Catch Me If You Can: Predicting Mobility Patterns of Public Transport Users

Stefan Foell, Santi Phithakkitnukoon, Gerd Kortuem, Marco Veloso and Carlos Bento

Abstract— **Direct and easy access to public transport information is an important factor for improving the satisfaction and experience of transport users. In the future, public transport information systems could be turned into personalized recommender systems which can help riders save time, make more effective decisions and avoid frustrating situations. In this paper, we present a predictive study of the mobility patterns of public transport users to lay the foundation for transport information systems with proactive capabilities. By making use of travel card data from a large population of bus riders, we describe algorithms that can anticipate bus stops accessed by individual riders to generate knowledge about future transport access patterns. To this end, we investigate and compare different prediction algorithms that can incorporate various influential factors on mobility in public transport networks, e.g., travel distance or travel hot spots. In our evaluation, we demonstrate that by combining personal and population-wide mobility patterns we can improve prediction accuracy, even with little knowledge of past behaviour of transport users.**

I. INTRODUCTION

In an era of rapid urbanization, public transport plays a key role in managing the balance between increasing demand for mobility and the environmental impact of mass transport. Nevertheless, in order to ensure that public transport is a viable option for many travellers, there is a constant need to stimulate its use. In particular, cars are still the most widely used mode of transportation valued for their comfort, ownership and controllability [12]. Therefore, identifying and overcoming barriers of transport use are key priorities of many public transport providers [23].

Information technology has great potential to improve the visibility and accessibility of public transport services [5]. Over the recent years, the wide-spread adoption of smart phones has provided transport providers new channels to engage with travellers [9]. As a result, transport users are able to request journey information regardless of their current location. In a survey with bus riders, various positive effects are attributed to enhanced information availability, e.g., better satisfaction and increased ridership [10]. Easy access to relevant travel information is therefore a decisive factor for the success and adoption of public transport systems.

In the future, public transport information systems could be turned into personalized recommender systems to provide even better support and guidance. For instance, in order to alert travellers about incidents or changes affecting their journeys, suggestions for better routes could be sent to them prior to their departures. Similarly, public transport users could receive recommendations about events and offers near by the transport stops that they visit. However, a main prerequisite for the development of such intelligent services is accurate knowledge of individual travel patterns. With the deployment of automatic fare collection systems, large-scale data becomes available about real-world transport usage [18]. However, studies of individual travel patterns are sparse in public transport research. In the past, research has mainly focused on aggregate demand forecast [7].

In order to fill this gap, we describe in this paper algorithms to extract and predict mobility patterns of public transport users with a specific focus on bus ridership. Bus networks in urban areas create complex mobility systems with a large number of stops and routes. Identifying and ranking the stops used by individual bus riders provides useful knowledge for information personalization. However, given the large variety of users with different mobility needs, ranging from frequent to occasional riders, an approach is required which can guarantee effective predictions for all rider types. To find an approach that exhibits these characteristics, we explore in this work a range of algorithms that can incorporate various influential factors on mobility decisions in public transport networks including a) personal travel habits and popular travel hot spots, b) geography and structure of the transport network and c) collective information of transport use from other travellers.

In our evaluation, we use large-scale bus ridership data from Lisbon, Portugal to analyse the prediction approaches. As our analysis shows, the performance of a predictor depends on its ability to deal both with or without much knowledge of a user's past rides. While knowledge from personal ride histories is valuable especially for more habitual riders, information from collective transport usage patterns of other riders is beneficial to new or infrequent riders for which data histories are limited. By showing that both beneficial features can be combined into a single approach, a powerful tool becomes available that can be applied to foresee mobility behaviours of any rider type. As a result, this work contributes important methods and insights for the design and development of more intelligent public transport information systems that are driven by accurate knowledge of mobility patterns of public transport users.

[1] Stefan Foell is with the Computing Department, The Open University, Milton Keynes, UK stefan.foell@open.uc.ak

[2] Santi Phithakkitnukoon is with the Department of Computer Engineering, Chiang Mai University, Thailand santi@eng.cmu.ac.th

[3] Gerd Kortuem is with Computing Department, The Open University, Milton Keynes, UK gerd.kortuem@open.uc.ak

[4] Marco Veloso is with Centro de Informtica e Sistemas da Universidade de Coimbra, Coimbra, Portugal mveloso@dei.uc.pt

[5] Carlos Bento is with Centro de Informtica e Sistemas da Universidade de Coimbra, Coimbra, Portugal bento@dei.uc.pt

The rest of the paper is organized as follows. In Section II, we report on prior studies of travel card data patterns. Then, in Section III, we introduce the datasets analysed in our work. The problem addressed in this paper is formally described in Section IV. In Section V, we present algorithms for the prediction of mobility behaviours of transport users. Subsequently, we describe in Section VI the results of our analysis. Finally, a conclusion is given in Section VII.

II. RELATED WORK

This work seeks to extract novel added values from the data generated by today's public transport systems. As more and more sensors have been integrated into public transport infrastructures, and electronic ticketing systems are widely deployed today, large-scale transport data is produced at high rates [24]. Especially, the data recorded by Automated Fare Collection (AFC) and Automated Vehicle Location (AVL) systems are valuable assets for strategic, tactical and operational planning of public transport systems [18]. In the following, we present prior work in this space.

Ferrari et al. have leveraged on AFC data to build a ridership demand model and investigate accessibility barriers for wheelchair users [8]. They discovered that measurable barriers are prevailing both in terms of travel time and number of required interchanges. Moreover, AFC data has been used to characterize passenger flows in intra-urban environments [21]. It was shown that some of the variation in the flows can be explained by gravity models, but other forces such as socio-economic factors are also relevant. Ceapa et al. have analysed temporal structures in AFC data to identify overcrowding patterns in public transport stations [6]. Their analysis revealed that overcrowding situations follow regular patterns associated with peak travel times that can be well predicted. Bejan et al. showed that AVL data can be exploited to estimate journey times experienced by road users [2]. This way, information about traffic conditions in urban areas can be provided without the need for additional costly sensing and monitoring infrastructure.

Over the recent years, personalized transport information systems have moved into the focus of research [9]. These systems benefit from predictions of travel decisions of individual travellers. Foell et al. have analysed temporal patterns in large-scale bus ridership data [11]. They identified temporal features relevant for the prediction of transport access, e.g., day of the week, but spatial movements patterns are not considered. Li et al. have analysed travel flow directions at peak hours in relation to stops classified according to surrounding land usage characteristics [15]. Furthermore, Liu et al. have mined spatial and temporal patterns of transport behaviour to quantify degrees of regularity inherent to travel [16]. However, the patterns explored in these works characterize aggregated usage, and cannot be used as effective tool to forecast mobility of individual riders.

Understanding human mobility patterns has been the subject of research in various domains beyond transport systems. For instance, Belik et al. explore the role of human movements in spatial epidemics and analyse how the spread of diseases is influenced by mobility patterns [3]. In a different setting, Noulas et al. have exploited check-in and movement patterns to predict new venues in location-based social networks [17]. Song et al. have evaluated the limits of predictability in human dynamics by analyzing mobility patterns of mobile phone users [22]. However, a detailed study into individual human mobility patterns in the context of public transport systems and an analysis of their predictability based on AFC and AVL data has not been reported in current literature.

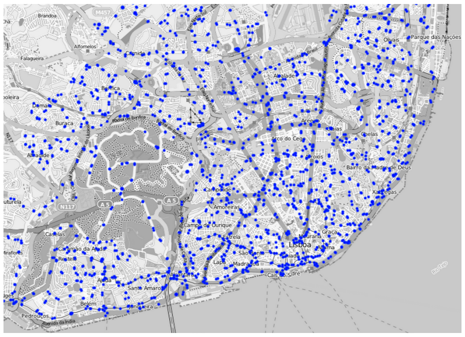

Fig. 1. Locations of bus stops in Lisbon.

III. DATASETS AND PREPARATION

The transport data used for our analysis has been collected from Lisbon. Lisbon is the capital and largest city of Portugal with a population of over half a million. Buses are an important part of the city's public transport infrastructure. In our analysis, we use datasets of the bus transportation in Lisbon. The data has been recorded between 1st of April and 31th of May 2010 (61 days). In the following, we describe the datasets in detail.

A. *Bus network data*

Our data provides geographic and topological information about the bus network in two sets of data. Dataset A contains all bus stops in Lisbon and their geographic locations as shown in Fig. 1. Formally, the set of stops is denoted as S. In total, $|S| = 2110$ stops are listed. Each stop $s_i \in S$ is associated with spatial coordinates given its latitude and longitude pair. This allows us to compute the geographic distance $dist(s_i, s_j)$ between two bus stops $s_i \in S$ and $s_j \in S$. Dataset B provides information about the bus network in Lisbon. To this end, bus routes are listed with information of the bus line id, direction, and the stops on the route. As routes are described as directional, they encompass different stops for each direction. Both datasets have been used to estimate geographic and network-based travel distances as explained later in this paper.

B. *Bus ridership data*

Ridership information has been scattered over two additional data sets which needed to be correlated.

Dataset C provides trip records collected by the AFC system deployed in Lisbon. Amongst other information, each record contains the id of the rider's travel card, time of bus

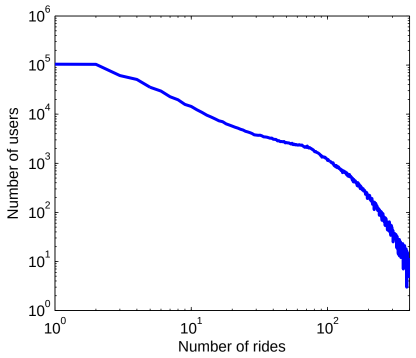

Fig. 2. The ridership distribution shows the count of users with a number of rides observed in our data

boarding and the id of the bus boarded. Moreover, dataset D provides AVL data from the buses. The data comprises the time-stamped bus arrivals of buses when dropping off passengers along their routes. Both datasets have been linked to gain complete bus ridership information that exposes the ids of the bus stops where the rides started. Note that only boarding information can be obtained as bus users in Lisbon are required to use their travel cards only at the beginning of their journey to get on the bus. In order to compensate for potential synchronization issues, we allowed for a small temporal deviation for a successful matching between bus arrival and bus boarding. Bus rides which could not be matched due to larger deviations or other inconsistencies (e.g. we observed some duplicate AFC entries) have been removed from our analysis. As the correlation was performed based on the time of a ride and unique bus id, unambiguous travel histories of individual riders have been obtained. Formally, the data can be described as $H = \{\langle u, s, t\rangle \ | u \in U, s \in S, t \in T\}$, where u is the rider, s is the bus stop where a ride was started and t is the time of a boarding. In total, we obtained $|H| = 24,257,353$ bus trips taken by $|U| = 809,758$ riders over the observation period.

IV. PROBLEM STATEMENT

In this paper, we address the problem of predicting the mobility patterns of bus riders travelling the bus network. Unlike prior work in public transport research [7], the focus of our study is not on aggregate demand patterns. Instead, we aim at personalized predictions which apply to individual travellers and their personal mobility behaviour. These predictions are much more useful when an understanding of the specific transport needs of a single person is required.

More precisely, we seek to anticipate the stops relevant for a rider u to access the transport network. For such a prediction, we make use of historic information about past rides from u's trip history $H_u = \{\langle u, s, t\rangle \ \in H | s \in S, t \in T\}$. While H_u provides useful knowledge about past rides, the accuracy of prediction depends on how u behaves in the future. In the future, u may access not only known stops, but also stops that u never used before. In addition, the relevance of the stops may change and certain stops may be used much more or less frequently by u in the future. To account for the fact that different bus stops are not equally relevant for a rider, the prediction problem is approached as a ranking task where a stop used more frequently by u in the future should receive a higher rank. Formally, the ranking results in a total order where a unique position $r_u(s) \in [1, |S|]$ is assigned to each stop $s \in S$ resulting in a prediction list. In this list, stop s_i is ranked higher than stop s_j with $i \neq j$ if it holds that $r_u(s_i) > r_u(s_j)$.

This problem definition naturally addresses different scenarios of real-world transport usage. On the one hand, the degree as to which the same stops are visited over and over again is determined by routine behaviour. There may be stops seen regularly as well as ones which are visited more occasionally. On the other hand, new transport users may be constantly joining the bus system. As a consequence, transport usage histories may contain only little information from which the prediction can benefit. However, accurate predictions should be also available for these users. Fig. 2 shows the distribution of ridership among all bus users over the entire observation period. It can be seen that a broad spectrum of different ridership demand exists which impacts on the amount of historic information available for prediction. In the following, we explore a set of algorithms which can be applied to riders with different characteristics to achieve accurate predictions.

V. PREDICTION ALGORITHMS

In the this section, we propose a set of algorithms to address the prediction problem introduced above. The algorithms make use of different features which imply mobility preferences among users travelling a bus transport network. We investigate: a) personal and global patterns of transport use as being encoded in travel card data, b) travel distance metrics which are either based on geographic distances or shaped by the layout of the network topology, and c) collaborative filtering algorithms that exploit similarities and commonalities in transport behaviour among different users. While the prediction algorithms and features are described next, a detailed evaluation and comparison of the approaches is given thereafter.

A. *Personal Mobility*

One straightforward way to predict future stop usage is to leverage on the information from the user's own trip history H_u. This approach is termed *Personal*. The idea is that those stops which have shown to be of high relevance in the past, will also be equally important in the future. For this purpose, we mine the user's transport history for stops that have been accessed in the past. Formally, we use

$$f_{u,i} = |\{\langle u, s_i, l, t\rangle \ \in H_u | l \in L, t \in T\}|$$

to determine the number of rides boarded by u at stop s_i. Knowledge of the past stop visits can then be employed to define a ranking among all stops. A higher rank is associated with stops that have been visited more often. This way, all stops $S_u = \{s \in S | \langle u, s, t\rangle \ \in H_u\}$ that have been

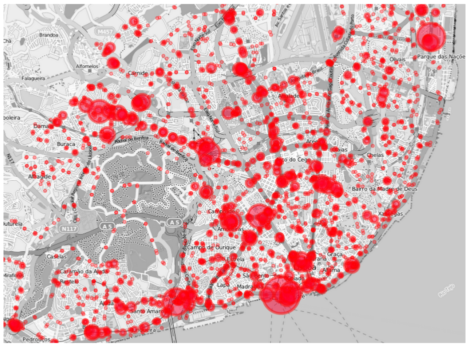

Fig. 3. Map showing the popularity of bus stops in Lisbon. A bus stop is represented with a circle whose radius is scaled according to the number of rides that have started the stop.

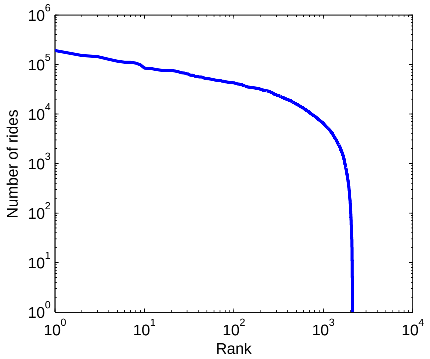

Fig. 4. Ranked distribution of the popularity of bus stops in Lisbon. The x-axis represents the rank of the stop (from most frequently to less frequently used), and the y-axis shows the number of rides starting at this stop.

visited before appear at the top of the list. As a tie breaker, stops that have been accessed the same number of times are included in a random order. However, this means that all stops $S \setminus S_u$ that have not been visited before cannot be weighted accordingly. In our evaluation, we will see that neglecting potential new behaviours leads to suboptimal predictions. This can be especially a problem for new or infrequent bus riders. Therefore, we seek to explore further solutions in the following which can operate on a greater variety of scenarios.

B. Global Mobility

When personal usage patterns superimpose each other, global patterns of transport usage emerge. These global patterns carry important information about which transport decisions are likely to be made by users. As a common phenomenon, popular hot spots exist in transport networks that are particularly attractive to travellers resulting in high levels of transport activity at specific locations. The emergence of such hot spots is due to manifold influences and forces in a city ecosystem, e.g., caused by transport hubs with access to different transport modalities, urban centres of social activity such as leisure and night-life districts, or skewed residential population distributions. Fig. 3 shows the popularity of bus stop visits in Lisbon based on our ridership data. The figure reveals a high skew in the popularity of different stops for attracting ridership. The most 20% frequently visited stops make up for 62.4% of total stop visits. This signifies that global mobility patterns are concentrated on the most popular bus stops. Fig. 3 shows that many popular stops in Lisbon are near to train stations, tourist spots and the city's harbour. We exploit this observation and define

$$g_i = \sum_{u \in U} f_{u,i}$$

as the global popularity of bus stop s_i among all riders. Knowledge of the global popularity patterns can then be used to influence the prediction. The approach termed *Global* simply ranks all stops S according to their global usage popularities. As a consequence, a universal ranking is established which is the same across all users. A more personalized approach is *Personal*$^+$ which applies the global popularity patterns only to stops $S \setminus S_u$ that have not been used by u before. With this approach, the top entries in the ranked stop list are derived from personal riding patterns as described in the previous section. The lower part of the list consists of all remaining stops ordered according to the their global usage popularities.

C. Geographic Mobility

Finding universal laws to model and explain the movement of people has been an active area of research in the past. In empirical studies, it has been shown that a close relationship exists between human movement and geographic distance [13]. Distance is seen as a barrier to travel, which is empirically proven by the emergence of skewed mobility patterns. More precisely, the probability of travelling to a destination decreases proportional to the distance involved in a trip [13]. As this has been described as a universal law which generally holds for human mobility behaviours, we seek to explore this feature also in the context of public transport usage. Therefore, we propose the *Geographic* approach that calculates personalized travel distances based on the stops S_u that have been previously visited by u. Based on these distances, we estimate the degree to which any other bus stop would be a relevant target for the user. Formally, this can be described as

$$d_{u,i} = \min_{1 \le j \le |S_U|} dist(s_j, s_i)$$

which yields the closest distance $d_i \in$ to a stop $s_i \in S \setminus S_u$ from any stop in S_u previously visited. A bus stop is assigned a lower rank if it is near to a bus stop that has been used before. Consequently, this approach is shaped both by the geographic layout of the bus network as well as the past bus rides of travellers. In our evaluation, we have tried different options to define a set of anchor points upon which the distance calculation is based. As one alternative, we have used the most popular stop as an approximation of a user's home location to centre the geographic search. However, the accuracy was higher when incorporating the user's entire mobility radius given by the full set of S_u.

Moreover, we have adapted the distance metric to account for variations in popularity among the bus stops in the city. The idea is that popularity represents a complementary factor which changes how distance is experienced by travellers. For this purpose, *Geographic*$^+$

$$d^+_{u,i} = (1 + log(\frac{\max_{1 \leq r \leq |S|} g_r}{g_i})) \cdot d_{u,i}$$

incorporates the global usage g_i as part of the weight factor to calculate the adjusted distance $d^+_{u,i}$. The weight factor is based on the inverse ratio of a stop's popularity to the highest stop popularity. We apply a logarithmically scaling to create a smoother weighing effect. The weight factor can be considered as a pulling or pushing force on the distance $d_{u,i}$. If the stop is unpopular, the distance is pushed further away, making it less reachable. In contrast, if the stop has a high popularity, the stop is pushed closer to the user. These factors therefore distort the geographic space to account for more realistic transport usage patterns. A similar technique has been applied in information retrieval where the tf-idf factor is used to quantify the degree of unique words in documents [1]. However, the relevance of popular stops is increased with our weighting scheme whereas information retrieval considers more popular documents as less important.

D. Network Mobility

While geographic distance is an unbiased distance measure in free spaces, public transport networks represent a planned and more constrained environment. Instead of arbitrary travel paths that can be followed through the city, public transport systems are based on predefined routes which guide the travel flows. The topology of a route network may significantly differ from relations found in geographic space: while stops may be geographically close to each other, short and direct connections may not always guaranteed in a public transport network. The existence of routes which seem more plausible and make certain destinations easier to reach than others can impact on mobility decisions. Beyond geographic distance, we therefore explore a more meaningful distance metric to predict more realistic travel paths in public transport networks.

To create this metric, we derive an adjacency matrix A that maps the neighborhood relations in the public transport network topology. Each entry $a_{ij} \in A$ of the matrix represents a binary variable which encodes whether stops s_i and s_j are connected through a direct bus route segment. More precisely, we set $a_{ij} = 1$ if there is at least one bus route which links stops s_i and s_j as successive stops in the same direction, and $a_{ij} = 0$ otherwise. Then, we use A as input to a shortest path algorithm (i.e., Dijkstra) to compute logical travel distances in the public transport network. The result is a matrix L whose entries l_{ij} denote the minimum number of hops required to travel between any stops s_i and s_j. Note that different refinements of this algorithm are feasible, e.g. penalizing interchanges or considering the actual travel time on route segments. However, in our work, we have focused on the basic network topology for a direction comparison with the geographic distance space.

For the *Network* approach we then calculate

$$n_{u,i} = \min_{1 \leq j \leq |S_U|} l_{ji}$$

to determine the minimum distance to reach stop $s_i \in S \setminus S_u$ from any stop in S_u previously visited. Hence, stops which are easily reachable through a path in the network topology from stops visited in the past receive a higher rank.

Following the same rationale as before, *Network*$^+$ take this idea one step further and adjusts the distances to account for the varying popularity of bus stops. Formally, we determine

$$n^+_{u,i} = (1 + log(\frac{\max_{1 \leq r \leq |S|} g_r}{g_i})) \cdot n_{u,i}$$

which is the hop-based travel distance to reach s_i from previously visited stops offset by its popularity. As a consequence, a stop is considered to be of high relevance if it has a good link to the user's stops visited in the past and if it is attracting a large number of rides.

E. Collaborative Filtering

On an abstract level, the prediction problem studied in this paper fits the purpose of recommender systems [19]. For suggesting relevant items to users, recommender system provide algorithms to analyse the users' past item ratings and find common patterns among the collective ratings of all users. In the following, we apply a similar strategy to capture travel decisions in public transport networks. The idea is that when bus stops are seen as items, stop visits define implicit ratings that can be mined to determine the strength as to which different stops are similar in usage among users. Knowledge of the similarity in stop usage then can be exploited to identify stops with strong relations that are likely to become relevant to the user in the future. To this end, we leverage on item-based recommendation [20] which allows us to manage the complexity of the recommendation algorithm despite the high number of users. Item-based recommendation is preferred over a user-based approach when the number of items outweighs the number of users. In the case of public transport networks, this premise is satisfied as the set of stops is much smaller than the large population of riders. According to this approach, we determine for each stop s_i a visit vector

$$t_i = \langle f_{u_1,i}, f_{u_2,i}, \ldots, f_{u_{|U|},i} \rangle$$

where the i-th component encodes the number of visits of the user $u_i \in U$ at this stop. Given two visit vectors t_i and t_j associated with stops s_i and s_j, a similarity score $sim(i, j)$ can be computed to indicate if both stops have similar usage patterns. In our work, we have used the Cosine similarity which measures the cosine of the angle between the vectors. According to this measure, two stops s_i and s_j are similar to each other if a rider visiting s_i implies that also s_j is visited.

The similarity scores can be incorporated into an approach *Collaborative* that implements collaborative filtering for predicting stops that will be visited by a user. For every user $u \in U$ a visit score

$$v_{u,i} = \sum_{s_j \in S_u} sim(j,i) \cdot f_{u,i}$$

is computed that quantifies the prospect that stop s_i will be used. The score is based on the similarities $sim(j,i)$ of s_i with all stops $s_j \in S_u$ found in a user's trip history. In addition, the similarity scores are weighted by the frequency of past usage of the stops in $s_j \in S_u$. As a consequence, those bus stops are ranked high that exhibit similarities to the ones that are frequently used by the rider. In contrast to classical item-based recommendation where a rating score is computed as the average rating from related items [20], we have customized our algorithm to accumulate evidence of potential stop usage.

F. Random Walk Approach

Random walks are employed in various domains to model and reason over uncertain behaviours. A random walk is based on a graph representation and can reveal the nodes in the graph with a high degree of attraction, e.g., the page rank algorithm [4]. For the purpose of this work, we apply a random walk approach to reason over the collective mobility patterns of bus users. The idea is to model the stop visit patterns of all users in a coherent graph structure, exposing the stops that the user is attracted to and therefore likely to visit in the future.

To implement this model, we define a directed graph $G = (V, E)$ whose nodes $V = (U \cup S)$ are the union of all users and stops, and the edges $E \subset V \times V$ represent usage relations observed in the data. For each user u_i who has used stop s_j two directed edges are introduced: $(u_i, s_j) \in E$ from the user to the stop as well as $(s_j, u_i) \in E$ in the reverse direction. Each edge $e \in E$ is associated with a probability $p(e)$ that models how likely the edge is to be traversed. For the definition of the edge probabilities, we incorporate the variation of stop visits to add weight to edges which lead to more frequently used stops. With this approach, $p(u_i, s_j)$ is defined as the fraction of u_i's rides that have started from stop s_j. On the other hand, $p(s_j, u_i)$ is defined as the fraction of all rides from s_j that have been taken by u_j. Hence, the graph encodes both structural relations and quantitative mobility information.

Given this bus usage mobility graph, we then perform a *RandomWalk* which has been devised for recommendation problems [14]. Initially, the random walk starts at node u representing the user whose mobility pattern is to be predicted. Then, in each iteration the graph is traversed according to the transition probabilities p assigned to the edges. With a restart probability of r, however, the random walk is taken back to node u. This is to direct the search in the direct neighbourhood of the user's node from which the graph is explored. As the random walk is continued, evidence is accumulated about the stops which are often encountered and therefore are more connected to the user.

TABLE I
AVERAGE PERCENTILE RANK (MEAN AND STANDARD DEVIATION)

Approach	**APR**	
	μ	σ
Random	0.500	0.174
Global	0.804	0.123
Personal	0.839	0.194
Personal$^+$	0.931	0.097
Geographic	0.942	0.116
Geographic$^+$	0.958	0.081
Network	0.948	0.105
Network$^+$	0.963	0.073
Collaborative	0.972	0.062
RandomWalk	0.973	0.055

In a matrix form, the solution of the random walk can be expressed by equation

$$s = (1 - r) \cdot p \cdot s + r \cdot q$$

where q encodes the user's node as a column vector, p is the transition probability matrix and s denotes the steady-state probabilities, i.e., the long-term rate that a random walk terminates in a node when followed infinitely. The steady-state probabilities associated with stops $S \subset V$ can thus be used as a ranking criterion. Consequently, those stops are ranked high which can be reached more easily from the user's position in the graph. This is influenced by the user's own usage pattern as well as that of all other riders which are represented in the graph.

VI. EVALUATION

For the evaluation of the prediction algorithms, we have relied on the large-scale bus usage data from Lisbon presented earlier. Given the large number of riders in the dataset, an analysis can be made about the predictability of mobility patterns of a large rider population, and the relation between prediction accuracy and different rider types. In the following, we first describe the methodology underlying our evaluation and then present the results from the analysis.

A. Methodology

In order to evaluate the prediction algorithms, the data has been split into a training and test set. The test set comprises the last two weeks of the bus usage data, while the training set spans all days before. This way, the travel histories of riders have been segmented into a historic part (training set) and future part (test set). For each algorithm, we created a ranked list of predicted bus stop usage specific to the individual traveller (using the user's own ride history and/or the histories of other riders depending on the algorithm). Then, we compared the predictions with the actual observations in the test data. Riders with not at least one ride in either the test or training set have been pruned from the evaluation.

As evaluation metric, we used the percentile rank as a measure of prediction accuracy. For any stop $s \in S'_u$

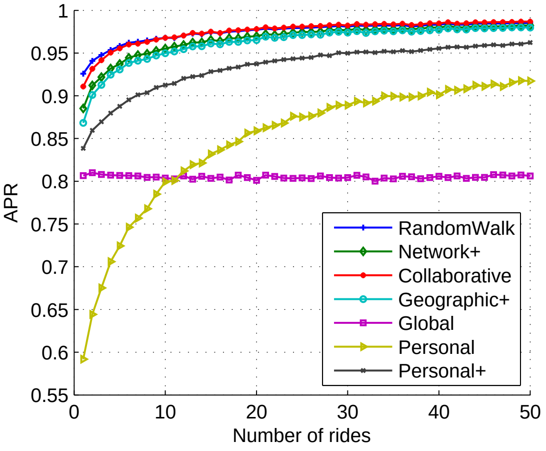

Fig. 5. APR depending on the size of the trip history

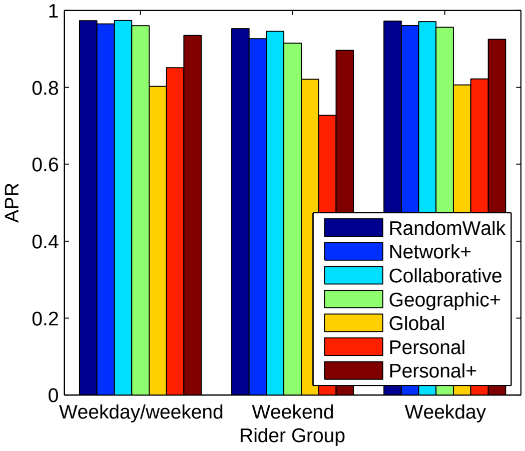

Fig. 6. APR depending on rider group

accessed by a user u, the percentile rank (PR) is given as

$$PR = \frac{|S| - r_u(s) + 1}{|S|}$$

The PR reveals the degree to which the ranked list of bus stops matches the real stop usage of a rider. A PR equal to 1 refers a perfect prediction where the stop used by the rider is ranked at the top of the list. In contrast, if the PR tends to be closer to 0, the prediction becomes worse as the stop is found more towards the end of the list. By applying this metric to all bus rides of a user, every stop use is counted towards the PR. Consequently, more frequently used stops have a bigger impact on the prediction accuracy as they are more relevant to the rider. The PR of all riders $u \in U$ is then averaged to obtain the average percentile rank (APR). The approach with the highest APR represents the best predictor which exhibits the highest predictive power.

B. Results

1) Performance Comparison: In Table I, we show the APR achieved by the different approaches. As a baseline, we generate a randomly ordered list (*Random*) which yields an APR of 0.5. It can be seen that all of the proposed approaches significantly outperform this baseline. The *Global* predictor shows already clear improvements, as stop visits are clustered around popular stops. The *Personal* approach shows that prior knowledge of the user's behaviour can improve the prediction. All other approaches outperform the basic approaches when combining personal and global patterns. *Personal*$^+$ improves the performance by a simple approach to fuse personal and global usage patterns. The improvement achieved by *Geographic* provides evidence that transport usage decisions emerge from regular spatial access patterns where closeness is an important criterion. The notion of closeness can be further enhanced with the *Network* approach. As our analysis shows, the topological travel distance within the bus network topology has a bigger influence on transport behaviour than geographic distance. Adapting the distances according to global usage patterns further lifts the accuracy of the predictions (both *Geographic*$^+$ and *Network*$^+$). Note that an increase of 0.01 in APR already corresponds to an improvement of 21 ranks in the list. The best results are achieved by the two approaches that can selectively combine collective usage patterns from travellers with similar behaviours: *Collab* achieves an APR of 0.977, and *RandomWalk* achieves the highest prediction accuracy with an APR of 0.973.

2) Dependency on Ridership: Fig. 5 shows the APR achieved by the different algorithms in relation to the number of rides taken by the rider. For this purpose, we have grouped all riders according to individual ridership demand observed in the training data, and then computed the APR over all riders in this group. Generally, it can be observed that more active users with a larger number of bus rides are better predictable. The *Personal* approach is particularly sensitive towards the amount of known past transport behaviour: While for more active users the rider's own history covers a large portion of the future stop usage, there is a high degree of uncertainty involved for low demand riders resulting in inaccurate prediction. The performance of the *Global* approach is largely independent from ridership demand. All riders tend to visit popular stops in the bus network in a similar way. The best performing approaches have the ability to adapt to both more active and less active riders. As these approaches are designed to interweave global usage patterns with personal ridership habits, a high prediction accuracy can be achieved across a spectrum of different transport behaviours. Among them, the *Personal*$^+$ approach is the most limited one as global usage patterns are only integrated but not correlated with the user's own behaviour. The *Random Walk* approach is the best approach, consistently across all rider groups. Notably, it also outperforms the *Personal* approach for the group of active riders, demonstrating that incorporating knowledge beyond the user's own travel history is beneficial for all riders. Consequently, the *Random Walk* approach can be regarded as the most generic predictor suitable for any level of ridership demand.

3) Different Rider Groups: We have further analysed the predictability of different rider groups with distinct temporal behaviours. To this end, riders have been assigned to one of three categories based on the temporal distribution of their bus usage (weekday, weekend or both). We found that weekday/weekend riders represent the largest group (63%) in the data, followed by weekday only riders (37%) and weekend only riders (7%). Then, we have measured the APR for each rider group. As Fig. 6 shows, weekday riders are most predictable, but only slightly more predictable than weekday/weekend riders. Across all predictors, both rider groups are almost indistinguishable. However, for weekend riders a different pattern can be observed. As our analysis, they constitute the group that is most difficult to predict. The reason for this can be found in the relevance of personal travel histories for prediction. This is reflected by the varying results for *Personal* and *Global* across the different rider groups. For weekend riders, personal histories are not very useful. Instead, the riders tend to use the the popular stops in the bus network. In contrast, for riders with weekday activity more regularity in behaviour is involved. As a result, the stops listed in their personal travel histories are accessed periodically. Notably, the *Random Walk* approach achieves the best APR across all different rider groups. This again demonstrates the effectiveness of combing personal usage data with related global mobility patterns. This way, accurate mobility predictions can be constructed for riders of different groups, such as weekday and weekend riders.

VII. CONCLUSION

In this paper, we have presented a large-scale data analysis of mobility behaviours of urban bus riders. By making use of travel card data from Lisbon, Portugal, we have analysed various approaches for predicting the bus stops used by individual riders for taking bus rides. In order to find relevant prediction variables, different features have been explored that represent influential factors on the rider's mobility choices, including spatial and topological travel distance, individual and popular stop usage as well as collective mobility behaviours. In our evaluation, we observed that the most accurate prediction takes into account personal ride histories and the mobility patterns of other riders. This work paves the way for a new generation of transport information systems which can take advantage of a better understanding of the mobility requirements in public transport scenarios, equally relevant for transport providers, third party application developers and finally the individual riders.

REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[2] A. I. Bejan, R. J. Gibbens, D. Evans, A. R. Beresford, J. Bacon, and A. Friday. Statistical Modelling and Analysis of Sparse Bus Probe Data in Urban Areas. In *Proc. of the 13th Intl. IEEE Conf. on Intelligent Transportation Systems (ITSC '10)* , 2010.

[3] V. Belik, T. Geisel, and D. Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Phys. Rev. X*, 1:011001, Aug 2011.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[5] T. Camacho, M. Foth, and A. Rakotonirainy. Pervasive Technology and Public Transport: Opportunities Beyond Telematics. *IEEE Pervaisive*, pages 18–25, 2013.

[6] I. Ceapa, C. Smith, and L. Capra. Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data. In *Proc. of the ACM SIGKDD Intl. Workshop on Urban Computing*, 2012.

[7] A. Chatterjee and M. M. Venigalla. Travel demand forecasting for urban transportation planning. In *Handbook of Transportation Engineering, Volume I: Systems and Operations*. AccessEngineering, 2011.

[8] L. Ferrari, M. Berlingerio, F. Calabrese, and B. Curtis-Davidson. Measuring public-transport accessibility using pervasive mobility data. *IEEE Pervasive Computing*, 12:26–33, 2013.

[9] B. Ferris, K. Watkins, and A. Borning. Onebusaway: A transit traveler information system. In *Mobile Computing, Applications, and Services*, pages 92–106. Springer Berlin Heidelberg, 2010.

[10] B. Ferris, K. Watkins, and A. Borning. OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit. In *Proc. of the 28th Intl. Conf. on Human Factors in Computing Systems (CHI '10)*, 2010.

[11] S. Foell, G. Kortuem, R. Rawassizadeh, S. Phithakkitnukoon, M. Veloso, and C. Bento. Mining Temporal Patterns of Transport Behaviour for Predicting Future Transport Usage. In *Proc. of the 3rd Workshop on Pervasive Urban Applications (PURBA '13)*, 2013.

[12] B. Gardner and C. Abraham. What drives car use? A grounded theory analysis of commuters' reasons for driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10:187 – 200, 2007.

[13] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[14] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *Proc. of the 32nd Intl. Conf. on Research and Development in Information Retrieval (SIGIR '09)*, 2009.

[15] M. Li, B. Du, and J. Huang. Travel patterns analysis of urban residents using automated fare collection system. In *Proc. of the 12th Intl. Conf. on Intelligent Transportation Systems Telecommunications (ITST '12)*, 2012.

[16] L. Liu, A. Hou, A. Biderman, C. Ratti, and J. Chen. Understanding individual and collective mobility patterns from smart card records: A case study in shenzhen. In *Proc. of International IEEE Conference on Intelligent Transportation Systems (ITSC '09)*, 2009.

[17] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Proc. of ASE/IEEE Intl. Conf.on Social Computing (SocialCom '12)*, 2012.

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

[19] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proc. of the ACM Conference on Computer Supported Cooperative Work (CSCW '94)*, 1994.

[20] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proc. of the 10th Intl. Conf. on World Wide Web (WWW '01)*, 2001.

[21] C. Smith, D. Quercia, and L. Capra. Anti-gravity underground? In *Proc. of the Second Workshop on Pervasive Urban Applications (PURBA '12)*, 2012.

[22] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, Feb. 2010.

[23] Transport for London. Understanding the travel needs of londons diverse communities, December 2011.

[24] Wilson, Nigel H.M., J. Zhao, and A. Rahbee. The potential impact of automated data collection systems on urban public transport planning. In *Schedule-Based Modeling of Transportation Networks*, volume 46, pages 1–25. Springer US, 2009.