Assessing the Potential of Ride-Sharing Using Mobile and Social Data: A Tale of Four Cities

Blerim Cici, Athina Markopoulou
University of California, Irvine, USA
{bcici, athina}@uci.edu

Enrique Frias-Martinez, Nikolaos Laoutaris
Telefonica Research, Spain
{efm, nikos}@tid.es

ABSTRACT

This paper assesses the potential of ride-sharing for reducing traffic in a city – based on mobility data extracted from 3G Call Description Records (CDRs), for the cities of Madrid and Barcelona (BCN), and from OSNs, such as Twitter and Foursquare (FSQ), collected for the cities of New York (NY) and Los Angeles (LA). First, we analyze these data sets to understand mobility patterns, home and work locations, and social ties between users. Then, we develop an efficient algorithm for matching users with similar mobility patterns, considering a range of constraints, including social distance. The solution provides an upper bound to the potential decrease in the number of cars in a city that can be achieved by ride-sharing. Our results indicate that this decrease can be as high as 31%, when users are willing to ride with friends of friends.

Author Keywords

Ride-sharing; mobile networks; social networks; geo-social data; call-description records (CDRs).

ACM Classification Keywords

H.2.8 Database Applications: data mining; H.4.m Information Systems Applications: Miscellaneous; C.2.5 Local and Wide-Area Networks; C.2.m Miscalleneous

INTRODUCTION

Ride-sharing is a promising approach for reducing car usage in a city, which is beneficial both for individuals [1], *e.g.* reducing gasoline, and for the city as a whole [2], *e.g.* reducing traffic and pollution. In recent years, a plethora of web and smartphone-based solutions have emerged for facilitating intelligent traffic management [3] and ride-sharing in particular. Early web-based systems, like `carpooling.com`, and `eRideShare.com`, provide matching of users for long distance travel as well as local ride-sharing, and have attracted a few million users across Europe and the US. More recently, companies like `Avego.com` or `Uber.com` offer smartphone apps that allow drivers and passengers to be matched; drivers offer cheaper peer-to-peer taxi services.

Smartphone-based ride-sharing technology gains momentum but still needs to deal with several issues including safety (traveling with strangers), liability (*e.g.* accidents), as well as the bootstrapping problem (the more users a ride-sharing service has, the more the ride-sharing opportunities). However, even if the above problems were completely solved, the opportunities for ride-sharing would still depend on the underlying human mobility patterns and the layout of a city, which ultimately determine the route overlap.

In this paper, we seek to understand what is the potential decrease in the number of cars in a city if people with similar mobility patterns are willing to use ride-sharing in their daily home/work commute. This is clearly an upper bound to the actual benefit of any practical system but it can be used to guide the deployment and policies regarding ride-sharing in a city. We assess this potential in four major cities using mobile and social data; we obtained two CDR data sets from a major cell provider (Madrid and BCN, in Spain, Europe), and we also collected data from Twitter (geo-tagged tweets) and FSQ (NY and LA, in US). A similar question has been asked before in [4], where the authors assumed a uniform distribution of home/work locations and concluded that ride-sharing has negligible potential. In contrast, we find that ride-sharing can provide significant benefits, depending on the spatial, temporal and social constraints for matching users. In particular, we take the following steps.

First, we infer home/work location of individual users, by adapting state-of-the-art techniques [5] to our CDRs and geo-tagged tweets. Also, we infer social ties among the users; we use phone calls in the CDR data and explicitly stated friendship in the Twitter data. These ties are later used for social filtering, to address concerns about riding with strangers.

Second, given a set of users with known home/work locations, we develop a framework for matching users that could share a ride. Our goal is to minimize the total numbers of cars and provide rides to as many users as possible. We consider several constraints including: spatial (ride-sharing with neighbors, *i.e.* someone within a certain distance from their home/work location), temporal (ride-sharing within a time window from the desired departure/arrival time) and social (ride-sharing with friends or friend-or-friends) constraints. We also consider two versions of the problem: `End-Points RS`, ride-sharing between home and work locations, and `En-Route RS`, allowing the possibility to pick up passengers along this route. Our formulation is rooted at the Capacitated Facility Location Problem with Unsplittable Demand. Since this is an NP-hard problem [6], and we want match more than 272K drivers and passengers, we develop efficient heuristic algorithms, namely `End-Points Matching` and `En-Route Matching` to solve the two aforementioned problems.

UbiComp '14, September 13 - 17, 2014, Seattle, WA, USA.

http://dx.doi.org/10.1145/2632048.2632055

Third, we use our framework to assess the inherent potential of ride-sharing to exploit the overlap in people's commute in a city. We find that there is significant potential for reducing traffic via ride–sharing, the exact magnitude of which depends on the constraints assumed for matching, as well as on the characteristics of the cities and the type of data set (CDR vs Twitter). For example, our study shows that traffic in Madrid can be reduced by 59% if users are willing to share a ride with people who live and work within 1 km; if they can only accept a pick-up and drop-off delay up to 10 minutes, this potential benefit drops to 24%; if drivers also pick up passengers along the way, this number increases to 53%. If users are willing to ride only with people they know ("friends" in the CDR and OSN data sets), the potential of ride-sharing becomes negligible; if they are willing to ride with friends of friends, the potential reduction is up to 31%. Albeit upper bounds to the actual benefit, these positive results encourage the deployment and policies in favor of ride-sharing.

The structure of the rest of the paper is as follows. Section 2 reviews related work. Section 3 presents our data sets, the methodology for inferring home and work location of individual users, and a characterization of the data sets that provides insight into the next steps. Section 4 provides the formulation of the `End-Points RS` problem, an efficient algorithm `End-Points Matching` for solving it, and results from applying it on the data sets. Section 5 provides the formulation of the `En-Route RS` problem, an efficient heuristic `En-Route Matching` for solving it, and results from applying it on the data sets. In Section 6 we first characterize the social ties in our data sets, then we further restrict the matching using social distance, and only allow users that know each other, or have common friends, to ride together. Section 7 provides a comparison across the four cities studied. Section 8 summarizes the results and concludes the paper.

RELATED WORK

Traditionally, carpooling studies focused in characterizing the behavior of carpoolers, identifying the individuals who are most likely to carpool and explaining what are the main factors that affect their decision [7]. Instead, in this paper we focus on assessing its potential for traffic reduction in a city. A similar study has been done before in [4], which assumed a uniform distribution of home/work locations in a city, and concluded that ride-sharing has little potential for traffic reduction. In contrast, we infer home/work locations from CDR and Twitter data and we find that they are far from uniform.

Some ride–sharing systems have been built over GPS [8, 9] data. He et al. [8] presents a route–mining algorithm that extracts frequent routes and provides ride-sharing recommendations based on these routes; they use the GPS traces of 178 individuals. Trasarti et al. [9] use GPS data to build mobility profiles for 2107 individuals, and match users with similar profiles; they also apply their algorithms to a GSM-like data set, which they synthesize by reducing the size of their GPS data. Bicocchi et al. [10] extract common routes from mobile traces and use them for ride-sharing recommendations. To the best of our knowledge, our work is the first attempt to study the potential of ride–sharing using CDR and OSN data. Although, our data have coarser granularity in terms of user trajectories (since we observe a user's location only when she makes a call or posts a geo-tagged tweet), they have information about orders of magnitude more users than previous carpooling studies and thus are better positioned to answer the question about the city-wide benefits of ride-sharing.

Compared to commercial ride-sharing systems, such as `Avego, Lyft, Uber`: our work is partly based on publicly available (*e.g.* geo-tagged tweets) as opposed to proprietary data; it has a larger number of users for the cities studied; it takes into account social ties for matching drivers and passengers; and it assesses offline the city-wide benefit of ride sharing, as opposed to online matching of passengers with a small set of dedicated drivers.

Our methodology on inferring home/work locations for individuals builds upon a recent work by Isaacman et al. [5, 11], on inferring important places from CDR. Social aspects of CDRs, *i.e.* the call graph, has been studied in [12], [13]. In this paper, we combine both aspects, namely inferred locations and social ties, to restrict ride sharing accordingly. We do the same with the Twitter data too.

Other related studies focus on characterizing crowd mobility and urban environments using information from Twitter or FSQ. Wakamiya et al. [14] and Fujisaka et al. [15] have used geo-tagged Twitter data to study crowd mobility, and Frias-Martinez et al. [16] to characterize land use. FSQ has been used by Noulas et al. [17], [18] for modeling crowd activity. To the best of our knowledge, Twitter and FSQ data have not been used for carpooling.

The most closely related work is our preliminary study [19]. Compared to [19], in this paper we make the following additional contributions: (1) we collect data from Twitter (geo-tagged tweets for NY and LA) in addition to CDRs, (2) we use CDRs from BCN, (3) we compare among the four cities, (4) we restrict ride sharing opportunities based on social ties, and (5) we estimate users' departure times from the data, instead of assuming a distribution.

INFERRING HOME/WORK LOCATIONS FROM DATA

The first step in assessing the benefits of ride-sharing is to infer where people live and work. To achieve this, we build on a state-of-the-art methodology that has been proven to infer important locations in people's lives with adequate accuracy [5]. We apply their methodology with some modifications in order to make it applicable to our scenario.

Data Sets

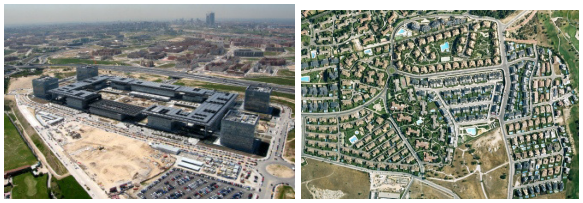
Tab. 1 summarizes our data sets, and the following subsections describe the data collection process.

Cell Phone Data

CDRs are generated when a cellphone makes or receives a call or uses a service, *e.g.* SMS. Information regarding the time/date and the location of the Base Transceiver Station (BTS), used for the communication, are then recorded. More specifically, the main fields of each CDR entry are the following: (1) the originating cellphone number (2) the destination cellphone number (3) a time-stamp (when call started) (4) the duration of the call and (5) the BTS tower used by

Data Set Name	Period	Number of Records	Total Users	Home/Work Users
CDR–Madrid	Sep 2013 - Dec 2013	820M	4.70M	272,479
CDR–BCN	Sep 2013 - Dec 2013	465M	2.98M	133,740
Twitter–NY	Nov 2012 - Feb 2013	5.70M	225K	71,977
Twitter–LA	Nov 2012 - Feb 2013	3.23M	155K	43,575
FSQ–NY	Nov 2012 - Feb 2013	362K	31.3K	–
FSQ–LA	Nov 2012 - Feb 2013	134K	13.6K	–
FSQ–US	Dec 2009 - Aug 2011	1.47M	40.1K	–

Table 1. Description of our data sets. The right column shows the number of users with inferred Home/Work locations, which is a subset of all the users. The Foursquare data sets were used to tune and validate the home/work inference methodology, for the Twitter data sets.



(a) Headquarters of Telefonica in Madrid (b) Home Area. Coordinates: 40.503736, -3.635469

Figure 1. Example of residential and working areas

one, or both if applicable, cellphones. CDRs' spatial granularity varies from a few hundred m^2, in urban areas, to up to a few km^2 in rural areas, therefore users' exact positions inside BTS areas are unknown. The CDRs used for this study are from the period of September 2009 – December 2009, excluding the last two weeks of December, which are holidays.

Twitter and Foursquare (FSQ)

Many users access Twitter from mobile apps and some of them choose to reveal their current location (typically as GPS coordinates) in their tweets, thus making Twitter an important source of human mobility information. We used the Twitter's *Streaming API* [20] in order to obtain individuals' mobility traces in large geographic areas. We collected geo-tagged tweets from the metropolitan areas New York and Los Angeles for a period of four months – from November 2012 until February 2013. This was possible thanks to Twitter's Public Stream Service where you can specify the geographic area that you are interested in. See Tab. 1 for more details.

Geo-tagged tweets contain location information, but they lack location semantics, which are crucial for inferring individuals' home/work locations and commuting routes. We collected this information from FSQ – a large location-based OSN with more than 30M users. FSQ does not provide an API for data collection but its users can post their check-ins in Twitter and other OSNs. We obtained FSQ check-ins from our Twitter data set. In addition, we exploited another FSQ data set that we obtained with the help of the authors of [21]. The latest data set was obtained by crawling publicly available tweets of check-ins in the US, and spans the period: December 2009 - August 2011. See Tab. 1 for more details.

Home/Work inference methodology

We apply the methodology of Isaacman et al. [5] for inferring important places for cell phone subscribers from (1) CDR data and (2) ground truth for a subset of subscribers. First the recorded cell towers of a user are clustered to produce the list of places that the user visits. Then, regression analysis is applied to the ground truth users (clusters and their true important locations) to determine the features of the clusters that represent important places. The used features are: (1) the number of days that the user appeared on the cluster; (2) the duration of user appearances on the cluster; and (3) the rank of the cluster based on number of days appeared. Once important locations have been inferred, and the algorithm chooses which of these are home and which are work locations. According to their results, the best features that characterize home and work are: (4) the number of phone calls between 7PM - 7AM, i.e. *Home Hour Events*, and (5) number of phone calls between 1PM - 5PM, i.e. *Work Hour Events*.

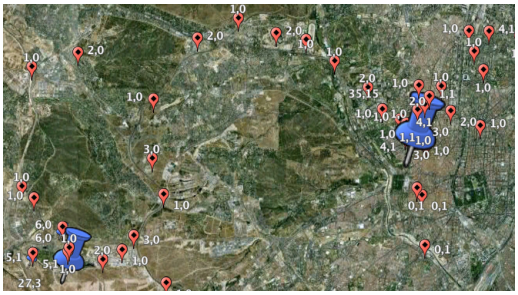
For our CDRs, first, we filter out users for whom we have too little data: *i.e.* users with less than 1 call per day on average, or less than 2 clusters with 3 days of appearance and 2 weeks of duration – the specific filtering parameters are consistent with [5]. Then, we tune the methodology of [5] to our needs. More specifically, we build two classifiers, one for home and one for work, and we train them using the 5 features described above and the ground truth, which is described in the following section. Once the training on the ground truth is done, we apply the classifiers to the rest of the users. Finally, after classification, we keep only the users who have only one inferred home location, and a different inferred work location, since we are interested only in commuters. Applying the home/work inference methodology to our CDR data, we are able to infer the home/work locations of more than 272K individual users in Madrid, and more than 133K users in BCN (See Tab. 1). Finally, we apply the same methodology in our Twitter data – FSQ data serves as ground truth – and we infer home/work locations for 71K users NY, and 43K users in LA.

Obtaining Ground Truth

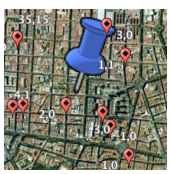
In [5], a set of 37 volunteers reported their most important locations, including home and work. This was used to tune their methodology, *i.e.* in the regression analysis, before applying it to the rest of their users – around 170K.

Ground Truth for CDR Data: For the CDR data, we obtained our ground truth for a select subset of users based on a previous study [22], which characterizes areas in Madrid. In particular, we exploited strictly residential and strictly industrial areas (see Fig. 1 for example), which offer a clear distinction between home and work. To this end, we selected 160 users that appeared for many days in only one such residential area during 7PM - 7AM ("home hours"), and only one such industrial area during 1PM - 5PM ("work hours"). Then, the location inside the residential area is pointed as the user's Home, while the location inside the industrial area is pointed as the user's work. For each one of the 160 users, we visually inspected their recorded locations through Google Earth. Fig. 2 shows a selected ground truth user.

Ground Truth for Twitter Data: We used the Foursquare data to build the ground truth for the geo-tagged Twitter data sets by selecting users who appear more than a week in a location tagged as Home(Private), and the same duration in a location containing one of the tags: Professional, Office, or Work. For



(a) A "ground truth" user



(b) Zooming in at home

(c) Zooming in at work

Figure 2. A ground truth example. The red paddles show the cell towers, while the blue pushpins the clusters. The numbers next to each mark indicate the number of weekdays and weekends she appeared in that location. Also, the size of each mark is proportional to the days of appearance.

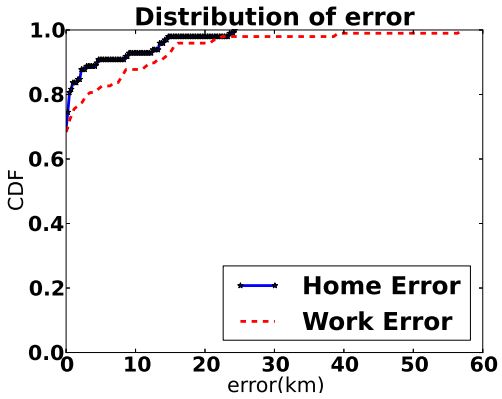


Figure 3. CDF of error for the home/work inference methodology. Inferred home/work locations from Twitter are compared against the declared locations in Foursquare.

Percentile	25^{th}	50^{th}	75^{th}	95^{th}
Our Home Error	0.0	0.01	0.49	13.62
Home Error in [5]	0.85	1.45	2.06	6.21
Our Work Error	0.1	0.03	1.52	16.09
Work Error [5]	1.0	1.34	3.7	34.17

Table 2. Comparing the home/work identification error to [5].

each one of these users we define their home to be the location tagged as home with highest number of days of appearance, and as work the location tagged as work with most days of appearance. We also manually inspect their Twitter account and, when possible, their LinkedIn accounts. In the FSQ-US data set, we found 481 such users, and in the FSQ-NY and FSQ-LA data sets we found 98.

Validation

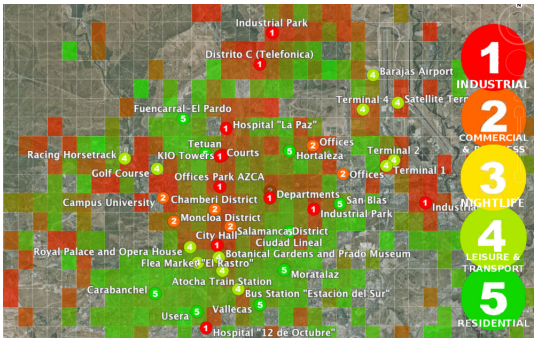
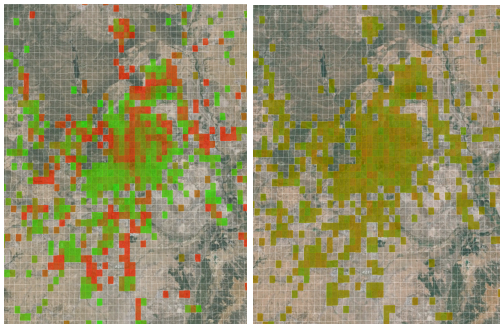


Figure 4. Characterizing Madrid based on the inferred home/work locations, and comparison to the characterization of [22]. We break the city into a grid, and color each square with a combination of green and red. Green squares have relatively more home than work locations; while red squares have relatively more work than home locations. We observe that the squares that we colored red contain contain more circles, indicating industrial and commercial zones, than residential zones. Also, squares colored green contain more residential than industrial zones.



(a) Inferred (b) Uniform

Figure 5. Inferred vs. uniformly distributed home/work locations. Fig 5(a) shows a city with segregated home and work areas, while Fig. 5(b) shows a city where all areas are the same.

Fig. 3 shows the accuracy of the home/work inference methodology for our Twitter data set. We use the FSQ-US data set to train the classifiers. Then, for the ground truth users that appear both in the Twitter data set and the FSQ data set, we infer their home/work locations using the geotagged tweets, and then we compare the inferred home/work locations to the ones in FSQ. In Tab. 2 we compare the accuracy of the home/work identification methodology with the reported accuracy in [5]. We see that in the case of the 75^{th} percentile the home error has decreased by 76% , and the work error has decreased by 59%. For a few cases, our error is higher. We attribute our overall higher accuracy to the more precise location information in the Twitter-Foursquare data sets. Finally, Fig. 4 shows a visual comparison between our results and the characterization of the Madrid's areas from [22], and indicates a strong agreement between our results and the related work.

Differences from the Uniform Distribution

We find that home/work distribution is far from uniform, which was assumed in [4], in the following aspects:

Segregation of home and work areas: According to Fig. 5, Madrid contains segregated home and work (*e.g.* industrial)

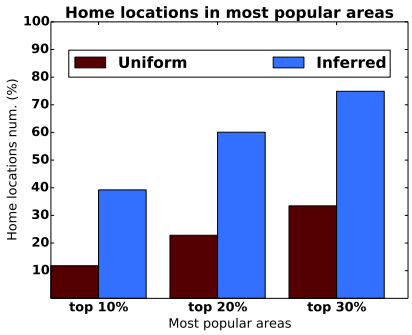
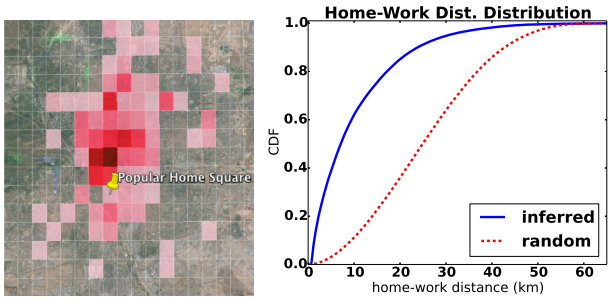


Figure 6. Number of home locations for the 10%, 20%, and 30% most popular areas. In the inferred distribution, areas vary in popularity (highly populated and sparsely populated ones), while in the uniform they are equally popular.



(a) Most popular home area (b) CDF of home-work distance

Figure 7. Distance between home and work locations. 7(a) shows the square grid with most homes (yellow paddle), and where are the corresponding work locations; stronger the colors indicate higher concentration of work locations. Users tend to work close to home.

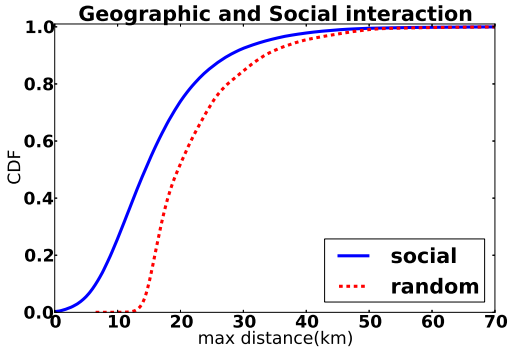


Figure 8. Distances between users who have social ties – inferred from the calls – vs. distance between random strangers. The distance between two users u and v is the maximum of their home and work distance. This figure indicates correlation between social and geographic proximity.

areas. In work areas, there is a relatively large number of working places, while in home areas there is a relatively large number of home locations Fig. 5(a). To illustrate the difference, we show how the city would look if the home/work distribution were uniform, Fig. 5(b).

Non-uniform density: The density of home and work locations in various areas is quite different from uniform, as shown in Fig. 6; 30% of most popular home areas – areas

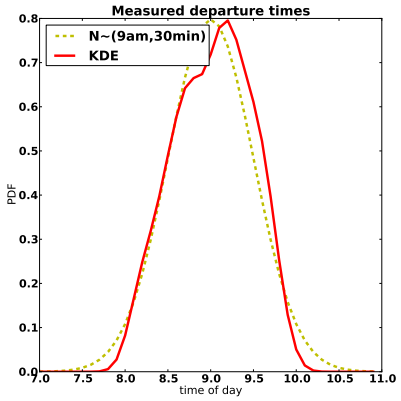


Figure 9. Distribution of home-departure Times. A normal distribution with mean at 9 am, and standard deviation 30 minutes, is a close approximation to the inferred departure times from our data. The continuous line is what we get via Kernel Density Estimation from our data.

with most home locations – contain 75% of the homes; if home/work distribution was uniform then the top 30% of home areas would contain only 30% of the homes.

Relatively short home-work distances: As seen in Fig. 7, users tend to work close to where they live. For the grid square with the highest number of users who have their home there, as shown in Fig. 7(a), the corresponding work locations tend to be close by. Also, according to Fig. 7(b), the home-work distances are shorter compared to what they would be if home and work were randomly distributed.

Geographic distances and social ties: In a later section, we will consider social ties among users, inferred from calls (CDRs), or declared relations (Twitter). In Fig. 8, we compare the average distance of each user u to her friends, vs. her geographic distance to randomly selected strangers (i.e., users who are not neighbors of u in the social graph). According to Fig. 8, the geographic distance between users who have social ties are shorter, on average, in comparison to strangers.

Departure Times

We estimate departure times of individual users from consecutive home/work calls. More specifically, we use pairs of calls where one is a home call, the other a work call, and the time difference between the calls is less than 2·trip_time, where trip_time is the time distance between home and work, as obtained from a popular Online Map service.

For each user, we find her departure time from home by taking the median of the calls, that: (1) were made between 8 am and 10 am from home, and (2) were followed by a work call no more than 2·trip_time later. Similarly, we find her work departure time, by taking the median of the calls, that : (1) were made from work between 4pm and 6pm, and (2) were followed by a home call no more than 2·trip_time later.

The distribution of home departure times for all individuals who had such calls is shown in Fig. 9 – each individual is required to have at least three such calls; there were 484 such

users in our data set. The departure time from work follows a similar distribution, which is omitted due to lack of space.

END-POINTS RIDE-SHARING

In this section, we formulate the problem of `End-Points RS`, *i.e.* ride-sharing among people that live and work close to each other. We develop a practical algorithm, we apply it to the users with inferred home/work locations, and we compute the number of cars that can be reduced under different scenarios.

Formulation

Let V denote a set of potential drivers and $c(v)$ the capacity, in terms of available seats, of the car of driver $v \in V$ and $p(v)$ a penalty paid if driver v is selected for driving her car and picking up passengers. Let $h(v,u)$ denote the geographic distance between the home locations of drivers v and u and $w(v,u)$ the corresponding distance between their work locations. Let δ denote the maximum acceptable distance between a driver's home/work and the home/work of passengers that she can pick up in her car, *i.e.*, v can have u as passenger only if: $max(h(u,v), w(u,v)) \leq \delta$

Let $d(v,u)$ denote a virtual distance between v and u, defined as follows:

$$d(v,u) = \begin{cases} h(v,u) + w(v,u), \\ \text{if} \quad \max(h(v,u), w(v,u)) \leq \delta \\ \\ \infty, \quad \text{otherwise} \end{cases}$$

Our objective is to select a subset of drivers $S \subseteq V$, and find an assignment $a : V \rightarrow S$, that minimizes $P(S) + D(S)$, the sum of penalty and distance costs, while satisfying the capacity constraints of cars. The two costs are defined as follows:

$$P(S) = \sum_{v \in S} p(v) \quad \text{and} \quad D(S) = \sum_{v \in V} d(a(v), v)$$

where $a(v) \in S$ is the driver in S that is assigned to pick up passenger v (can be himself if v is selected as a driver). By setting $p(v) > 2\delta \cdot c(v)$ we make sure that an optimal solution will not increase the number of cars in order to decrease the (pickup) distance cost between a driver and its passengers [1]. The above problem is an NP-hard *Capacitated Facility Location Problem with Unsplittable Demand* in metric distance: the set of potential drivers corresponds to the set of locations; the set of chosen drivers corresponds to opened facilities; car capacity corresponds to facility capacity; distance $d(v,u)$ corresponds to the cost of assigning a location v to the facility u. Efficient approximation algorithms are known for this type of facility location problem [6].

The above formulation includes spatial constraints only. Next, we refine our formulation to include time. In the previous section (see Fig. 9), we showed that departures from home and work can be approximated by a normal distribution, centered at 9 am and 5 pm respectively, with standard deviation σ. We introduce the delay tolerance τ that captures the maximum amount of time that an individual can deviate from her normal schedule in order to share a ride. More specifically, if $LH(u)$ denotes the time a person u leaves home to go to work, and $LW(u)$ expresses the time she leaves work in order to return to home. Then, two people u and v, can share a ride only if:

$$max(|LH(u) - LH(v)|, |LW(u) - LW(v)|) \leq \tau$$

The introduction of the temporal constrains will change the virtual distance between v and u to the following :

$$d(v,u) = \begin{cases} h(v,u) + w(v,u), \\ \text{if} \quad \max(h(v,u), w(v,u)) \leq \delta \\ \text{AND} \quad |LH(u) - LH(v)| \leq \tau \\ \text{AND} \quad |LW(u) - LW(v)| \leq \tau \\ \\ \infty, \quad \text{otherwise} \end{cases}$$

A Practical Algorithm

In this section, we modify the existing approximation algorithm [6] for the facility location problem described above and design a heuristic that can cope with the size of our matching needs, the biggest of which has 272K users.

The algorithm in [6] starts with a random solution and improves it iteratively via local search. At each iteration, there are $O(n^2)$ candidate solutions, where n corresponds to the number of potential drivers. For each one of them, it finds the assignment (passengers to drivers) that minimizes the cost; this is done in polynomial time by solving an appropriately defined instance of the *transportation problem*. The algorithm terminates when local search cannot find a better solution.

We modify the algorithm in three ways. First, since the quality of the solution depends mostly on the number of drivers, we try to keep that number as low as possible. Therefore, we use the b-matching [23] algorithm to generate the initial solution, instead of generating it randomly. The input to the b-matching algorithm consists of the set of potential drivers V, a function $o(v)$ that defines the set options for a potential driver v *i.e.* $o(v) = \{u | d(u,v) < \inf\}$, and a global ordering of the potential drivers, O. The global ordering will be based on the number of options; the fewer the options, the higher the position in O. By using b-matching with a global order we are guaranteed to find a solution in $O(n)$ time [23]. For each match generated by b-matching, we assign the potential driver with the most occupied seats to drive; we make sure that every user in V appears in only one car. This solution has much lower cost than the random one by paying $O(nlog(n))$ for sorting the users to generate the global preference list and $O(n)$ for the matching.

Second, solving a transportation problem with 272K users is computationally expensive. Therefore, we need to modify the local search steps of the approximation algorithm. Given an initial solution we leave the users commuting in cars of four as they are and search for better assignments only for the rest. This reduces the size of the transportation problem and speeds up the process of generating the assignment.

[1]For all (u,v) pairs withing constraints, $d(v,u) \leq 2\delta$, therefore in worst case a full car v can increase the total cost by $2\delta \cdot c(v)$. We set the penalty for every car, to be higher that the worst case scenario.

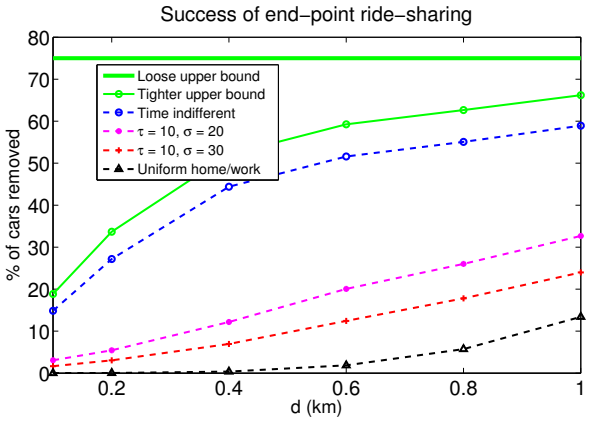


Figure 10. Benefits of `End-Points RS`.

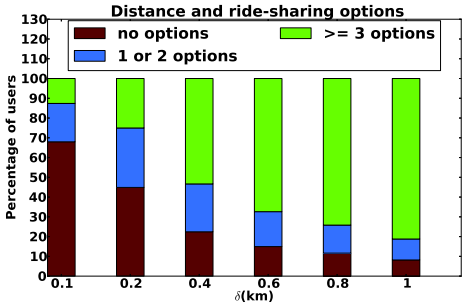


Figure 11. How δ affects the ride-sharing options

Third, reducing the size of the transportation problem is not enough; we also need to reduce the neighborhood of candidate solutions. Given an initial set of drivers, S, we create a fixed size neighborhood, where each solution S' is created by doing random changes in S. The reason why we do that is because considering all potential solutions that differ from S only by one, means that we have to examine $O(n^2)$ candidate solutions; that makes each iteration very expensive. Therefore, the fixed size solution helps us speed up the time we spend in each improvement step.

Without the above modifications it would be impossible to solve the problem in real time. Solving an instance of the transportation problem for 270K users required a couple of hours for $\delta = 0.6$ km, and even more when $\delta = 0.8$ or $\delta = 1.0$ km. Therefore, solving $O(n^2)$ such problems for a single iteration becomes too expensive. Moreover, in our experiments, we observed that the improvement steps would add little value to the solution offered by the b-matching.

Results

We now calculate the effectiveness of `End-Points RS` based on our data sets. For ease of exposition, we will focus on the Madrid metropolitan area (we cover the remaining cities in a later section). We reduce the size of our data set by randomly selecting only 60% of the users. We do that to capture the fact that only 60% of the population has a car in the area of Madrid [24]. We also show results for the case that half of the car owners use their car at their daily commute (the results are quantitatively similar). For the remaining of the section, we will refer to users who can share rides with a specific user v, as *options of* v. We compute the reduction of cars, as % of the initial number:

$$\text{success} = \frac{\#(\text{init. cars}) - \#(\text{ride-sharing cars})}{\#(\text{init. cars})} \cdot 100$$

using the following algorithms:

Loose upper bound: Given our definition of success, we cannot do better than 75%, when all cars carry 4 people.

Tighter upper bound: Assuming that all users with at least one ride-sharing option commute in cars of 4.

Time-indifferent matching ($\tau = \infty$): This is the practical algorithm described in the previous section.

Time-aware matching: This is the version of the algorithm that considers timing constraints under the assumption of normally distributed departure times.

Uniform home/work: Ride-sharing assuming that home/work locations are distributed uniformly.

Fig. 10 presents what happens when the users are willing to tolerate a detour of δ km and deviate τ minutes from their departure times, in order to share the same car with another individual. The results show that with a modest delay tolerance of 10 minutes and a detour distance of 1.0 km (a couple of city blocks) more than 20% of the cars can be saved. The success ration improves when δ or τ increase. The diminishing improvement with increasing δ can be explained by the number users' options, given the distance δ. In Fig. 11, the red color represents the users with no options, the blue color the users with 1 or 2 options, and the green color the users with 3 or more options. We see that the success of ride-sharing is proportional to the number of users with 3 or more options.

Fig. 10 also shows that the potential of `End-Points RS` is quite small in the case of uniformly distributed home/work locations; note that no time constraints were applied in this case. If we apply time constrains too, then the success of `End-Points RS` is even smaller, e.g. for $\delta = 1$ km, $\tau = 10$ min, and $\sigma = 30$ min, its potential becomes 0.2%

EN-ROUTE RIDE-SHARING

The effectiveness of ride-sharing can be greatly improved by picking up additional passengers en-route; a driver that lives in a sparsely populated area might not have any neighbors to fill her seats, but once she enters the city she can pick several passengers that are on her way. In order to quantify the benefits of en-route ride-sharing we obtain routes from a popular Online Map service for the 272K users with inferred home/work locations, and we extend the algorithm of the previous section. Again, we will focus on Madrid.

En-Route Algorithm

We use an iterative algorithm with the following steps:

1. Run the basic `End-Points RS` algorithm.
2. Exclude from the solution cars that get fully packed (a car of 4). Then order cars in decreasing order of passengers and start "routing" them across the urban environment (*e.g.* Madrid) using data from the Online Map service.

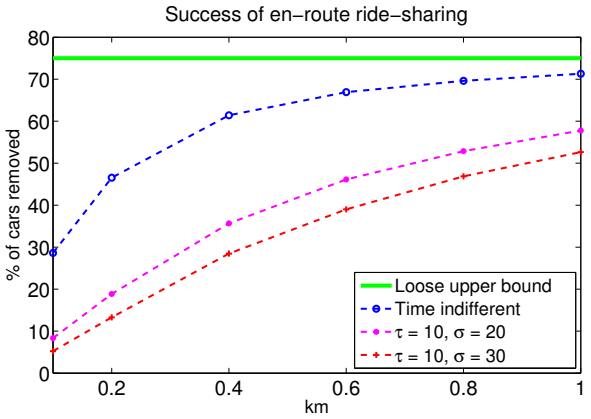


Figure 12. Benefits of `En-Route RS`.

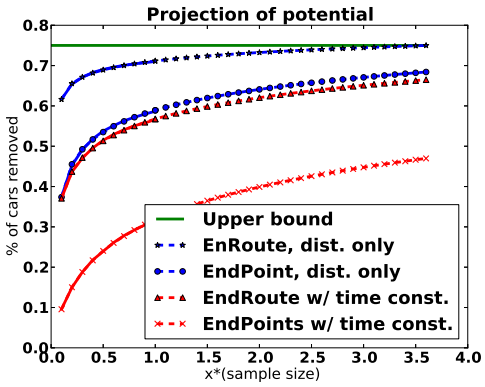


Figure 13. Extrapolation to commuters' size. "Sample" refers to the 272K users with inferred home/work locations in Madrid. The solid lines correspond to values generated from our data set, while the dashed lines correspond to values generated through extrapolation.

3. When the currently routed car v meets a yet un-routed car v', then v is allowed to steal passengers from v' as long as it has more passengers than v' (a rich-get-richer strategy). Whenever a routed car gets fully packed it is removed from further consideration. Whenever a car with a single passenger is encountered the number of cars is reduced by one.

4. The algorithm finishes when no change occurs.

These steps are repeated until there is no possible improvement. The rich-get-richer rule leads to convergence, since it forces cars to either become full or to stay home (the driver becomes a passenger in another car). The algorithm converged in every single execution.

Results

Fig. 12 shows the performance of `En-Route RS`. To make the comparison with `End-Points RS` easier we summarize our results in Tab. 3. One can see the significant improvement obtained through `En-Route RS`, which in several cases comes within 10% of the optimal performance.

Projection to the entire commute population: All previous results have been produced based on the 272K users with inferred home/work location in Madrid. This, however, represents only roughly 8% of the total population of the city. To get a feeling of the ride-sharing potential based on the entire population, for which we do not have location information, we extrapolate to a larger number of users. We repeat the calculation of ride-sharing with different subsets of our total 272K users, and fit numerically these data points to a logarithmic function (in order to capture the diminishing effect). Then we extrapolate the potential of ride-sharing for larger population sizes. The results are summarized in Tab. 3 and shown in Fig. 13, where we see that the population size has a progressively diminishing results on the ride-sharing potential. In the remainder of the paper we will report results for both our 8% sample, and extrapolation to the commuters' population.

Sample (%)	δ (km)	τ (min)	σ (min)	`End-Points RS` (%)	`En-Route RS` (%)
30	1.0	–	–	54	65
30	1.0	10	30	17	47
60	1.0	–	–	59	70
60	1.0	10	30	24	53
100	1.0	–	–	62	71
100	1.0	10	30	30	56
360	1.0	–	–	70	75
360	1.0	10	30	44	65

Table 3. Effect of population size on the performance of `End-Points RS` and `En-Route RS` in Madrid. "Sample" refers to the 272K users with inferred home/work locations in Madrid. 100% means using all of them. 30% and 60% means using a random subsets, while 360% means projecting the potential to the entire commuters' population of Madrid.

Graph	Nodes #	Edges #	Mean degree	Median degree
call graph Madrid	4M	21M	6.0	1
twitter graph NY	132K	725K	10.95	5

Table 4. Graph sizes

city	filter	`End-Points RS` (%)	`En-Route RS` (%)	`En-Route RS` extrapolation (%)
Madrid	no filter	30	56	65
Madrid	1-hop	0.26	1.1	–
Madrid	2-hop	3.7	19	31
NY	no filter	20	44	68
NY	1-hop	0.18	1.2	–
NY	2-hop	2.1	8.2	26

Table 5. Social Filtering. The potential or `End-Points RS` and `En-Route RS` for $\delta = 1.0$ km (distance constr.), τ=10, σ = 30 (time constr.). The third and the forth column show the potential for sample size, while the last column shows the potential of ride-sharing extrapolated to the commuters' population.

SOCIAL FILTERING - RIDING WITH FRIENDS OF FRIENDS

In this section, we present how social filtering affects the potential of ride-sharing. Instead of assuming that anybody is willing to share a ride with anybody else, we introduce social constraints in selecting ride-sharing partners. The social constraints are represented by graphs, e.g. as shown in Fig. 15: the nodes correspond to users, and the edges correspond to social ties. A user considers sharing a ride with a one-hop neighbor, or with a two-hop neighbor (a friend of a friend).

Given that we have two different types of data sets – CDR and geo-tagged tweets – we need to use two different definitions of edges. In the case of CDR data [25] [26], choosing a threshold condition for an edge between two users involves

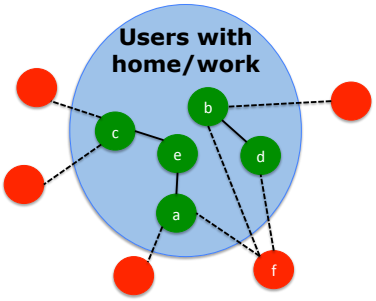
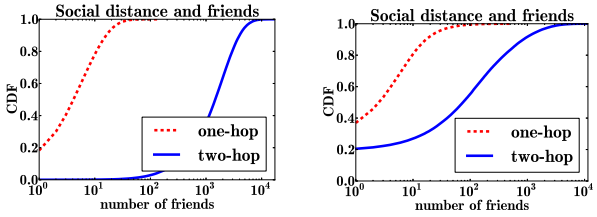


Figure 14. How social filtering works. Green nodes represent users with inferred home/work locations. Red nodes represent their neighbors (without inferred home/work locations). We only consider ride-sharing among the green nodes. In one-hop filtering, *a* can share a ride only with *e*. In two-hop filtering, *a* can share a ride with *e*,*c*,*b*, and *d*.



(a) call friends in Madrid (b) declared friends in NY

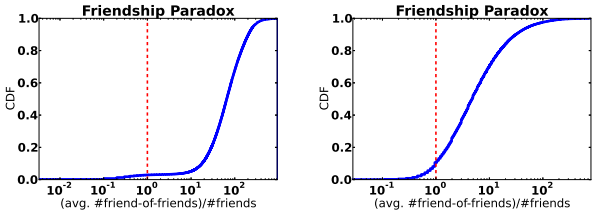
Figure 15. Number of friends for the users with home and work address.

a trade-off between the strength of the tie and the number of edges. When choosing a threshold one needs to take into account the needs of the application [27]. In this study, we create an edge in the social graph between two users when there is at least one call between them. We experimented with various definitions, and we found that – due to the small number of users with inferred home/work locations – higher thresholds would result in extremely sparse, thus useless,[2] graphs.

In the case of Twitter, we crawl the friends and the followers of the users for with inferred home/work locations, and we create an edge in the social graphs if there is a bidirectional edge on Twitter. See Tab. 4 for graph details. Moreover, to be sure that the friend nodes in our Twitter graph represent real people we considered only users who had at least one geo-tagged tweet. Finally, in both CDR and Twitter cases, we filtered out nodes with more than 1000 friends, in order to exclude popular phone services, or celebrities, respectively.

Now, we examine how social filtering affects the potential of ride-sharing. Fig. 14 illustrates the social filtering process. Lets start with Madrid. As we can see from Tab. 5 the potential of ride-sharing is quite low when users are willing to share a ride only with their one-hop friends. This is expected, since the graph shows only a small portion of a user's friends, and the users for whom we have home/work addresses are only a small subset of all users. From Fig. 15(a), we can see

[2]Using a reciprocal call, as a threshold, would result in a graph with 2.4M nodes, and 3.7M edges. In that case, 92% of the users had zero one-hop neighbors with whom they could share a ride. As a result, the ride-sharing potential was 2% (5.1% with extrapolation) for `En-Route RS` with 2-hop social filter, and $\delta = 1.0$ km (dist. constr.), τ=10, $\sigma = 30$ (time constr.)



(a) Call graph (b) Twitter graph

Figure 16. The CDF of the ratio between average number of friends-of-friends over number of friends. Friendship paradox holds when this ratio is greater than one (over 90% of the users both in figures).

scenarios	`End-Points RS` ratio. (%)	`En-Route RS` ratio. (%)
τ=10, σ = 30	3.3	1.8
Social constr.	68	14

Table 6. Madrid vs. BCN. This table shows the difference in ride-sharing potential between BCN and Madrid, for both `End-Points RS` and `En-Route RS`, in two different scenarios : (1) $\delta = 1.0$ km, τ=10, and σ= 30, and (2) $\delta = 1.0$ km, τ=10, σ= 30, and two-hop friends. The ratio is computed as : $((BCN - Madrid)/Madrid) * 100$

scenarios	`End-Points RS` ratio. (%)	`En-Route RS` ratio. (%)
τ=10, σ = 30	-33	-9
Social constr.	-50	-46

Table 7. NY vs. LA. This table shows the difference in ride-sharing potential between New York and Los Angeles, for both `End-Points RS` and `En-Route RS`, in two different scenarios : (1) $\delta = 1.0$ km, τ=10, and σ=30, and (2) $\delta = 1.0$ km, τ=10, σ= 30, and two-hop friends. The ratio is computed as: $((LA - NY)/NY) * 100$

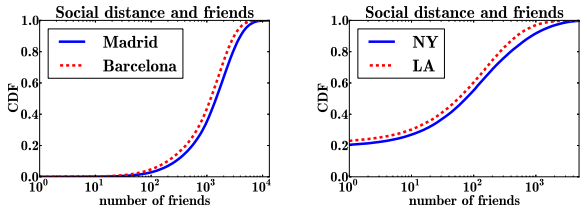
that 80% of the nodes in the call graph have no more than 10 one-hop friends, whose home/work addresses have been identified. However, if users are willing to share rides with friends of friends, then from Tab. 5 we can see that, even with a sparse social graph, there can be considerable gain from `En-Route RS`. This can be explained from Fig. 15(a), in which we can see the much higher number of two-hop than one-hop friends. In all data sets, there is a considerable improvement; e.g., in Madrid, ride-sharing has a potential of 19% (or 31% if extrapolated to the entire population of Madrid).

In general, the number of nodes and edges in the social graph is crucial for any ride sharing application that wants to exploit social filtering. Moreover, the difference between the large increase in the ride-sharing potential when using friend-of-friends can be attributed to the friendship paradox ("on average your friends have more friends that you do", [28, 29] that also holds in our data sets as illustrated in Fig. 16.

A TALE OF FOUR CITIES

In this section, we compare the potential of ride-sharing in the four cities (Madrid, BCN, NY, and LA).

We start by comparing Madrid and BCN. The first row of Tab. 6 shows that, for spatio-temporal constraints only, the potential or ride-sharing in the two cities is very similar, with the potential of `En-Route RS` being slightly higher in BCN. In the second row, we show that when also considering social constraints, the relative difference in ride-sharing benefit between the two cities becomes becomes much higher: the



(a) Barcelona vs. Madrid (b) NY vs. LA

Figure 17. Comparing the CDF of 2-hop friends for Madrid vs. Barcelona, and NY vs. LA.

potential of `End-Points RS` in BCN is 68% higher, and the potential of `En-Route RS` in BCN is 14% higher. This difference cannot be explained by the social graph, since, as we can see from Fig. 17(a), the users in both cities have almost the same number of friends. We attribute the better potential in BCN to its higher population density: Madrid has a density of 5,390 $people/km^2$, while BCN has a density of 15,926 $people/km^2$.

The same observation holds in the comparison between the two US cities. The potential or ride-sharing in NY is higher that the potential of ride-sharing in LA – see Tab. 7. The difference gets even higher when time or social constraints are included – see Tab. 7. Again, the difference in the potential of ride-sharing can be explained by the densities of the two cities: LA has a density of 3,124 $people/km^2$, and NY has a density of 10,429 $people/km^2$.

We obtained the mobility data for Madrid and BCN from CDRs, and we obtained the mobility data for NY and LA from geo-tagged tweets, therefore a comparison between European and US cities may lead to incorrect conclusions. However, both comparisons (Madrid *vs.* BCN and NY *vs.* LA) show that ride-sharing is more beneficial in denser cities, especially when time and social constraints are considered.

SUMMARY AND CONCLUSION

We used mobile and social data to demonstrate that there is significant overlap in people's commute in a city, which indicates a high potential benefit from ride-sharing systems. This is clearly an upper bound to any practical ride-sharing system, but the positive result motivates the deployment of such systems and policies. Our results indicate that en-route ride-sharing with up to two-hop social contacts offers a good trade-off between technological feasibility, people's security concerns, and a substantial impact on traffic reduction. A more detailed summary of our findings is as follows.

We started by considering `End-Points RS` in which rides can be shared only with neighbors with nearby home and work. Even with a modest detour of 1 km we observed a great potential reduction of cars. In the case of Madrid, this reduction is 59%, based on our location data set that captures close to 8% percent of the total population. Our estimation of the ride-sharing potential extrapolated to the total commuting population of the city is significantly higher. The distribution of home/work locations, which is far from uniform is crucial to the success of ride-sharing: if Madrid had a uniform home and work distribution then the reduction would be 13% assuming only spatial constraints, and 0.2% assuming time constraints too. This is in agreement with [4] and shows that ride-sharing has negligible benefit in a city with uniform home/work distribution.

Adding time constraints, the effectiveness of ride-sharing becomes proportional to the driver/passenger waiting time for a pick-up, and inversely proportional to the standard deviation of the distribution of departure times. With a standard deviation of 30 min, a wait time up to 10 min and a δ of 1km there is a 24% reduction of cars in Madrid.

`En-Route RS`, i.e., allowing passenger to be picked up along the way, yields a great boost in ride-sharing potential with or without time constraints. In the case of Madrid, `En-Route RS` increases the savings from 24% to 53%.

Then, since people are often hesitant to ride with strangers, we decided do add social constrains too. Social ties can be inferred from calls (CDRs), or declared friendship (Twitter). First, we consider ride-sharing only with one-hop friends. Then `En-Route RS` in the city of Madrid using CDR and Twitter friendship provides only a tiny traffic reduction of 1.1% and 1.2% respectively. This dramatic decrease is attributed to the low density of the social graphs and to the fact that only a small portion of the graphs' nodes have known home/work addresses – each user has the opportunity to share a ride only with a small portion of her neighbors. However, if we relax the social constraints and permit ride-sharing with friends-of-friends, the ride-sharing potential increases significantly, especially in `En-Route RS`. The corresponding numbers are 19% and 8.2% for friendship based on CDRs and Twitter data, respectively. Furthermore, if we project the potential of ride-sharing to the total commuting population of the city (much larger than the number of users with inferred home/work locations), the benefit increases up to 31% for call based filtering and 26% for OSN based filtering.

Finally, we compared the four cities and observed that the population density of a city has a profound effect on its ride-sharing potential, especially when strict social filtering is applied. For example, BCN is denser and has a 14% higher ride-sharing potential than Madrid; LA, on the other hand, has 46% lower ride-sharing potential than NY.

Directions for future work include designing real-time matching algorithms (motivated by the offline analysis in this paper) and implementing a prototype ride-sharing system. The methodology developed in this paper can potentially be used on other cities and different data sets to assess the potential of ride-sharing and guide related deployment and policies.

ACKNOWLEDGMENTS

This work has been supported by AFOSR MURI award FA9550-09-0643, and NSF CDI award 1028394. Blerim Cici and Athina Markopoulou are affiliated with EECS, NetSys and CPCC at UC Irvine. Blerim Cici was visiting Telefonica Research Labs in Barcelona, Spain, when part of this work was conducted. Special thanks to Vijay Erramilli for his help with the Twitter/FSQ data collection.

REFERENCES

1. "The true cost of a car over its lifetime: http://www.doughroller.net/smart-spending/true-cost-of-a-car-over-its-lifetime."
2. A. M. Amey, "Real-Time Ridesharing: Exploring the Opportunities and Challenges of Designing a Technology-based Rideshare Trial for the MIT Community," Master's thesis, Massachusetts Institute of Technology, 2010.
3. A. Thiagarajan, L. S. Ravindranath, K. LaCurts, S. Toledo, J. Eriksson, S. Madden, and H. Balakrishnan, "Vtrack: Accurate, energy-aware traffic delay estimation using mobile phones," in *Proc. of SenSys*, 2009.
4. H.-S. J. Tsao and D. Lin, "Spatial and temporal factors in estimating the potential of ride-sharing for demand reduction." California PATH Research Report, UCBITS-PRR-99-2, 1999.
5. S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in People's Lives from Cellular Network Data.," in *Proc. of Pervasive Computing*, 2011.
6. M. R. Korupolu, C. G. Plaxton, and R. Rajaraman, "Analysis of a local search heuristic for facility location problems," *Journal of Algorithms*, vol. 37, pp. 146–188, 2000.
7. R. Teal, "Carpooling: Who, how and why," *Transportation Research*, vol. 21A, no. 3, pp. 203–214, 1987.
8. W. He, D. Li, T. Zhang, L. An, M. Guo, and G. Chen, "Mining regular routes from gps data for ridesharing recommendations," in *Proc. of UrbComp*, 2012.
9. R. Trasarti, F. Pinelli, M. Nanni, and F. Giannotti, "Mining mobility user profiles for car pooling," in *Proc. of UrbComp*, 2011.
10. N. Bicocchi and M. Mamei, "Investigating ride sharing opportunities through mobility data analysis," *Pervasive and Mobile Computing*, 2014.
11. S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human Mobility Modeling at Metropolitan Scales," in *Proc. of MobiSys*, June 2012.
12. E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proc. of SIGKDD*, ACM, 2011.
13. N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data.," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15274–15278, 2009.
14. S. Wakamiya, R. Lee, and K. Sumiya, "Urban area characterization based on semantics of crowd activities in twitter," in *GeoSpatial Semantics*, Lecture Notes in Computer Science, 2011.
15. T. Fujisaka, R. Lee, and K. Sumiya, "Exploring urban characteristics using movement history of mass mobile microbloggers," in *Proc. of Hotmobile*, 2010.
16. E. Frias-Martinez, V. Soto, and H. Hohwald, "Characterizing urban landscapes using geolocated tweets," in *Proc. of SocialCom*, 2012.
17. A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks," in *Proc. of SMW*, 2011.
18. A. Noulas, C. Mascolo, and E. Frias-Martinez, "Exploiting foursquare and cellular data to infer user activity in urban environments," in *Proc. of MDM*, 2013.
19. B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris, "Quantifying the Potential of Ride-Sharing using Call Description Records," in *Proc. of HotMobile*, 2013.
20. "https://dev.twitter.com/docs/streaming-apis."
21. C. J. Riederer, A. Chaintreau, J. Cahan, and V. Erramilli, "Challenges of keyword-based location disclosure," in *Proc. of WPES*, 2013.
22. V. Soto and E. Frias-Martinez, "Automated Land Use Identification using Cell-Phone Records," in *Proc. of HotPlanet*, 2011.
23. K. Cechlárová and T. Fleiner, "On a generalization of the stable roommates problem," *Transactions on Algorithms*, vol. 1, no. 1, pp. 143–156, 2005.
24. "Instituto de estadistica de la comunidad de madrid: http://www.madrid.org/iestadis/."
25. K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. a. Nanavati, and A. Joshi, "Social ties and their relevance to churn in mobile telecom networks," *in Proc. of EDBT*, 2008.
26. J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and a L Barabási, "Structure and tie strengths in mobile communication networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 7332–7336, 2007.
27. M. D. Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts, "Inferring Relevant Social Networks," in *Proc. of WWW*, 2010.
28. N. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you," in *Proc. of ICWSM*, 2013.
29. S. L. Feld, "Why your friends have more friends than you do," *American Journal of Sociology*, vol. 96, no. 6, pp. 1464–1477, 1991.