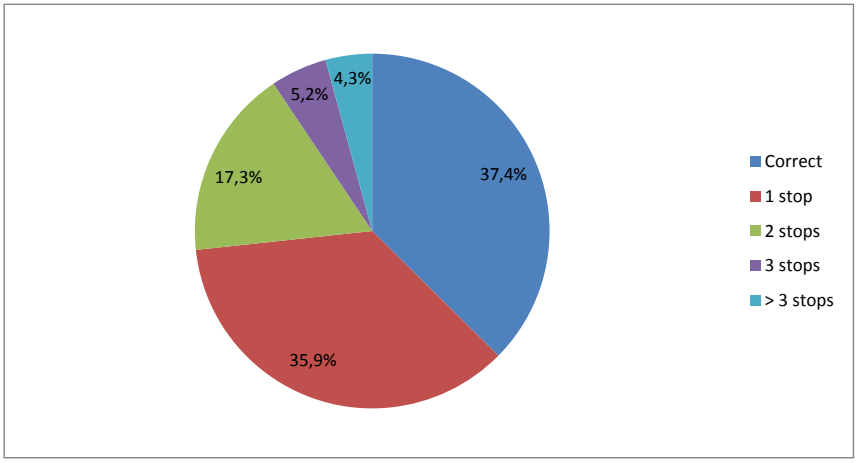

Figure 3.11: Results from the inference algorithm on the simulated dataset

With the inference of the simulated sample of nearly 2 million validations, and matching the inference with the foreseen destination for each record, we obtain a average of 37,4% of right predictions, 73,3% right if considering a tolerance of one stop away from the inferred destination, and over 90% if considering a tolerance of two stops. Only 4,3% of predictions exceed a distance of three stops from the expected destination. Following the assumptions described, these results show that our proposed algorithm has a significant probability of, at least, give an accurate prediction for the destination of one travel.

Through these observations, we concluded that this part of the work provided good results that can be explored in the final stage of implementation, which is the creation of networks of travellers based on their travels and on its relevance among each other.

3.8 Summary and Conclusions

Public transport open systems require journey validation only at entrance, and consequently datasets which gather these validations aren't able to provide full travel paths. Regarding this problem, an origin-destination inference algorithm was proposed with main objective of estimate full travel paths for each passenger on the dataset, identified by his card serial number.

In order to prematurely discard all the corrupt records, a data pre-processing procedure was studied and implemented, manipulating data with the goal being the improve of results efficiency. Thus, records with missing key-information, with irrelevant information or with incomplete mapping were discarded.

The results obtained, checked against a simulated dataset created with base on controlled random passenger behaviour, show accurate estimations with an average of around 90% inferences at least than 2 stops from distance, with prevalence of right inferences.

Concluding, the computed dataset with inferred travel paths gives way to the creation of temporary networks, the second stage of development in this work, that follows on the next chapter.

Chapter 4

Temporary networks

This chapter details the second and last stage of development on this thesis, the creation of temporary networks based on the travel paths inferred on the previous section. This is the implementation of a concept described in section 2.2 of the state of the art [NGCP11][NGC12].

The next sections provide deeper explanation on the interconnection between the two stages of work. The concept of relevance of journeys that was developed in the context of the project this work relates to [Nun12] will be described. Measurement of such relevance between journeys is used to obtain the aforementioned temporary networks [NDGC14a].

Finally, the temporary networks created are analysed through two different experiences that show promising results for future development on the area.

4.1 Introduction

Public transport users, during the course of their day, can have multiple travels through different paths and have those paths in common with other passengers, either in the same vehicle or others, sharing the same bus route or similar ones. Transports networks have multiple routes that share different road segments and bus stops, resulting in riders travelling in the same path of others at the same time, at previous journeys or other passengers who still are going to begin their journey.

Resulting from the inference of bus travels from boarding stop to ending destination, travellers' full paths for each trip segment are estimated. This allows comparison between their travels and the clustering of the most similar ones through its relevance among each other.

One travel can be similar to another in different ways: their paths can have a great distance in common, their routes can have a large number of stops in common or they can be complimentary routes, i.e, have alternative paths that share start and ending points.

Passengers whose travels are significantly relevant to each other must, thus, be clustered in the same networks. The implementation of such methods on the *smartphone* application developed to exchange information among public transport travellers [Gon12], described in 2.2, allows

users to exchange real-time and relevant information among each other, receiving through their own temporary network and sending to those for whom they are relevant in a similar way to the one represented below in figure 4.1. Here, "User 4" has relevance to "User 1", and, this way, is connected to his temporary network in order to feed information to it, along with "User 2", "User 3" and "User 5". At the same, "User 1" and "User 3" are relevant to "User 4", and are connected to his own temporary network. This is an example of how one user feeds the others' temporary networks, at the same time that his own receives information from its relevant connections.

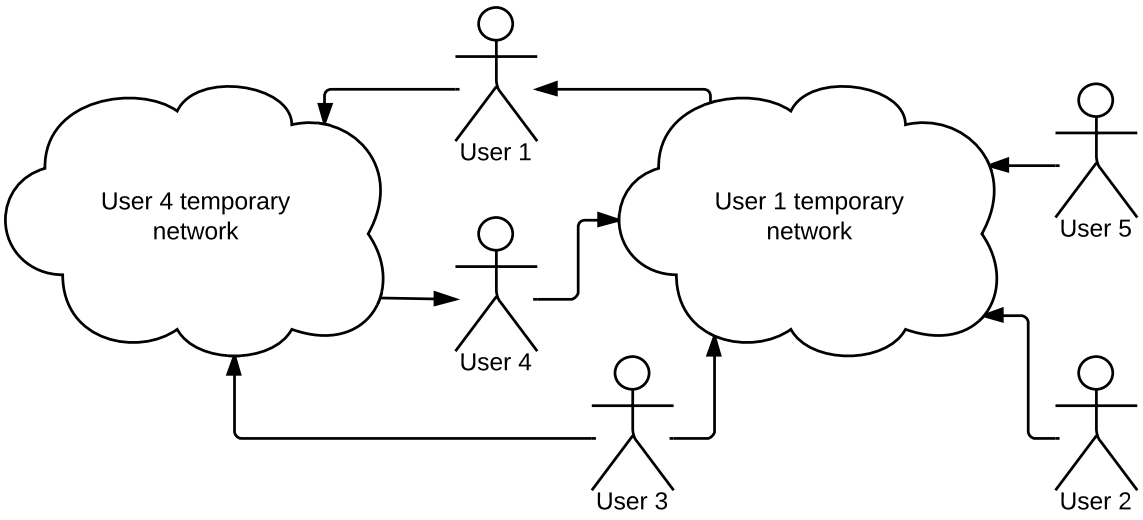

Figure 4.1: Temporary network basic architecture

These concepts of travel relevance and grouping passengers are detailed in the next sections, starting with an architecture description to understand how these processes interconnect to create the aforementioned networks.

4.2 Architecture

Starting from the origin destination matrix inference, the creation of the temporary networks follows the architecture seen in figure 4.2 below. Along with the dataset with the inferred journeys, the data from Andante network collected in the context of the project [Nun12] is used as support to compute the full travel paths. Since the origin-destination inference data provides route and bus stops of boarding and destination, this data is an essential asset to obtain the full travel paths for each traveller.

Similarly to the approach to the travel inference algorithm, the network creation was also implemented in the Java programming language.

In this phase the relevance between the different travels, at a given time, is computed. This concept is detailed on the section 4.4 that follows on this document. This computation results in a set of connections between passengers that originates temporary networks centred in each of the passengers, i.e, each of the computed travellers at a given time has his own set of temporary connections.

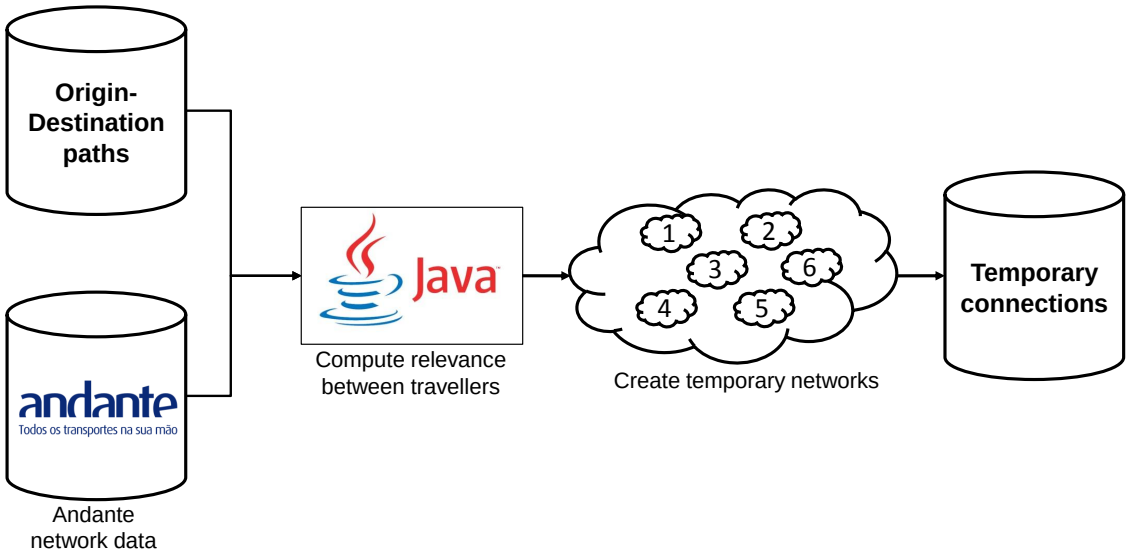

Figure 4.2: Temporary network creation architecture

The created sets of connections are finally stored in the local database in order to visualize and evaluate the results for the intended instant of the day. Concluding, in order to better understand how the architectures from the travel inference and temporary network creation interconnect, the following section provides one overview on the full process of this mechanisms when integrated with a *smartphone* application to exchange information among public transport travellers [Gon12].

4.3 Full-process overview

As detailed previously on chapter 2, this work follows a previous prototype application developed to allow passengers on public transports to exchange travel information such as external events (accidents, interferences on the bus route) or bus-relative information such as vehicle conditions, fullness or driver skills [Gon12][NGG13].

The full process can be followed in figure 4.3 [NDGC14b]. The first stage comes with the passenger notifying the application of the journey start. This can be made through checking-in on the application, notification from an identified travel pattern or from the journey planner present on the application. Conceptually, in the future there is a real possibility of interconnection with a *smartphone* ticketing validation system, using its advantages to automatically check-in on the application when validating via *smartphone* [NDGC14a][NGG13][FND13].

Once identified the check-in boarding location, the destination is estimated from the passenger's travel history. This history is created as the passenger checks in the application and the destination of the previous journey is estimated using the proposed inference method. In case the passenger uses the journey planner and checks in through the application notification, the travel is automatically added to his history.

Starting from the path estimated for the traveller's current trip segment, a temporary network is created based on its relevance with other passengers' journeys at that specific time. Each time one user checks in or checks out, a new user-centred network cloud is created, meaning that passengers' connections to travellers relevant to them in one specific time may differ after other passengers join or previous ones leave.

As soon as the journey ends, the temporary network created for that user is destroyed - resulting on a updated network cloud for all the other passengers. One journey may end through several ways: passenger checks out directly on the application or his GPS position matches either the foreseen destination location or the previously established journey ending on the application journey planner [NGG13].

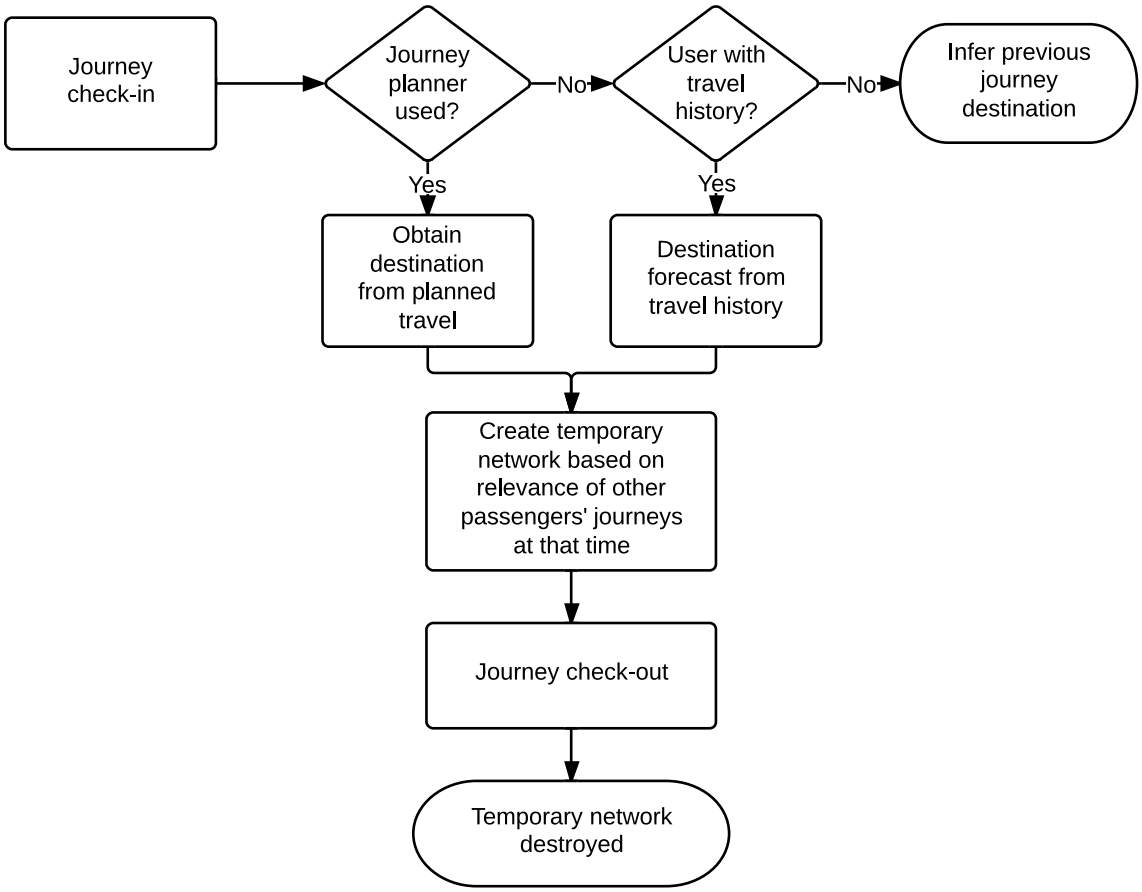

Figure 4.3: Real-time full process overview for the passenger process when using the application [NDGC14b]

Knowing how the processes are connected, it is mandatory to conceptualize and define relevance. The following section describes what is relevance in the context of this project [NDGC14a] and how it can be measured in order to create the passengers' temporary networks.

4.4 Relevance among travels

In information retrieval, relevance measures how well the information matches the one the user needs. In the context of this project [NDGC14a], relevance represents how well one passenger's travel path meets another passenger's journey on a specific time, meaning that this relevance depends on the travellers checked in on the application at every time. This results on temporary and highly dynamic scores for relevance among the travellers, and thus this relevance has to be measured each time one passenger enters the system, does another boarding or checks out from the network.

Knowing this, relevance among travellers has to be measured in a way that its score defines how much one traveller's path information can impact another's behaviour and, consequently, users must be connected with those travellers with the most important real-time information for them at that specific time.

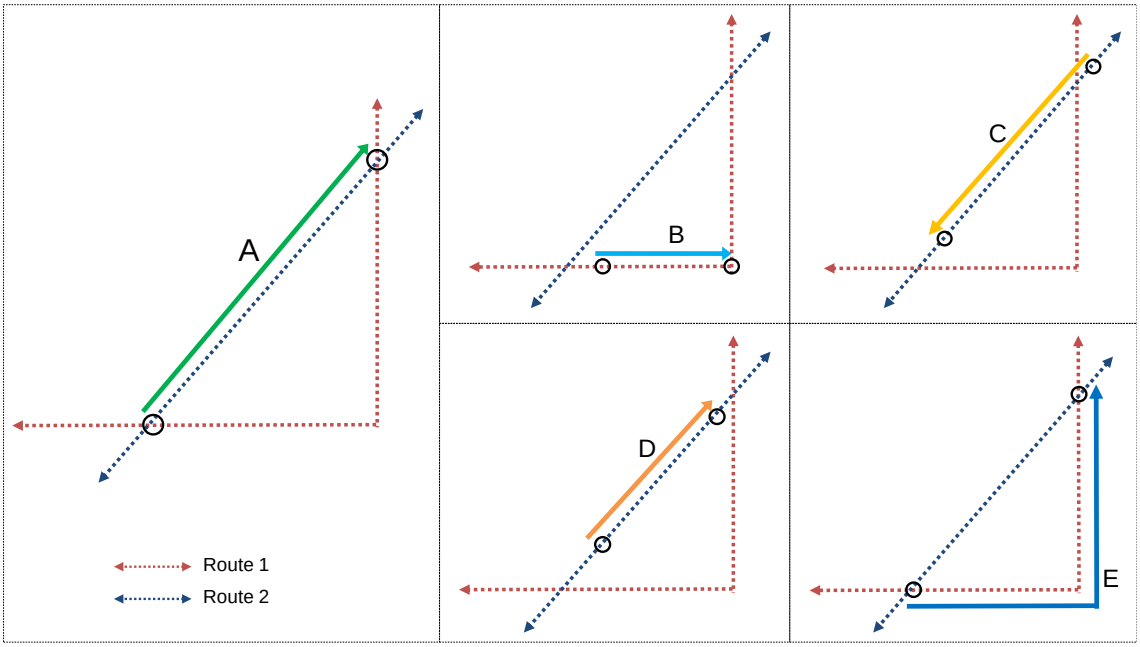

Figure 4.4: Comparison between relevant and non-relevant travels for one passenger [NDGC14b]

In figure 4.4 [NDGC14b] we can see several simple different travel paths and compare how much they have in common to the passenger *A*. In the example, the travel path of *B* has nothing in common with the first one (*A*), since he travels on a completely different route. That passenger journey information, thus, should not be relevant enough to the passenger *A*'s current travel.

Case *C* represents another case where relevance is not considered significant. Although their vehicles share the same path, the direction is not the same. Consequently, there could be little information significant to the first passenger, with the exclusion of possible events which resulted in road blocks, occupying the entire road. Giving the low frequency of these events, connecting these travellers in the same network would lead to great quantities of possibly irrelevant information exchange among them, and thus relevance in this context is not considered.

Example D contains a path that overlaps the first one over some distance. This case represents, in this context, the basic type of relevance between two travels: the number of bus stops or distance between them that those travel paths have in common. The relevance among them, thus, must be considered taking into account that path similarity.

Case E contains a path that has little relevance when looking for path similarity. However, this path contains both origin and destination from the first one, and thus can be seen as an alternative route. Hereupon, if there is information on the passenger from A's network that informs of incidents, delays or another situation like fullness of the vehicle, this passenger might want to receive information from users currently on vehicles on the alternative route represented by case E, eventually changing his planned behaviour.

Finally, through this analysis we concluded that the following detailed main types of relevance provide significant results and are those considered for this context. The minimum relevance defined for one passenger's travel to another's, in a specific time, was 50%, being positive if equal or above this value or negative if below. If one passenger is relevant through both types, the largest one is considered the true value.

Journey similarity

The most basic case of relevance between two passengers' paths. This case considers the similarity between the estimated paths for each of the passengers' journeys, defined in one of two criteria: distance and number of bus stops in common.

First, distance is defined as the distance in meters between two bus stops, given that both of those stops are common to both paths in consecutive locations. In other words, distance considers the total distance from all the road segments between two or more stops that are shared by the two passengers' paths.

The formula to obtain the score the relevance from the passenger A's path to passenger B's in an instant t, through this criterion, is the following:

$$R_t(A,B) = \frac{\sum_{i=1}^{n} s_{i,t}(A,B)}{d(B)} \tag{4.2.2.1}$$

n is the total number of road segments in common between both paths and $s_{i,t}(A,B)$ is the total distance of the $i-est$ shared road segment among them in the instant t. Function $d(B)$ returns the total distance for the estimated path of the passenger B.

The number of stops criterion, as the name indicates, stands for the number of bus stops the two paths have in common. In this case road segments between consecutive bus stops are not mandatory and only the total number of stops has relevance. Assuming that for two vehicles need to cross the same location for some distance to pass through the same bus stops, this criteria makes sense and results in a total of locations (and its roads segments) that the different passengers cross in common at a specific time.

Similarly to the first case, the calculation of the relevance for passengers A and B paths, in an instant t, is given by:

$$R_t(A,B) = \frac{n_t(A,B)}{q(B)} \tag{4.2.2.2}$$

$n_t(A,B)$ is the total number of bus stops in common between the paths of A and B in an instant t and $q(B)$ is the total number of stops for the passenger B's path.

In both criteria for this relevance type the indicator is a score from 0% to 100% of relevance among the passengers' journey paths, being 100% fully relevant and 0% not relevant at all. Furthermore, relevance is not symmetric between two passengers. Figure 4.5 [NDGC14b] helps to understand why this happen. Using the function (4.2.2.2), and knowing that the path of passenger A fully overlaps the path of passenger B, the result is that B is 100% relevant to him, whereas the inverse doesn't happen since B's path only covers a portion of A's, in this case 70%. This reinforces the choice of user-centred temporary networks, since each user has his own set of temporary connections at a given time whereas those users may not have him on their most relevant connections.

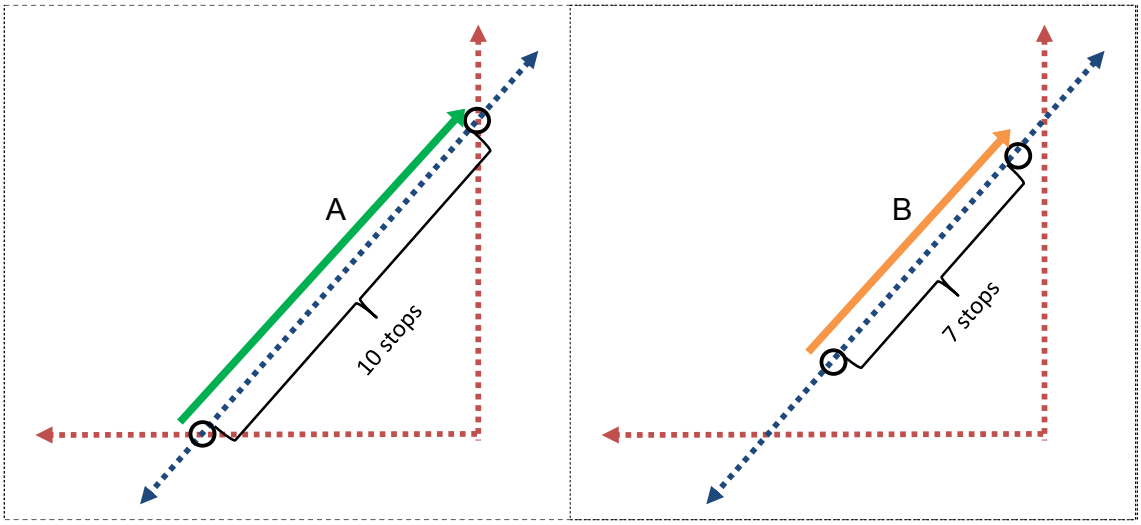

Figure 4.5: Asymmetry between travel relevance for two passengers [NDGC14b]

Journey complementarity

This type of relevance considers two complementary travel paths in case they share both boarding and destination locations. In other words, passengers' journeys that do not share any intermediate bus stops but have the same boarding and estimated destination are considered complementary, and are relevant to each other since they provide alternative information to their journey. If the passenger receives information that his predicted journey has any impediment or delays he must receive information for the alternative routes and change his behaviour if it reveals a better decision.

However, this scenario does not apply only when the journeys share exactly the same boarding and ending locations. Another situation might occur: the passengers do not share both boarding and ending location, but one of them or both are close to the other passenger's, in a way that allows the rider travel the distance between them on foot, i.e, the distance between them is inferior to the assumed maximum distance on foot, 640 meters.

Three possible cases are shown on the below figure 4.6. In example 1, passenger *A*'s journey matches *B*'s on the boarding location but not in the destination, yet both destinations are 300 meters away. Following the same assumption made to the inference method, this is a *walkable* distance (since it is inferior to 640 meters) and thus the *B*'s travel path is semi-complementary to *A*'s. The same goes for example 2, with the difference that now it's the destination that is the same and the boarding stop the one close enough and the same rule as the previous example applies. Example 3 shows another situation that might happen, with both boarding and destination stops are only close for each passenger. Provided that both distances are small enough to walk on foot, this still represents a semi-complementary route and it may also be relevant enough because the travel on foot can be advantageous in case of reported problems on the other alternative routes.

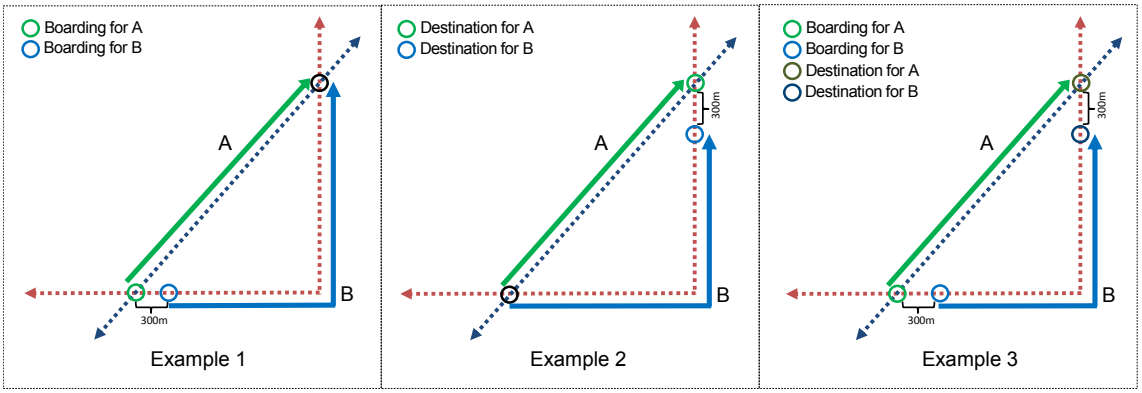

Figure 4.6: Different levels of complementarity between two passengers' travel paths

The complementarity property can be complete or semi-complete: one journey is fully complementary if it shares both boarding and ending locations and thus the complementary passengers are 100% relevant for each other; if the journey is semi-complementary, with boarding and/or destination stops being different but within a *walkable* radius, then a calculation must be made to score the relevance among them. Since a fully complementary route is 100% relevant, the function to obtain relevance for the semi-complementary must return a score never equal to 100, and must be lower as higher the distance between them.

In order to make the closest semi-complementary journeys more relevant, the relevance had to decrease as the distance was bigger. Through

$$\frac{d(|A_{stop}, B_{stop})|}{K}, \tag{4.2.2.3}$$

being $d(A_{stop}, B_{stop})$ the distance between the close stops A_{stop} and B_{stop} and K the constant maximum distance on foot (640 meters), we obtain the ratio between both distances, provided that $d(A_{stop}, B_{stop}) < K, K > 0$.

Furthermore, this ratio can then be used to decrease relevance as the distance is bigger, for two complementary paths A and B in an instant t, through

$$R_t(A,B) = 100(1 - \frac{|d(A_{stop}, B_{stop})|}{K}) \tag{4.2.2.4}$$

Because the minimum score for two paths being relevant was defined as 50% and the previous function lead to values below the minimum relevance, there was the need to adjust the distance ratio function to ensure minimum relevance. Thus, since the farthest walking distance on foot, 639 meters, returned a ratio of 99,8% and a relevance of $100(1-0,998) = 0,2\%$, the distance ratio function was adjusted by finding a c multiplier that ensured the minimum 50% relevance:

$$100(1 - \frac{(639)}{640}c) = 50 \Leftrightarrow c \approx \frac{1}{2} \tag{4.2.2.6}$$

Finally, once defined a distance ratio function that ensured minimum of 50% distance, the following formula was defined to return this score for the passengers A and B's travel paths in an instant t, when having one of the locations at walking distance on foot:

$$R_t(A,B) = 100(1 - \frac{|d(A_{stop}, B_{stop})|}{2K}) \tag{4.2.2.8}$$

Concluding, to define the relevance when both boarding and destination locations are complementary within walking distance, the mean value for boarding and destination relevance was calculated and through this an average relevance was obtained considering both distances. This way, relevance score is define by:

$$R_t(A,B) = 100(1 - \frac{|d(A_{board}, B_{board})| + |d(A_{dest}, B_{dest})|}{4K}) \tag{4.2.2.8}$$

where $d(A_{board}, B_{board})$ and $d(A_{dest}, B_{dest})$, respectively, define the boarding and destination locations distances between both passengers. In case they share boarding, $d(A_{board}, B_{board}) = 0$ and if they only share destination, $d(A_{dest}, B_{dest}) = 0$.

Summarizing, through this calculations the relevance among riders with similar, complementary or semi-complementary travel paths is calculated, and only the ones with positive relevance (i.e $relevance >= 50\%$) are accepted to connect to the passenger for whom they are relevant. Between both similarity and complementarity relevance, the largest one is attributed to the relevant passenger, since it indicates the real score of importance from him to the traveller to whom he is relevant.

The next sections detail the network creation process through travel relevance and analyse the results using a case study for different times, evaluating them through different experiences regarding the behaviour of one passenger's temporary network at the course of the day and at the course of the week.

4.5 Network creation structure

Following the structure seen in figure 4.7, to train the network creation process the first step was to obtain a sample of validations which provided travellers behaviour on a given time.

Thus, a sample from one-hour time interval was obtained from the full inferred dataset, simulating the Andante network at the end of that hour. The chosen interval of one hour is due to the legal travelling time with one validation (described in A.1): because the validations from the Andante dataset do not provide detail regarding how much time that passenger still has to travel, and because one hour is the minimum total available time for one validation, this was the assumption for the network creation. Furthermore, even with the inference of the arrival time, it's only possible to infer time of destination when the next travel begins. Thus, by choosing one-hour time interval we get all the travellers that could still be on their vehicles and the end of that hour, following the maximum total available travel time of one hour.

Regarding the journey path of each traveller's validation on the sample, we had only boarding and ending locations, and in latter stages it became mandatory to pre-compute the full journeys before further advance. Thus, the computation of the full journey paths followed, recurring to the gathered Andante network information for route stops, resulting on a set of bus stops estimated for each traveller's journey path. From this choice two scenarios may appear for each user: finding one unique validation performed by that user or more than one. This could happen due to transshipments made by the passenger on the available one-hour interval. Because we were considering the passengers on the network at the end of that hour, the last validation made by the user was the travel to be considered, since he no longer was on the other one(s)'s path.

The next step, thus, was to select randomly one of the users that is part of the sample. From this user, it was chosen the last (or unique) validated travel, followed by obtaining the full travel path relative to that validation. Here, the decision of pre-calculating the full path was important, since it provided optimal search behaviour for the latter network computing.

Finally, the last stage was to create the user-centred temporary network based on relevance. After this step, the algorithm was repeated as long as there was still "new" users to process, ending when all of them were processed and their temporary networks obtained.

4.6 User-centred temporary networks

At a given time, the behaviour of each traveller riding in a public transport is seconded or preceded by many others also on that path. Furthermore, each passenger travelling at a given time has his

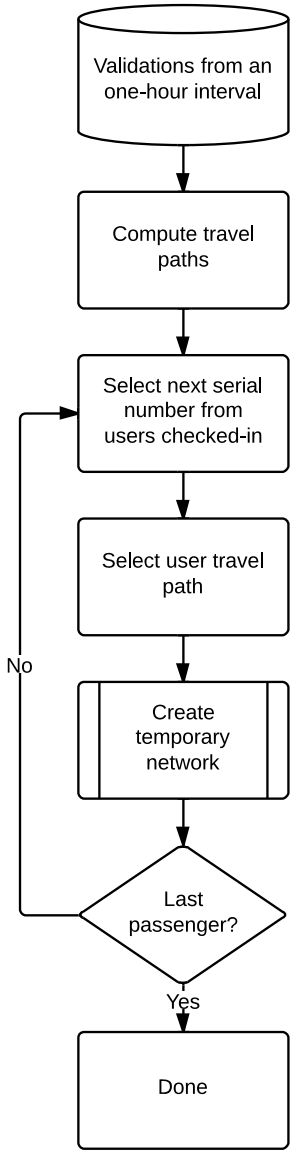

Figure 4.7: Structure overview for the network creation process

own set of "*seconders*" and "*preceders*". Based on this premise, we conclude that each traveller has his own set of connections during the day, or, in other words, has multiple temporary connections.

This leads to the definition of user-centred temporary networks: sets of connections that a given traveller has at a given time. Centring each temporary network on the passenger to whom the connections are relevant results in dynamic behaviour and depends on the travellers riding at that exact time.

From the passenger's full path provided by the one-hour sample we could obtain the needed journey route and boarding and destination locations, that would match against the other users paths.

Once the full path for one passenger is obtained, the next step was to match it against the other travellers' paths circulating at the same time. The process can be seen below, in figure 4.8. The

algorithm runs that single passenger's path for all the other travellers, and the decision of pre-calculating the full paths for each traveller became important for the speed of the process. In this work, the similarity type of a journey path was based on common bus stops, while the distance between them was not considered relevant. Thus, the full set of travels is enough to score each traveller relevance for the others.

Using the two main types of relevance described previously on this chapter - journey similarity and complementarity - the next step was to calculate the relevance score using the appropriate defined functions and matching the full sets of bus stops on each traveller's path with the others'.

The score obtained for each candidate user can be negative (less than 50%), or positive (equal or more than 50%). If the result is positive, that passenger is added to the current passenger's temporary connections, otherwise he would be discarded.

Concluding, done all the calculations for each passenger the result was a cloud network of travellers' temporary connections, each one connected to his most relevant users on that specific time as seen below in figure 4.1. The next section analyses the results obtained for several different time instants and evaluates the relevance of those results for this thesis work.

4.7 Results and Evaluation

4.7.1 Single-day experience

In the study performed for January validations, presented on Appendix A, it was identified that a weekday (from Monday to Friday) has on average more than 200000 validations. Due to this scenario, to help the reader to visualize the results of the temporary networks creation we chose to use, to the first experience, one day from the weekend, in which the number of travellers has a significant drop, to around 20000 validations.

Knowing this, and using the set of inferred destinations, five samples from January 10^{th} of 2010 were collected to use on the temporary network creation algorithm: 07:00-08:00, 07:15-08:15, 07:30-08:30, 07:45-08:45 and 08:00-09:00. The selected intervals were between 08:00 and 09:00 since it was identified in the study on the Appendix A (in figure A.9) as one of the peak hours in number of passengers on the network, which increases the variety in number of different routes being travelled at a given moment.

Each sample was fed to the algorithm in order to obtain, for each passenger circulating at that time, the ones that could provide the most relevant information considering their travel path. The result was a total of 2395 connections between 332 unique travellers (average of 7 connections per passenger).

The full computed temporary network cloud and the score of relevance between its passengers can be seen in the Appendix B. In order to visualize and cluster the resulting connections it was used the open source software *Gephi* [1], a flexible tool for graph and network analysis with main

[1] https://gephi.org/

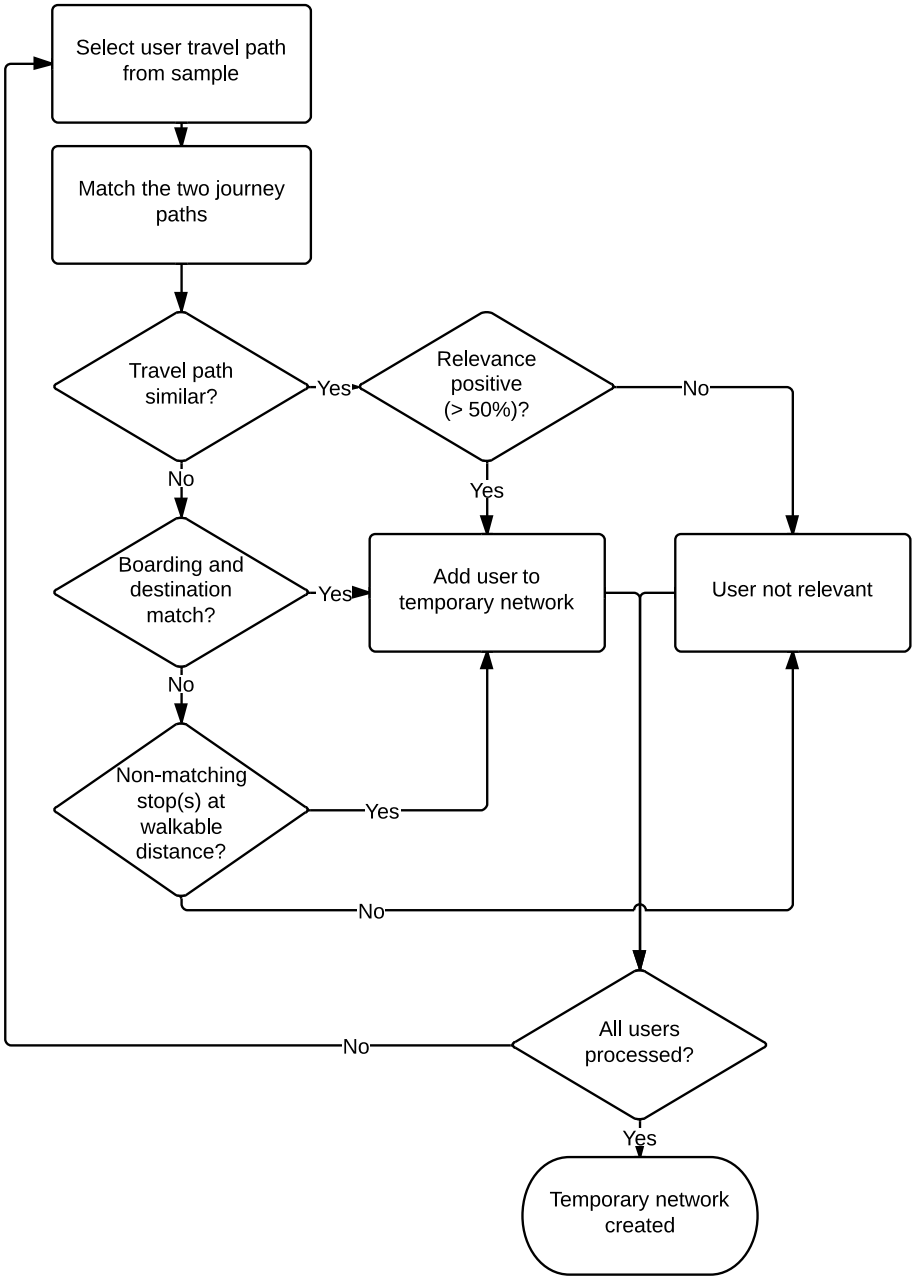

Figure 4.8: Temporary network creation process for one passenger

features to allow exploration and manipulation of large networks [BH09]. Through this framework, the visualization of either the temporary networks or the large temporary network cloud was easy, and its clustering functionalities allowed the distinction between the main different clouds of temporary networks.

Figure 4.9 shows the full cloud existent at 08:00. This is composed by several temporary networks that connect to each other through its passengers' own user-centred temporary networks. Two important attributes must be detailed:

- The different colors represent the clustering of the networks by its modularity, i.e, measuring

the structure of the network by the degree in which its elements inter-connect, with dense connections belonging to the same modules (groups or clusters).

- Each network is represented by a directed graph, meaning that each connection is from one passenger to another with some bi-directional connections when both travellers are relevant to each other. The width of the connection (edge) is related to the weight of the relevance: higher relevance gets a wider edge, and vice-versa.

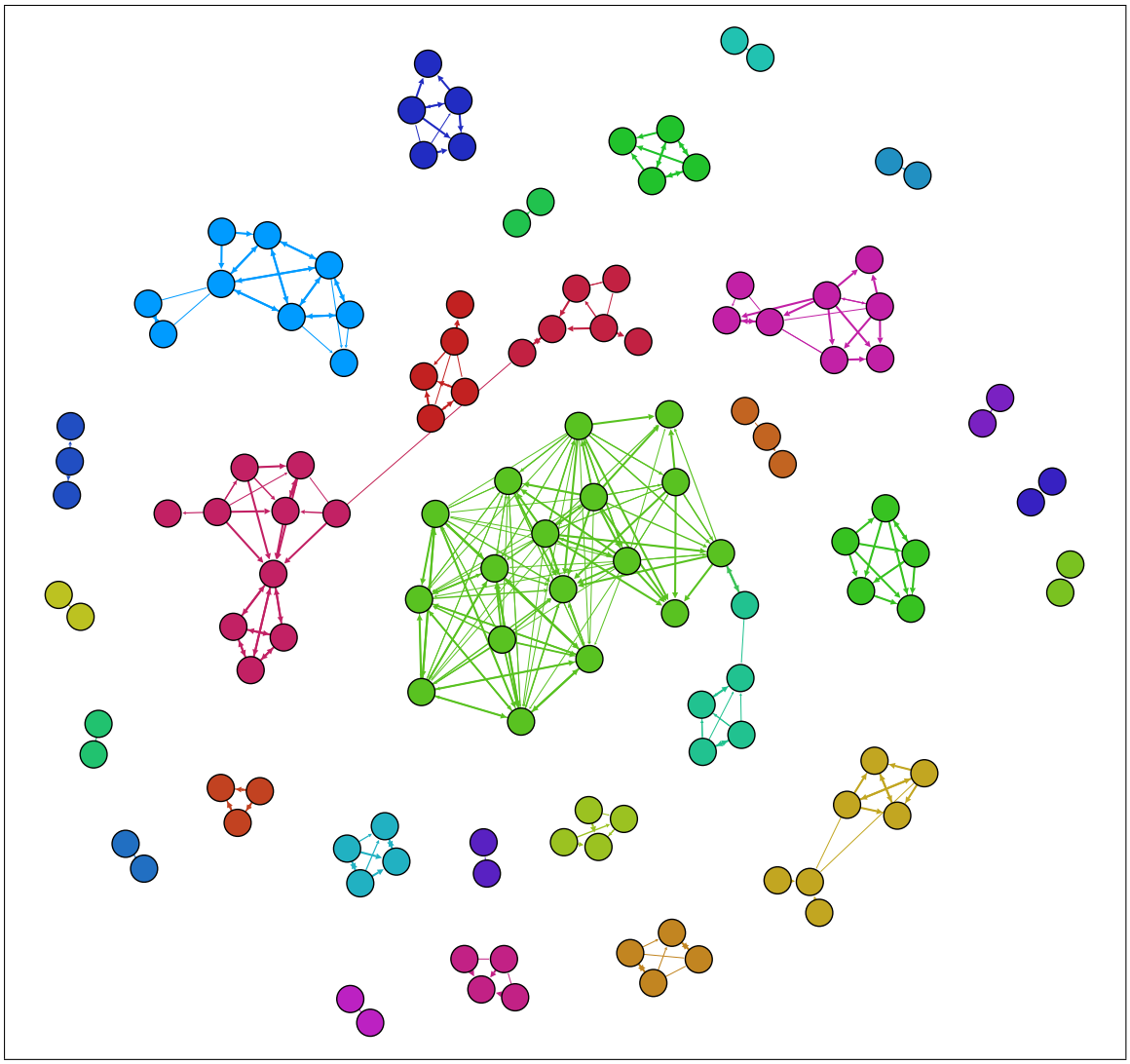

Figure 4.9: Cloud of inter-connected temporary clusters at 08:00

Through this example we can see that different passenger temporary networks can inter-connect, creating bigger cluster of temporary networks. To examine with greater detail the behaviour of one passenger's temporary network from 08:00 to 09:00 from here on we will look to one of them in particular, here named traveller **1** - figure 4.10 in darker blue. Each number only represents the traveller as a way to properly identify them.

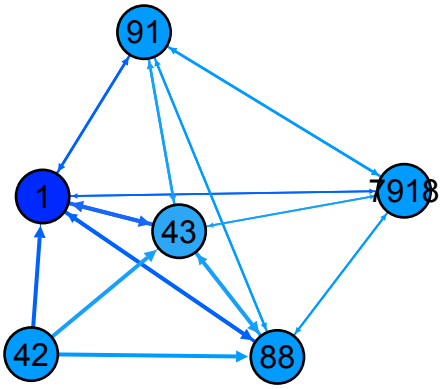

Figure 4.10: Initial temporary network for traveller 1 at 08:00

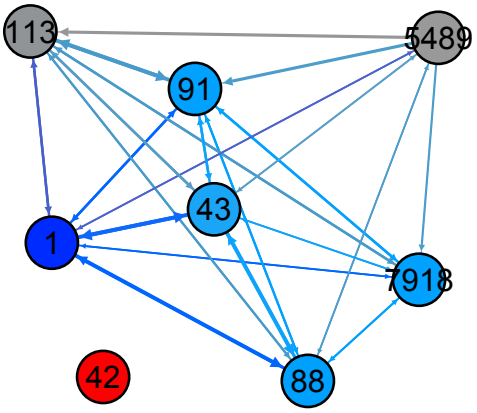

Figure 4.11: Traveller 1's temporary network with member relevance at 08:15

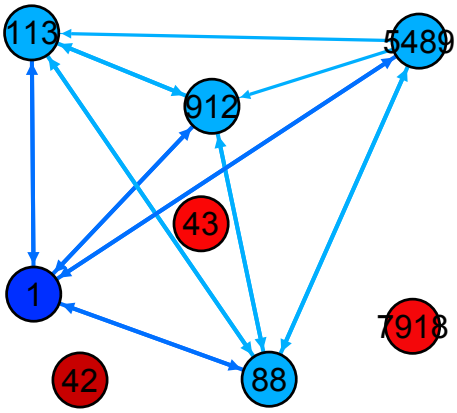

Figure 4.12: Traveller 1's temporary network with member relevance at 08:30

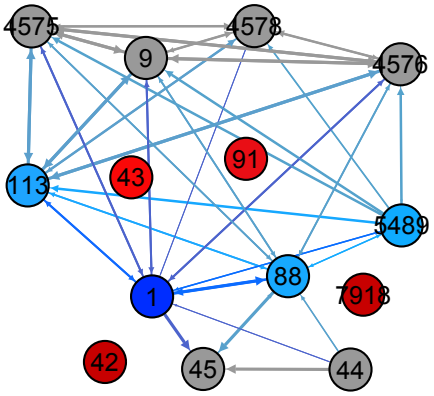

Figure 4.13: Traveller 1's temporary network with member relevance at 08:45

The network of passenger 1, at 08:00, is composed by 5 travellers, numbered randomly on the inferred travels dataset. It's important to refer that, although we are only examining passenger 1's network, each other passenger has his own temporary network with its temporary connections. This way, we also represent in the figure the connections among the traveller 1's relevant passengers at each sample. For example, figures 4.14, 4.15 and 4.16 show, respectively, the temporary networks at this time for travellers 91, 88 and 43. We can see that each one of these travellers is connected to other passengers relevant to them, who can be, or not, also relevant to traveller 1. Table 4.1 shows the members' travels of passenger 1's temporary network, with their respective relevance.

Figure 4.17 represents the paths of the members of Passenger 1 network at 8:00. As it can be visually observed, the estimated travel paths are according to the values obtained and all the travellers are either relevant through path similarity or complementarity and the type of relevance

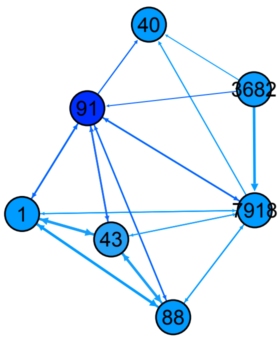

Figure 4.14: Traveller 91's temporary network at 08:00

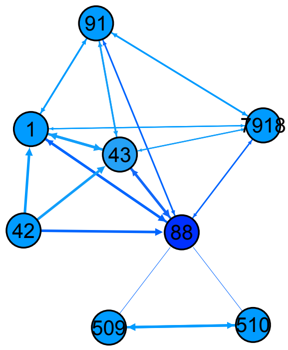

Figure 4.15: Traveller 88's temporary network at 08:00

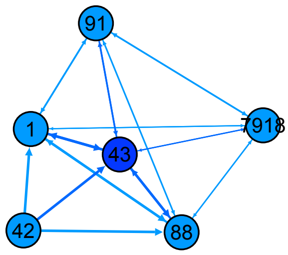

Figure 4.16: Traveller 43's temporary network at 08:00

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
91	502	BCM1	FTM4	0%	82,1%
42	202	TRD1	FTM2	100%	0%
88	202	BCM3	BRP1	88,89%	0%
7918	203	BCM1	SRV1	0%	72,94%
43	202	BCM3	BRV3	100%	0%

Table 4.1: Traveller 1's temporary network with member relevance at 08:00

is easily distinguishable. 2

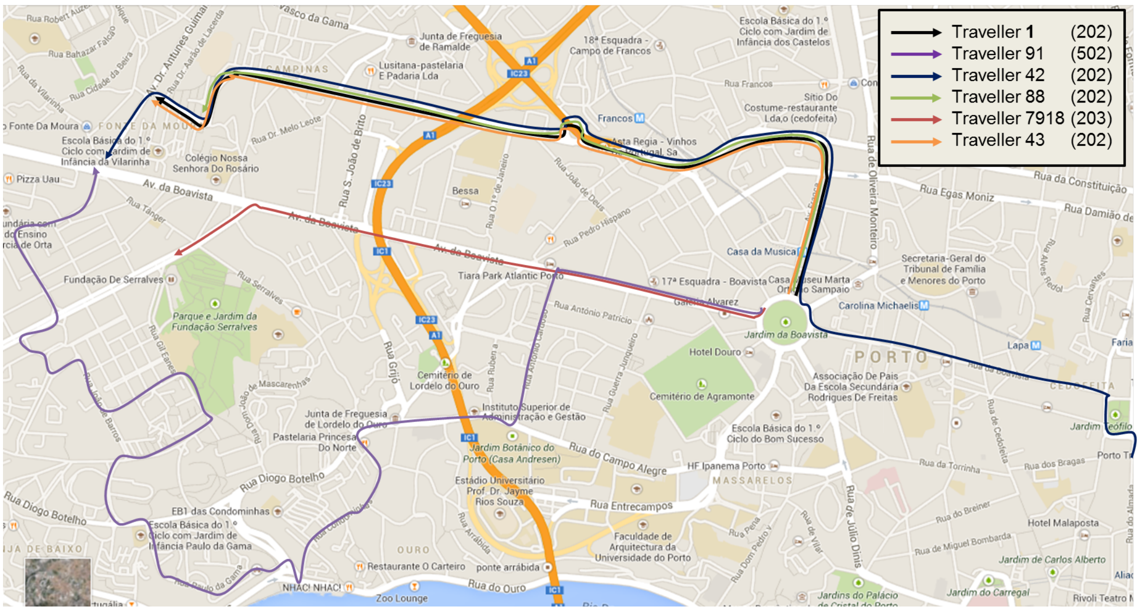

Figure 4.17: Map representation of traveller 1's temporary network connections at 08:00

Considering the sample for 15 minutes later, we find that the network suffered some changes. Analysing figure 4.11, we can notice that some new travellers joined the cluster (in gray) and that

traveller 42 has left. This happens because traveller 42 expired the maximum travel time (1 hour), so it is assumed that he left the transportation system. Table 4.2 presents the members of the temporary network for Passenger 1 at 8:15, and, as in the previous sample, it is represented on a map in figure 4.18.

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
91	502	BCM1	FTM4	0%	82,1%
88	202	BCM3	BRP1	88,89%	0%
7918	203	BCM1	SRV1	0%	72,94%
43	202	BCM3	BRV3	100%	0%
113	502	BCM1	FTM4	0%	82,1%
5489	201	BS1	LDD1	0%	73,1%

Table 4.2: Traveller 1's temporary network with member relevance at 08:15

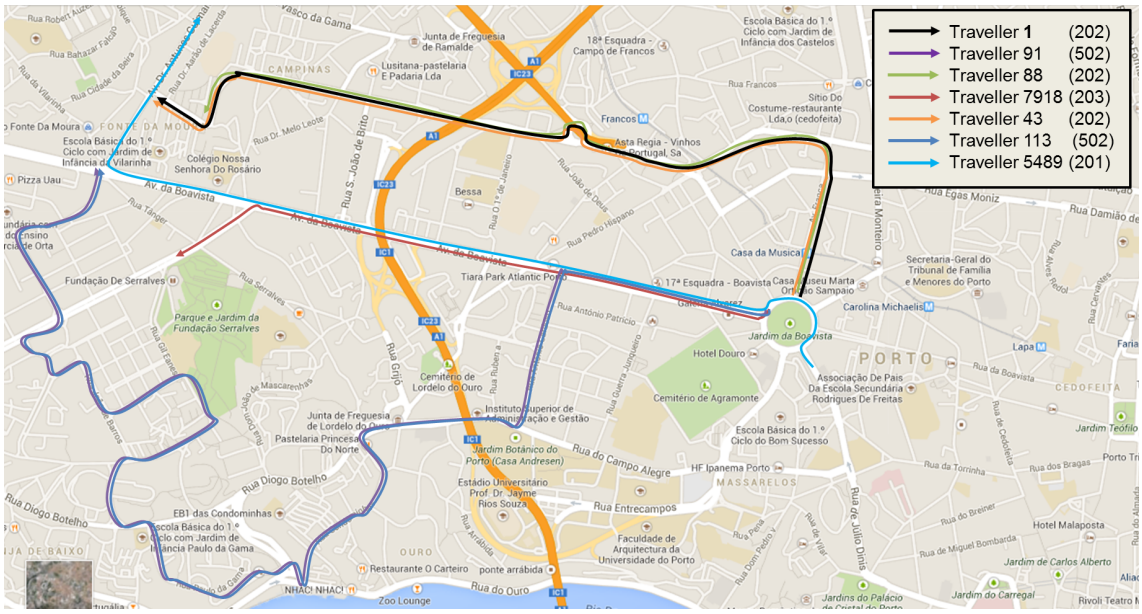

Figure 4.18: Map representation of traveller 1's temporary network connections at 08:15

At 08:30 the scenario has changed again, with travellers 7918 and 43 leaving the network (figure 4.12). At this time there is no new connection to passenger 1. Table 4.3 presents all the relevant connections at this time, visually represented in figure 4.19.

At 08:45 we encounter the scenario shown in figure 4.13. Here, some major changes can be seen, with another traveller leaving and six new travellers checking in on the network. At this moment we can see, supported by table 4.4, the current network is very different from the initial one.

In this temporary network we can observe that almost all the relevant passengers are on alternative routes, which shows the importance of the relevance type based on journey complementarity, providing alternative information to passenger 1. This can be easily observed in figure 4.20, as

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
91	502	BCM1	CART	0%	82,1%
88	202	BCM3	BRP1	88,89%	0%
113	502	BCM1	FTM4	0%	82,1%
5489	201	BS1	LDD1	0%	73,1%

Table 4.3: Traveller 1's temporary network with member relevance at 08:30

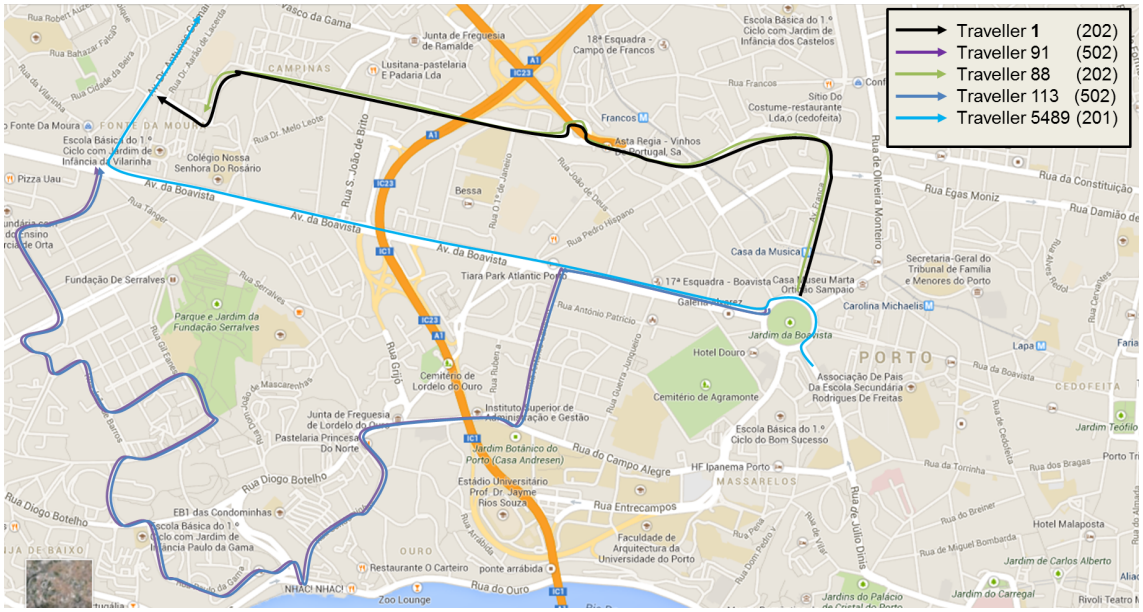

Figure 4.19: Map representation of traveller 1's temporary network connections at 08:30

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
88	202	BCM3	BRP1	88,89%	0%
113	502	BCM1	FTM4	0%	82,1%
5489	201	BS1	LDD1	0%	73,1%
5477	502	BCM1	FTM4	0%	82,1%
45	202	BCM3	SJB1	66,67%	0%
4575	502	BCM1	FTM4	0%	82,1%
4578	502	GJQ3	LGO2	0%	51,5%
4576	502	BCM1	FTM4	0%	82,1%
44	202	GCRT1	SJB1	66,67%	0%

Table 4.4: Traveller 1's temporary network with member relevance at 08:45

only three of the total 9 connections share road segments with traveller 1 and all the others are divided in different routes.

Finally, at 09:00, the final sample on this analysis, traveller 1 is already out of circulation, and so his temporary network is discarded and his connections destroyed. The travellers relevant to him, who remained in circulation, still have their own temporary networks with their connections,

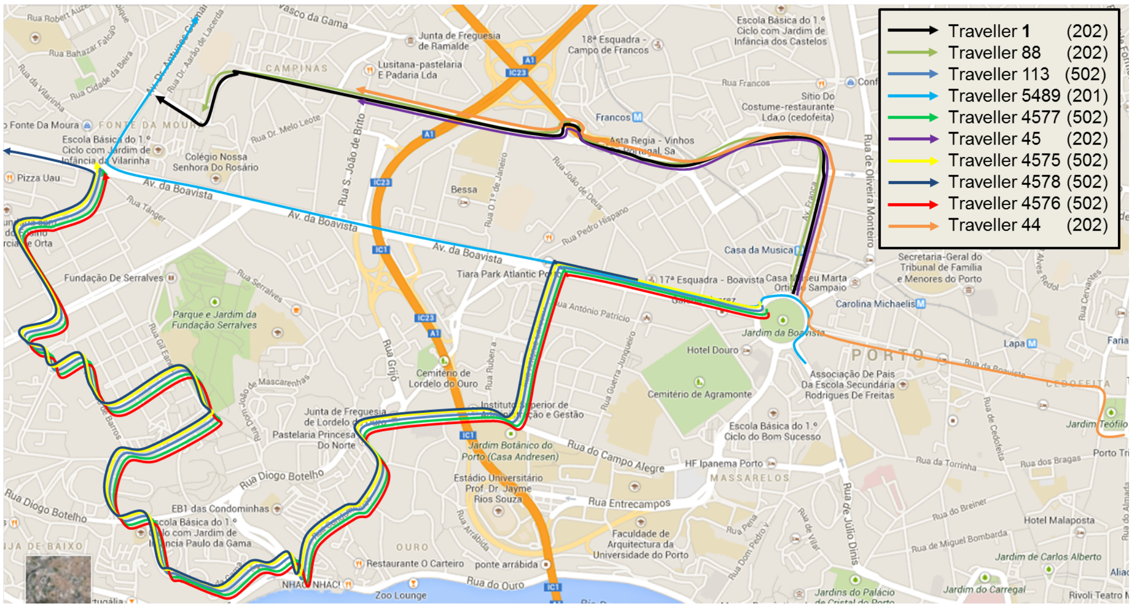

Figure 4.20: Map representation of traveller 1's temporary network connections at 08:45

since this only destroys 1's connections to them.

Concluding, from this analysis we can show that the user-centred temporary networks are highly dynamic and time dependent, since through the course of 15 minutes time stamps we could see some travellers connected to different ones in circulation at that time, with detailed demonstration for all the connections for "traveller 1".

From 08:00 to 09:00, for the analysed cluster almost the entire initial network was replaces by the new travellers, with a total of 14 travellers having similar travel paths, either by path similarity or route complementarity, until passenger 1 leaves the transportation network and his temporary network connections are discarded.

4.7.2 Five-day experience

In addition to the experience performed for one day on the weekend, which was focused on the visualization of the results, another experience was made. Starting on the day that followed the first experience - January 11^{th} - the objective was to identify if, during one weekday (from Monday to Friday), the same travellers can be found on "traveller 1"'s temporary network, in order identify patterns on temporary networks. The final results of the observation can be seen on table 4.5, at the end of this section.

Using the same set of inferred travels, one of them was arbitrarily chosen to feed to the algorithm in order to obtain, for each of the passengers circulating at that time, the ones that could provide the most relevant information considering their travel paths. The result was a total of 146440 connections between 3430 unique travellers (average of 42 connections per passenger). As in the previous experience, we selected arbitrarily one passenger (passenger 1) in order to illustrate the results for the temporary networks.

From the five days used, 3 different routes were identified from this passenger's travels, which can be visualized on the map in figure 4.21. We can see that, although running on different routes and/or destination stops, all the travels can be considered similar through the week, since the boarding and destination locations are always very close.

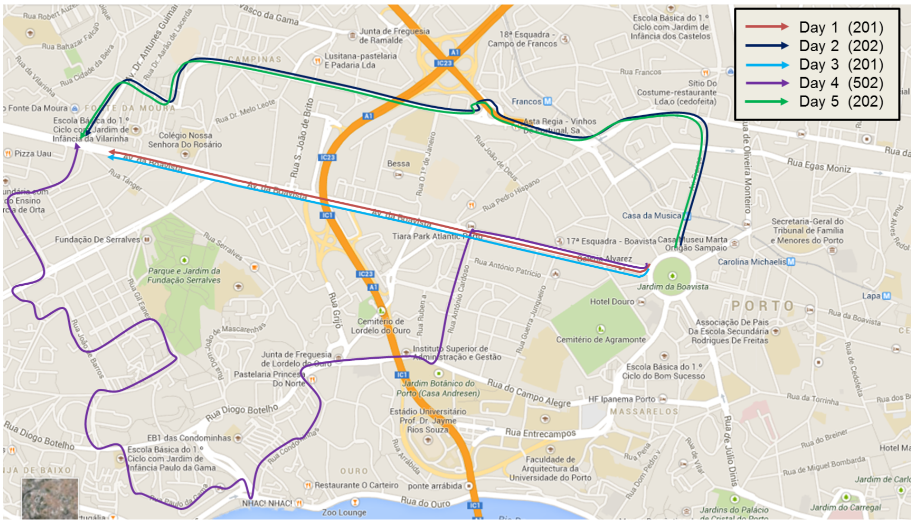

Figure 4.21: Map with travels performed by passenger 1 before 8:00 during the five days

Regarding day 1(out of 5), in this network, seen in figure 4.22, we have a set of 103 relevant travellers. As previously, each one of this connections has his own temporary network with its temporary connections. However, in this case due to the high number of interconnections we will only present the connections to traveller 1.

Advancing for the next day, 2 out of 5, we find that the network is composed by some travellers of the previous day, with addition of several new ones. Although the path travelled by passenger 1 is different from that of the previous day, some similarity can be found in it, since the new path is still complementary to the previous one. 31% of the passenger 1 network connections are the same of previous days, as seen in figure 4.23.

Figure 4.24 shows the temporary network on the 3^{rd} day, in which we find that, although several of the connections present in previous days no longer exist, a big part of them are still included in passenger 1 temporary network, which also contains several new connections. In total, around 56% of the network is composed by the travellers of the previous two days.

The day 4, Thursday, we can find several new passengers, along with a significant percentage of the same travellers from the three previous days - figure 4.25. On this day, 77% of passenger 1 temporary network is composed of passengers from the previous three days.

On the fifth and last day of this analysis, similarly to the previous observations, new users are found on the temporary network together with a large number of passengers from the previous days. From this day, 72% of the total connections were also present on the previous days. Figure 4.26 summarizes these results for the temporary networks of passenger 1, during 5 consecutive days at 8:00.

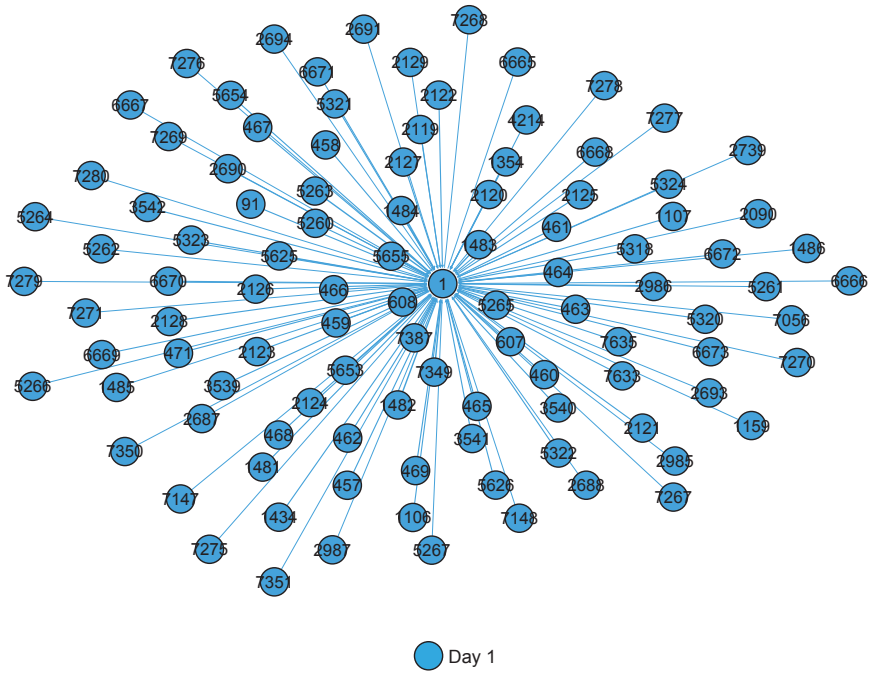

Figure 4.22: Temporary temporary network for traveller 1 at 08:00 on January 11st

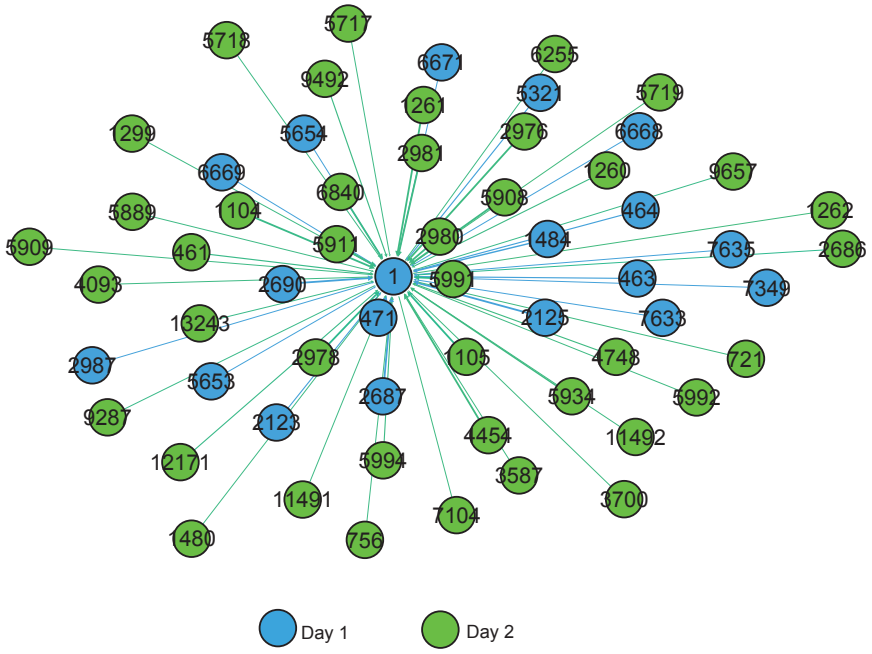

Figure 4.23: Temporary temporary network for traveller 1 at 08:00 on January 12nd

Concluding, through this analysis we were able to see that, during the course of five consecutive days, passenger 1 temporary networks at 08:00 were composed in a great part by connections that were also present in previous days. This suggests that finding travel patterns using temporary networks is a promising idea, that will allow that people who usually share the same travel paths

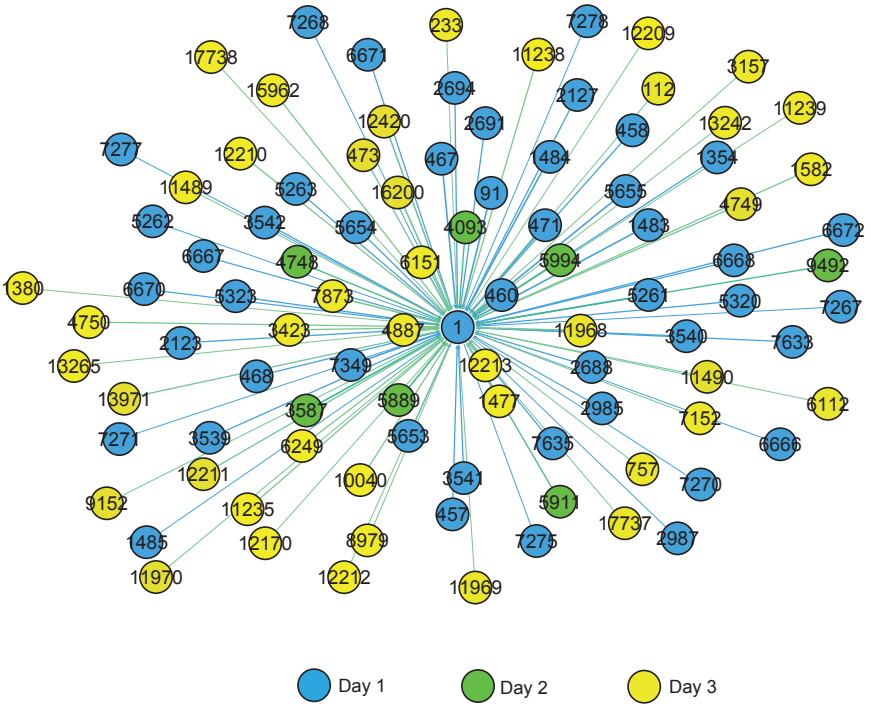

Figure 4.24: Temporary temporary network for traveller 1 at 08:00 on January 13rd

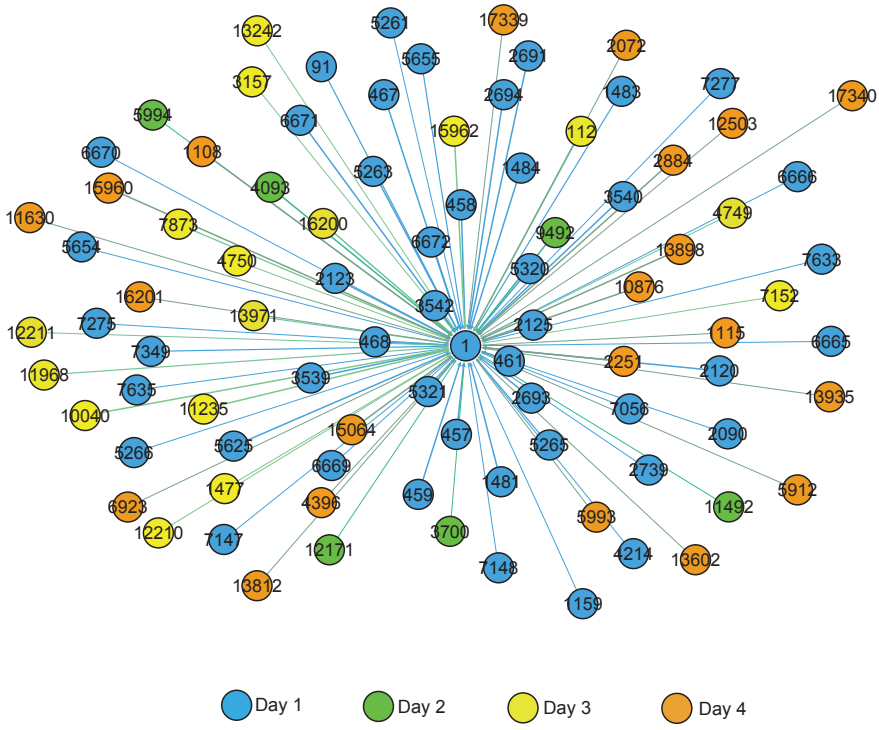

Figure 4.25: Temporary temporary network for traveller 1 at 08:00 on January 14th

on the same network at the course of time, could be sharing information through the *smartphone* [2] application prototype created for this project [Gon12].

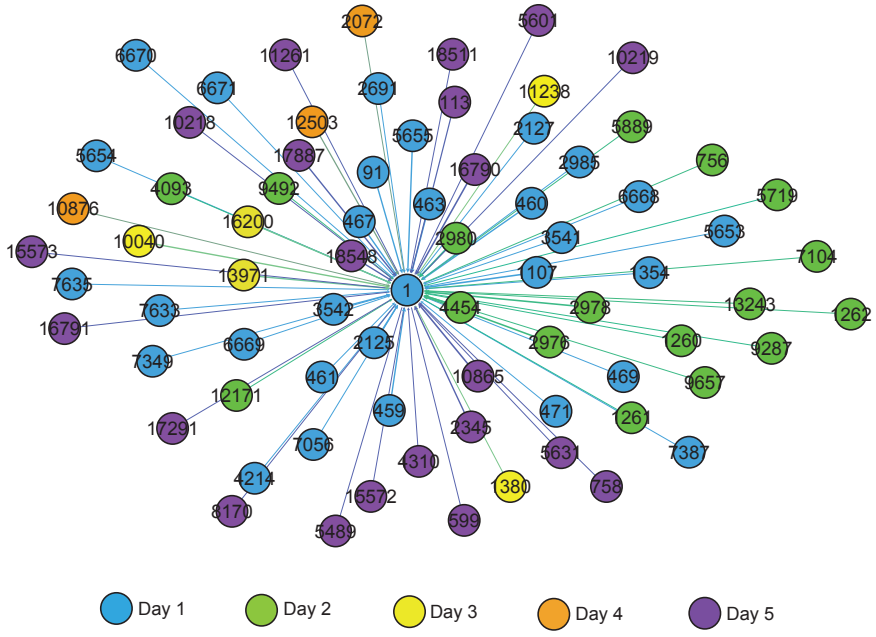

Figure 4.26: Temporary temporary network for traveller 1 at 08:00 on January 15^{th}

	% same passengers of day 1	% same passengers of day 2	% same passengers of day 3	% same passengers of day 4	Total % from previous days
Day 2	17,5%	-	-	-	31%
Day 3	45%	17,5%	-	-	56%
Day 4	45%	15%	50%	-	77%
Day 5	29%	42,5%	12%	14,3%	72%

Table 4.5: Temporary network patterns for the five-days experience

4.8 Summary and Conclusions

During each travel in public transports, several different passengers are using the same routes, at the same time, before or after. This way, bus riders can be, at a given time, having similar journeys, travelling in the same bus routes or in alternative ones.

These travellers can have important information related to these routes, and this knowledge may be shared with other travellers whose journeys are considered similar. This network of travellers must, thus, be composed by travellers relevant to each other.

The relevance among two travellers' journeys is measured through its similarity to each other, either because both bus routes have bus stops in common or because the two passengers are riding in alternative paths. When sharing bus stops or road segments, the relevance is measured through the number of stops (or total road segment distance) in comparison to the total number of stops (or total road distance) in the passenger's route. If the route is an alternative one, the relevance is measured on the boarding and destination bus stops. If the two passengers have both in common,

the journey is 100% relevant, otherwise the relevance is calculated proportionally to the distance between the stops - giving that they are at a walking distance, defined in this work as 640 meters or 8 minutes [fL10].

Knowing that each passenger has his own individual group of relevant passengers at a given time, each one's temporary network of connections must be unique to him. From here, the concept of user-centred temporary network is defined. Using the measurement of relevance, these networks are created, resulting in a set of temporary networks belonging to its respective travellers, connected to his most relevant users.

In the end, two different experiences were studied. First, analysing one passenger's temporary network at the course of one hour, in 15 minutes time spans, the conclusion was that his network was highly dynamic, with with new passengers joining the network and/or older ones leaving at every new 15 minute sample analysed. These results provide good perspectives for the future integration in a *smartphone* application to share public transport information among travellers [Gon12](described on section 2.2 on the state of the art).

Through the second experience, the behaviour of the same passenger's network at the course of one week, from Monday to Friday, was analysed. The results obtained lead to conclude that through the course of the five days a large amount of connections are preserved, and thus, through future work in this project [Nun12], travellers that usually share the same path could be automatically connected to the temporary networks in order to share information.

Chapter 5

Conclusions and Future Work

Mobility and the rising concept of always on-line social networking allow many possibilities in areas like public transports, since the travellers spend most of their travel time with their *smartphones*, connected to one or more social networks.

As urban areas increase in population, improving user's experience gains more importance, providing them easiness of use of public transports. At the same time, public transport companies must be encouraged to maintain and increase transport quality and quantity, and also rewarded for providing good services to the travellers.

In the context of a on going project [Nun12][NGCP11], a *smartphone* application was developed to allow sharing information on public transports [Gon12], which presents a change in the public transport information system, providing to travellers data that is fed to the system by the ones who know it better: themselves. The time and space efficiency of this information must be one of the main concerns, and so this work develops mechanisms that pretend to improve both time and space relevance of the information shared on the application through the creation of temporary networks composed by the most relevant passengers at a given time.

Using bus riders travel validations data provided by STCP for this research purpose, in a first phase this work aims to infer the travellers destinations in order to be able to create the passenger's travel history. To achieve this, an algorithm to infer the traveller path was implemented, with base on the work from Barry et al. [BNRS02], with innovative steps developed the context of this project [Nun12], and implemented as part of this thesis, such as zonal verification [NDGC14a]. Through the inference of a simulated dataset with foreseen destinations, the results show that, considering the assumptions made for the algorithm implementation, the travellers' destinations can be inferred with good accuracy.

From the inferred travels, passengers journeys could be extracted to begin the second phase of this work. Using the travellers boarding and destination locations, the goal was to create temporary networks [NGCP11] of travellers with similar travel paths. To this effect, the concept of relevance among travellers was studied and was decided as the measurement of how much one traveller

journey was similar to another. Based on these ideas, temporary networks were created for each of the travellers, since each of them has his own set of most relevant connections, distinct from the other travellers both in number and relevance score.

The final results were analysed and the conclusion was that temporary networks are highly dynamic, adapting to the travellers circulating at a given time and their journeys. In addition, the networks observed for one passenger at the course of one week are composed in a great part by the same travellers from previous days, giving good perspectives for future work of joining travellers on the same network with base on their usual patterns.

Finally, due to the identification of two types of relevance, based on similarity and complementarity, the networks created are enriched with travellers both from the same route and from alternative ones. Considering the future of this project [Nun12], this feature may be interesting, as way for the users to obtain information regarding alternative routes, in the event of problems that may happen, like high traffic, vehicle delays, among others.

To conclude, there is still a long way to work following this thesis results. The destination inference algorithm is featured to build the user travel history, only being able to estimate the destination of one journey from the previous travel data, and thus when checking in on a bus stop the application must infer the most likely destination from that user travel history.

Furthermore, the resulting temporary networks were created assuming that each passenger travelled for a maximum of one hour, but it could be less if another validation was made in the meanwhile. However, in the future, mechanisms that allow checking in and checking out on the application can be used to know if the passenger is still on the network, and thus the problem of not knowing if the traveller is still riding his vehicle can be solved.

Concerning the estimation of the passenger destination, another of the purposes of this work was to infer the arrival, in order to have the time each passenger left the system to improve the creation of the temporary networks. However, the results showed that only a small percentage of the inferred destinations had also the time of arrival, and so this stage was discarded from the algorithm. In the future, other possibilities must be studied to improve this work and accurately estimate, not only the destination, arrival time to it.

At the same time, integrating the temporary network creation mechanism with the application's journey planner may allow the passenger to receive relevant and real-time information before even checking in on the boarding stop, improving the traveller's final decision of going on the predicted journey or taking another route in case that the information received is recommends that change.

Finally, in the context of this project [Nun12], fully integrate these concepts with the public transport systems is one of the goals. Taking advantage of technologies like NFC [1], the check-in procedures on the application could be done automatically with the validation [FND13], providing immediate information regarding the boarding location and vehicle. This would allow integration

[1] Near Field Communication

of the zonal verification proposed on chapter 3, since only through this integration the passenger check-in can provide the number of zones allowed to travel.

References

[AFSG+08] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, and F. Herrera. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318, May 2008.

[BAMH85] Moshe Ben Akiva, P.P. Macke, and P.S. Hsu. Alternative Methods to Estimate Route-Level Trip Tables and Expand On-Board Surveys. *Transportation Research Record: Journal of the Transportation Research Board*, (1037):1–11, 1985.

[BH09] Mathieu Bastian and Sebastien Heymann. Gephi : An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media*, pages 361–362, 2009.

[BM01] Indranil Bose and Radha K. Mahapatra. Business data mining — a machine learning perspective. *Information & Management*, 39(3):211–225, December 2001.

[BNRS02] James J Barry, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board*, 1(1817):183–187, 2002.

[Dic14] Oxford Dictionaries. Social Network. Oxford Dictionaries, available in http://www.oxforddictionaries.com/definition/english/social-network, January 2014.

[fL10] Transports for London. Measuring Public Transport Accessibility Levels PTALs Summary. London Datastore, available in http://data.london.gov.uk/documents/PTAL-methodology.pdf, 2010.

[FND13] Marta Campos Ferreira, Henriqueta Nóvoa, and Teresa Galvão Dias. A Proposal for a Mobile Ticketing Solution for Metropolitan Area of Oporto Public Transport. *Lecture Notes in Business Information Processing*, 143:263–278, 2013.

[FS97] F Famili and WM Shen. Data pre-processing and intelligent data analysis. *International Journal on Intelligent Data Analysis*, 1(1), 1997.

[FS10] G Friedland and R Sommer. Cybercasing the joint: on the privacy implications of geo-tagging. *Proc. USENIX Workshop on Hot Topics in Security*, 2010.

[GB13] Moshe Givoni and David Banister. *Moving Towards Low Carbon Mobility*. Edward Elgar Publishing, 2013.

REFERENCES

[Gon12] Tiago Gonçalves. A smartphone application prototype for exchanging valuable real time public transport information among travellers, 2012.

[Gor12] Jason B Gordon. *Intermodal Passenger Flows on London's Public Transport Network: automated inference of full passenger journeys using fare-transaction and vehicle-location data*. PhD thesis, Massachusetts Institute of Technology, 2012.

[Jia06] Jian Pei Jiawei Han, Micheline Kamber. *Data Mining : Concepts and Techniques Second Edition*. Morgan Kaufmann, 2006.

[Jor14] Stcp perde nove milhões de passageiros em seis meses. Jornal de Negócios, available in http://www.jornaldenegocios.pt/empresas/detalhe/stcp_perde_nove_milhoes_de_passageiros_em_seis_meses.html, January 2014.

[JRK13] Krzysztof Janowicz, Martin Raubal, and Werner Kuhn. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, (2):29–57, 2013.

[Kan11] Mehmed Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons, 2011.

[LCL13] I Lee, G Cai, and K Lee. Mining points-of-interest association rules from geo-tagged photos. *System Sciences (HICSS), 2013 46th Hawaii International Conference on System Sciences*, (12):1580–1588, 2013.

[LPSZ10] A Lenhart, K Purcell, A Smith, and K Zickuhr. Social media & mobile internet use among teens and young adults. *Washington, DC: Pew internet & american life project*, 2010.

[MH09] HJ Miller and J Han. *Geographic data mining and knowledge discovery*. CRC Press, 2009.

[Mit69] J.C. Mitchell. Social networks in urban situations: analyses of personal relationships in Central African towns. *Manchester University Press*, 1969.

[MWW+13] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining Smart Card Data for Transit Riders' Travel Patterns. *Transportation Research Board 92nd Annual Meeting*, 13(3460), 2013.

[NDGC14a] António A. Nunes, Luís Dias, Teresa Galvão, and João Falcão E Cunha. Reimagining the Social Network for Urban Public Transport. *IEMS '14 - 5th Industrial Engineering and Management Symposium*, 2014.

[NDGC14b] António A. Nunes, Luís Dias, Teresa Galvão, and João Falcão E Cunha. Reimagining the social network for urban public transport.[powerpoint slides]. *IEMS '14 - 5th Industrial 22 Engineering and Management Symposium*, January 2014.

[NGC12] António A. Nunes, Teresa Galvão, and João Falcão E Cunha. Public Transport Service Co-Creation: a Social Media Based Framework. *The Art & Science of Service Meeting*, 2012.

REFERENCES

[NGCP11] António A. Nunes, Teresa Galvão, João Falcão E Cunha, and Jeremy V Pitt. Using Social Networks for Exchanging Valuable Real Time Public Transport Information among Travellers. *2011 IEEE 13th Conference on Commerce and Enterprise Computing*, pages 365–370, September 2011.

[NGG13] António A. Nunes, Tiago Gonçalves, and Teresa Galvão. A Prototype for Public Transport Service Co-Creation Using Social Media: Results from Usability Testing. *International Conference on Exploring Service Science*, (1.3), 2013.

[Nun12] António A. Nunes. *An Innovative Model for Pervasive Mobile Computing Based Services to Improve Urban Public Transport Experience: PhD Research Proposal*. PhD thesis, Faculdade de Engenharia da Universidade do Porto, 2012.

[PRA10] GS Prasad, NVS Reddy, and UD Acharya. Knowledge discovery from web usage data: A survey of web usage pre-processing techniques. *Information Processing and Management*, pages 505–507, 2010.

[PYKL08] Park, Jin Young, Dong-Jun Kim, and Yongtaek Lim. Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea. *Trans- portation Reserach Record*, (2063):3–9, 2008.

[SCH08] A SCHWERING. Approaches to semantic similarity measurement for geo-spatial data—a survey. *Transactions in GIS 12*, (1):5–29, 2008.

[Sim10] Jesse. Simon. Origin/Destination Applications from Smart Card Data. *Los Angeles County Metropolitan Transportation Authorit*, 2010.

[SjRkG13] N Suthar, I jeet Rajput, and V kumar Gupta. A Technical Survey on DBSCAN Clustering Algorithm. *International Journal of Scientific & Engineering Research*, 4(5), 2013.

[SMS+11] Mitsuhisa Sato, Satoshi Matsuoka, Peter M. Sloot, G. Dick van Albada, Jack Dongarra, Michal Munk, and Martin Drlík. Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System. *Procedia Computer Science*, 4:1640–1649, 2011.

[STC14] Preços - stcp. STCP, available in www.stcp.pt/pt/viajar/tarifas/precos/, January 2014.

[Sto08] Peter R. Stopher. The Travel Survey Toolkit: Where To From Here. *International Conference on Travel Survey Methods, Annecy*, 2008.

[ZLZC11] YT Zheng, Y Li, ZJ Zha, and TS Chua. Mining travel patterns from GPS-tagged photos. *Advances in Multimedia Modeling*, 2011.

[ZN14] Hugo Zaragoza and Marc Najork. Web search relevance ranking. Microsoft Research, available in www.stcp.pt/pt/viajar/tarifas/precos/, January 2014.

[ZRW07] Jinhua Zhao, Adam Rahbee, and Nigel H. M. Wilson. Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):376–387, July 2007.

REFERENCES

Appendix A

Data analysis - Porto public transport system

Porto's bus transport system is operated by STCP [1], servicing a total of 74 bus lines with two routes for each of them (one for each direction), excluding the circular ones (same station for start and end of journey). This service is one available public transport modes, with trains, metro and bus being the main ones, with bus sustaining a total of 41,2 millions of travellers in the first semester of 2013 [Jor14]. Porto's metropolitan area ticketing system is Andante [2], managed by TIP-Transportes Intermodais do Porto, being an entry-only system in which travellers only validate their journey at the start of the trip.

This work's validations data for the periods of January, April and May of 2010 was provided by STCP and OPT for research purposes, with nearly 30 million records. The first one, January, was prepared and used for the implemented algorithms, with around 8,3 million validations during this month. The Andante network data, including bus stops and routes, was collected in the project to which this thesis work relates to [Nun12].

A.1 Zonal system

Andante is designed on top of a zonal system, meaning that the entire area of use of this system is divided in smaller areas called zones. The entire Andante network is composed by 46 zones, and the STCP bus network works in 17 of them - seen in A.1.

Each traveller has one Andante card, occasional or signature, and to begin their trip travellers start with one validation of their card at the entry point, with a minimum time of travel available of one hour. This minimum travel is denominated Z2, corresponding to a restriction of a maximum of two zones of travel. The needed travel type is calculated with base on the entry zone, forming a ring of zones around it A.2, being a Z2 the needed title for the ring adjacent to the entry zone, Z3 for the next adjacent ring, etc.

[1] STCP
[2] http://www.linhandante.com/

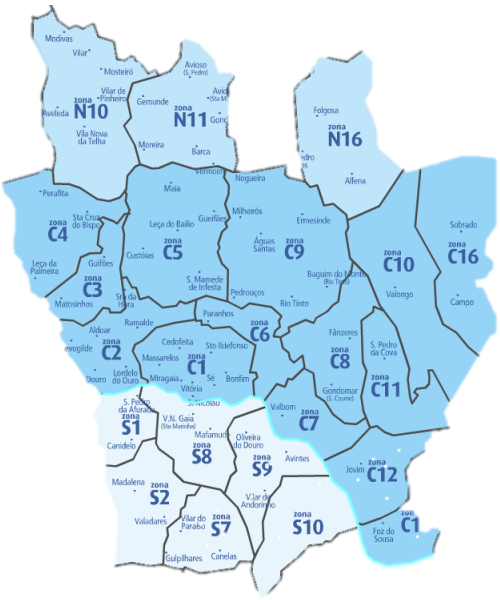

Figure A.1: Andante zones worked by STCP [STC14]

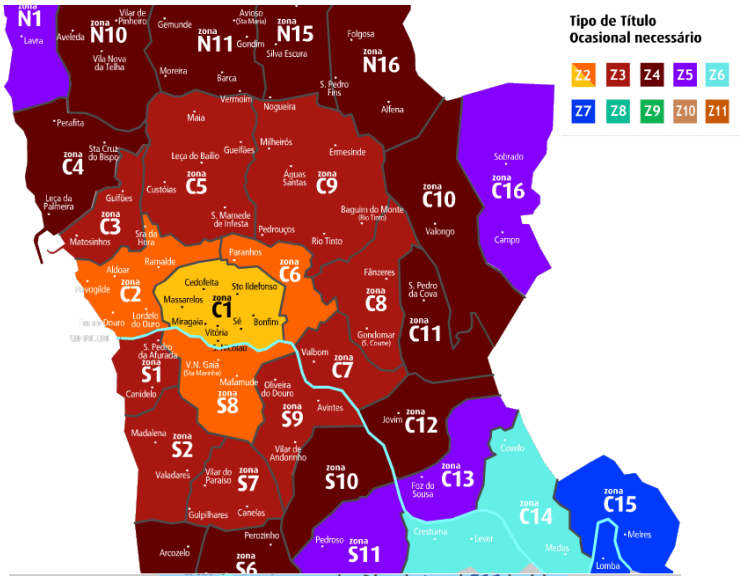

Figure A.2: Andante zoning rings from zone C1 [STC14]

A.1.1 Statistics and data analysis

Before deeper analysis on January's data, it's important to understand the real impact of the non-Andante validations on the set. Only 35% of January's validations are Andante - meaning that 65% of validations are from ticketing systems that are no longer run on the STCP bus system (figure A.3). Looking at real numbers, on figure A.4 we can see that the Andante zone **C1** has predominance on validations over all the other Andante zones. However, if non-Andante data was present it would overshadow this statistic, since it contains almost 3.5 million validations (opposing nearly 800000 validations from **C1**).

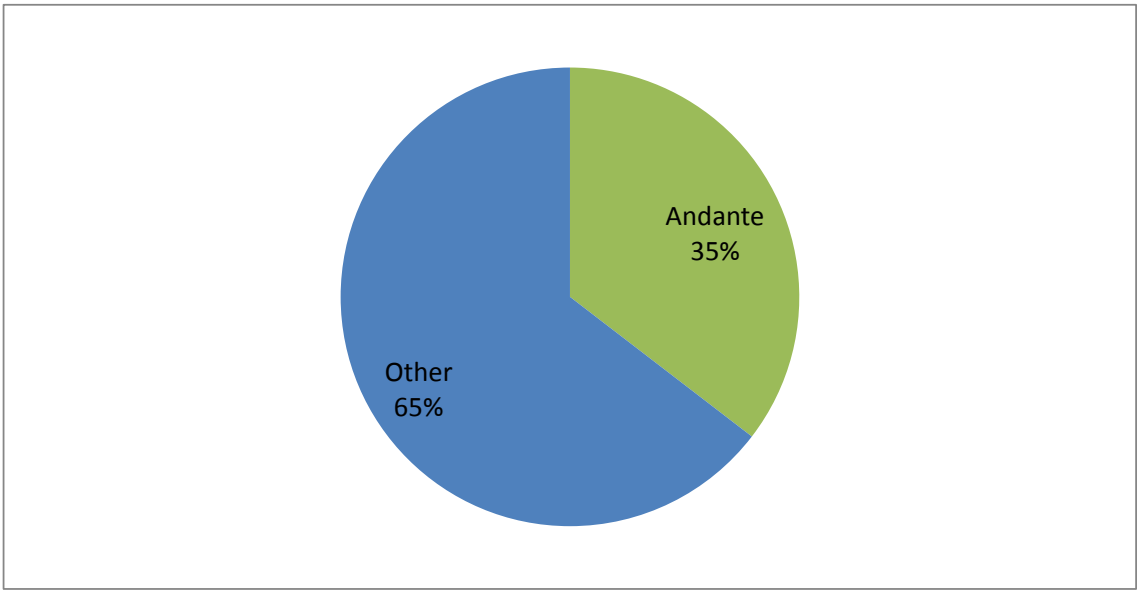

Figure A.3: Validations with Andante cards versus other cards on January

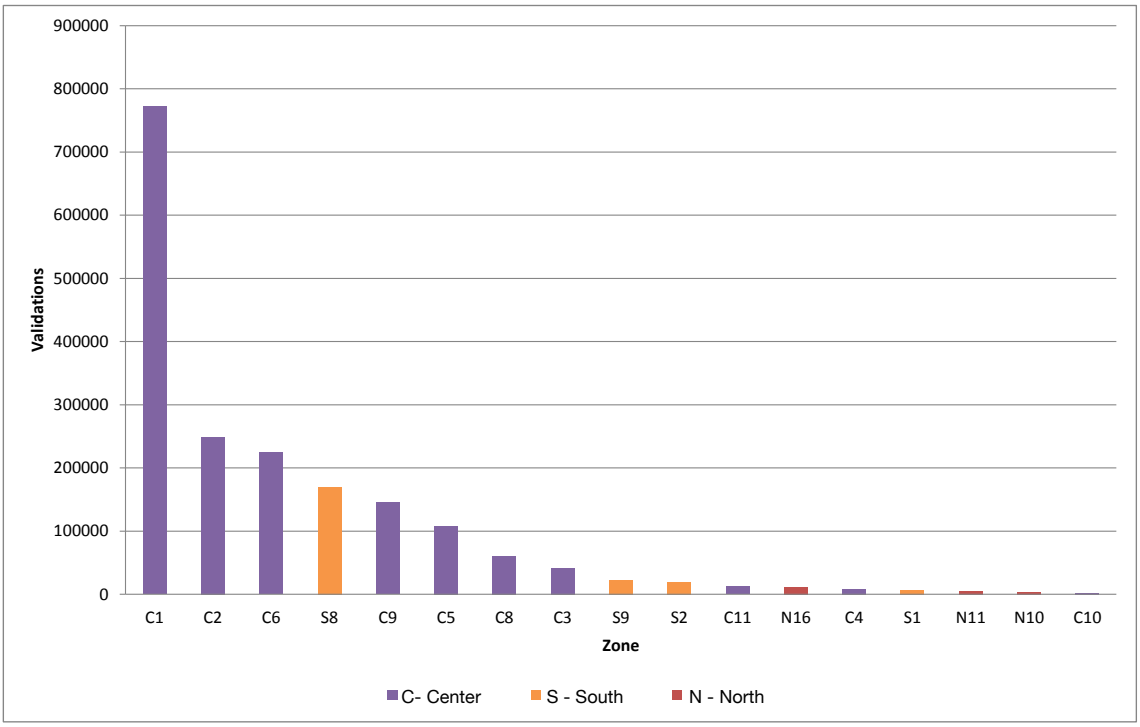

Figure A.4: Validations per zone on January

Figure A.5 shows how many validations occurred per day on January, both for all the dataset and Andante-only. As expected, the amount of validations has little variation during the week, except on weekends, where a big drop of usage occurs, specially on Sunday. It's also noteworthy that the first three days of the month, after the new years eve, have a considerably low amount of travellers. Through this comparison it's visible that behaviour is approximately the same, showing

that from this perspective we can rely on the subset of Andante validations to use on this thesis work.

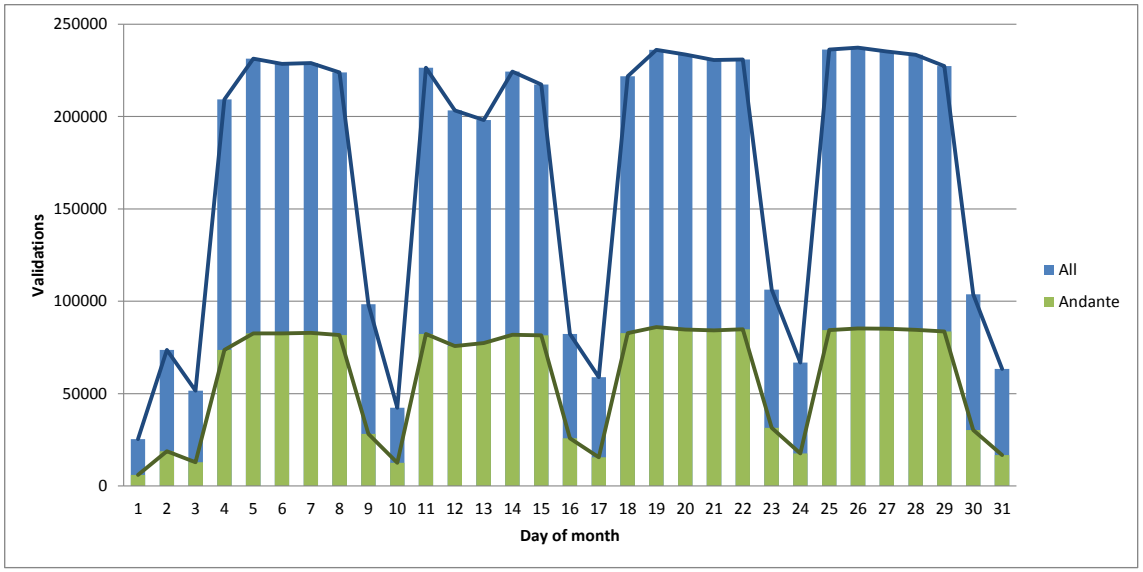

Figure A.5: Validations from January

Regarding user variety, a similar experience was made. Comparing the results of the number of users per day on all the dataset with Andante only users (figure A.6), we concluded that the behaviour throughout the month is similar between the two, with a natural decrease regarding the number of users, from an average of 56217 users per day on all validations, to 21946 on the Andante system.

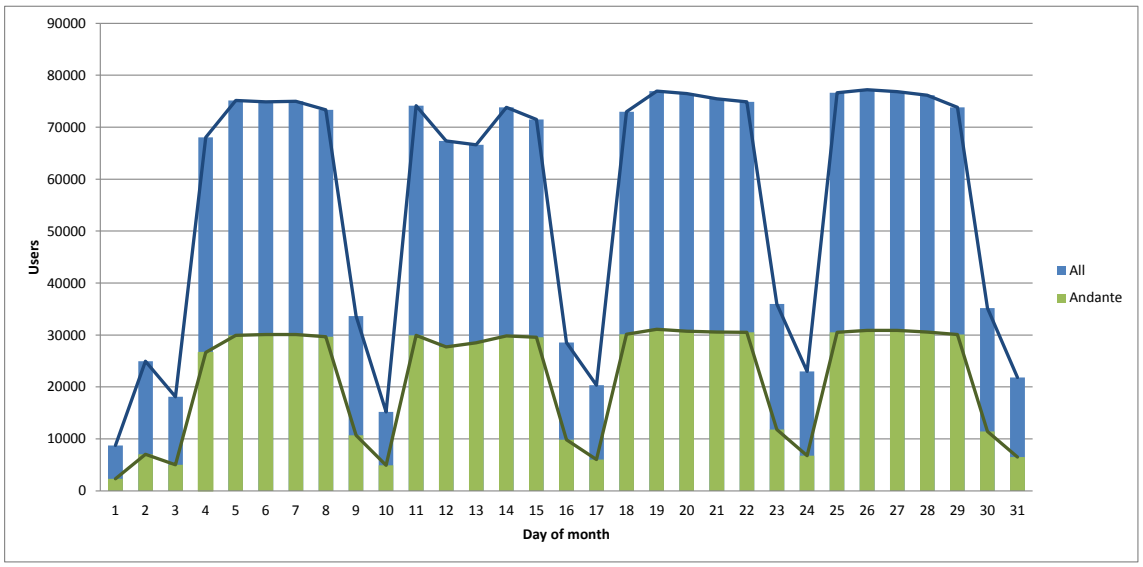

Figure A.6: Unique users per day on January

Furthermore, to present the variety necessary to provide good data for our algorithms, another necessary analysis was concerning the number of different lines and stops covered by January's

data. As seen on figures A.7 and A.8, there's similarity both in terms of bus lines and bus stops when comparing between all the validations and Andante cards only, with a slight decrease in number of stops in general.

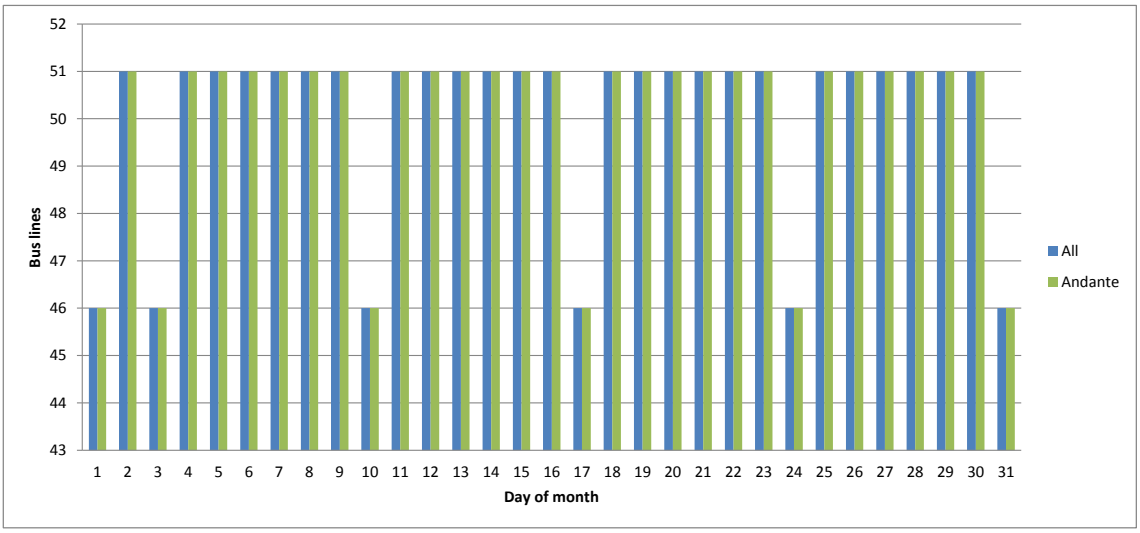

Figure A.7: Travelled routes on January

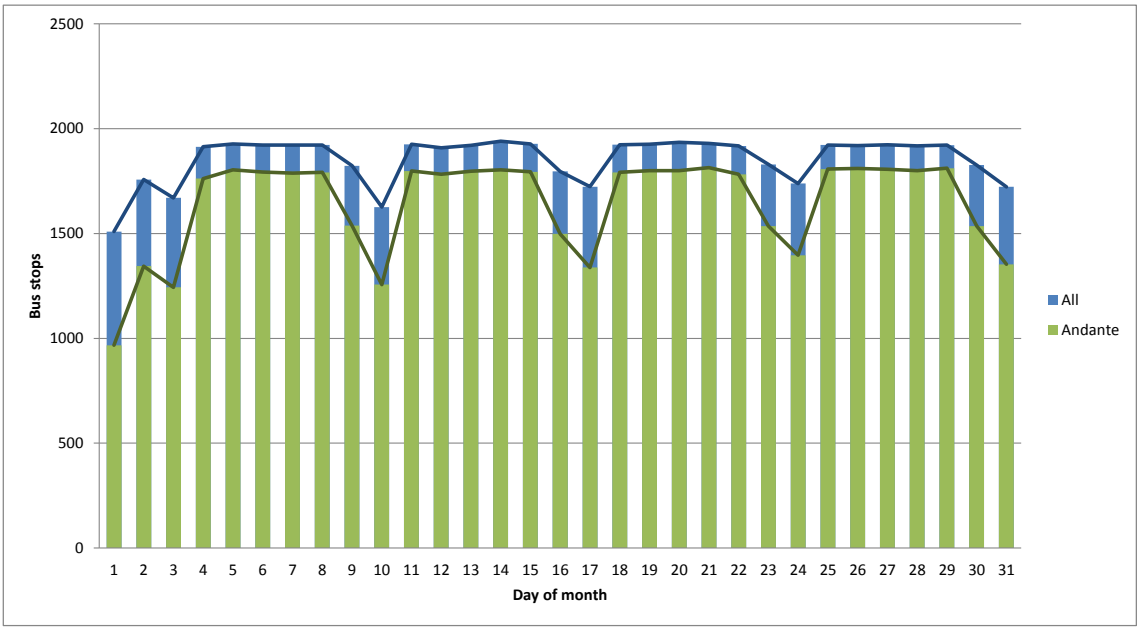

Figure A.8: Bus stops with validations on January

Hereupon, and because our final goal is to create networks of users with similar travel routines, the amount of users connected through the different stages of the day can impact the creation of those networks - poorly crowded time intervals suggested a reduction in the range and diversity of the networks. Therefore, further analysis on this behaviour was one of the major concerns. In this case, presented in figure A.9, the behaviour changes on some of the hour intervals, with reduced significance on the afternoon (*14h-17h*). However, the significance in as nearly all the

other time intervals remains proportional, suggesting that the impact of the significance reduction on the afternoon wouldn't have serious consequences on the inference quality.

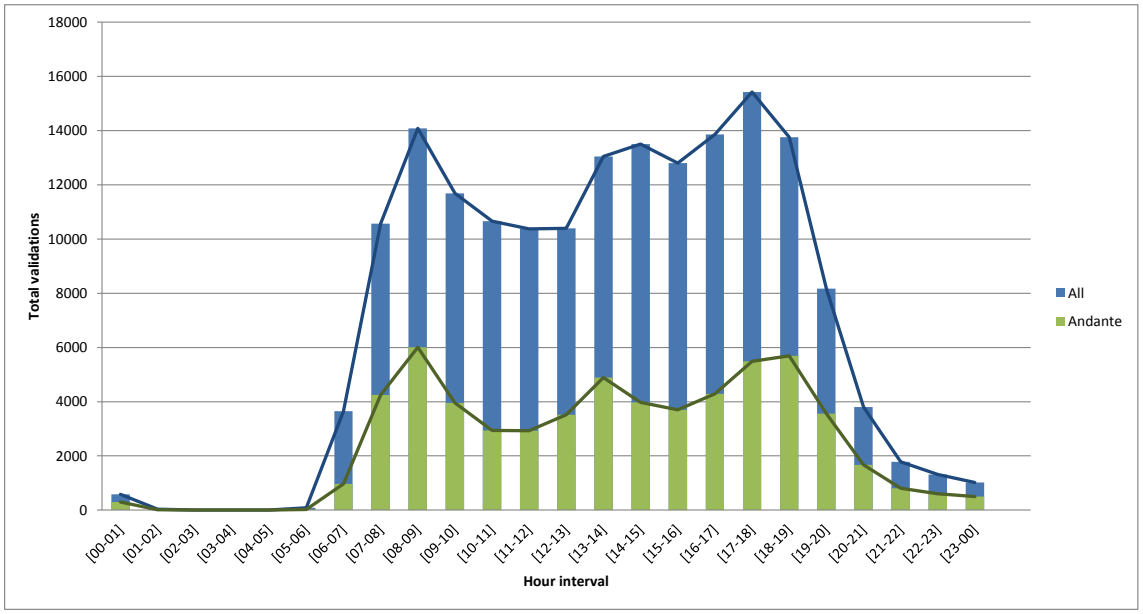

Figure A.9: Average validations per hour on January

Summarizing, through several analysis and comparison it was concluded that the usage of only Andante validations in this work was the most reliable option. The passengers' behaviour was found similar to the full set, and the zoning verification would become severely less reliable when using the full dataset. Thus, 65% of the collected data was considered was discarded from further processing. The result was a subset of Andante-only validations, used on this thesis work.

Appendix B

Temporary networks

In order to visually evaluate the travellers networks in this document one temporary network was studied with more detail in chapter 4.

The first experience, detailed in section 4.7.1, presents the case of one travellers temporary network and its set of connections. Since all the connections to the traveller analysed have also their own temporary connections, the behaviour of those travellers own temporary networks and its own connections was also studied.

The five-days experience, in section 4.7.2, the behaviour of the same traveller's temporary network was also analysed, but this time to the same hour instante (08:00) in five different days, from Monday to Friday. In that experience we only present a visualization of the data, due to the long set of connections, but detailed results on those connections relevance score were also obtained, and are here presented.

Concluding, this appendix provides deeper analysis on each of the previously seen travellers own networks and, for each of them, its connections' relevance scores, both for similarity and complementarity, with travel paths and route indications. Following the same behaviour from the previous analysis, this will compose details for samples 08:00, 08:15, 08:30, 08:45 and 09:00 on the single-day experience and samples from 08:00 for the five days 11, 12, 13, 14 and 15 of January.

B.1 Single-day temporary networks

B.1.1 Temporary networks at 08:00

Traveller 1 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
91	502	BCM1	FTM4	0%	82,1%
42	202	TRD1	FTM2	100%	0%
88	202	BCM3	BRP1	88,89%	0%
7918	203	BCM1	SRV1	0%	72,94%
43	202	BCM3	BRV3	100%	0%

Table B.1: Traveller 1's temporary network with member relevance at 08:00

Traveller 43 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	100.0%	0.0%
42	202	TRD1	FTM2	100.0%	0.0%
91	502	BCM1	FTM4	0.0%	82.1%
7918	203	BCM1	SRV1	0.0%	72.94%
88	202	BCM3	BRP1	88.89%	0.0%

Table B.2: Traveller 43's temporary network with member relevance at 08:00

Traveller 91 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	82.1%
43	202	BCM3	BRV3	0.0%	82.1%
7918	203	BCM1	SRV1	70.0%	81.53%
4396	203	MPL3	PRI1	70.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.3: Traveller 91's temporary network with member relevance at 08:00

Traveller 42 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	52.94%	0.0%
43	202	BCM3	BRV3	52.94%	0.0%

Table B.4: Traveller 42's temporary network with member relevance at 08:00

Traveller 7918 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	72.94%
43	202	BCM3	BRV3	0.0%	72.94%
91	502	BCM1	FTM4	77.78%	81.53%
4396	203	MPL3	PRI1	100.0%	0.0%
88	202	BCM3	BRP1	0.0%	75.52%

Table B.5: Traveller 7918's temporary network with member relevance at 08:00

Traveller 88 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
511	507	BCM3	FMAI	50.0%	0.0%
1	202	BCM3	BRV3	100.0%	0.0%
510	507	BCM3	EXP6	50.0%	0.0%
509	507	BCM3	EXP2	50.0%	0.0%
43	202	BCM3	BRV3	100.0%	0.0%
42	202	TRD1	FTM2	100.0%	0.0%
91	502	BCM1	FTM4	0.0%	78.06%
7918	203	BCM1	SRV1	0.0%	75.52%

Table B.6: Traveller 88's temporary network with member relevance at 08:00

B.1.2 Temporary networks at 08:15

Traveller 1 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
43	202	BCM3	BRV3	100.0%	0.0%
113	502	BCM1	FTM4	0.0%	82.1%
91	502	BCM1	FTM4	0.0%	82.1%
5489	201	BS1	LDD1	0.0%	73.11%
7918	203	BCM1	SRV1	0.0%	72.94%
88	202	BCM3	BRP1	88.89%	0.0%

Table B.7: Traveller 1's temporary network with member relevance at 08:15

Traveller 43 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	100.0%	0.0%
113	502	BCM1	FTM4	0.0%	82.1%
91	502	BCM1	FTM4	0.0%	82.1%
5489	201	BS1	LDD1	0.0%	73.11%
7918	203	BCM1	SRV1	0.0%	72.94%
88	202	BCM3	BRP1	88.89%	0.0%

Table B.8: Traveller 43's temporary network with member relevance at 08:15

Traveller 113 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
43	202	BCM3	BRV3	0.0%	82.1%
91	502	BCM1	FTM4	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
7918	203	BCM1	SRV1	70.0%	81.53%
4396	203	MPL3	PRI1	70.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.9: Traveller 113's temporary network with member relevance at 08:15

Traveller 91 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
43	202	BCM3	BRV3	0.0%	82.1%
113	502	BCM1	FTM4	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
7918	203	BCM1	SRV1	70.0%	81.53%
4396	203	MPL3	PRI1	70.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.10: Traveller 91's temporary network with member relevance at 08:15

Traveller 5489 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	73.11%
112	502	BCM1	PRCD3	64.29%	0.0%
43	202	BCM3	BRV3	0.0%	73.11%
113	502	BCM1	FTM4	64.29%	0.0%
91	502	BCM1	FTM4	64.29%	0.0%
7918	203	BCM1	SRV1	50.0%	0.0%
4396	203	MPL3	PRI1	50.0%	0.0%
88	202	BCM3	BRP1	0.0%	72.11%

Table B.11: Traveller 5489's temporary network with member relevance at 08:15

Traveller 7918 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	72.94%
112	502	BCM1	PRCD3	77.78%	0.0%
43	202	BCM3	BRV3	0.0%	72.94%
113	502	BCM1	FTM4	77.78%	81.53%
91	502	BCM1	FTM4	77.78%	81.53%
5489	201	BS1	LDD1	77.78%	0.0%
4396	203	MPL3	PRI1	100.0%	0.0%
88	202	BCM3	BRP1	0.0%	75.52%

Table B.12: Traveller 7918's temporary network with member relevance at 08:15

Traveller 88 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
511	507	BCM3	FMAI	50.0%	0.0%
1	202	BCM3	BRV3	100.0%	0.0%
510	507	BCM3	EXP6	50.0%	0.0%
509	507	BCM3	EXP2	50.0%	0.0%
43	202	BCM3	BRV3	100.0%	0.0%
113	502	BCM1	FTM4	0.0%	78.06%
91	502	BCM1	FTM4	0.0%	78.06%
5489	201	BS1	LDD1	0.0%	72.11%
7918	203	BCM1	SRV1	0.0%	75.52%

Table B.13: Traveller 88's temporary network with member relevance at 08:15

B.1.3 Temporary networks at 08:30

Traveller 1 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
113	502	BCM1	FTM4	0.0%	82.1%
91	502	BCM1	FTM4	0.0%	82.1%
5489	201	BS1	LDD1	0.0%	73.11%
88	202	BCM3	BRP1	88.89%	0.0%

Table B.14: Traveller 1's temporary network with member relevance at 08:30

Traveller 113 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
91	502	BCM1	FTM4	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
4396	203	MPL3	PRI1	70.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.15: Traveller 113's temporary network with member relevance at 08:30

Traveller 91 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
113	502	BCM1	FTM4	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
4396	203	MPL3	PRI1	70.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.16: Traveller 91's temporary network with member relevance at 08:30

Traveller 5489 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	0.0%	73.11%
112	502	BCM1	PRCD3	64.29%	0.0%
113	502	BCM1	FTM4	64.29%	0.0%
91	502	BCM1	FTM4	64.29%	0.0%
4396	203	MPL3	PRI1	50.0%	0.0%
88	202	BCM3	BRP1	0.0%	72.11%

Table B.17: Traveller 5489's temporary network with member relevance at 08:30

Traveller 88 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
511	507	BCM3	FMAI	50.0%	0.0%
1	202	BCM3	BRV3	100.0%	0.0%
510	507	BCM3	EXP6	50.0%	0.0%
509	507	BCM3	EXP2	50.0%	0.0%
113	502	BCM1	FTM4	0.0%	78.06%
91	502	BCM1	FTM4	0.0%	78.06%
5489	201	BS1	LDD1	0.0%	72.11%

Table B.18: Traveller 88's temporary network with member relevance at 08:30

B.1.4 Temporary networks at 08:45

Traveller 1 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
4575	502	BCM1	FTM4	0.0%	82.1%
3279	502	GJQ3	LGO2	0.0%	51.5%
45	202	BCM3	SJB1	66.67%	0.0%
113	502	BCM1	FTM4	0.0%	82.1%
44	202	GCRT1	SJB1	66.67%	0.0%
5489	201	BS1	LDD1	0.0%	73.11%
4577	502	BCM1	FTM4	0.0%	82.1%
4576	502	BCM1	FTM4	0.0%	82.1%
88	202	BCM3	BRP1	88.89%	0.0%

Table B.19: Traveller 1's temporary network with member relevance at 08:45

Traveller 4575 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
3279	502	GJQ3	LGO2	80.0%	0.0%
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
113	502	BCM1	FTM4	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
4577	502	BCM1	FTM4	100.0%	0.0%
4576	502	BCM1	FTM4	100.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.20: Traveller 4575's temporary network with member relevance at 08:45

Traveller 3279 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
4575	502	BCM1	FTM4	88.89%	0.0%
1	202	BCM3	BRV3	0.0%	51.5%
112	502	BCM1	PRCD3	100.0%	0.0%
113	502	BCM1	FTM4	88.89%	0.0%
5489	201	BS1	LDD1	77.78%	0.0%
4577	502	BCM1	FTM4	88.89%	0.0%
4576	502	BCM1	FTM4	88.89%	0.0%

Table B.21: Traveller 3279's temporary network with member relevance at 08:45

Traveller 45 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	100.0%	0.0%
44	202	GCRT1	SJB1	100.0%	0.0%
88	202	BCM3	BRP1	100.0%	0.0%

Table B.22: Traveller 45's temporary network with member relevance at 08:45

Traveller 113 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
4575	502	BCM1	FTM4	100.0%	0.0%
3279	502	GJQ3	LGO2	80.0%	0.0%
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
4577	502	BCM1	FTM4	100.0%	0.0%
4576	502	BCM1	FTM4	100.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.23: Traveller 113's temporary network with member relevance at 08:45

Traveller 44 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
1	202	BCM3	BRV3	50.0%	0.0%
45	202	BCM3	SJB1	50.0%	0.0%
88	202	BCM3	BRP1	50.0%	0.0%

Table B.24: Traveller 44's temporary network with member relevance at 08:45

Traveller 5489 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
4575	502	BCM1	FTM4	64.29%	0.0%
3279	502	GJQ3	LGO2	50.0%	0.0%
1	202	BCM3	BRV3	0.0%	73.11%
112	502	BCM1	PRCD3	64.29%	0.0%
113	502	BCM1	FTM4	64.29%	0.0%
4577	502	BCM1	FTM4	64.29%	0.0%
4576	502	BCM1	FTM4	64.29%	0.0%
88	202	BCM3	BRP1	0.0%	72.11%

Table B.25: Traveller 5489's temporary network with member relevance at 08:45

Traveller 4577 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
4575	502	BCM1	FTM4	100.0%	0.0%
3279	502	GJQ3	LGO2	80.0%	0.0%
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
113	502	BCM1	FTM4	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
4576	502	BCM1	FTM4	100.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.26: Traveller 4577's temporary network with member relevance at 08:45

Traveller 4576 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
4575	502	BCM1	FTM4	100.0%	0.0%
3279	502	GJQ3	LGO2	80.0%	0.0%
1	202	BCM3	BRV3	0.0%	82.1%
112	502	BCM1	PRCD3	100.0%	0.0%
113	502	BCM1	FTM4	100.0%	0.0%
5489	201	BS1	LDD1	90.0%	0.0%
4577	502	BCM1	FTM4	100.0%	0.0%
88	202	BCM3	BRP1	0.0%	78.06%

Table B.27: Traveller 4576's temporary network with member relevance at 08:45

Traveller 88 network

Traveller	Route	Origin	Destination	Relevance (similar)	Relevance (complementar)
4575	502	BCM1	FTM4	0.0%	78.06%
1	202	BCM3	BRV3	100.0%	0.0%
45	202	BCM3	SJB1	75.0%	0.0%
113	502	BCM1	FTM4	0.0%	78.06%
44	202	GCRT1	SJB1	75.0%	0.0%
5489	201	BS1	LDD1	0.0%	72.11%
4577	502	BCM1	FTM4	0.0%	78.06%
4576	502	BCM1	FTM4	0.0%	78.06%

Table B.28: Traveller 88's temporary network with member relevance at 08:45

B.2 Five-day temporary networks

B.2.1 Traveller 1 temporary network on 11/01/2010

Table B.29: Traveller 1's temporary network with member relevance on 11/01/2010

Table B.29 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
2090	203	BCM1	PRI1	70.0%	0.0%
1106	202	BCM3	FTM2	0.0%	88.78%
3540	502	BCM1	MAL2	90.0%	0.0%
1107	202	BCM3	FTM2	0.0%	88.78%
3541	502	BCM1	PRCD3	90.0%	0.0%
7349	201	PAL3	PRO1	100.0%	0.0%
7351	201	ACRD1	EZC1	60.0%	0.0%
608	502	BLFZ1	MAL2	90.0%	0.0%
607	502	BLFZ1	MAL2	90.0%	0.0%
7350	201	BCM1	ACRD1	50.0%	0.0%
1481	201	CMO	GC1	90.0%	0.0%
1482	201	BS1	LDD1	100.0%	0.0%
1483	201	BS1	LDD1	100.0%	0.0%
1484	201	BCM1	PNV2	100.0%	0.0%
3539	502	JN3	LGO2	90.0%	0.0%
1486	201	BCM1	ACRD1	50.0%	0.0%
7280	203	GJQ3	GC11	50.0%	67.15%
1485	201	BCM1	FOCO1	70.0%	0.0%

Continued on next page

Table B.29: Traveller 1's temporary network with member relevance on 11/01/2010

Table B.29 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
4214	203	BCM1	GC11	70.0%	88.1%
3542	502	BCM1	FTM4	90.0%	89.81%
5653	201	BS1	GC1	90.0%	0.0%
471	502	PNV2	FTM4	10.0%	89.81%
1434	502	BCM1	FTM4	90.0%	89.81%
7633	203	BCM1	GC11	70.0%	88.1%
91	502	BCM1	FTM4	90.0%	89.81%
5626	502	PRR1	FTM4	90.0%	0.0%
5625	502	BLFZ1	FTM4	90.0%	0.0%
6667	203	BCM1	SGS1	70.0%	0.0%
6666	203	BCM1	PRI1	70.0%	0.0%
6669	203	BCM1	GC11	70.0%	88.1%
6668	203	BCM1	GC11	70.0%	88.1%
6673	203	BCM1	GC11	70.0%	88.1%
6670	203	BCM1	GC11	70.0%	88.1%
6671	203	BCM1	GC11	70.0%	88.1%
6672	203	BCM1	GC11	70.0%	88.1%
7056	201	AGM1	FTM1	80.0%	0.0%
5260	201	GGF	GC1	90.0%	0.0%
5263	201	BCM1	FTM1	100.0%	0.0%
5264	201	BCM1	ACRD1	50.0%	0.0%
5261	201	BS1	FOCO1	70.0%	0.0%
5262	201	BCM1	FOCO1	70.0%	0.0%
5324	502	AGM1	MAL2	80.0%	0.0%
5267	201	ACRD1	FTM1	60.0%	0.0%
6665	203	BCM1	PRI1	70.0%	0.0%
5265	201	BCM1	FTM1	100.0%	0.0%
5266	201	ACRD1	FTM1	60.0%	0.0%
2120	502	FIG	FTM4	90.0%	0.0%
5320	502	PRR1	PRCD3	90.0%	0.0%
5321	502	BCM1	LGO2	90.0%	76.19%
2122	502	BCM1	FTM4	90.0%	89.81%
5322	502	BCM1	FTM4	90.0%	89.81%
2121	502	CVA1	MAL2	90.0%	0.0%
5323	502	BCM1	FTM4	90.0%	89.81%

Continued on next page

Table B.29: Traveller 1's temporary network with member relevance on 11/01/2010

Table B.29 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
7635	203	BCM1	GC11	70.0%	88.1%
7387	201	CMO	PRO1	100.0%	0.0%
2987	202	BCM3	ABM1	0.0%	72.97%
5318	502	BLRB2	FTM4	90.0%	0.0%
2985	202	BCM3	BRP1	0.0%	84.47%
2119	502	PRR1	LGO2	90.0%	0.0%
2986	202	BCM3	BRP1	0.0%	84.47%
7147	203	MPL3	PRI1	70.0%	0.0%
7148	203	MPL3	SGS1	70.0%	0.0%
7279	203	BCM1	SGS1	70.0%	0.0%
7278	203	BCM1	SGS1	70.0%	0.0%
457	502	BLFZ1	NEV2	90.0%	0.0%
7275	203	BSB	PRI1	70.0%	0.0%
458	502	CVA1	MAL2	90.0%	0.0%
5654	201	BCM1	GC1	90.0%	0.0%
7277	203	BSB	GC11	70.0%	0.0%
5655	201	BCM1	PNV2	100.0%	0.0%
7276	203	BSB	SGS1	70.0%	0.0%
2129	502	AGM1	FTM4	80.0%	76.83%
7271	203	MPL3	PRI1	70.0%	0.0%
7270	203	MPL3	SGS1	70.0%	0.0%
2127	502	BCM1	PRCD3	90.0%	0.0%
459	502	BCM1	LGO2	90.0%	76.19%
2128	502	BCM1	LGO2	90.0%	76.19%
2125	502	BCM1	FTM4	90.0%	89.81%
2126	502	BCM1	FTM4	90.0%	89.81%
2123	502	BCM1	FTM4	90.0%	89.81%
2124	502	BCM1	LGO2	90.0%	76.19%
1159	201	BSS1	LDD1	50.0%	0.0%
2690	202	BCM3	FTM2	0.0%	88.78%
2691	202	BCM3	LGO4	0.0%	75.58%
2694	202	BCM3	LGO4	0.0%	75.58%
461	502	BCM1	PRCD3	90.0%	0.0%
2739	203	MPL3	GC11	70.0%	0.0%
2693	202	BCM3	LGO4	0.0%	75.58%

Continued on next page

Table B.29: Traveller 1's temporary network with member relevance on 11/01/2010

Table B.29 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
460	502	BCM1	FTM4	90.0%	89.81%
462	502	BCM1	PRCD3	90.0%	0.0%
463	502	BCM1	LGO2	90.0%	76.19%
7269	203	MPL3	SRV1	70.0%	0.0%
464	502	BCM1	MAL2	90.0%	0.0%
7268	203	MPL3	PRI1	70.0%	0.0%
2687	202	BCM3	LGO4	0.0%	75.58%
465	502	BCM1	FTM4	90.0%	89.81%
7267	203	MPL3	PRI1	70.0%	0.0%
2688	202	BCM3	LGO4	0.0%	75.58%
466	502	BCM1	PRCD3	90.0%	0.0%
467	502	BCM1	LGO2	90.0%	76.19%
468	502	BCM1	LGO2	90.0%	76.19%
469	502	BCM1	LGO2	90.0%	76.19%
1354	502	BCM1	PRCD3	90.0%	0.0%

2

B.2.2 Traveller 1 temporary network on 12/01/2010

Table B.30: Traveller 1's temporary network with member relevance on 12/01/2010

Table B.30 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
9287	501	BCM3	BFTM1	60.0%	0.0%
4454	202	BCM3	RCRT2	100.0%	0.0%
5994	201	BCM1	GC1	0.0%	80.28%
1104	202	TRD1	LNEV2	100.0%	0.0%
5911	202	BCM3	FTM2	100.0%	0.0%
5992	202	TRD1	SJB1	60.0%	0.0%
1105	202	BCM3	NEVG	100.0%	0.0%
5991	202	BCM3	FTM2	100.0%	0.0%
7349	501	PAL3	BFTM1	60.0%	0.0%
9492	502	BCM1	LGO2	0.0%	81.48%
1480	502	GJQ3	LGO2	0.0%	60.69%
5889	201	PRG1	FTM1	0.0%	73.81%

Continued on next page

Table B.30: Traveller 1's temporary network with member relevance on 12/01/2010

Table B.30 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
9657	501	BCM3	BFTM1	60.0%	0.0%
1484	201	BCM1	PNV2	0.0%	88.78%
5321	502	BCM1	LGO2	0.0%	81.48%
11492	203	BCM1	GC11	0.0%	76.91%
11491	203	BCM1	GC11	0.0%	76.91%
6255	202	TRD1	ABSS1	50.0%	0.0%
4748	201	BCM1	GC1	0.0%	80.28%
6840	202	TRD1	RCRT2	100.0%	0.0%
5653	201	BS1	GC1	0.0%	71.61%
471	202	BCM3	RCRT2	100.0%	0.0%
13243	203	BCM1	GC11	0.0%	76.91%
5934	201	BCM1	PNV2	0.0%	88.78%
2981	202	TRD1	LGO4	100.0%	0.0%
7635	203	BCM1	GC11	0.0%	76.91%
2987	202	BCM3	ABM1	70.0%	0.0%
3587	201	BS1	BRV1	0.0%	74.29%
5717	202	BCM3	ABM1	70.0%	0.0%
5718	202	SDP1	LGO4	70.0%	0.0%
5719	202	SJB1	FTM2	50.0%	0.0%
7104	501	BS1	PDVD4	60.0%	0.0%
12171	201	BCM1	GC1	0.0%	80.28%
2980	202	TRD1	LGO4	100.0%	0.0%
7633	203	BCM1	GC11	0.0%	76.91%
3700	502	GJQ3	LGO2	0.0%	60.69%
2976	202	TRD1	MSLD	100.0%	0.0%
5654	203	BCM1	GC11	0.0%	76.91%
721	501	BS1	IGAL2	60.0%	0.0%
2125	202	BCM3	FTM2	100.0%	0.0%
2978	202	TRD1	LNEV2	100.0%	0.0%
4093	203	BCM1	GC11	0.0%	76.91%
2123	201	BCM1	GC1	0.0%	80.28%
1261	202	BCM3	RCRT2	100.0%	0.0%
1262	202	SJB1	LGO4	50.0%	0.0%
5908	202	TRD1	LNEV2	100.0%	0.0%
5909	202	PRR1	ABM1	70.0%	0.0%

Continued on next page

Table B.30: Traveller 1's temporary network with member relevance on 12/01/2010

Table B.30 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
1260	202	CSBV	LNEV2	80.0%	0.0%
1299	202	SDP1	FTM2	70.0%	0.0%
2690	202	BCM3	FTM2	100.0%	0.0%
461	502	BCM1	LGO2	0.0%	81.48%
756	501	BCM3	PDVD4	60.0%	0.0%
6669	203	BCM1	GC11	0.0%	76.91%
6668	203	BCM1	GC11	0.0%	76.91%
2686	202	TRD1	SJB1	60.0%	0.0%
2687	202	BCM3	LGO4	100.0%	0.0%
6671	203	BCM1	GC11	0.0%	76.91%
469	502	BCM1	LGO2	0.0%	81.48%
1354	502	BCM1	LGO2	0.0%	81.48%

2

B.2.3 Traveller 1 temporary network on 13/01/2010

Table B.31: Traveller 1's temporary network with member relevance on 13/01/2010

Table B.31 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
1477	502	BCM1	FTM4	90.0%	89.81%
112	201	AGM1	FTM1	80.0%	0.0%
12210	201	BCM1	GC1	90.0%	0.0%
3540	502	BCM1	MAL2	90.0%	0.0%
5994	502	BCM1	FTM4	90.0%	89.81%
12211	201	BCM1	GC1	90.0%	0.0%
3541	502	BCM1	PRCD3	90.0%	0.0%
12212	201	BCM1	ACRD1	50.0%	0.0%
5911	202	BCM3	FTM2	0.0%	88.78%
12213	201	BCM1	FTM1	100.0%	0.0%
7349	201	PAL3	PRO1	100.0%	0.0%
12420	201	BCM1	FTM1	100.0%	0.0%
1582	502	BCM1	MAL2	90.0%	0.0%
12209	502	AGM1	PRCD3	80.0%	0.0%
11489	203	BCM1	SGS1	70.0%	0.0%

Continued on next page

Table B.31: Traveller 1's temporary network with member relevance on 13/01/2010

Table B.31 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
9492	502	BCM1	LGO2	90.0%	76.19%
5263	502	BCM1	PRCD3	90.0%	0.0%
5261	201	BS1	LDD1	100.0%	0.0%
5262	201	BCM1	FOCO1	70.0%	0.0%
1483	201	BS1	LDD1	100.0%	0.0%
5889	201	PRG1	FTM1	100.0%	0.0%
1484	201	BCM1	PNV2	100.0%	0.0%
3539	502	JN3	PRCD3	90.0%	0.0%
16200	201	BCM1	FTM1	100.0%	0.0%
11238	502	BCM1	LGO2	90.0%	76.19%
4749	201	BCM1	GC1	90.0%	0.0%
11235	502	BLRB2	PRCD3	90.0%	0.0%
1485	201	BCM1	FOCO1	70.0%	0.0%
6151	502	BCM1	FTM4	90.0%	89.81%
5320	502	PRR1	PRCD3	90.0%	0.0%
6112	502	GJQ3	LGO2	70.0%	55.23%
8979	201	BCM1	FOCO1	70.0%	0.0%
5323	502	BCM1	FTM4	90.0%	89.81%
473	502	PNV2	FTM4	10.0%	89.81%
11490	203	BCM1	SGS1	70.0%	0.0%
11239	502	ACRD1	LGO2	60.0%	0.0%
1380	201	AGM1	FTM1	80.0%	0.0%
4748	201	BCM1	GC1	90.0%	0.0%
4887	201	BCM1	FTM1	100.0%	0.0%
15962	201	BCM1	GC1	90.0%	0.0%
9152	203	BCG	PRI1	70.0%	0.0%
3542	502	BCM1	FTM4	90.0%	89.81%
5653	201	BS1	GC1	90.0%	0.0%
6249	502	BLFZ1	LGO2	90.0%	0.0%
471	201	BCM1	FTM1	100.0%	0.0%
4750	201	BCM1	GC1	90.0%	0.0%
13242	203	MPL3	PRI1	70.0%	0.0%
233	203	MPL3	GC11	70.0%	0.0%
7635	203	BCM1	GC11	70.0%	88.1%
2987	202	BCM3	ABM1	0.0%	72.97%

Continued on next page

Table B.31: Traveller 1's temporary network with member relevance on 13/01/2010

Table B.31 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
3587	201	BS1	BRV1	100.0%	0.0%
2985	202	BCM3	BRP1	0.0%	84.47%
13971	502	BCM1	FTM4	90.0%	89.81%
12170	201	BCM1	ACRD1	50.0%	0.0%
7633	203	BCM1	GC11	70.0%	88.1%
7278	203	BCM1	SGS1	70.0%	0.0%
11968	201	BS1	PRO1	100.0%	0.0%
11969	201	BCM1	ACRD1	50.0%	0.0%
7275	203	BSB	PRI1	70.0%	0.0%
457	502	BLFZ1	LGO2	90.0%	0.0%
458	502	CVA1	MAL2	90.0%	0.0%
7277	203	BSB	GC11	70.0%	0.0%
5654	201	BCM1	GC1	90.0%	0.0%
5655	201	BCM1	FTM1	100.0%	0.0%
7271	203	MPL3	PRI1	70.0%	0.0%
10040	502	BCM1	FTM4	90.0%	89.81%
7270	203	MPL3	SGS1	70.0%	0.0%
2127	502	BCM1	FTM4	90.0%	89.81%
4093	502	BCM1	FTM4	90.0%	89.81%
7873	201	BCM1	FTM1	100.0%	0.0%
2123	502	BLFZ1	FTM4	90.0%	0.0%
7152	203	BCM1	PRI1	70.0%	0.0%
13265	203	BCM1	SGS1	70.0%	0.0%
91	201	BCM1	PNV2	100.0%	0.0%
3157	203	BCM1	SGS1	70.0%	0.0%
6667	203	BCM1	GC11	70.0%	88.1%
17737	201	BCM1	ACRD1	50.0%	0.0%
6666	203	BCM1	PRI1	70.0%	0.0%
2691	502	BCM1	LGO2	90.0%	76.19%
17738	201	ACRD1	VIS5	60.0%	0.0%
2694	502	BCM1	LGO2	90.0%	76.19%
6668	201	BCM1	GC1	90.0%	0.0%
460	502	BCM1	LGO2	90.0%	76.19%
757	201	BCM1	BSS1	60.0%	0.0%
7268	203	MPL3	PRI1	70.0%	0.0%

Continued on next page

Table B.31: Traveller 1's temporary network with member relevance on 13/01/2010

Table B.31 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
7267	203	MPL3	PRI1	70.0%	0.0%
2688	502	BCM1	LGO2	90.0%	76.19%
6670	203	BCM1	GC11	70.0%	88.1%
467	201	BCM1	FTM1	100.0%	0.0%
6671	203	BCM1	GC11	70.0%	88.1%
468	502	BCM1	PRCD3	90.0%	0.0%
6672	201	BCM1	GC1	90.0%	0.0%
11970	201	BCM1	ACRD1	50.0%	0.0%
1354	502	BCM1	FTM4	90.0%	89.81%
3423	201	BCM1	GC1	90.0%	0.0%

2

B.2.4 Traveller 1 temporary network on 14/01/2010

Table B.32: Traveller 1's temporary network with member relevance on 14/01/2010

Table B.32 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
13812	201	PAL3	BSS1	50.0%	0.0%
2090	203	BCM1	PRI1	70.0%	0.0%
1477	201	BCM1	GC1	80.0%	81.15%
112	502	BCM1	PRCD3	100.0%	0.0%
1108	502	BCM1	FTM4	100.0%	0.0%
12210	201	BCM1	PINM1	70.0%	0.0%
3540	502	BCM1	MAL2	100.0%	0.0%
5994	201	BCM1	GC1	80.0%	81.15%
12211	201	BCM1	GC1	80.0%	81.15%
5993	203	BCM1	PRI1	70.0%	0.0%
2251	502	BCM1	LGO2	100.0%	0.0%
5912	203	MPL3	JBR1	70.0%	0.0%
7349	201	PAL3	PRO1	90.0%	0.0%
4396	203	MPL3	PRI1	70.0%	0.0%
13898	502	PRR1	FTM4	100.0%	0.0%
11630	203	MPL3	SRV1	70.0%	0.0%
9492	502	BCM1	LGO2	100.0%	0.0%

Continued on next page

Table B.32: Traveller 1's temporary network with member relevance on 14/01/2010

Table B.32 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
5263	502	BCM1	PRCD3	100.0%	0.0%
17340	203	MPL3	PRI1	70.0%	0.0%
5261	201	BS1	FOCO1	60.0%	0.0%
6665	203	BCM1	SGS1	70.0%	0.0%
1481	502	BCM1	FTM4	100.0%	0.0%
6923	201	BS1	BSS1	50.0%	0.0%
1483	201	BS1	EZC1	90.0%	0.0%
5265	201	PRG1	FTM1	90.0%	70.68%
1484	502	BCM1	FTM4	100.0%	0.0%
3539	502	JN3	PRCD3	100.0%	0.0%
5266	502	ACRD1	LGO2	70.0%	0.0%
16200	502	BCM1	LGO2	100.0%	0.0%
4749	201	BCM1	GC1	80.0%	81.15%
16201	502	GJQ3	FTM4	80.0%	0.0%
11235	502	BLRB2	PRCD3	100.0%	0.0%
2120	502	FIG	FTM4	100.0%	0.0%
5320	502	PRR1	PRCD3	100.0%	0.0%
4214	203	MPL3	GC11	70.0%	0.0%
5321	502	BCM1	MAL2	100.0%	0.0%
11492	201	BCM1	BSS1	50.0%	0.0%
15960	502	HML1	LGO2	100.0%	0.0%
15962	502	BCM1	FTM4	100.0%	0.0%
2884	502	HML1	MAL2	100.0%	0.0%
3542	502	BCM1	FTM4	100.0%	0.0%
4750	201	BCM1	GC1	80.0%	81.15%
13242	203	MPL3	PRI1	70.0%	0.0%
7635	203	MPL3	GC11	70.0%	0.0%
15064	502	BCM1	LGO2	100.0%	0.0%
13971	201	BCM1	PNV2	90.0%	89.81%
13602	201	BS1	BSS1	50.0%	0.0%
10876	502	BCM1	LGO2	100.0%	0.0%
2072	502	GJQ3	LGO2	80.0%	0.0%
12171	201	BCM1	GC1	80.0%	81.15%
7633	203	MPL3	GC11	70.0%	0.0%
7147	203	MPL3	PRI1	70.0%	0.0%

Continued on next page

Table B.32: Traveller 1's temporary network with member relevance on 14/01/2010

Table B.32 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
7148	203	MPL3	SGS1	70.0%	0.0%
3700	502	GJQ3	FTM4	80.0%	0.0%
11968	201	BS1	EZC1	90.0%	0.0%
7275	203	BSB	PRI1	70.0%	0.0%
457	502	BLRB2	NEV2	100.0%	0.0%
458	502	CVA1	MAL2	100.0%	0.0%
13935	203	BCM1	PRI1	70.0%	0.0%
7277	203	BSB	GC11	70.0%	0.0%
5654	203	MPL3	GC11	70.0%	0.0%
12503	202	BCM3	BRV3	0.0%	82.1%
5655	201	BCM1	PNV2	90.0%	89.81%
17339	203	MPL3	GC11	70.0%	0.0%
10040	502	BCM1	PRCD3	100.0%	0.0%
459	502	BCM1	LGO2	100.0%	0.0%
2125	502	BCM1	FTM4	100.0%	0.0%
4093	502	BCM1	FTM4	100.0%	0.0%
7873	201	BCM1	FTM1	90.0%	94.97%
2123	502	BCM1	FTM4	100.0%	0.0%
7152	203	BCM1	PRI1	70.0%	0.0%
1159	502	BSS1	LGO2	60.0%	0.0%
91	201	BCM1	PNV2	90.0%	89.81%
3157	203	BCM1	SGS1	70.0%	0.0%
5625	502	BLFZ1	FTM4	100.0%	0.0%
1115	502	GJQ3	PRCD3	80.0%	0.0%
6666	203	BCM1	PRI1	70.0%	0.0%
2691	502	BCM1	LGO2	100.0%	0.0%
6669	203	BCM1	GC11	70.0%	78.29%
2694	502	BCM1	LGO2	100.0%	0.0%
461	502	BCM1	LGO2	100.0%	0.0%
2693	502	BCM1	LGO2	100.0%	0.0%
2739	203	MPL3	PRI1	70.0%	0.0%
6670	201	BCM1	GC1	80.0%	81.15%
467	502	BCM1	LGO2	100.0%	0.0%
6671	502	BCM1	FTM4	100.0%	0.0%
468	502	BCM1	PRCD3	100.0%	0.0%

Continued on next page

Table B.32: Traveller 1's temporary network with member relevance on 14/01/2010

Table B.32 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
6672	502	BCM1	FTM4	100.0%	0.0%
7056	201	AGM1	FTM1	80.0%	81.99%

2

B.2.5 Traveller 1 temporary network on 15/01/2010

Table B.33: Traveller 1's temporary network with member relevance on 15/01/2010

Table B.33 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
9287	501	BCM3	BFTM1	60.0%	0.0%
113	201	BCM1	PNV2	0.0%	88.78%
17887	201	BCM1	FTM1	0.0%	95.26%
4454	202	BCM3	RCRT2	100.0%	0.0%
1107	202	BCM3	MSLD	100.0%	0.0%
3541	502	BCM1	LGO2	0.0%	81.48%
7349	501	PAL3	BFTM1	60.0%	0.0%
9492	502	BCM1	LGO2	0.0%	81.48%
18548	502	BCM1	FTM4	0.0%	94.24%
16790	202	TRD1	LGO4	100.0%	0.0%
16791	202	BCM3	SJB1	60.0%	0.0%
4310	201	BCM1	GC1	0.0%	80.28%
5889	201	PRG1	FTM1	0.0%	73.81%
9657	501	BCM3	BFTM1	60.0%	0.0%
5601	502	AGM1	LGO2	0.0%	68.58%
16200	201	BCM1	FTM1	0.0%	95.26%
11238	502	BCM1	LGO2	0.0%	81.48%
4214	203	BCM1	GC11	0.0%	76.91%
10865	202	BCM3	LGO4	100.0%	0.0%
5489	501	BCM3	BFTM1	60.0%	0.0%
1380	201	GJQ3	FTM1	0.0%	74.48%
17291	501	BCM3	IGAL2	60.0%	0.0%
3542	502	BCM1	FTM4	0.0%	94.24%
5653	201	BS1	GC1	0.0%	71.61%
471	502	BCM1	LGO2	0.0%	81.48%

Continued on next page

Table B.33: Traveller 1's temporary network with member relevance on 15/01/2010

Table B.33 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
13243	203	BCM1	GC11	0.0%	76.91%
7635	203	BCM1	GC11	0.0%	76.91%
7387	501	GGF	BFTM1	60.0%	0.0%
15572	202	SDP1	LGO4	70.0%	0.0%
2985	202	BCM3	BRP1	80.0%	0.0%
15573	202	SJB1	MSLD	50.0%	0.0%
5719	202	SJB1	FTM2	50.0%	0.0%
13971	502	BCM1	FTM4	0.0%	94.24%
7104	501	CMO	PDVD4	60.0%	0.0%
10876	502	BCM1	LGO2	0.0%	81.48%
2072	502	GJQ3	LGO2	0.0%	60.69%
12171	201	BCM1	GC1	0.0%	80.28%
2980	202	TRD1	LGO4	100.0%	0.0%
7633	203	BCM1	GC11	0.0%	76.91%
2976	202	TRD1	MSLD	100.0%	0.0%
5654	203	BCM1	GC11	0.0%	76.91%
12503	202	BCM3	BRV3	90.0%	0.0%
5655	201	BCM1	FTM1	0.0%	95.26%
10040	502	BCM1	FTM4	0.0%	94.24%
2127	502	BCM1	LGO2	0.0%	81.48%
459	502	BCM1	FTM4	0.0%	94.24%
2125	502	BCM1	FTM4	0.0%	94.24%
2978	202	TRD1	LNEV2	100.0%	0.0%
4093	203	BCM1	GC11	0.0%	76.91%
1261	202	SDP1	LGO4	70.0%	0.0%
1262	202	SJB1	LGO4	50.0%	0.0%
18511	201	BCM1	GC1	0.0%	80.28%
599	501	BCM3	BFTM1	60.0%	0.0%
10218	501	PRG1	BFTM1	60.0%	0.0%
10219	501	BCM3	BFTM1	60.0%	0.0%
2345	502	BCM1	FTM4	0.0%	94.24%
1260	202	CSBV	LNEV2	80.0%	0.0%
91	502	BCM1	FTM4	0.0%	94.24%
758	501	BCM3	PDVD4	60.0%	0.0%
2691	502	BCM1	LGO2	0.0%	81.48%

Continued on next page

Table B.33: Traveller 1's temporary network with member relevance on 15/01/2010

Table B.33 – continued from previous page

Traveller	Route	Origin	Destinonion	Relevance (similar)	Relevance (complementar)
461	502	BCM1	LGO2	0.0%	81.48%
756	501	BCM3	PDVD4	60.0%	0.0%
6669	201	BCM1	GC1	0.0%	80.28%
460	502	BCM1	FTM4	0.0%	94.24%
6668	203	BCM1	GC11	0.0%	76.91%
463	201	BCM1	FTM1	0.0%	95.26%
5631	502	BCM1	LGO2	0.0%	81.48%
467	201	BCM1	FTM1	0.0%	95.26%
6670	203	BCM1	GC11	0.0%	76.91%
6671	201	BCM1	GC1	0.0%	80.28%
8170	501	BCM3	BFTM1	60.0%	0.0%
469	502	BCM1	LGO2	0.0%	81.48%
1354	502	BCM1	LGO2	0.0%	81.48%
11261	501	BCM3	BFTM1	60.0%	0.0%
7056	201	AGM1	FTM1	0.0%	82.37%