The GRAAL of carpooling: GReen And sociAL optimization from crowd-sourced data

Michele Berlingerio*[1], Bissan Ghaddar†[1,2], Riccardo Guidotti‡[1,3], Alessandra Pascale§[1] and Andrea Sassi¶[1,4]

[1]IBM Research Ireland
[2]University of Waterloo
[3]University of Pisa
[4]University of Modena and Reggio Emilia

Abstract

Carpooling, i.e. the sharing of vehicles to reach common destinations, is often performed to reduce costs and pollution. Recent work on carpooling takes into account, besides mobility matches, also social aspects and, more generally, non-monetary incentives. In line with this, we present GRAAL, a data-driven methodology for GReen And sociAL carpooling. GRAAL optimizes a carpooling system not only by minimizing the number of cars needed at the city level, but also by maximizing the *enjoyability* of people sharing a trip. We introduce a measure of enjoyability based on people's interests, social links, and tendency to connect to people with similar or dissimilar interests. GRAAL computes the enjoyability within a set of users from crowd-sourced data, and then uses it on real world datasets to optimize a weighted linear combination of number of cars and enjoyability. To tune this weight, and to investigate the users' interest on the social aspects of carpooling, we conducted an online survey on potential carpooling users. We present the results of applying GRAAL on real world crowd-sourced data from the cities of Rome and San Francisco. Computational results are presented from both the city and the user perspective. Using the crowd-sourced weight, GRAAL is able to significantly reduce the number of cars needed, while keeping a high level of enjoyability on the tested data-set. From the user perspective, we show how the entire per-car distribution of enjoyability is increased with respect to the baselines.

*michele.berlingerio@gmail.com
†bghaddar@uwaterloo.ca
‡riccardo.guidotti@di.unipi.it
§apascale@ie.ibm.com
¶andrea.sassi@unimore.it

1 Introduction

Carpooling is a scheme in which people share a vehicle in order to reach common or nearby destinations. Despite its clear advantages in reducing costs, pollution, and time spent in finding a car park, there are still a few obstacles that prevents it from being the preferred way to move: safety of passengers, sub-optimal mobility matches, and time flexibility, among others.

A common underlying aspect across many such obstacles is a hidden psychological barrier that makes carpooling less attractive. However, due to the increasing popularity of online social networks in the last few years, there are some social aspects that people intentionally decide to share with the outside world, including strangers. In fact, sharing interests, pictures, and visited locations, are the basis of the success of services such as Facebook, Twitter, and Foursquare. The availability of such information allows external services and people to use this data for third party applications. As a result, such social aspects can be now *measured*, and *exploited* to overcome this invisible psychological barrier in the context of carpooling.

Inspired by the literature on carpooling [35, 41, 11, 13, 27], and by the recent work on data-driven analysis in urban networks [31] and data-driven optimization of urban transit networks [30, 2], we present a mathematical formulation of the carpooling problem taking into account the above factors, and a data-driven methodology to automatically derive mobility and social matches to be used as recommendations for the carpooling system. The main goal of our work is to present a "what-if analysis" in which we measure, from sources available online, how users would *enjoy* sharing a trip with other people, and to devise a new methodology for carpooling driven by these measurements. Our contribution is mainly methodological, rather than a carpooling system tested on the field. Thus, we focus in this paper on the theoretical core of carpooling, i.e. the data-driven multi-objective optimization problem.

In contrast with on-demand carpooling setting, where the user typically opens a mobile application to select origin, destination and departure time, and find matching drivers, we process data in temporal batches and focus on recurring trips. In turn, well known results on human behavior analytics [17, 34] show that our mobility is largely predictable, i.e. processing data in batches, rather than in an on-demand basis, covers a large portion of our demand. Moreover, this allows us to gain more room for optimization, as we treat space, time, and interest patterns of users all at once.

Based on all the above, we build GRAAL, a methodology for GReen And sociAL carpooling. GRAAL optimizes a carpooling system, at the city level, not only by minimizing the number of cars needed, but also by maximizing the *enjoyability* of people traveling together. Starting from the concept presented by the authors in [21] we introduce a measure of enjoyability based on people's interests, social links, and tendency to connect to people with similar or dissimilar interests. Specifically, our enjoyability measure takes into account two factors: (i) what we call *like-mindness*, i.e. a topic similarity between any two users; and (ii) what we define as *homophily*, i.e. the tendency of a person to

group with similar ones. Previous attempts to use social context in carpooling include putting together in a car people who are friends [8]. However, by looking only at the direct (or even the two-hop) friends, we may loose other good matches from the optimization model, as the set of potential drivers (or co-passengers) is usually much larger than the typical number of friends pairs in a social network.

In GRAAL, we introduce a multi-objective optimization based on a weighted linear combination of two components: i) number of cars (which is minimized) and ii) total enjoyability of the users in the system (which is maximized). In our experiments, we vary this weight, which we refer to as α, between 0 and 1. Moreover, we learn a crowd-sourced value for α by means of an online survey. The survey has the double effect of both confirming the interest of potential carpooling users to a more social solution, and providing us with a realistic estimation of α to use in the optimization model. We present the results of applying GRAAL on real world crowd-sourced data from Twitter, geo-located in the cities of Rome and San Francisco. Results are presented from both the city-wide and the user perspective, and we compare them with different baselines: a random model; a heuristic model aimed at maximizing the user enjoyability; GRAAL model with a value of α equals to one that makes GRAAL minimize only the number of cars (which is derived from the state of art of carpooling); GRAAL model with α set to zero, such that only the maximization of enjoyability is performed. Results show that with the crowd-sourced α, GRAAL is able to reduce the number of cars needed compared to using private vehicles (i.e., each user driving his/her own car), while keeping a high level of total enjoyability. From the user perspective, we show how the per-car distribution of enjoyability is increased with respect to our baselines. We also compare our algorithm with the methodology described in [8]. Although the computational results are based on real-data, however the outcome and the analysis are theoretical as they depend on the assumption that all users in the system would accept our recommendations. In line with the literature of theoretical models for carpooling, a field test with an evaluation involving the end users is out of scope in this paper. To summarize, the main contributions of this paper are:

- we formulate the carpooling problem as a multi-objective optimization of number of cars and enjoyability;
- we learn the weight for the multi-objective optimization by means of a user study;
- we build a methodology extracting enjoyability and mobility demand from data like Twitter;
- we show the results of the application of our methodology on real world data, from the perspectives of the city, and the users, and we evaluate against different baselines.

The paper is organized as follows: related work is presented in Section 2; we define the carpooling problem and the formulation in Section 3; the methodology

is described in Section 4; the user study is presented in Section 5; experiments on Rome and San Francisco data are presented in Section 6; limitations and future work are presented in Section 7; the conclusion is summarized in Section 8.

2 Related work

Our presneted work is positioned within the theoretical framework of human mobility and recommendations for carpooling, and finds motivation in the work focusing on psychological barriers of carpooling [13]. The method proposed exploits systematic behaviors in human daily travels to produce carpooling recommendations. With the increasing availability of data, several research directions started in this area. Using fine mobility data, [40] and [22] extract not only patterns of presence in locations, but also systematic routes, from GPS trajectories. These routes are used to find user-to-user matches and to provide ride-sharing recommendations. In [24], a methodology working on GPS data that organizes trajectories in a tree to speed up geographical carpooling queries, is developed.

Mobility has been used in conjunction with other types of information to improve recommendations. For example, [7] proposes a location prediction model that combines users movements with a social network structure, highlighting how the social dimension may improve significantly the performances. Geosocial data is used in [14], recommending individuals to join their friends during trips by using home location models and users' similarities. In [8] the authors derive home and work locations using Twitter and Foursquare data, then social ties are used to develop an algorithm for matching users with similar mobility patterns. Twitter data is used also in [15] as a complementary source of information for urban planning applications. Similarly, [10] proposes a model to find compatible matches for traditional groups of users and also to find rides in alternative groups. In [3] the authors develop an application for car sharing by exploiting a clustering algorithm applied to labeled trajectories. Finally, [6] introduces a Facebook-based carpooling, with Twitter-based traffic monitoring and Flickr-based incident reporting applications.

Carpooling is often modeled as an optimization problem [16], where it finds several solutions. In [9], the problem is reduced to the *chairman assignment problem* [38], while [29] use an instance of the *transportation problem*. In [42], the authors propose a carpooling based on taxicab. That is, they analyze the reduction of circulating taxis in presence of ride sharing. In [26], a dynamic ride-sharing problem is proposed to efficiently serve real-time requests sent by taxi users and to generate ride-sharing schedules that reduce the total travel distance. The authors of [23] present an algorithm that provides carpooling advises by maximizing the expected value for negotiation success. Other related work [37, 19, 20], study the impact of using network analysis for better matches in location-based services and applications. With a spirit similar to GRAAL, [33] proposed a journey planner that maximizes (a different concept of) the enjoyability of the journey. Besides solving a different problem (route planning,

as opposed to carpooling), their concept of enjoyability differs from ours, as it is based on characteristics of the journey itself (presence of trees, landscape, etc.), rather than on the relation among users.

3 Problem formulation

The objective of the carpooling optimization problem proposed in this paper is the minimization of the total number of cars in the system, together with the the maximization of the *enjoyability* experienced by the users traveling together. Our goal is to follow the main advantage of the carpooling idea, i.e. lowering the number of cars on the road, while ensuring that the passengers will enjoy traveling together. We believe this may serve as an additional, non-monetary, incentive to motivate people to share a car. Our user study presented in Section 5 shows how the potential users that we polled are actually sensitive to these two functions. This section presents the needed preliminaries and definitions to formulate the model in Section 3.1, while the optimization model itself is presented in Section 3.2

3.1 Preliminaries

Enjoyability. We define a measure of enjoyability that takes into account not only whether two users share the same interests, but also whether they tend to connect to people with similar or dissimilar interests. Let U be a set of users. Every user $i \in U$ may consider other users in U as friends, or interesting in general, and we denote such set of users as F_i. Each user i generates, or is interested in, a set of articles or documents D_i. Given i and D_i, we can build a vector of topics $\vec{t_i}$,where each topic is weighted by its relative importance, i.e. frequency, within the documents. We define a measure, which we call *like-mindness*, of how much two users are interested in the same topics, as follows.

Definition 1 (Like-mindness) *Given two users i, j we call their* like-mindness *the number:*

$$lm_{ij} = 2\frac{\vec{t_i} \cdot \vec{t_j}}{\|\vec{t_i}\|\|\vec{t_j}\|} - 1$$

We say i and j are like-minded, i.e. they share a set of interests, if $lm_{ij} \approx 1$, not-like-minded if $lm_{ij} \approx -1$. We want to take into account two different categories of people: those who are more prone to be in contact with other people with similar interests (*homopilous* people), and those who tend to connect with people with dissimilar interest (*heterophilous* people). For this reason we evaluate a user's tendency to connect with people with whom he/she has a high or low like-mindness. In social networks, the concept of homophily is well known [28].

Definition 2 (Homophily) *Given a user i we compute his/her* homophily *as the median of the like-mindness between i and other users in F_i:*

$$h_i = \underset{j \in F_i}{\text{median}} \; lm_{ij}$$

If $h_i \approx 1$, we say that i tends to be homophilous, while if $h_i \approx -1$ we say that i tends to be heterophilous. Our objective is to relate the like-mindness of a pair of users with the homophily/heterophily of the single user. Thus, we define the enjoyability as:

Definition 3 (Enjoyability) *Given two users i, j, their like-mindness lm_{ij} and their homophily values h_i, h_j, we define the* enjoyability *of them being together as:*

$$e_{ij} = \frac{lm_{ij} h_i + lm_{ij} h_j}{2}$$

We denote the set E to be the set of the enjoyabilities computed between each pair of users. Note that $e_{ij} \approx 1$ if either: i) both i and j are homophilous and like-minded; or ii) i and j are heterophilous and not like-minded. In the other cases, $e_{ij} \approx -1$. The added value of social diversity has been studied in social science, and finds applications also in the scientific community. Socio-cultural diversity is often considered fundamental [32] to make people enjoying a discussion.

The objective function presented in Section 3.2 is a linear combination of two components: number of cars and total enjoyability. As we minimize the number of cars, we take into account the *unenjoyability* of the system, rather than the enjoyability, to minimize this as well. The unenjoyability is computed as $\bar{e_{ij}} = 1 - \frac{1}{2}\left(e_{ij} + 1\right)$.

Mobility demand. Another important step when considering carpooling is the analysis of the users' mobility demand. To avoid lack of generality, we define a *location* l as any geo-referenced format. Depending on the available data, a location may be a pair of (lat, lon) GPS coordinates, a geo-hashed area belonging to a geo-index, or any shape with associated geographical information used in a GIS system. In our experiments, we divide the areas of interest in a grid of cells of either 500m or 70m of width. Each user i can have a different *location* l over time. We call *time-stamped location* a pair $tsl = (l, ts)$ where l is a location and ts is an associated relative time-stamp. Two time-stamped locations are defined to be close in space and time as follows:

Definition 4 (Close time-stamped locations) *Given two time-stamped locations $tsl_1 = (l_1, ts_1)$ and $tsl_2 = (l_2, ts_2)$, we say that tsl_1* is close to *tsl_2 ($tsl_1 \simeq_{\delta,\tau} tsl_2$) iff*

$$\text{space-dist}(l_1, l_2) \leq \delta \;\; \textit{and} \;\; \text{time-dist}(ts_1, ts_2) \leq \tau$$

where space-dist$(\cdot,\cdot)$ *and* time-dist$(\cdot,\cdot)$ *are two functions of spatial and temporal distance.*

The choice of the specific functions is left for the specific application. Examples for distance calculation include the euclidean, spherical, or Manhattan, and for time function one can consider simply the time difference. In this work, we use the spherical distance between two cells of the grid defined above, and the time difference for computing the time. If ts_1 or ts_2 are undefined, then the $\simeq$ operator considers only space-dist$(l_1, l_2) \leq \delta$. We refer to a *trajectory* as the sequence $tr = \{tsl_1, \ldots, tsl_n\}$ of time-stamped locations. We associate, to each user i, a set of trajectories, which constitutes the *mobility demand* $T_i = \{tr\}$ for that user. We indicate with $\mathcal{T}_U = \{T_i\}$ the mobility demand of all the users.

For the sake of carpooling, we have to define a match between two trajectories, i.e. the trajectory of the user who will be the candidate driver, and one who will be the candidate passenger. Several options are possible here, but we chose to force a matching of the two initial time-stamped locations of the two trajectories, and allow for a match of the final time-stamped location of the trajectory of the candidate passenger with any of the locations of the trajectory of the candidate driver, including (where possible) the final time-stamped one. In carpooling terms, this means that the driver-passenger pair should depart from their initial locations (the first on their trajectories), but the driver is allowed to drop the passenger on any of the locations along the associated trajectory which are close enough. More formally, we define the following condition.

Definition 5 (Trajectory containment) *Given two trajectories* $tr' = \{tsl'_1, \ldots, tsl'_{\bar{n}}\}$ *and* $tr'' = \{tsl''_1, \ldots, tsl''_{\bar{m}}\}$*, we say that* tr' contains tr'' *(*$tr' \sqsubseteq_{\delta,\tau} tr''$*) iff*

$$tsl'_1 \simeq_{\delta,\tau} tsl''_1 \text{ and } \exists n, 1 < n \leq \bar{n} \text{ s.t. } tsl'_n \simeq_{\delta,\tau} tsl''_{\bar{m}}$$

Note that this definition can be extended to any origin and destination, if required by the final application. In practice we fix the maximum walking distance from the passenger's departure/arrival locations to pick-up/drop-off points (set by the driver) as δ and the maximum time difference in departure and arrival times as τ. Given the above definition, two users i and j having trajectories tr_i and tr_j in their mobility demand, respectively, generate a recommendation for carpooling if tr_i is contained in tr_j or viceversa. More formally, we define the *recommendation* as follows.

Definition 6 (Recommendations) *Given a set of users* U*, we define* R_U *as the set of* recommendations *with respect to the users in* U*.* $R_U = \{r_{ij}\}$ *where* $i, j \in U$ *are users and* $r_{ij} = (i, j, tr_i, tr_j)$ *denoting that* passenger j *is recommended to* driver i *because*

$$\exists\, tr_j \in T_j \text{ and } tr_i \in T_i \text{ s.t. } tr_i \sqsubseteq_{\delta,\tau} tr_j$$

where T_i *and* T_j *are the mobility demands of* i *and* j*, and* j *is the* passenger *and* i *is the* driver*.*

By Definition 5, a passenger has to walk no more than δ, and wait no more than τ. We can group all the recommendations in a set R_U, containing all the possible recommendations between any pair of users in U. Following the

recommendations in R_U we call D the set of possible drivers and P the set of possible passengers. For each recommendation we define the variable m_{ij} that is computed as the sum of the walking distances for pick-up and drop-off point and then normalized by the maximum. This is referred to as *normalized distance between trajectories*. Note that m_{ij} exists within the interval $[0, 1]$ only if a recommendation between i and j exists

Given all the definitions above, the objective of the optimization method is to find a set A_{R_U} of *assignments* containing a subset of recommendations of R_U, such that the total number of cars required to satisfy $\mathcal{T}_U$ is minimized and the total enjoyability of the system is maximized and the following constraints are satisfied: i) no user is both passenger and driver; ii) each vehicle holds no more than γ passengers; iii) each user can be found in only one vehicle.

3.2 Optimization problem

Given the enjoyability and mobility patterns described above we formulate the problem using an integer linear program. We start from a set of users that can be potentially grouped together into cars. Within each car only one of the users is a driver while the other ones are defined as passengers. The number of drivers in the system indicates the number of cars allocated by the algorithm for the entire set of users. The grouping process is regulated by two aspects: i) trajectory containment; ii) enjoyability between users. The optimization procedure takes as input the enjoyability values and the set of recommendations and generates the optimal assignment A_{R_U}. From the recommendation set R_U we can build three sets: D, the set of candidate drivers in the system; P, the set of candidate passengers (D and P may overlap in the recommendations, but not in an assignment); C, the set of possible couples (i, j) driver-passenger. We define the following parameters:

- a parameter m_{ij} describing the normalized trajectory distance, with $m_{ij} \in [0, 1]$ if driver i can give a ride to passenger j. We set $m_{ij} > 1$ otherwise. We call M the set of all m_{ij} with $i, j \in U$;
- a parameter $\bar{e}_{ij}$ that describes the unenjoyability of two users traveling together, $\bar{e}_{ij} \in [0, 1]$ where 1 indicates that users i and j are not prone to travel together and 0 indicates that users i and j are prone to travel together. Further, $\bar{e}_{ii} = 1$ so as to indicate that a user will not enjoy traveling alone.

Additionally, we also define the following variables:

- a binary variable x_{ij} that describes the assignments between drivers and passengers, specifically $x_{ij} = 1$ if i is the driver of passenger j, $x_{ii} = 1$ if i is a driver and zero otherwise;
- a binary variable y_{jki} indicates whether two passengers share the same car, specifically $y_{jki} = 1$ if passengers j and k share the same car with driver i, and zero otherwise;

The optimization model finds the minimum, over x_{ij}, of the following objective function:

$$\alpha\rho\sum_{i\in D} x_{ii} + (1-\alpha)(\sum_{(i,j)\in C} \bar{e}_{ij}\cdot x_{ij} + \sum_{i\in D}\sum_{(i,j)(i,k)\in C, j\neq k} \bar{e}_{jk}\cdot y_{jki}) \quad (1)$$

where the parameter ρ is the cost of adding a new car to the system. The purpose of the scale factor ρ is to sum the two objectives and have a comparable scale. Furthermore, the weight α is used to give a preference for minimizing the number of cars and minimizing the total unenjoyability in the system. The data-driven method to compute α and ρ, and is explained in Section 6.

The optimization is subject to:

$$\sum_{j\in P} x_{ij} \leq \gamma x_{ii}, \forall i \in D \quad (2)$$

where the maximum number of passengers per car is set to γ.

$$\sum_{i\in D} x_{ij} = 1, \forall j \in P \quad (3)$$

where one driver has to be assigned to only one car.

$$m_{ij}\cdot x_{ij} \leq 1, \forall (i,j) \in C \quad (4)$$

a limit different than 1, within $[0,1]$ may be taken instead, to restrict the set of recommendations to take into account. For the sake of broader optimization, we take them all.

$$y_{jki} \leq x_{ij} \quad (5)$$

$$y_{jki} \leq x_{ik} \quad (6)$$

$$y_{jki} \geq x_{ij} + x_{ik} - 1 \quad (7)$$

$$\forall i \in D, j \in P, k \in P : (i,j) \in C, (i,k) \in C, j \neq k$$

that are used to linearize the relation $y_{jki} = x_{ij}\cdot x_{ik}$.

The algorithm proposed aims at minimizing the number of cars jointly with maximizing the enjoyability of the system (formulated as minimization of the unenjoyability for convenience here). The output is to group passengers in cars and at the same time ensure that they will enjoy the ride in each car.

4 Methodology

In this Section, we present the GRAAL methodology (as well as some baselines), to derive an optimal assignment starting from Twitter data, to come-up with the relevant parameters. While the problem formulation was intentionally left generic and agnostic to the real dataset used, this methodology assumes Twitter as sole source of data, although we discuss the applicability of GRAAL to other types of data in Section 7.

4.1 Assumptions

Twitter may be not the perfect source of data for any of the three dimensions (text, trajectory, and co-presence) that we need for the optimization model. However, it is among the few public ones providing some information in all of three areas. We tackled the problems arising from not having ideal data as follows:

- Co-presence: we estimate the co-presence of two users in a cell at the same time, and thus the mobility demand of users, by aggregating several days of data.
- Trajectory: as geo-tagged tweets are too sparse to track users between origin/destination pairs, we assume every user is following the best path between them, which we compute by running the same journey planner for every pair of user locations.
- Topic mining: tweets are short, and the typical usage of Twitter include typos, abbreviations and slang. However, topic extraction via *Latent Dirichlet Allocation* [4] is typical on documents, and is shown to provide insightful results also from Twitter [43].

Moreover, we work under the following assumptions, most of which are common in this context: *i)* we assume all the users in the system travel by car; *ii)* we assume all the cars moving from A to B follow the trajectory returned by a journey planner (the same planner is used by all the cars); *iii)* we assume users accept the recommendations; *iv)* we assume to be working on frequent, recurring mobility, rather than solving the on-demand carpooling problem; *v)* having divided the space into a grid of cells, we perform the geo-match on the center of the cells. Section 7 discusses some limitations arising from our assumptions.

4.2 Algorithm

Algorithm 1 shows the steps performed by our methodology to solve the socially-optimal carpooling problem. The algorithm takes five parameters as input, and we explain in Section 6 how to tune the last two: i) the bounding box where to perform carpooling, ii) a spatial threshold δ, iii) a temporal threshold τ to define the time-stamped locations, and to compute trajectory containment, iv) α, to balance enjoyability and number of cars, and v) ρ, the cost of adding a car to the system. Lines $1-3$ are used to get a geo-tagged corpus of tweets from the bounding box, to derive a set of users from it, and to filter those users with poor data. Namely, we remove users with an average tweet per day ratio below a certain threshold (see Section 6 for details), and with a ratio between average number of distinct words and number of tweets below 1. This last step aims at removing automated tweets, and spammers. In lines $4-10$, for each user, we get his/her tweets (not necessarily geo-located) to build a larger corpus (geo-located tweets constitute a small fraction of the entire set of tweets), which we clean by removing stopwords and performing stemming. Then we get the users'

friends list, i.e. the other users that the user is following. In line 7 we compute the vector of most visited (systematic) time-stamped locations of a user, given δ and τ. In particular, we define a spatial grid over the *boundingbox*, consisting in cells of width δ, while we slice time in non overlapping slots of duration τ. From this set, in line 8 we query a journey planner to derive trajectories connecting any two time-stamped locations in each users' L_u. Finally, in line 11, we compute the vector of topics contained in the users' documents. This is done by running a *Hierarchical Dirichlet Process* (*HDP*) [36] on the users' tweets texts. *HDP* is a parameter-free version of *Latent Dirichlet Allocation* (*LDA*) [4] that automatically infers the number of topics. Lines $12-13$ compute the like-mindness between any two users, and then, for each user, use the median value of it to compute the homophily in lines $14-15$. In lines $16-18$, we compute the enjoyability values between any two users. In lines $19-21$ we generate the recommendations from the set of mobility demands. In line 22, we build the matrix of mobility matches from the recommendations. Finally, we solve the multi-objective optimization in line 23 to find a set of assignments minimizing the objective function described in Section 3.

To clarify what happens to each user in the system, we consider the user's perspective: assuming the user has passed the filter in line 3 (i.e. we have enough data about this user - this filter may be applied once for all, and could be lifted for different input data like mobile phone records, user-generated input, etc.), spatio-temporal as well as social and topic analytics are performed in line 4-18 and the computed parameters are associated to this user. In lines 19-22 an implicit "labeling" of users as possible passengers and drivers is done. In fact, we review all the trajectories mined above, and we find matches between them. If, for a given user, there are no matches at all, this user will not be in the R_U set, and will be driving a single occupancy vehicle on his/her own. These users are not considered in the optimization at all, as no recommendations are possible for them. For every other user, generally speaking, it is true that they may be considered as either passengers or drivers. For instance, if user A has a trajectory including one of the trajectories of user B, and user B has a trajectory including one of the trajectories of user C, then A can potentially become a driver, B can potentially become either a driver or a passenger, and C can potentially become a passenger. However, the optimization in line 23 takes all these possibilities into account, and a user is finally either labeled as a passenger, or a driver, but cannot be both. In other words, we do not pre-select who are the drivers, and who are the passengers, but this is rather automatically discovered by the optimization model.

4.3 Complexity

The complexity of GRAAL is dominated by the optimization step. Optimization problems involving discrete decision variables are NP-Hard in general [39]. However, as this may be optionally replaced by heuristic approaches, and for the sake of completeness, we report also the complexity of the other relevant steps: *computeTimeStampedLocations* and *computeTrajectories* are linear in the num-

Algorithm 1 GRAAL ($boundingbox$, δ, τ, α, ρ)

1: $\mathcal{G} \leftarrow getTweets(boundingbox)$
2: $U \leftarrow getUsers(\mathcal{G})$
3: $U \leftarrow filterUsers(U)$
4: **for** $i \in U$ **do**
5: $\quad D_i \leftarrow getTweets(i)$
6: $\quad F_i \leftarrow getFriends(i)$
7: $\quad L_i \leftarrow computeTimeStampedLocations(D_i, \delta, \tau)$
8: $\quad T_i \leftarrow computeTrajectories(L_i)$
9: $\quad \mathcal{T}_U \leftarrow \mathcal{T} \cup T_i$
10: $\quad \mathcal{D}_U \leftarrow \mathcal{D} \cup D_i$
11: $\{\vec{t_i}\} \leftarrow computeTopics(\mathcal{D}_U)$
12: **for** $i, j \in U$ **do**
13: $\quad lm_{i,j} \leftarrow computeLikemindness(\vec{t_i}, \vec{t_i})$
14: **for** $i \in U$ **do**
15: $\quad h_i \leftarrow computeHomophily(i, F_i)$
16: **for** $i, j \in U$ **do**
17: $\quad e_{i,j} \leftarrow computeEnj(lm_{i,j}, h_i, h_j)$
18: $\quad E \leftarrow E \cup e_{i,j}$
19: **for** $tr_i, tr_j \in \mathcal{T}$ **do**
20: $\quad$ **if** $tr_i \sqsubseteq_{\delta,\tau} tr_j$ **then**
21: $\quad\quad R_U \leftarrow R_U \cup (i, j, tr_i, tr_j)$
22: $M \leftarrow computeMobilityMatches(R_U)$
23: $A_{R_U} \leftarrow optimize(\alpha, \rho, E, M)$
24: **return** A_{R_U}

ber of locations; regarding HDP, with large amounts of data, the time to process individual documents increases due to increased density, leading in the worst case to a super-linear increase (cubic in the number of terms) [25]; *computeLikemindness* is constant, but it is executed in lines $12-13$ which are quadratic in the number of users; in the same way, the computation of homophily in line 15 is constant but is repeated linearly in the number of users; lines $16-18$, computing the enjoyability which takes constant time for each pair of users, is quadratic in the number of users; line 21 is executed in a nested **for** loop which is quadratic in the number of trajectories.

4.4 Baselines

We compared GRAAL to a number of baselines, which we briefly describe in the previous section. We compared with a random approach, a heuristic approach maximizing the enjoyability, and against GRAAL used with two particular values of α. Additionally, we used an approach based on the same rationale behind [8], maximizing the number of friends in a car. However, as the goal of the latter is different, and as their method also solves a different version of the carpooling

problem, we present different types of results for it in Section 6.

All the baselines start from a set of recommendations R_U computed as described in this Section. Then, they each return a (potentially different) subset of it, together with the recommendations on the single occupancy vehicles that constitute different sets of assignments A_{R_U}. To describe the first two baselines, consider the set of recommendations R_U as a directed graph G_{R_U} built by having a directed edge (i, j) if j can get a ride from i. Then:

- *Random*: we rank randomly the edges of G_{R_U}, then we take the first edge (i, j) in the rank and, if i has not been already selected as a passenger and there are less than $\gamma = 4$ assignments (see Sec. 3) with i as driver, then we flip a coin: with probability 0.5, we thus remove all the edges linked to j and produce the assignment (i, j). Otherwise, we proceed to the next edge, and repeat the procedure for all subsequent edges in the ranking. If, at the end of the procedure, there are nodes (passengers) for which no final recommendation was made, they become drivers of single-occupancy vehicles.

- *Heuristic*: we maximize the enjoyability with a greedy approach. We proceed like in *random* with the only difference that the edges of G_{R_U} are ranked by descending enjoyability e_{ij}.

- *Social*: this is basically GRAAL with $\alpha = 0$, i.e. we optimize only for total enjoyability (which is maximized).

- *Green*: this is basically GRAAL with $\alpha = 1$, i.e. we optimize only for total number of used cars (which is minimized).

Lastly, we compare also with [8] in terms of user impact, in Section 6. We additionally present results for: i) GRAAL with α varying from 0 to 1 with 0.05 increments (this thus include the Social and Green baselines); ii) GRAAL with a particular value for α as learned as described in Section 5; iii) all the baselines as described above. The experiments were conducted for two possible values for δ ($500m$ and $70m$), and two possible values for τ ($60min$ and $30min$).

5 User study

In order to assess the effect of enjoyability in carpooling compared to other factors like sustainable mobility, we conducted a survey with potential end-users. The goal of this user study is to learn a crowd-sourced value for the weight α, and to better understand the potential impact of a carpooling system based on the GRAAL methodology, if this were to be implemented. This study, hence, is not meant to replace an on-field validation of the proposed carpooling methodology but rather to obtain an estimate of the weight (i.e., the user preference) of the two objective functions for the multi-objective optimization model of Section 3.2. This section presents the design of the study, while Section 6 presents the numerical results obtained.

The survey was sent via direct Twitter messages, other social networks (e.g. Facebook, LinkedIn, etc., including dedicated carpooling groups), and direct e-mail and mailing lists. The webpage containing the survey is shown in Figure 1. To generate the landing page, we picked a user $i \in U$ from our data, and computed which cars he/she would be assigned to using the two approaches (one minimizing the number of cars and the other maximizing enjoyability). The two solutions presented contain the following: i) a bar indicating the average enjoyability among the occupants of the car; ii) a bar indicating the "greenness" of the solution, computed as the global amount of cars saved by the city-wide system if all the users were to click on this choice. The two cars were presented in random order, to minimize the probability of clicks performed on a given column. The two presented solutions are referred to as "social choice" and "green choice". The first one is the car with higher enjoyability but lower greenness value (obtained by Social), while the second choice is the car with lower enjoyability but higher greenness value (obtained by Green). Note that, while the enjoyability is a local property of the car, the greenness is a global, city-wide, property. That is, there are only two values of greenness for a city: the one obtained if every user were to click on the social choice, and the one obtained if every user were to click on the green one. After this step, the users were directed to a subsequent set of general questions on carpooling, including the following: "which of the following would make carpooling more attractive to you? Savings, sharing the car with interesting people, or sustainability of the solution?"

We collected 237 answers, with 39% in favor of a social solution. After collecting the answers, the values are exploited to learn the weight α in the multi-objective optimization model which represents how much the users are more likely to prefer the Social car with respect to the Green one. As mentioned, the page presents two cars with their enjoyability values of e_S (the enjoyability of the Social car) and e_G (the enjoyability of the Green car). If their difference

Figure 1: Part of the landing web page of the survey

$(e_S - e_G)$ is high, meaning that the social car has a high value of enjoyability, while the green car has a low value for it, we may expect the user to be tempted to click on the social car, rather than the green one. As the greenness values of the Social and Green car for a given city are fixed (i.e., they do not change if a different pair of solutions is displayed), we do not take them into account in the learned weight. Instead, we consider the difference of enjoyabilities between the green and the social car, which depends on the pair of solutions displayed. We define the following two values: for the Green car, the value v_G is given as:

$$v_G = e_S - e_G$$

while for the Social car, the value v_S is computed as:

$$v_S = 1 - (e_S - e_G).$$

Given S, the set of the social choices that were obtained from the survey and G the set of the green choices, the values v_S and v_G are computed on their elements and the weight α is defined as the following ratio:

$$\alpha = \frac{\sum_{j=1}^{|G|} v_{G_j}}{\sum_{i=1}^{|S|} v_{S_i} + \sum_{j=1}^{|G|} v_{G_j}} \tag{8}$$

6 Experiments

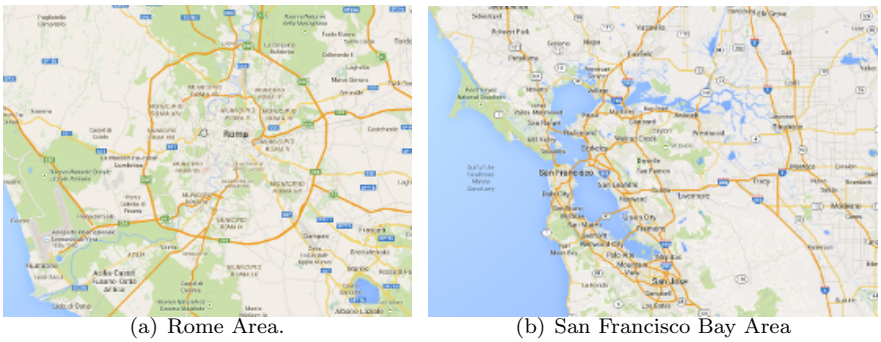
This Section presents the results of running GRAAL and the baselines on real Twitter data, using different sets of parameters. We present here the data and tools used, the parameter tuning (this includes the results from the user study), the results of computing the social measures, and the results of GRAAL and all the baselines. The results of optimization were assessed from a city-wide perspective (i.e. looking at the total values of the components of the objective function), and from a user perspective (i.e. looking at the distribution of the enjoyability of single cars, and impact on the user).

6.1 Data

We used the Twitter's Streaming API[1] to obtain two large datasets of geo-tagged tweets. We queried the API using two bounding boxes on the area of *Rome* (Fig. 2(a)), and the bay of *San Francisco*, hereafter referred to as San Francisco (Fig. 2(b))[2], for 50 days from the beginning of October 2014. As a result, we collected 558,000 geo-tagged tweets from 17,600 users in Rome, and 3,286,000 geo-tagged tweets from 113,000 users in San Francisco. We chose

[1] https://dev.twitter.com/docs/streaming-apis

[2] GPS coordinates bounding box: Rome (12.234498, 41.655642, 12.85576, 42.141028), San Francisco (-122.667, 36.8378, -121.2949, 38.0771)



(a) Rome Area. (b) San Francisco Bay Area

Figure 2: Geographical areas analyzed.

Rome as it is the city with the largest population in Italy, while San Francisco was chosen for its popularity in carpooling studies [7, 12, 1]. By applying user and tweet filtering, we ended up with the statistics reported in Table 1. We wanted to consider around 1000 users in each city. We then filtered out the users in Rome having less than 40 tweets, and users in San Francisco having less than 300 tweets. This resulted in 1106 users in Rome, and 1052 users in San Francisco.

6.2 Tools

GRAAL was written in Java and C, making use of external libraries for some specific tasks. We used a publicly available Java implementation of HDP[3], to perform non-parametric topic modeling. To execute route planning we used OpenRouteService[4], a public Java library. As *space-dist* and *time-dist* we used the geo-spherical distance and the absolute difference respectively. Finally, to perform the optimization steps, we used the C APIs of IBM *CPLEX*[5].

6.3 Parameters

To run GRAAL on our data, besides the parameters of Algorithm 1, we had to choose a sample of the data (in number of days) and a number of topics to put in the topic vectors. We decided to leave the bounding box, and the spatio-temporal parameters δ and τ as choices available to the analysts to conduct different analyses. This allows to run the optimization model with different input parameters as a function of different temporal and spatial resolutions. In our experiments, we report results for δ set to 500 and 70 meters, and for τ set to 30 or 60 minutes. Note that the combination $\delta = 500m$ and $\tau = 30min$ agrees with common sense, or best practice, in journey planning: users are typically willing to walk distances up to 500 meters, and have a flexibility of waiting up to 30 minutes to find a means of transport [18]. In terms of number of most frequent locations, we chose three as it typically covers home, work, and the so

[3]https://github.com/arnim/HDP
[4]http://openrouteservice.org/
[5]http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/

Dataset	**Rome**	**San Francisco**
Users	1,106 (06.53%)	1,052 (00.93%)
Tweets	237,351 (42.19%)	681,597 (20.74%)

Table 1: Users and tweets statistics after filtering. Numbers in brackets are the percentages over the initial unfiltered data.

called "third place". Home and work place detection were out of the scope for this paper. To decide the number of days of data to take, we saw that the ratio of people which have at least one change in their top three locations if we take more than x consecutive days drops dramatically after $x = 40$. We thus chose to take 40 consecutive days of data in our sample.

Next, we describe how we learned a suitable number of topics, α and ρ.

6.3.1 Number of topics

As stated above, we used a nonparametric HDP algorithm to estimate the number of topics automatically. Since HDP is nondeterministic, we ran it 2,000 times on our data, obtaining on average 25.48 topics ($\sigma = 1.56$) on Rome and 25.61 ($\sigma = 1.54$) on San Francisco. According to this, we selected the results relative to a number of topics of 25, to construct our vectors $\vec{t_i}$.

6.3.2 Tuning ρ

The ρ parameter is defined as the cost of adding a car to the system within the optimization model. We studied the effects of varying this parameter, in terms of number of cars saved while keeping α fixed to 0.5 (i.e., there is no preference between minimizing the number of cars and enjoyability). We varied ρ from 0 to 10, and observed that $\rho = 2$ had the largest impact on the number of cars saved (see Figure 3 for the case of Rome). The results were both similar in Rome and San Francisco. Hence, $\rho = 2$ is used to scale the first part of the objective function (i.e., minimizing the number of cars) to have an objective function with comparable scales which is critical in multi-objective optimization.

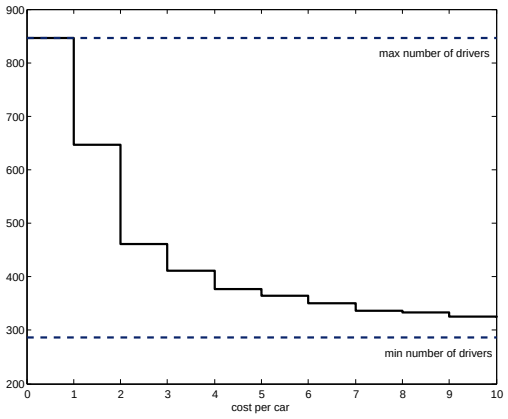


Figure 3: Plot of ρ vs number of cars.

6.3.3 Tuning α

The α parameter was learned looking at the results of the user study conducted as described in Section 5. We collected 237 responses coming from three different sources: 2% came from direct messages sent via Twitter; 12% came from sharing the survey in other social networks; 86% came from direct e-mail or mailing lists sharing. In total, 39% of people clicked on the social choice. This is encouraging, as it confirms the need for a social-aware carpooling system. Another encouraging result was provided by the answers to the additional survey question: 24% of the people were more attracted by sharing the car with interesting people, while 41% by the savings provided by carpooling, and 35% considered the sustainability to be the most attractive aspects of carpooling. We consider these numbers as a measure of the potential impact of a carpooling system that takes into account the enjoyability of a car as an additional factor, rather than just minimizing the number of cars. Based on Equation (8), the final value obtained for α is 0.36 which is an estimate of the preference of the number of cars saved compared to that of enjoyability as given by the users. This value is used for balancing the two objectives in the optimization model.

6.4 Results on Social Measures

Figure 4 presents the distributions of like-mindness (top left), homophily (top right) and enjoyability between pairs of users (bottom row) for all the users. We report no significant differences in like-mindness and homophily between Rome

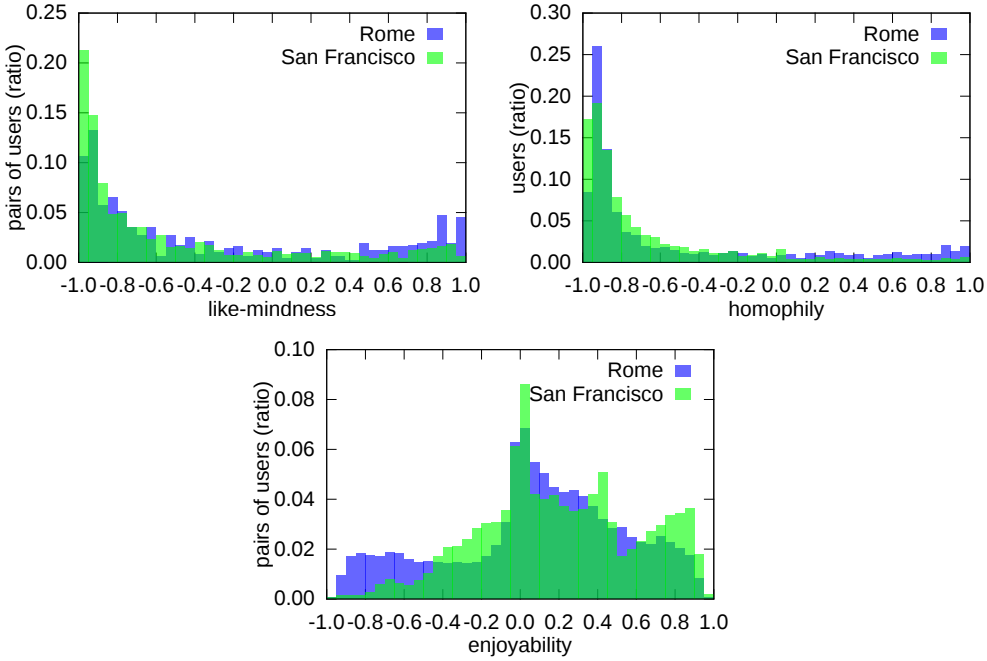


Figure 4: Social measures for all the couples of users.

δ	τ	Rome				San Francisco			
		$\lvert R_U \rvert$	$\lvert S_1 \rvert$	$\lvert S_2 \rvert$	$\lvert Z \rvert$	$\lvert R_U \rvert$	$\lvert S_1 \rvert$	$\lvert S_2 \rvert$	$\lvert Z \rvert$
500	60	6,883	81.56%	76.04%	18.44%	2,298	68.63%	57.41%	31.37%
500	30	5,870	79.84%	73.51%	20.16%	1,106	54.37%	36.88%	45.63%
70	60	349	26.85%	15.46%	73.15%	245	16.73%	9.60%	83.27%
70	30	309	24.68%	13.92%	75.32%	250	16.44%	9.79%	83.56%

Table 2: Statistics on user recommendations by δ and τ for GRAAL. $S_1 \subseteq U$ is the set of users with one or more recommendations, $S_2 \subseteq U$ is the set of users with two or more recommendations, $Z \subseteq U$ contains the users with no recommendations.

and San Francisco. As we can see from the first plot, computing a similarity based only on the like-mindness may end up recommending connections in a limited number of pairs of users. On the other hand, from the second plot, we learn that most of the people are heterophilous. If we combine the two things into the enjoyability, we see, in the third plot, that there is broader space for recommendations based on this measure, rather than the like-mindness. Moreover, the combination of the first two measures produces two different distributions for Rome and San Francisco, highlighting that the enjoyability is capturing a different phenomenon as opposed to just like-mindness.

6.5 Results on Recommendations

Table 2 reports some statistics for the recommendations using different combinations of spatio-temporal resolutions. The first column reports the number of recommendations, in column S_1 we see the percentage of users with one or more recommendations, in column S_2 we see the percentage of users with two or more recommendations, (for which the optimization has more impact), while in column Z we report the percentage of users with no recommendations (these will end up being drivers of single occupancy vehicles in all the models). From this table, we see the clear effects of taking the same number of users in the two cities having very different geographical structure. In particular, San Francisco Bay Area is a much larger area than Rome. As carpooling in San Francisco works actually across the entire area, it would not make much sense to keep the same user density per area and reduce the area over San Francisco, we decided not to take any corrective actions. In this way, we could also assess the effects of having different recommendation densities on the performances of the optimization model. Thus, we report a larger room for optimization in Rome in general, and for $\delta = 500m$ in general as well. In San Francisco, only $\delta = 500m$ provides significant room for optimization. We expect this to be seen in the results we have, at the city level.

6.6 City-wide perspective

All the parameters tuned and the recommendation calculated as explained in previous sections are used as an input to the optimization model. We consider the results as applied to a single day of trips. The results are here presented at the city level, i.e. at the level of the entire optimization model. GRAAL

with $\alpha = 0.36$, as learned from the user survey, theoretically saves up to 57% of the cars needed in Rome and 40% in San Francisco for the subset of users under consideration as compared to users taking their private cars, while having a high level of total enjoyability. In addition to $\alpha = 0.36$, we studied the effects of varying $\alpha \in [0, 1]$ (with steps of 0.05), on the total number of cars saved and the total enjoyability of the system. We compared GRAAL for $\alpha = 0.36$ with all the baselines.

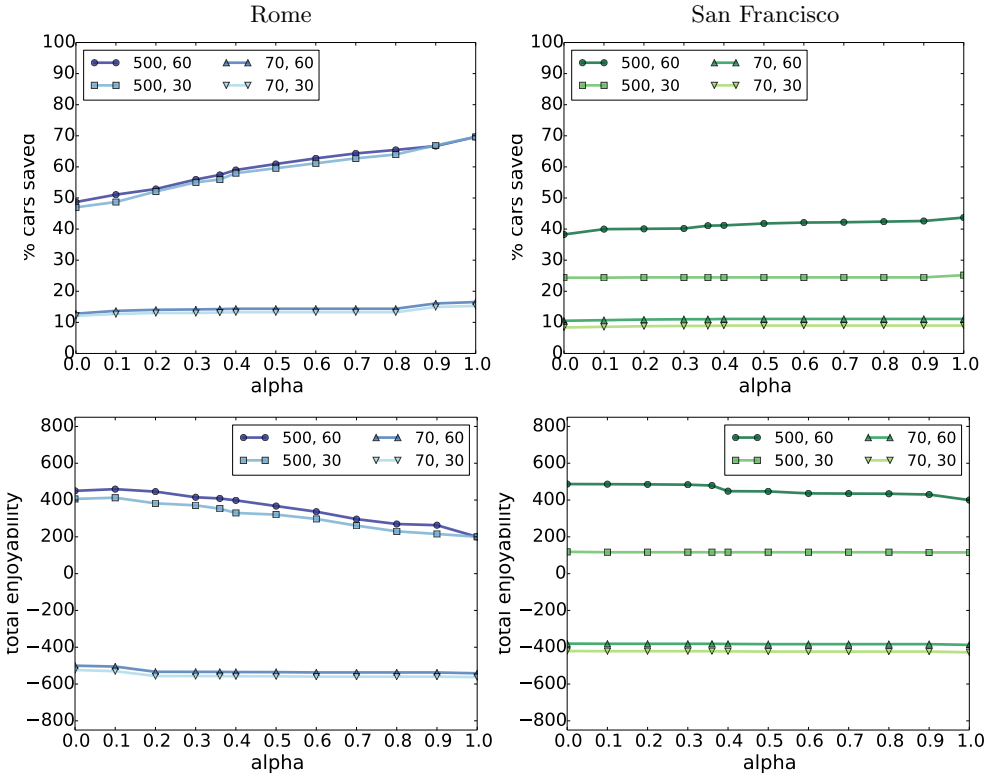


Figure 5: Cars saved (top row) and total enjoyability (bottom row), in Rome (left column) and San Francisco (right column) by running GRAAL with 20 values of α and different values of δ and τ. For all the plots, higher is better.

Figure 5 reports the number of cars saved (in the top row) and the total enjoyability (in the bottom row) for Rome (in the left column) and San Francisco (in the right one). As we can see, the best performance are reached for the city of Rome with $\delta = 500m$ and for any values of τ. In all the other cases (and in San Francisco as well) we see a mostly flat behavior, which means that with less room for optimization, α can not make a big difference in the results. Moreover, $\delta = 70$ provides the lowest number of cars saved and the lowest enjoyability in both cities. This agrees with the lowest numbers in the S column of Table 2. Moreover, where the Z column of Table 2 is very high, we see negative values

of enjoyability. This is due to the large number of people going alone, for which we assign an enjoyability score of -1, as described in Section 3.

Figure 6 shows the percentage of cars saved (in the top row) and total enjoyability (in the bottom row), in Rome (left column) and San Francisco (right column) by running GRAAL with $\alpha = 0.36$ and all the baselines. As expected, the highest number of cars is saved by the Green approach. One encouraging result is that Social saves a significantly higher number of cars with respect to Random and Heuristic. This is due to the choice of assigning -1 as enjoyability to a person traveling alone. As a consequence, even if Social does not directly minimize the number of cars, it tends to put more people together anyway. The GRAAL approach with $\alpha = 0.36$ is a trade-off between Social and Green (which are basically GRAAL with the two possible extreme values for α).

Consider now the bottom row of Figure 6, with $\delta = 500m$. The negative total enjoyability confirms that in those cases there is a significant number of people going alone. This is avoided by GRAAL with all alpha values (as reported in Figure 5). In accordance with Table 2, reporting a high number of single occupancy vehicles for $\delta = 70m$, we have only negative total enjoyability for all models for this value of δ.

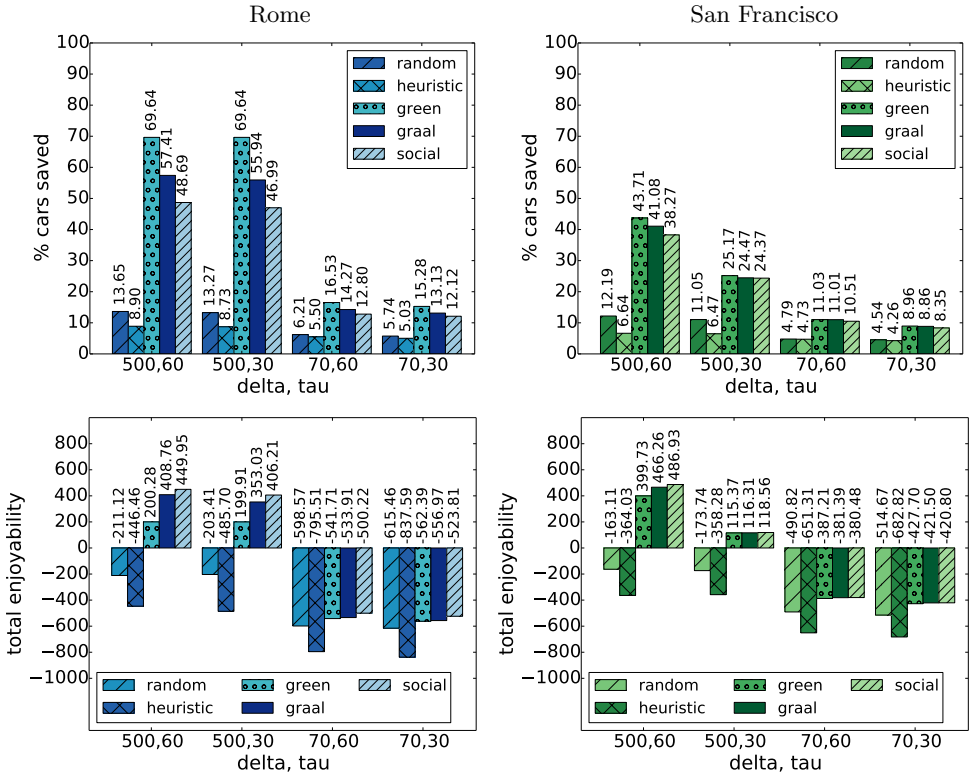


Figure 6: Cars saved (top row) and total enjoyability (bottom row), in Rome (left column) and San Francisco (right column) by running GRAAL with $\alpha = 0.36$ and all the baselines. For all the plots, higher is better.

Finally, we report the results of comparing GRAAL with the Green model minimizing the number of cars, in terms of two KPIs: additional cars used, and additional km traveled by the cars in the system. Table 3 reports these values for each city and combination of δ and τ. In the "% cars" cell, we report the percentage of additional cars used by GRAAL with respect to Green, normalized by the number of cars needed if all the users were taking a car. In the "% km" cell, we report the percentage of additional km traveled by the GRAAL drivers with respect to Green, normalized by the total amount of km traveled if all the users were taking a car. The first column can be seen as a way to measure the cost of adding a car to the system (for example, in terms of parking slots needed), while the second column can be seen as a way to measure the overall cost of the system (for example, in terms of CO_2 emissions). As we see, although we add up to 13% of cars into the system with GRAAL, they are typically used to cover short distances, as the additional km traveled, in percentage, are well below the percentage of cars added. We note that in our model, drivers are not allowed to detour to pick up passengers. That is, giving a lift to someone will always reduce the total distance traveled.

δ	τ	**Rome**		**San Francisco**	
		% cars	% km	% cars	% km
500	60	12.23	3.67	2.63	0.02
500	30	13.70	4.39	0.70	0.43
70	60	2.26	0.35	0.02	0.01
70	30	2.15	0.38	0.10	0.32

Table 3: Percentages of additional cars and additional km needed by GRAAL with respect to the Green solution.

6.7 User perspective

We assess the results from the user perspective, in terms of enjoyability in the single cars. As aggregates, we report minimum, maximum, average, 90th, 75th, 50th, 25th, and 10th percentiles of the distribution of the enjoyability across vehicles, in order to understand the improvement introduced for a user, as shown in Figure 7. For this assessment, we considered only the users who received a recommendation. That is, we remove most of the effects of considering an enjoyability equal to -1 for a high number of people in these plots. The first clear result is that there is a globally higher enjoyability in San Francisco, compared to Rome. This is coherent with the results on the distribution of the enjoyability per city reported in Figure 4, which shows both a higher negative tail in Rome for the enjoyability, and a higher positive tail for San Francisco. Despite the globally higher enjoyability, there is again the problem of the results being more flat. Consider now the results in Rome, with $\delta = 500m$. In the Green model, where the optimization disregards the enjoyability, the results are inline with Random, while Heuristic does a better job. This is also true for the other value of δ, although less evident.

Table 4 is used to compare our approach with the method described in [8], which tries to put friends (i.e., direct Twitter links) together, as their concept

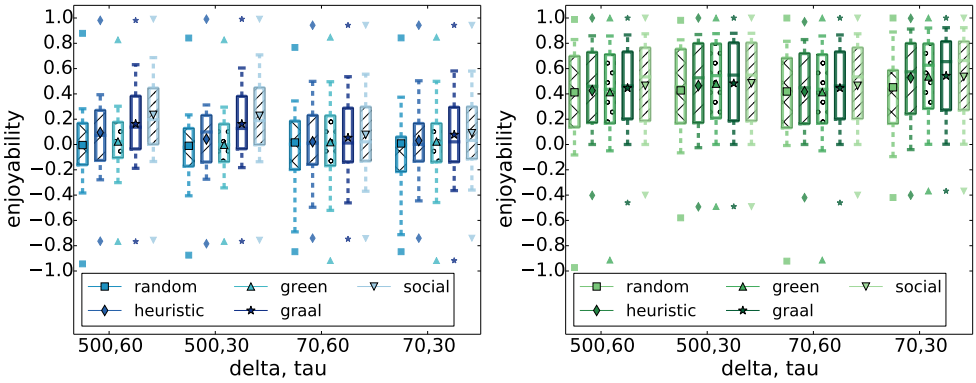


Figure 7: Enjoyability per cars (min, max, 10th, 25th, 50th, 75th, 90th percentiles, and average across all cars), for Rome (left) and San Francisco (right). Higher is better.

of enjoyability. We evaluate against [8] in terms of impact on users, reported in Table 4, which contains the same columns as Table 2. To produce this table, we first ran Green, then we applied a brute force approach to optimize according to friendship (friends are put together in a car). As we clearly see, the number of recommendations between friends is much smaller than what we can achieve in GRAAL reported in Table 2, due to the sparsity of the friendship connections in Twitter (and in the real world, too), as opposed to the fact that we could compute the enjoyability between any two users in GRAAL. Thus, the room for optimization here is much smaller, with numbers in S_2 not reaching two digits. We include the S_1 column in the two tables to give more chances to this approach. In fact, even if we can optimize less, with at least one recommendation we can still put friends together. Nevertheless, numbers go up to slightly more than 11%. We did not compare with the same approach ran with a 2-hop network for the following reason: 2-hop friends (i.e., friends of friends), when they are not direct friends, are people with whom we can not give any guarantee on the enjoyability from a topic perspective, and neither they are direct friends. On the other side, 2-hop friends could be at least more trustworthy than unknown (but enjoyable) people. However, neither our methodology, nor the one in [8] are meant to be seeking a higher trust in the system, which is then left as future work.

δ	τ	**Rome**				**San Francisco**			
		$\lvert R_U\rvert$	$\lvert S_1\rvert$	$\lvert S_2\rvert$	$\lvert Z\rvert$	$\lvert R_U\rvert$	$\lvert S_1\rvert$	$\lvert S_2\rvert$	$\lvert Z\rvert$
500	60	189	11.36%	7.42%	88.64%	148	7.57%	4.89%	92.43%
500	30	183	11.12%	7.02%	88.88%	120	7.50%	4.23%	92.50%
70	60	53	9.37%	4.40%	90.13%	46	5.82%	2.31%	94.18%
70	30	51	9.40%	4.32%	90.60%	45	4.35%	2.25%	95.65%

Table 4: Statistics on user recommendations by δ and τ for [8]. $S_1 \subseteq U$ is the set of users with one or more recommendations, $S_2 \subseteq U$ is the set of users with two or more recommendations, $Z \subseteq U$ contains the users with no recommendations.

6.8 Running times

GRAAL ran in around 2 minutes with each of the α values under each of the δ and τ combinations, for both cities. Exceptions were $\delta = 500m$ in Rome, where a higher number of recommendations brought the running times up to 1 hour.

7 Limitations and future work

Here, we discuss some limitations of our methodology, hints for how to overcome them, and possible future work.

Customized optimization. GRAAL optimizes at the level of the city, not of a single user. This means we use one single value for α for all the users, under the assumption that, with this, we can achieve good performances at the level of the entire system. However, different people may have different preferences. Taking into account this would require a different model, which was out of the scope of this paper, as we were looking for a system-wide perspective. Along these lines, future work will include: customized optimization; traffic-dependent optimization (i.e. higher cost of adding a car during peak hours); distance-dependent optimization (i.e., relaxing the optimization model for shorter trips).

On-demand carpooling. We worked under the assumption that GRAAL may be the basis for a system which can run once in a while, to support systematic mobility. However, nothing prevents us from setting up the system using historical data in batches, then satisfying requests for rides on demand. A real-time, on-demand carpooling will be within the future work to investigate. However, it would be hard to reach the same levels of system-wide total enjoyability with the same model, as the room for optimization might be smaller.

Applicability to, and comparison with, other data. The choice for Twitter data was supported by a number of considerations. One could easily replicate our experiments on Call Detail Records (CDR) data used in conjunction with phone calls transcripts. This mix would improve the data quality and quantity in all the dimensions of our problem: better location data, better textual content, weighted friendship information. However, this type of data is not public. Alternative public data includes Flickr (whose typical usage is tourism [5]), Foursquare (for which we would need to find an external source for textual data for the same users), or Facebook (whose APIs do not expose the same type or amount of data. In addition, we used only one estimation of the mobility demand: the top three locations computed on the geo-located tweets. Other data (like the above mentioned CDR, or ride-sharing data, taxi-sharing data, journey planner request logs) may be used to estimate the mobility demand as well. However, we only needed one estimation to assess feasibility, while accuracy against a better demand estimation was out of scope for this paper. Future work will include such comparisons.

Field test. Although the user study gave us good input and motivation for this work, a field study, with the system put at work, has not been implemented. We are exploring the possibility of partnership with the mobility

agencies of a few cities, to test this solution with end users. The scope of this paper was to devise a theoretical, data-driven, methodology, starting from data available online, and ending at the recommendations. This is inline with the other theoretical carpooling papers presented in Section 2.

Extension to other objectives. As mentioned in Section 1, we acknowledge that there may be several obstacles for a larger adoption of carpooling: safety of passengers, mobility mismatches, social incompatibility, and so on. GRAAL aims at analyzing only one of them, and under only one definition of enjoyability. More possibilities will be investigated in the future, together with the inclusion of other factors like gender match, trust, network similarity, and others.

8 Conclusions

We have described GRAAL, a multiobjective model that optimizes carpooling recommendations for a weighted linear combination of number of cars used (which is minimized) and total enjoyability (which is maximized). GRAAL takes Twitter data in input, as this contains information on spatio-temporal, text, and social dimensions of geo-located user tweets. We conducted a survey to tune the weight of the linear combination in the optimization function. We received 237 answers, 39% of which were in favor of the Social solution, motivating this work, and providing the needed weights of the objectives for the multi-objective optimization model. We have then presented the theoretical results of the multiobjective optimizationin terms of cars saved and enjoyability for a subset of users in Rome and San Fransico. We have presented the results from the city and the user perspective. With the crowd-sourced α of 0.36, GRAAL has the potential of saving a significant number of the cars needed, while keeping a high level of the total enjoyability. From the user perspective, we have shown how the entire per-car distribution of enjoyability is increased with respect to the baselines. Future works include: customized optimization, on-demand optimization, the usage of different data for demand estimation and comparison, a field test with mobility agencies in different cities, and an extension of the optimization function to include also other factors like trust, safety, or network-based similarity.

Acknowledgment This work has been partially supported by the EC under the FET-Open Project n. FP7-ICT-609042, PETRA.

References

[1] Antonio Bento, Jonathan Hughes, and Daniel Kaffine. Carpooling and driver responses to fuel price changes: Evidence from traffic flows in los angeles. *Journal of Urban Economics*, 77:41–56, 2013.

[2] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. Allaboard: a system for exploring

urban mobility and optimizing public transport using cellphone data. In *Machine learning and knowledge discovery in databases*, pages 663–666. Springer, 2013.

[3] Nicola Bicocchi and Marco Mamei. Investigating ride sharing opportunities through mobility data analysis. *Pervasive and Mobile Computing*, 14:83–94, 2014.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[5] Igo Brilhante, Jose Antonio Macedo, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. Where shall we go today? planning touristic tours with tripbuilder. In *CIKM*, pages 757–762, 2013.

[6] Young-Ji Byon, Young Seon Jeong, Said Easa, and Joonsang Baek. Feasibility analysis of transportation applications based on apis of social network services. In *ICITST*, pages 59–64, 2013.

[7] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[8] Blerim Cici, Athina Markopoulou, Enrique Frias-Martinez, and Nikolaos Laoutaris. Assessing the potential of ride-sharing using mobile and social data: a tale of four cities. In *Ubicomp*, pages 201–211, 2014.

[9] Don Coppersmith, Tomasz Nowicki, Giuseppe Paleologo, Charles Tresser, and Chai Wah Wu. The optimality of the online greedy algorithm in carpool and chairman assignment problems. *TALG*, 7(3):37, 2011.

[10] Gonçalo Correia and José Manuel Viegas. Carpooling and carpool clubs: Clarifying concepts and assessing value enhancement possibilities through a stated preference web survey in lisbon, portugal. *Transportation Research Part A: Policy and Practice*, 45(2):81–90, 2011.

[11] Gonalo Correia and Jos Manuel Viegas. Applying a structured simulation-based methodology to assess carpooling time–space potential. *Transportation Planning and Technology*, 33(6):515–540, 2010.

[12] Joy Dahlgren. High occupancy vehicle lanes: Not always more effective than general purpose lanes. *Transportation Research Part A: Policy and Practice*, 32(2):99–114, 1998.

[13] Gonçalo Homem de Almeida Correia, João de Abreu e Silva, and José Manuel Viegas. Using latent attitudinal variables estimated through a structural equations model for understanding carpooling propensity. *Transportation Planning and Technology*, 36(6):499–519, 2013.

[14] Ahmed Elbery, Mustafa ElNainay, Feng Chen, Chang-Tien Lu, and Jeffrey Kendall. A carpooling recommendation system based on social vanet and geo-social data. In *SIGSPATIAL*, pages 546–549, 2013.

[15] Vanessa Frias-Martinez, Victor Soto, Heath Hohwald, and Enrique Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *PASSAT, SocialCom*, pages 239–248, 2012.

[16] Masabumi Furuhata, Maged Dessouky, Fernando Ordez, Marc-Etienne Brunet, Xiaoqing Wang, and Sven Koenig. Ridesharing: The state-of-the-art and future directions. *Transportation Research Part B: Methodological*, 57:28 – 46, 2013.

[17] Marta Gonzalez, Cesar Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[18] Erick Guerra, Robert Cervero, and Daniel Tischler. The half-mile circle: Does it best represent transit station catchments? *Transportation Research Record: Journal of the Transportation Research Board*, 2276:101 – 109, 2012.

[19] Riccardo Guidotti and Michele Berlingerio. Where is my next friend? Recommending enjoyable profiles in location based services. *Complex Networks 2016*, 2016.

[20] Riccardo Guidotti, Mirco Nanni, Salvatore Rinzivillo, Dino Pedreschi, and Fosca Giannotti. Never drive alone: Boosting carpooling with network analysis. *Information Systems*, pages –, 2016.

[21] Riccardo Guidotti, Andrea Sassi, Michele Berlingerio, Alessandra Pascale, and Bissan Ghaddar. Social or green? a data-driven approach for more enjoyable carpooling. In *IEEE ITSC 2015, to be presented*, 2015.

[22] Wen He, Deyi Li, Tianlei Zhang, Lifeng An, Mu Guo, and Guisheng Chen. Mining regular routes from gps data for ridesharing recommendations. In *KDD*, pages 79–86, 2012.

[23] Luk Knapen, Daniel Keren, Ansar-Ul-Haque Yasar, Sungjin Cho, Tom Bellemans, Davy Janssens, and Geert Wets. Estimating scalability issues while finding an optimal assignment for carpooling. *Procedia Computer Science*, 19:372–379, 2013.

[24] DongWoo Lee and Steve HL Liang. Crowd-sourced carpool recommendation based on simple and efficient trajectory grouping. In *SIGSPATIAL*, pages 12–17, 2011.

[25] Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. Reducing the sampling complexity of topic models. In *KDD*, pages 891–900, 2014.

[26] Shuo Ma, Yu Zheng, and Ouri Wolfson. T-share: A large-scale dynamic taxi ridesharing service. In *ICDE*, pages 410–421, 2013.

[27] Vittorio Maniezzo, Antonella Carbonaro, and Hanno Hildmann. *New Optimization Techniques in Engineering*, chapter An ANTS Heuristic for the Long — Term Car Pooling Problem, pages 411–430. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[28] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[29] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1):32–38, 1957.

[30] Joe Naoum-Sawaya, Randy Cogill, Bissan Ghaddar, Shravan Sajja, Robert Shorten, Nicole Taheri, Pierpaolo Tommasi, Rudi Verago, and Fabian Wirth. Stochastic optimization approach for the car placement problem in ridesharing systems. *Transportation Research Part B: Methodological*, 80:173 – 184, 2015.

[31] Alessandra Pascale, Thanh Lam Hoang, and Rahul Nair. Characterization of network traffic processes under adaptive traffic control systems. *Transportation Research Part C: Emerging Technologies*, 2015.

[32] Dino Pedreschi. Big data, social mining, diversity, and wellbeing. In *SIS*, pages 1–6, 2014.

[33] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Hypertext*, pages 116–125, 2014.

[34] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[35] Roger F Teal. Carpooling: who, how and why. *Transportation Research Part A: General*, 21(3):203–214, 1987.

[36] Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

[37] Michael Terry, Elizabeth Mynatt, Kathy Ryall, and Darren Leigh. Social net: using patterns of physical proximity over time to infer shared interests. *Conference on Human Factors in Computing Systems*, pages 816–817, 2002.

[38] Robert Tijdeman. The chairman assignment problem. *Discrete Mathematics*, 32(3):323–330, 1980.

[39] Craig Tovey. Tutorial on computational complexity. *Interfaces*, 32(3):30–61, 2002.

[40] Roberto Trasarti, Fabio Pinelli, Mirco Nanni, and Fosca Giannotti. Mining mobility user profiles for car pooling. In *KDD*, pages 1190–1198, 2011.

[41] Hai Yang and Hai-Jun Huang. Carpooling and congestion pricing in a multilane highway with high-occupancy-vehicle lanes. *Transportation Research Part A: Policy and Practice*, 33(2):139 – 155, 1999.

[42] Desheng Zhang, Ye Li, Fan Zhang, Mingming Lu, Yunhuai Liu, and Tian He. coRide: carpool service with a win-win fare model for large-scale taxicab networks. In *SenSys*, page 9, 2013.

[43] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. 2011.