Investigating Ride Sharing Opportunities through Mobility Data Analysis

Nicola Bicocchi\textsuperscript{1}, Marco Mamei\textsuperscript{2}

\textsuperscript{1}Dipartimento di Ingegneria “Enzo Ferrari”, \textsuperscript{2}Dipartimento di Scienze e Metodi dell’Ingegneria
University of Modena and Reggio Emilia, Italy

Abstract
Smart phones and social networking tools allow to collect large-scale data about mobility habits of people. These data can support advanced forms of sharing, coordination and cooperation possibly able to reduce the overall demand for mobility. Our goal is to develop a recommender system - to be integrated in smart phones, tablets, and in-vehicle platforms - capable of identifying opportunities for sharing cars and rides. We present a methodology, based on the extraction of suitable information from mobility traces, to identify rides along the same trajectories that are amenable for ride sharing. We provide experimental results showing the impact of this technology and we illustrate a Web-based platform implementing the key concepts presented.

Keywords: Mobility Patterns, Ride Sharing, Pervasive Computing.

1. Introduction
Travel has become an indispensable aspect of our lives. The current level of personal mobility was unheard of just 50 years ago, and it has shaped the way in which we build our communities, where and how we work and spend our leisure time. People travel more often and over longer distances than in the past - whether for commuting between home to work or school, shopping or going on holidays. But the freedom of personal mobility has brought it into a collision course with the boundary of finite resources. On a global scale, personal mobility is now responsible for 26\% of carbon dioxide (CO2) emissions [1]. In Europe, mobility has the fastest growing energy demands of all sectors and, despite international agreements, is the only sector with consistently increasing emissions in most countries [1].
In response to these problems, improvements are expected from vehicle-to-vehicle or vehicle-to-infrastructure coordination schemes. Moreover, autonomous cars currently being developed promise to further improve fuel efficiency by removing the human factor. Another, more radical solution, could be based on a radical shift to renewable energy sources and the development of fully electric vehicles.

In this work we investigate a social, more than a technical, solution tackling the issue by reshaping the demand for mobility by supporting car sharing and ride sharing practices. Such an approach looks promising in that it can provide viable and high-impact solutions in a short time frame, and could address several issues related to personal mobility at the same time [2]. As reported in [1], mobility demand spans all aspects of our life and by no means it reduces to home-work commute. Accordingly, effective systems to reshape mobility demand have to consider all aspects of our life, including free time and leisure.

The increasing adoption of smart phones and social networking tools allow to collect large-scale information about the mobility habits of people, and can support advanced forms of sharing, coordination and cooperation that can drastically reduce the overall demand for mobility [3, 4, 5]. Specifically, our proposal is to develop a recommender system - to be integrated in smart phones, tablets, and in-vehicle platforms - capable of identifying opportunities for sharing cars and rides. In particular, such a system would:

1. Collect mobility information
2. Identify routine behaviours
3. Identify sharing opportunities (e.g., two users who do not know each other but live and work close by - familiar strangers [6])
4. Recommend users about mobility alternatives

The key novelty of our proposal is that our system is fully autonomous. Users do not enter their availabilities and needs explicitly. Sharing opportunities are automatically identified by analysing people mobility patterns, and are recommended when suitable conditions arise. This kind of unobtrusive and proactive approach could facilitate the application diffusion and improve its effectiveness [7].

The contribution of this paper is twofold: on the one hand, we present a methodology to extract suitable information form mobility traces. On the other hand, we describe the implementation of a prototypical system offering ride sharing opportunities on the basis of the extracted information.
The rest of the paper is structured as follows: Section 2 discusses related work in the area both in terms of data mining mechanisms for mobility data, and in terms of ride sharing applications. Section 3 introduces our approach for automatically discovering and labelling routine behaviours and describes some experiments conducted with two complementary real-world datasets. Section 4 discusses our proposals for identifying possible shared rides on the basis of the discovered routines, illustrates our implementation and discusses about privacy issues associated with these systems. Section 5 concludes the paper and illustrates further developments.

2. Related work

An increasing number of research proposals applies advanced data mining techniques on mobility data with the goal of improving mobility services in smart city scenarios. In this section, we analyse related work in the areas of both mobility analysis and innovative mobility services.

2.1. Mobility Analysis

The availability of affordable localisation mechanisms and the recognition of location as a primary source of context information has stimulated a wealth of work trying to extract high-level information from raw mobility traces. While a complete survey is outside the scope of this paper, we present some exemplary researches trying to emphasize the novel aspects of our work.

The CitySense project (www.citysense.com) uses GPS and WiFi data to cluster people whereabouts and discover hotspots of activity in the city area. In a similar work based on extremely large anonymized mobility data coming from Telecom operators authors were able to extract the spatio-temporal dynamics of the city, highlighting where people usually go during the day. The authors were able also to identify the most visited areas by tourists and the typical time of the visit (see for example [8], [9]). While these works focus mainly on hot-spot identification, our approach goes further and is able to identify and label patterns and routine behaviours that will be useful to identify ride-sharing opportunities.

The approach proposed in [10] uses Principal Component Analysis (PCA) to identify the main components structuring daily human behaviour. The main components of the human activities, which are the top eigenvectors of the PCA decomposition are termed *eigenbehaviours*. Similarly, the work
presented in [11] compares different data mining techniques to extract patterns from mobility data. In particular, they found that Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are best suited to the task of identifying daily patterns. In comparison with these techniques, the topic model we propose has the advantage of capturing characteristic trends occurring over parts of the day (such as lunch time only), whereas eigenbehaviours tend to capture features over the entire day (see Section 3.4 for more discussion with this regard).

In [12, 13] authors propose the use of probabilistic topic models to capture human routines from cell tower connections. Our work uses a more complex dataset, thus allowing to analyse the topic models method at a finer-grain scale with a higher number of places. As mentioned above, the geographic coordinates provided by our GPS and CDR datasets allow to enrich the location vocabulary with a higher number of places (in contrast with the ‘home’, ‘work’ and ‘elsewhere’ labels used in [12]). In addition, in our work we also present algorithms to automatically label topics in order to make them more understandable and usable.

The M-Atlas approach is a recent proposal to extract patterns from mobility data [14]. M-Atlas creates an origin-destination matrix counting the number of repeated trips in the dataset. High-values in this matrix represent recurrent patterns. The proposed topic model is more flexible in that it can identify patterns even if no direct trajectories between places are present.

The evaluation of most of the researches in this area has been mainly qualitative, largely because of lack of ground-truth. Novel methodologies to quantitatively assess results are needed. In particular, the idea of using multiple data sources – observing user daily life from multiple perspectives – to cross validate results seems very promising. We presents some experiments evaluating our proposal in Section 3.6 and in Section 4.1.

2.2. Mobility Services and Car Sharing Applications

The idea of reshaping mobility demand by supporting ride sharing and car pooling activities is not new [15]. For example, systems like - www.mitfahrzentrale.de provide advanced car pooling solutions since 1998. However, the vast majority of systems that are in place today supports ride sharing for planned and long-distance (> 100km) travels. The novel opportunities offered by pervasive technologies (and exploited in our work) lies in bringing ride sharing into routine drives within the city. Pervasive technologies and
mobile computing allow in fact to automate the process of identifying and negotiating matching rides among users.

BlaBlaCar [16], a French startup, is a fully working web service enabling peer-to-peer ride sharing among users. Users can buy and sell rides on an eBay-like platform. Ride fares are established by the car owner and eventually negotiated in person. UberPop [17], is a peer-to-peer ride sharing service offered by Uber. It extends the Uber taxi-like service by enabling users to contact directly other users. Rides have a time-based fare. Flinc [18], is another ride sharing service gaining popularity. It has been designed to deal both with occasional trips and frequent routines. It embeds a social network to keep users connected and enforce trust-based policies. It is also integrated with widespread navigation platforms allowing users to offer and search rides from the car deck. However, while these systems use recent Web technologies, they still do not automate the process of identifying sharing opportunities. Users have to explicitly enter their availability and needs. The system we present, instead, performs the latter step in an automatic way on the basis of past user’s routines. Accordingly, our mechanisms could be effectively integrated in systems like BlaBlaCar, UberPop or Flinc for providing users with an automatic way to participate.

The work presented in [19] describes a multi modal journey planner that, relying on a number of mobility data, computes the probability distributions of end-to-end travel times, taking into account uncertain departure, travel and transfer times. While the goal of this system is rather different from the one we present, the basic mechanisms to extract information from underlying mobility data are similar.

The research presented in [20] analyses mobile phone location data to measure the amount of overlapping in people daily mobility as a pre-condition to enable ride sharing opportunities. They show that simple ride-sharing among people having neighbouring home and work locations can reduce the number of cars in the city at the expense of a relatively short detour to pick up/drop of passengers. Our work is very similar in spirit with this research, in addition the mobility analysis we propose can improve and generalise mobility routine identification. The proposed topic model, in fact, can identify also routine behaviours other than home-work commute.

Another example of using social interactions to ease traffic problems is described in [21]. Authors propose to use crowd-sourcing to collect, and keep updated the information about parking slots in the city. Although they focus on a different case study, we think that some of the algorithms
we propose to automatically extract places and routine behaviours could effectively complement crowd-source information collection.

The work presented in [15] use a probabilistic model to analyse the likelihood that a person will be successful in finding a ride-match, given a pool size of potential ride matches. In this work, the authors also observe that there are many obstacles, primarily in terms of communication and social norms, preventing ride sharing from actually happening. The work in [22] analyses this latter issue further, trying to identifying the social enablers for ride sharing. In this direction is also the work on persuasive technologies [7, 23] that analyses social and psychological aspects related to behaviour-influencing technologies and crowd-source coordinated activities. Although our work does not deal with this issues yet, these aspects are very important for the practical applicability and success of ride sharing systems. In our future work we will investigate these aspects trying, for example, to leverage on the social links among people as an information to foster sharing practices. For example, a user might be more inclined to share a ride with a friend-of-a-friend rather than with a complete stranger. In this line of inquiry is also the analysis of incentives and reputation mechanisms [23] supporting ride sharing activities.

3. Discovering and Labeling routine behaviours

The architecture in Figure 3, 6, 7 summarises the proposed approach. We analysed mobility data using two different datasets described in Section 3.1. Our approach works as follows: the most visited places by the user are automatically identified. Moreover, they are labeled either via reverse geo-coding, or by names provided by the user (see Figure 3). Then, we show how a probabilistic data mining technique, namely Latent Dirichlet Allocation (LDA), can both successfully extract routine behaviours in an unsupervised manner and make mobility data more meaningful. Eventually, extracted routines are automatically labeled via semantic descriptions (see Figure 6). The final result (see Figure 7) is the input of the ride sharing module described in Section 4.

3.1. Dataset

The research has been based on two complementary datasets.

1. We recorded the daily whereabouts of three persons over the period of almost one year. One subject is one of the authors of this paper, the
Figure 1: GPS traces collected via Google Latitude (upper two lines). CDR data collected from the Telecom operator (lower two lines).

other two subjects are not part of our research group. Subjects recorded data by continuously running Google Latitude in background. Figure 1 illustrates the schema of this dataset.

2. We obtained a large set of mobility data from an Italian telecom operator. In particular, we analysed anonymized CDR data (Call Detail Records) for 1000 persons living in the same city spanning 1 month. For each user, CDR data logs the approximate location of the user whenever (s)he uses the phone to send or receive calls or text messages, or access the network. The user’s location is given in terms of the cell network antenna the user was connected with. The area covered by a given antenna can be approximated by a circle with a given center and radius. Figure 1 illustrates the schema of this dataset.

Despite the structural similarity, these datasets are rather different. The former dataset is very fine grained with information on users’ GPS locations every few minutes. We have ground-truth information and we can ask the users to validate results of data analysis. The latter dataset is much more approximate. For each user we have about 10 locations per day, we do not have ground-truth information, nor we can interact with the users. On the other hand, we have a lot of users to experiment with. Figure 2 illustrates some key distributions for the telecom dataset. In particular, we present the users’ distributions both in terms of the number of trips being conducted and the number of kilometres. These distributions show that we are focusing on a rather large number of short movements. As discussed in the related work section, while traditional ride sharing systems provide solutions for planned long distance trips, one innovation of our approach is to focus on shorter routine movements, typically happening within a city.
3.2. Key Places Identification

The first step to process data is to identify the places most visited by the user (see Figure 3-left). Mainstream approaches are based on segmenting and clustering location-traces to infer what are the places relevant to the user [24]. More in detail, we created a grid of non-overlapping cells (500x500m size) over the area visited by the user. Cells in which on average the user spends more than an hour per week are marked as relevant (see Figure 4).

Once key places have been identified, it is important to name them. Asking the users to label key places is a practical way to get concise and understandable descriptions. This is basically the approach used by Foursquare and Google Now which tries to predict where people might travel to on the basis of their past activities. In addition, it could be also possible to reverse-geocode a given location to discover what is in there, and use that information for labelling. This latter approach tends to produce noisier labels, but would be completely unsupervised. We adopted the first mechanism for the first dataset. On the contrary, we adopted the second mechanism for the CDR data in that we could not contact the original subjects involved in the data collection. We automatically name Home “H” the place most visited at night and Work “W” the place most visited during the day. We named other places with the name of the grid cell overlaid to the area (see Figure 4).

An informal survey conducted with the Google Latitude users in our study verified that the mechanism correctly labels their home and workplace. We
applied the same algorithm to the CDR dataset extracting few key places for each user. In this case we did not have ground-truth information to validate the results. However, some evaluation with regard to this dataset is presented in Section 3.6.

3.3. Bag of Words Representation

In the second step, following the approach proposed in [12, 10], we organised the dataset into a sequence of days each consisting of 24 time-slots lasting 60 minutes each. For each time-slot, if the associated location records fall within 500 meters from a place identified in the previous step, then we mark the time-slot with that place. Otherwise, we mark the location as NULL. Following this process, each day is then represented by a string of 24 symbols. To capture transitions between locations, in the final step of the pre-processing phase, we run a sliding window over each day. The sliding window takes 3 consecutive symbols and concatenates them with another label capturing the time of the day where these locations have been visited. In particular, we considered the following 4 time labels: 0-6am (1 - night), 7am-1pm (2 - morning), 2pm-6pm (3 - afternoon), 7pm-11pm (4 - evening). These time segments were chosen to capture common events in daily life, such as lunch time (transition form time labels 1 and 2), dinner time (time label 4), or morning and afternoon work times. The result is a set of words each containing 3 location letters and a time label (see Figure 3-right). Each word also keeps a reference to the actual geographic coordinates in order to localise the place.

Figure 3: General architecture of our approach (part 1). We start from a log of GPS traces. We apply a place identification algorithm to discover the places most visited by the user. Places are annotated either by user input or by reverse geo-coding. Places are organized in a bag of words representation.
Figure 4: Grid-based key place identification technique. GPS points are accumulated over a grid (500x500m size). Cells reaching a certain threshold, are marked as relevant. In the example above, a day of a single commuter has been processed. Home and work places have been correctly identified. The other marked cells are those that accumulated GPS points but did not reach the threshold.

The resulting bag of words summarises the original dataset and is the input data structure for the algorithm to extract routine behaviours (i.e., topics) and add semantics.

3.4. Routine Identification

Discovering routine behaviours is an important step to add semantics to users’ whereabouts. On the one hand, patterns and routines represent a step further in describing information about the user. On the other hand, they represent also information about how a place is “used” by a given user. So that the “Fox pub” can be a place where to go after work for some users while it is the workplace itself for the bar tender.

LDA is a probabilistic generative model [25] used to cluster documents according to the topics (i.e., word patterns) they contain. LDA is an unsupervised learning mechanism that do not require a labeled (difficult to be acquired) training set. Moreover, LDA has two key advantages compared to other clustering mechanisms (such as k-means): (i) The LDA model results in probabilistic distributions of days given all topics whereas other clustering algorithms (e.g., k-means) assigns only one cluster per day. (ii) Meaningful word distributions as the representation of topics. Topics are based on discriminative location sequences characterising routines [12].

More in detail, LDA is based on the Bayesian network depicted in Figure 5. A word $w$ is the basic unit of data, representing user location at a given
time-label. A set of \( N \) words defines a day of the user (i.e., a document). Each user has a dataset consisting of \( M \) documents. Each day is viewed as a mixture of topics \( z \), where topics are multinomial distributions over words (i.e., each topic can be represented by the list of words associated to the probability \( p(w|z) \)). For each day \( i \), the probability of a word \( w_{ij} \) is given by 
\[
p(w_{ij}) = \sum_{t=1}^{T} p(w_{ij}|z_{it})p(z_{it})
\]
where \( T \) is the number of topics. \( p(w_{ij}|z_{it}) \) and \( p(z_{it}) \) are assumed to have multinomial distributions. Mixture parameters are assumed to have Dirichlet distributions with hyper-parameters \( \alpha \) and \( \beta \) respectively. Hyper-parameters \( \alpha \) and \( \beta \) were both set to 1 representing an uninformative uniform distribution. LDA can use Gibbs sampling to learn the model parameters. In our implementation we use the library Mallet (http://mallet.cs.umass.edu) to perform these computations. Once the model parameters have been found, Bayesian deduction allows to extract the topics best describing the routines of a given day (rank \( z \) on the basis of \( p(d|z) \)).

However, as already introduced, since \( z \) are just distributions over words, it is difficult to give them an immediate meaning useful in applications.

We want to emphasise that we are not extending the LDA model, we take the model as it is. On top of LDA we propose mechanisms to give labels to topics \( z \) and to predict which topic \( z \) the user is in. Moreover, in Section 4 we use the extracted topics to identify ride sharing opportunities.

In Figure 6-left, we illustrate two exemplary topics extracted from one of the user of the Google Latitude dataset (topic 0 and topic 21). In particular, we present each topic by listing the top words (ranked by \( p(w|z) \)).

- **Topic 0** captures the *home-work-gym-work* routine. The most probable words for such topic are WWW2 and HHH3, which are respectively being at work in **time-slot 2** (7am-1pm) and being at home in **time-slot 3** (2pm-6pm). They are followed by working in **time-slot 2** (7am-1pm) and going from work to gym and then back to work in **time-slot 2** (7am-1pm). From the distribution of the routine over the days of the week we can see that it corresponds to a weekdays trend (Tuesdays and Thursdays in particular).

- **Topic 21** captures the *pub-home-pub* routine. The corresponding top words illustrate that the user is in a pub from 0am to 6am (PPP1), then he remains at home during the day (HHH2, HHH3) and finally he moves from home to the same pub in **time-slot 4** (7pm-11m). From the distribution of the routine over the days of the week we can see that it corresponds to a weekends trend.
Figure 5: Plate notation representing the LDA model. $\alpha$ is the parameter of the uniform Dirichlet prior on the per-document topic distributions. $\beta$ is the parameter of the uniform Dirichlet prior on the per-topic word distribution. $\theta_i$ is the topic distribution for document (day) $i$. $\phi_j$ is the topic distribution for word $j$. These variables are modelled with a Dirichlet distribution. $z_{ij}$ is the topic for the $j$-th word in document $i$, and $w_{ij}$ is the specific word. These variables are modelled as multinomial distributions. The $w_{ij}$ are the only observable variables, and the other variables are latent variables.

These results illustrate that the LDA model applied to GPS data successfully reveals different types of patterns.

The above results illustrate also one of the key advantages of LDA compared to other clustering mechanisms (e.g., k-means). While most other clustering algorithms group together days that are similar for the whole 24 hours (i.e., they associate a given day - feature vector with a single cluster), LDA can cluster days that are similar only in a given time interval (i.e., they associate a given day - feature vector with multiple topics). For example, LDA can cluster the days in which the user went to a given place in the afternoon, even if those days have very different signatures in the morning. Other clustering mechanisms are not able to identify that cluster since they consider whole days only [25, 12].

3.5. Routine Labeling

As from the previous examples, automatically understanding topics can be difficult, in that they are raw probability distributions over words representing places and time of visit [13]. A further step in the meaningful description of user whereabouts would be to automatically attach to each topic a concise and meaningful label describing the user routine (e.g., the user is in the “Gym at lunch break” routine). Applications and services could then
Figure 6: General architecture of our approach (part 2). We use the Latent Dirichlet Allocation (LDA) algorithm to extract patterns and routines from user whereabouts. Results are LDA Topics representation and associated distribution over the days of the week. Letters in words represent places: W = Workplace, H = Home, G = Gym, P = Pub and they are all associated to actual geo coordinates. Patterns are semantically described via automatic labels. The graph presents tf-idf weights of all the words. Highlighted words are those constituting the topic label.

simply take context-aware decisions on the basis of the label without the need of processing distributions like those in Figure 6-left.

Our approach is based on two ideas:

- Words comparing with high probability are those best describing the topic
- Words that are present in many topics are not much characteristic of any topic (e.g., Home at 3am is something really common and not peculiar of a specific routine).

Accordingly, our algorithm works as follows:

1. The probability of each word $w$ in each topic is multiplied by the inverse document frequency (idf) of that word in the corpus of user days, i.e., $\log(\frac{\#\text{words}}{\#w})$. This term complies with the second idea in the list above.
2. For each topic, we identify its peculiar words, that are those that are not present in several other topics. In particular, for each time slot, we computed the median probability (corrected by point 1) of that word in all the topics. Words that are outliers with respect to this median are considered peculiar for this topic. Outliers detection have been conducted using the MAD median rule [26].
Figure 7: General architecture of our approach (part 3). Graphic representation of two discovered routines: spatial distribution of the routine displayed on a map. The larger the circle, the higher is the probability associated to the given place. These results will be used to find users with similar routines to identify ride sharing opportunities – see next Section.

3. We create the label for each topic by ordering the topic’s peculiar word according to their time stamp. For each word we take the place in which the user spends most of the time (e.g., WHH = ‘Home’).

As an example, assuming that all the words of the topics reported in Figure 6 are peculiar (e.g., they are characteristic of that topics), topic 0 would be described by the label “Home - Work - Gym - Work”. Topic 21 would be described by the label “Pub - Home - Pub”.

To evaluate the appropriateness of the label to describe the topic, we computed the tf-idf weight (term frequency – inverse document frequency) of each word with respect to the identified topic. Tf–idf is a weight, often used in information retrieval and text mining, to evaluate how important a word is to a document in a collection or corpus. Figure 6-right shows the tf-idf weights of all the words. Words considered for the label are indeed those having the highest weight. Time stamps having no words with a spike (i.e., having all words with low tf–idf weights) are not considered for the label, since they are not peculiar for the given topic.

3.6. Validation and Discussion

Given the lack of accurate ground-truth information, and the fact that topics cluster data in classes that are not defined a priori, it is difficult to provide sound measures on the accuracy of the obtained results (this is a general limitation of the current state of the art). As a partial solution, we collected Google Latitude dataset in order to evaluate the outcome of the proposed approach. Talking with the Google Latitude subjects in our study
the topics being extracted are reasonable and describe parts of the actual users’ routine behavior.

To better understand the result of our topic modelling, we investigate the contribution of different topics in describing the days of the user. In particular, we show results from one of the Google Latitude users (the one selected also to generate results in Figure 3, 6, 7).

We started by looking at days for which \( p(z_i|d) \leq T_L, \forall i \in [1, T] \) where \( T \) is the number of topics and \( T_L = 0.5 \). These days are not strongly associated to any topic and thus can be described only by combining some topics together. For each topic \( i \) we computed the probability of \( z_i \) given a day which is not highly probable for any topic (i.e., the above set of days). Such a distribution is depicted in Figure 8(a). Topics having a high value in this distribution are topics that do not explain a complete day, but parts of a lot of days. For example, a topic with words associated only to being at home at night would belong to this category since it does not explain the whole day.

Vice Versa, we identified days which are best represented by few topics by collecting days for which \( \exists i \in [1, T] : p(z_i|d) > T_H \), where \( T_H = 0.5 \). In Figure 8(b) we plot the probability distribution of topics given days well represented by few topics. Topics having a high value in this distribution are topics that describe a complete day. For example, a topic associated to being at home early morning and at night, and being at work during the day would belong to this category since it explains a complete day.

Finally, in Figure 8(c) we show an histogram of the number of dominating topics per day. We compute the number of topics composing at least 90% of the probability mass of each day in the study. The histogram highlights that
the subject under investigation follows a consistent routine and few topics are enough to explain most of his days.

In general it is possible to see that topic modelling works best, like in this case, when the number of relevant locations for the user is limited. In this situation, there is enough repetitiveness in the user routine to let the LDA algorithm effectively identify relevant topics. It is worth noticing that most related work [10, 12, 13] deal in fact with only 3 kind of places (Home, Work and Elsewhere). In our experiments we considered users having up to 10 different locations, that are suitable to identify most of their routine behaviours.

In addition to these experiments on the Google Latitude dataset, we run also some analysis on the larger telecom dataset. In these experiments we tried to measure the fraction of daily movements that are represented in the top words of the most probable topic (see Figure 9(a)). Vice versa we tried to measure the fraction of the topics’ top words that are represented by the days associated with that topic. (see Figure 9(b)). The two graphs show the distribution of such a fraction over all the users. It is possible to see that even considering only the most probable topic’s top words we are able to represent a large fraction of the user routine movements. However, from Figure 9(b), it is possible to see that some movements described in a topic are not always present in the actual data. In the lack of ground-truth information, we speculate that some of that movements did actually take place and where simply just not recorded by the CDR data.

4. Ride Matching

Once routine movements have been identified, the next step is to discover similarities among them. In particular, given a user $A$ and one of her routines $r_u$ we want to identify a set of $n$ routines from other users “covering” $r_u$.

More in detail, the proposed algorithm works as follows: each routine $r_u$ that is typically followed by a user $u$ on some days of the week can be factored in a set of transitions described by the travel’s source $s$ and destination $d$ coordinates, and the time $t$ in which the transition happens.

The algorithm selects all the routines $r$ performed by users other than $u$ on days of the week that overlaps with the set days. Then, for each transition $t_x \in r_u$, the algorithm scans all the transitions $t_y \in r$ looking for matches [27]. In particular – calling $t_x.t$ the time frame in which the transition happens,
Figure 9: Evaluation of topic modeling approach. a) fraction of daily movements that are represented in the top words of the most probable topic. b) fraction of the topics’ top words that are represented by the days associated with that topic.

\[ t_x.s \text{ the transition starting coordinates and } t_x.d \text{ the transition destination coordinates – we define that two transitions } t_x \text{ and } t_y \text{ match if:} \]

\[ t_x.t = t_y.t \land \text{dist}(t_x.s, t_y.s) < \delta \land \text{dist}(t_x.d, t_y.d) < \delta \]

Where \( \text{dist} \) is the geographic distance between the points and \( \delta \) is a threshold parameter. It is worth noticing that the current approach is rather simple: two transitions match if they happen in the same time window, and if their starting and ending points are close to each other. More advanced matching could rely on the fact that a user could pick up/leave another user in between of a transition. Our current algorithm does not consider this possibility since it would require a detailed road map of the area under analysis to understand if the pick up/leave is actually compatible with user route. We will address this extension in our future work.

4.1. Experimental Evaluation

To evaluate the impact of the proposed approach we conducted a set of experiments with the dataset from the telecom operator described in Section 3.1. Specifically, we wanted to show the number of both movements and km that could be saved.

As described in Section 3.4, we applied the LDA algorithm and selected the first 1000 most active users. The mobility routines of each user are de-
Figure 10: Ride matching approach. The routine of user A is “covered” by the routines of users B and C. Successful ride sharing would allow user A to leave the car at home and going to work with user C and coming back home with user B.

scribed by a set of LDA topics. Each topic is essentially a list of movements described as a tuple like the following: (source time, source longitude, source latitude, destination time, destination longitude, destination latitude, probability). To make our investigation more meaningful, we removed all the tuples with the following features:

- distance between geographic source and destination < 100m (i.e., the user spending time in the same place)
- probability < 0.1 (i.e., topics that are likely to be poorly informative about actual habits of the user)
- duplicated tuples (i.e., the same movement included in different topics)

By applying this filtering stage to the initial set of about 50000 movements, we obtained 7700 unique movements (characterised in Figure 2). Then, we spatially and temporally aggregated all movements. Specifically, we aggregated movements with the same time values and with spatial sources and destinations comprised within a search radius variable between 100m and 2000m. Furthermore, we considered the capacity boundary of cars and avoided to aggregate more than 4 movements together. What we achieved is depicted in Figure 11. The number of possibly saved trips halves with a radius of 1300m, while exceeds 60% with a radius of 2000m. Despite, these results are based on the assumption that all users participate to our sharing
system, they are encouraging especially in light of the fact that more complex and effective matching algorithms could be used.

4.2. Implementation

We developed a first prototype of our system as a Web application (http://lica.lab.unimo.it). The web application connects to a mobile app developed for both Android and iOS recording GPS data for each user and allowing ride searches. The application runs the workflow we have presented, extracting users’ mobility routines and storing key places and movements in a database.

We still did not implement opportunistic (i.e., proactive) recommendations, as we are still trying to engineering the user interface for such advices. In the current implementation, the user has to access a Web interface to query for available ride sharing options. The system applies the described ride matching algorithm to select users with similar routine behaviours.

The system connects via email to the selected user indicating the “ride sharing” request. On acceptance, the system notifies both users about the complete transaction. Figure 12 illustrates a screenshot of the application. In the current system, we did not implement any compensation/payment schema for a successful ride share.
4.3. Privacy Issues

Privacy issues are fundamental for the applicability of the idea presented in this work. In fact, even if the practical advantages (both at the individual and social level) would be consistent, only few people would allow to share detailed mobility data because of privacy concerns. To mitigate these issues we propose a layered approach. In general, adding layers increases both the complexity of the system and the level of anonymity.

The first layer regards the physical devices in which raw data are stored and LDA models computed. While, at the time of writing, computation is done in a centralised fashion, it would be possible to let each user run the LDA algorithm locally (e.g., in a smart phone) and transmit to a central provider the resulting LDA model. Although the LDA model still contains privacy sensitive information, it is less detailed (i.e., more privacy-compliant) that the raw traces. In fact, it does not contain every spatial movements but only the recurrent ones.

Furthermore, as a second layer, it is worth considering that the matching algorithm among users’ traces naturally tolerates approximate locations. In
fact, a user $A$ can share a ride with user $B$ even though the source and the destination of the trips do not perfectly overlap. This would allow to deliberately approximate users’ locations within LDA models thus enabling the system for k-anonymity [28].

Finally, as a third layer, LDA models could be exchanged in a peer-to-peer fashion by making use of technologies such as distributed hash tables. As for other well-established networks such as FreeNet [29], both requests and replies (i.e., matching LDA models), could be rooted through the overlay. Communication messages between nodes are ciphered and the actual source and destination IP addresses are obfuscated. Thus for an eventual attacker, it would be difficult to collect and reconstruct LDA models of specific users.

As a final general remark, the proposed topic-modelling approach and its inherent tolerance in identifying ride sharing opportunities already support some basic privacy issues. In our future work we will try to actually implement these aspects in our system.

5. Conclusion

A number of problems in our society, from urban management to environmental issues, require highly collaborative approaches in which people are asked to gradually change their behaviours and habits. Pervasive and mobile computing technologies allow to support such collaborative activities by automatically monitoring people behaviours and suggesting possible improvements. In this direction, the paper presents the idea of a recommender system capable of identifying opportunities for sharing cars and rides by analysing mobility data. In particular, we presented a methodology to automatically extract suitable information from mobility traces and identify sharing opportunities. Experimental results show the feasibility of the approach and the potential to mitigate several traffic-related issues. In our future work, other than better assessing the accuracy and generality of the approach, we will conduct user studies to evaluate the contribution of the identified whereabouts patterns in the development of behaviour-influencing applications.

Acknowledgements: Work supported by the ASCENS (Autonomic Service Components Ensembles) project (FP7-FET, Contract No. 257414). We would like to thank Laura Ferrari who realized the LDA approach. We would
also like to thank our students: Dario Stabili, Andrea Borghi, Marco Frigeri and Francesco Bartolacelli who developed the Web site of the application.


